Geo372
Vertiefung GIScience

Simple error models for spatial data

Herbstsemester

Ross Purves

Your projects

- This week is the last one where we will give you **specific practicals**
- We said we would be happy to look at your **project proposals this week** – we can do it next too!
- You need to **start preparations** – read the materials we gave you, and do some background reading
- Come up with a **question**, and **sketch out how you wish to address it**
- Remember, we won't grade your ideas – we are trying to help

Last week

- We explored the notion of **viewsheds**
- I emphasised the idea that **simple binary viewsheds** were only one possibility, and demonstrated the **influence of algorithms** on the results
- We saw some **applications of viewsheds** for **practical and scientific questions**

Some reminders

- We defined **data quality** in the second lecture, and emphasised the idea of **fitness for use**
- We saw examples where data were clearly **not fit for a particular use**
- Recall, **uncertainty** can be introduced at **multiple stages** (e.g. concepts, measurement and representation and analysis)
- The **spatial data transfer standard** allows us to document aspects of **data quality**

Outline

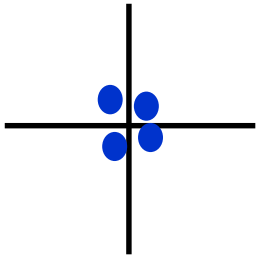
- Firstly, we'll look briefly at **error sources** and their **types**
- Then we will explore the STDS and some **simple error models**
- We'll see how we can document data quality using error models, whether data are **numerical or nominal**
- Our models will all be relatively **simple**, and we will see that complexity can **quickly increase**
- Using **OpenStreetMap** as an **example**, we'll look at some practical examples of measuring data quality

Learning objectives

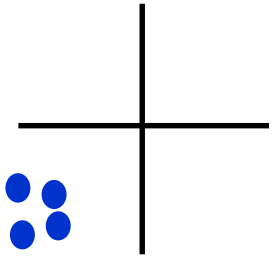
- You can give examples of **potential blunders, systematic** and **random errors** arising in spatial data
- You can critically use **simple error models** to describe **positional** and **attribute accuracy**
- You can show how **errors in lines an polygon positions** can be modelled by **epsilon bands**
- For given data you can propose ways of **documenting data quality** and explain **potential implications**

Precision and accuracy

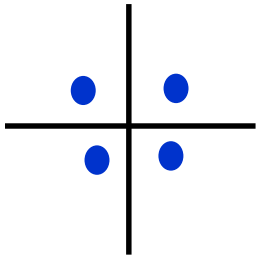
- You should understand the difference between:
  - **Accuracy** – the difference between a recorded value and its true value
  - **Precision** – the detail with which a measurement is made



Accurate and precise



Inaccurate but precise



Accurate but imprecise

What value could you use to **describe the precision?**

What do we need to **describe accuracy?**

In practice, precision and accuracy usually both influence uncertainty (why?)

Back to the SDTS

Positional accuracy

Attribute accuracy

Logical consistency

Completeness

Lineage

- If we want to use the SDTS we need a way of **measuring accuracy**
- Remember that implies that we need to know something about **"truth"**
- Where our data are the "best" we have two choices
  - We can **sample** some **ground truth data** to test accuracy
  - At a minimum we can **report** on the **precision** of the measurements – but this **doesn't** tell the whole story

Error types - blunders

- **Blunders** result from some gross, usually human, error (e.g. using the instrument wrongly, transposing values when entering them in a database)
- Blunders are often (but not always) easy to identify (e.g. measuring someone's height as 3m)
- The **likelihood of blunders** might be guessed at through completeness and lineage e.g.:

  "*…many of the depths in these areas have not been systematically surveyed. Depths in these areas are from miscellaneous lines of passage sounding or old leadline surveys. Uncharted dangers may exist."*

Error types - blunder

- Blunders cannot be described statistically or removed systematically
- If a dataset contains many blunders (rendering it unfit for purpose) the data must be collected again
- Most **people assume** data to be blunder free (dangerous assumption)

gadgets

Woman follows GPS into lake

MAY 16, 2016 7:59PM

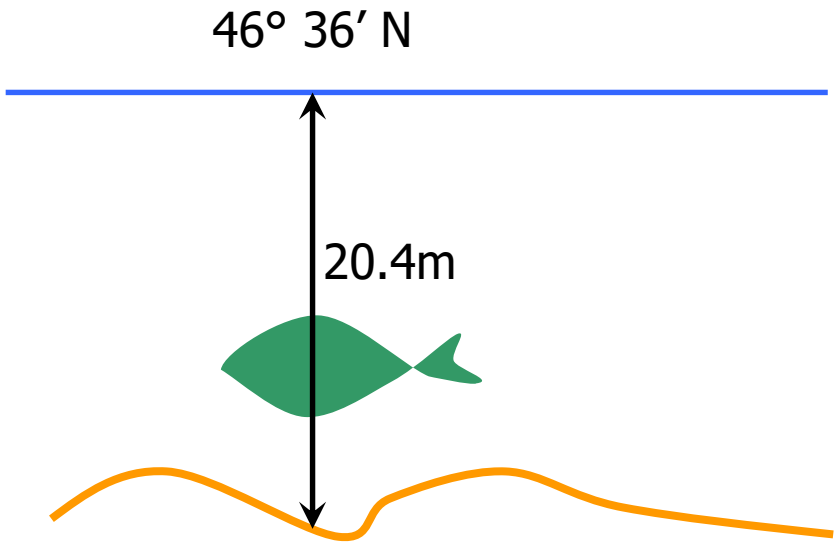
A 23-year-old woman managed to drive into a lake during a foggy Canadian night.

news.com.au

IT'S official, we rely too much on technology.

In years past motorists were at the mercy of road maps, sign postings and the view out their windshield. These days, we have a computer with a selection of ethereal yet authoritative voices to tell us which turn to take.

But as a 23-year-old Canadian woman found out, that method doesn't always pan out so well.



Position	**26° 36′ S**
Depth	**10.9m**

Source: http://www.news.com.au/technology/gadgets/woman-follows-gps-into-lake/news-story/a7d362dfc4634fd094651afc63f853a1

Types of error – systematic error

- **Systematic errors** follow some **sort of pattern**
- They are **usually** easy to spot by **comparing** two datasets
- Where **full metadata** are available, it is sometimes easy to fix systematic errors (since **mismatch** between **intention and data** becomes apparent)

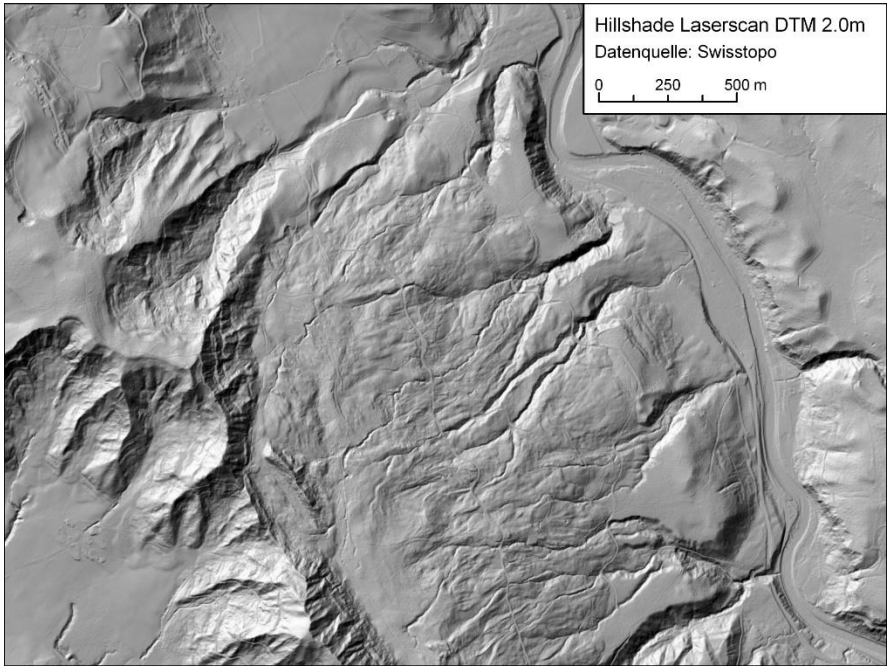
Types of error – systematic error

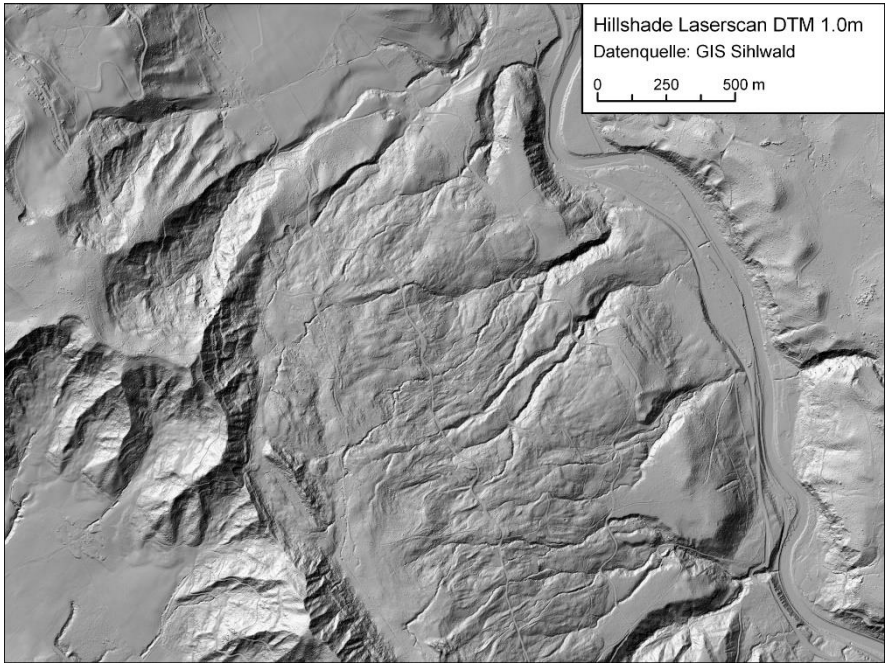
Quantitative systematic errors

- These can be **very simple** (e.g. reporting all latitudes as positive instead of negative)
- They are **often constant** (e.g. a false offset applied to calculate some difference value)
- They can also **vary** (systematically) over space (for example as a function of gradient)

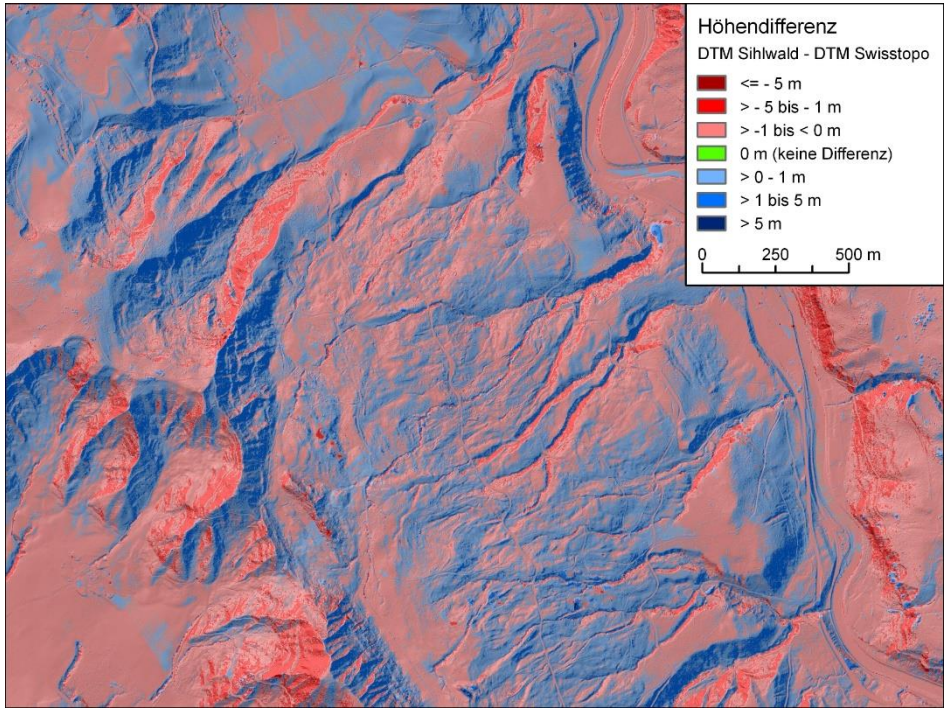
Qualitative systematic errors

- Such errors might occur when an observer **consistently misapplies** a classification (e.g. classifies all parks as football pitches)
- Could also be the result of an **conceptual error** in an algorithm (e.g. setting all white in a visual satellite image as snow…)





Two lidar surveys of Sihlwald – clear systematic error



Types of error – random errors

- Random errors are those attributable to precision **and** non-systematic (unpredictable) differences in the measurements made

- These errors are **always present** and are **well suited** to being **modelled**
- **Positional accuracy** and **attribute accuracy** are often described in terms of random error associated with a known value

Error models

Error models are a tool for the documentation of **data quality and uncertainty** and the influence of this on further work with data

Multiple uses, including:

1) Formal, compact **documentation** of errors in our data
2) Tools to **evaluate quality** and **suggest improvements**
3) Inputs to models of error propagation (e.g. how do **multiple errors influence results of analysis**)

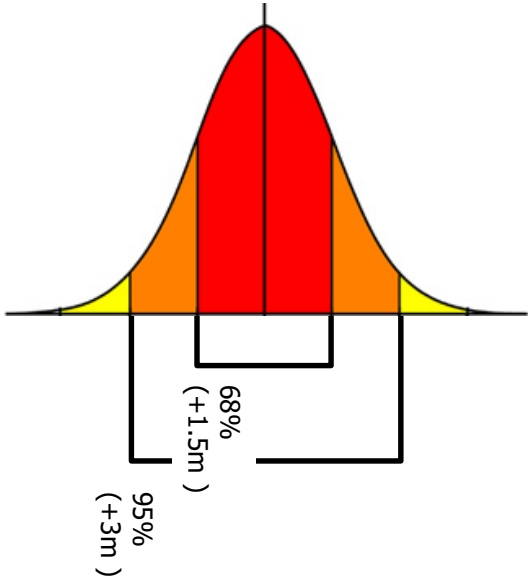
A first error model: Root Mean Square Error

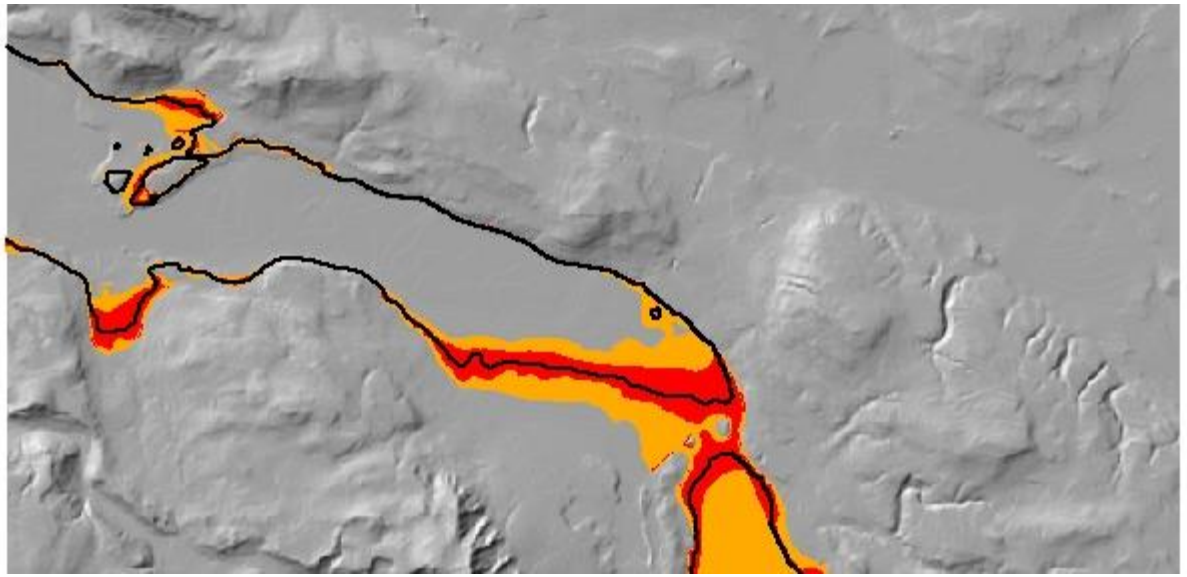
- Random errors can be quantified by repeated measurements and comparison with some "truth"
- Classic measure is the Root Mean Square Error (RMSE)
- Easy to remember what this is -> the **Root** of the **Mean** of the **Squared Error**

$$\sqrt{\frac{\Sigma(e_i - e_r)^2}{n}}$$

- If we make repeated comparisons, then we also know something about the error distribution
- Classic **Gaussian distribution** implies **68%** of errors between **±1 RMSE** and **95%** between **±2 RMSE**

Using the RMSE



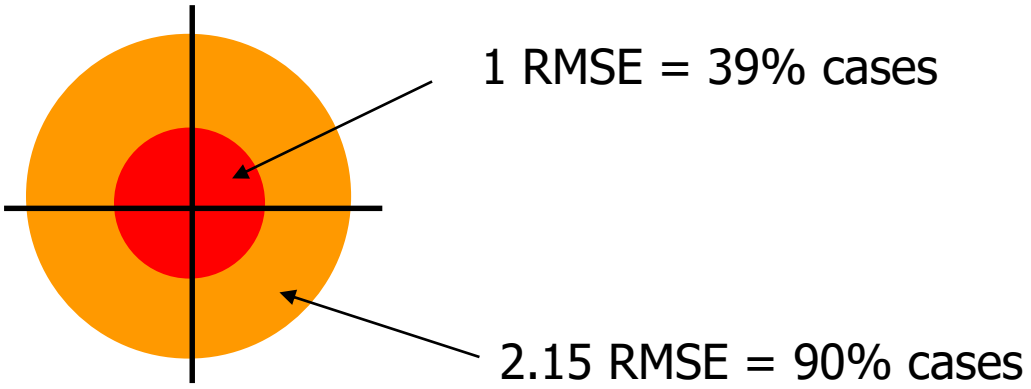


- SwissTopo say for their 25m DEM "**Die Genauigkeit der Höhenangaben beträgt im Mittelland rund 1,5 m und in den Alpen etwa 5 bis 8m**"
- We assume this is an RMSE with a random (Gaussian) distribution and calculate the following map:
  - **Black line:** 408m contour derived from DHM25
  - **Red area** is ±1.5m (68% probability our contour lies within)
  - **Orange area** is ±3m (95% probability our contour lies within)
- For what **applications** has this **implications?**
- For what **applications** has it **no real implications?**

An error model for position

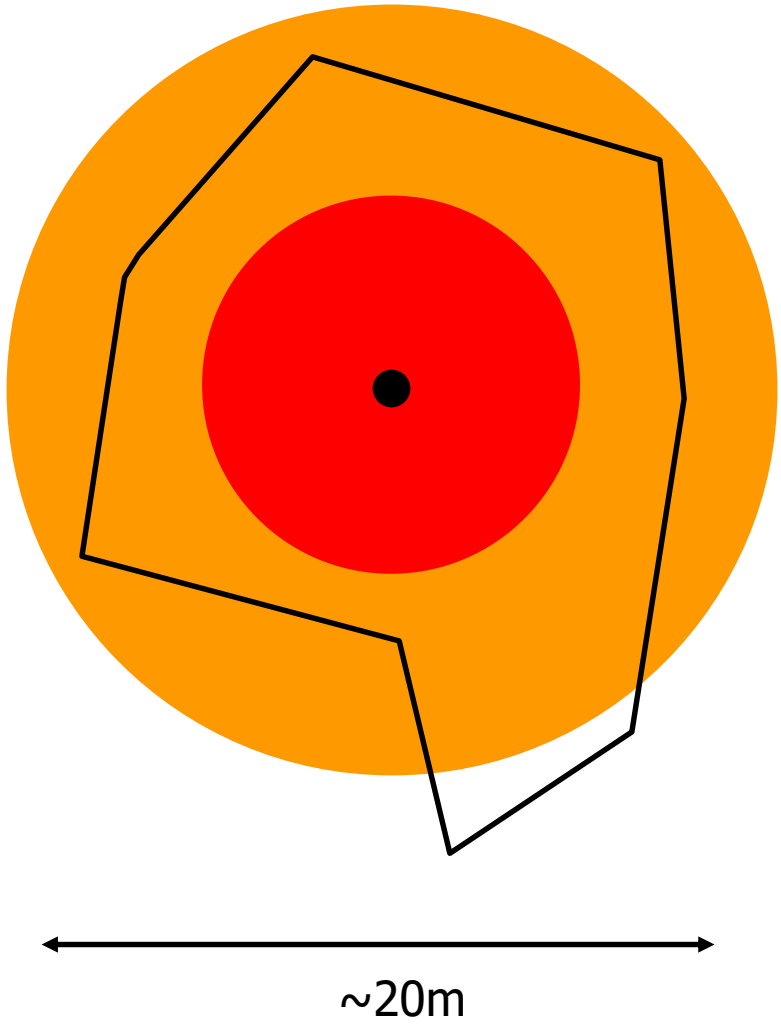
- Errors in position take the form (in 2D) $x \pm \delta x$, $y \pm \delta y$
- For a given RMSE (say from a GPS position) we can describe the **Circular Standard Error (CSE)** using a **bivariate Gaussian distribution**

Bivariate distribution is less "optimisitic" than a univariate (c.f. our height values)



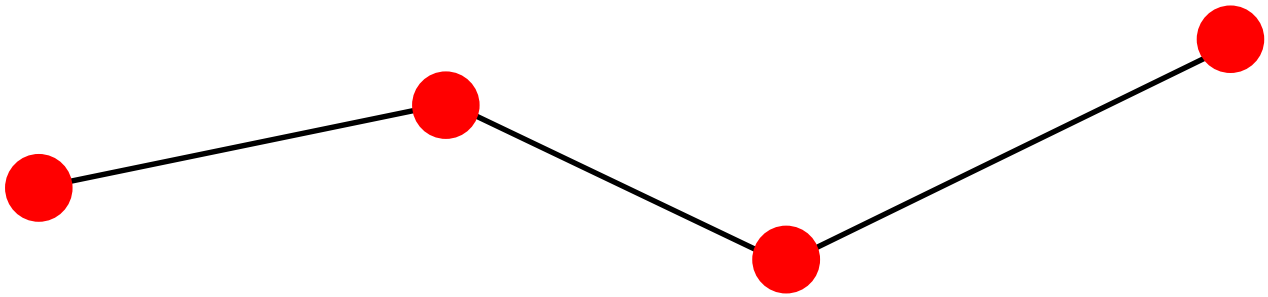
Using Circular Standard Error

- Testing for "**point in polygon**" -> e.g. Rabbit in a clearing
- Rabbit wearing a **GPS** – RMSE≈5m[1]
- **Red circle** shows **p=39%**
- Orange circle shows **p=90%**
- 0.9 * $A_{outside}/A_{total}$ is the probability of a point at the centre of the polygon actually lying outside
- **Assuming polygon boundaries certain**



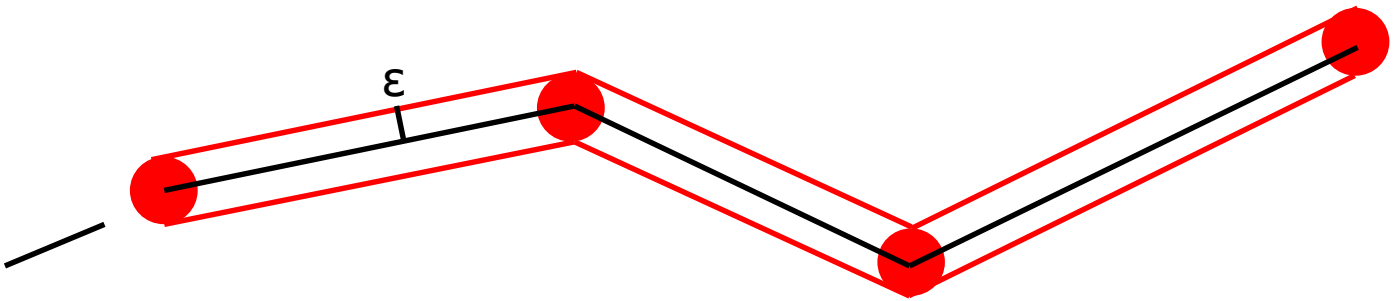
Modelling errors for lines

- **Lines in GIS** are represented by **collections of points (vertices)**
- How do the errors in point positions affect the accuracy of the line?
- Trivial representation – each vertex has an associated **circular standard error (CSE)**



Errors for lines: Epsilon or Perkal bands

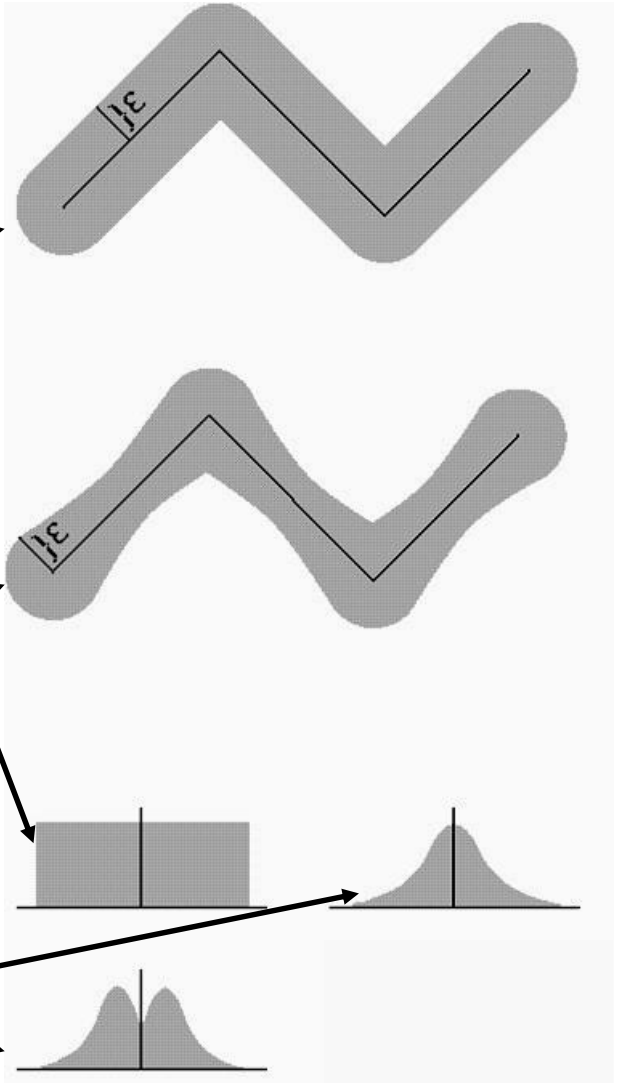
- Assumption: Error in line is a **function of CSE**
- Error represented by a **constant error band**, with width ε equivalent to CSE
- What does this error band **represent** (worst/ best case)?
- Why – and what other representation would be possible?



N.B: Ends are not square!

Form of Epsilon band

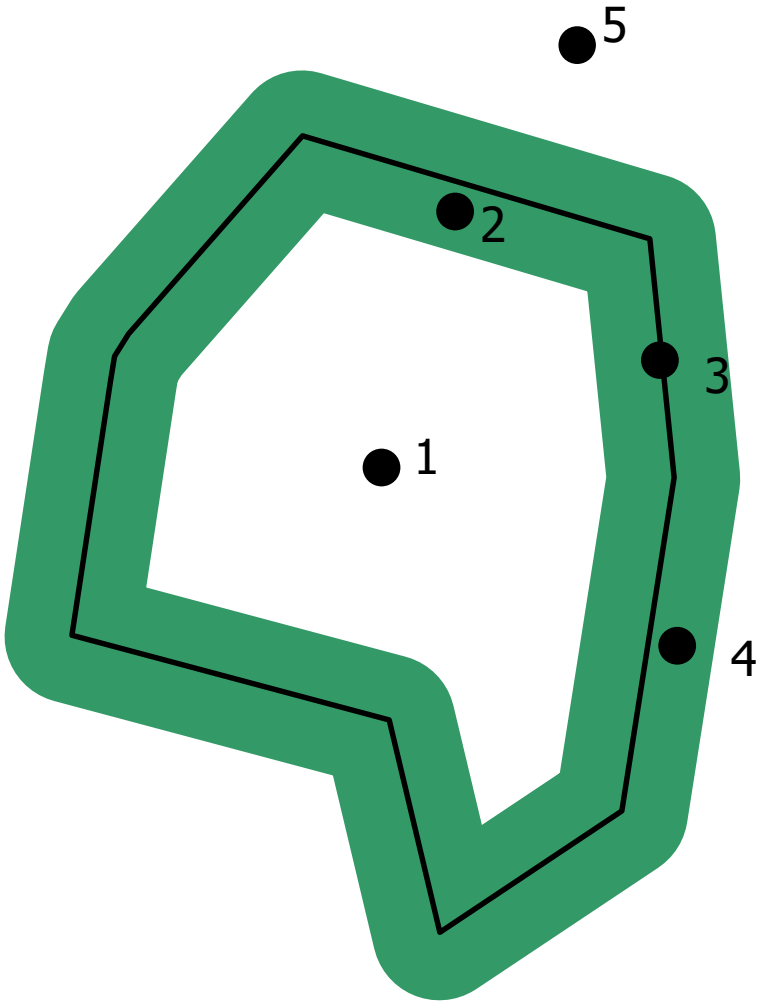
- Original:
  - Constant around the digitised line
  - Constant distribution
- Better (?):
  - Form more 'dumb-bell' like
  - **Distribution** in cross-section **not constant**: normal or even bi-modal (why?)



(Veregin 1999)

Point in polygon uncertainty

- We can apply our **epsilon bands** to **polygons**
- Blakemore (1984) looked at what this implied for **point in polygon** (where the point had **no** uncertainty) and identified the **following cases**:
  - 1: Definitely inside
  - 2: Probably inside, but could be outside
  - 3: On the boundary - indefinite
  - 4: Probably outside but could be inside
  - 5: Definitely outside



Modelling nominal attribute uncertainty: confusion matrix

	Agricultural	Parkland	Building	Transport	Total
Agricultural	10	15	1	1	27
Parkland	3	17	2	1	23
Building	1	2	40	10	53
Transport	2	1	8	30	41
Total	16	35	51	42	144

Rows show **classification**

Columns show **ground truth**

We can **compare** the classification for every case

Using the confusion matrix

- The **diagonals** give the number of **correctly classified parcels**
- Other cells show **disagreement** (confusion)
- Normally, **some classes** are more **easily confused** than others (i.e. disagreement is **not random**)
- For example here:
  - 51 cells with buildings in ground truth – 40 are correct
  - 27 cells with agriculture in our database – in reality only 16, and only 10 are correct in database
  - Parkland is incorrectly classified as agriculture 15 times

Error measures for the confusion matrix

- **Overall accuracy**: Diagonal/ Total (97/144)
- **Error of omission** – the proportion of values in reality, which were interpreted as something else: Sum of column's non-diagonal elements/ column total
  - e.g.: For parkland 18/35 parcels were omitted
- **Error of commission** – proportion of values which were in reality found to belong to another class: Sum of row's non-diagonal elements/ row total
  - e.g.: For parkland 6/23 parcels were falsely assigned to another class
- How could you **use this information to improve** the accuracy of the **data collection**?

What's wrong with accuracy?

- **Overall accuracy** is not really a very useful measure…
- …because it doesn't take account of the fact that a **random classification** will have accuracy > 0
- The **Kappa Index** (also sometimes called Cohen's Kappa) takes this into account by including an **estimation of agreement due to chance**…

$$\kappa = \frac{\sum_{i=1}^{n} c_{ii} - \sum_{i=1}^{n} c_{i.} c_{.i} / c_{..}}{c_{..} - \sum_{i=1}^{n} c_{i.} c_{.i} / c_{..}}$$

where c_{ii} is the value on the diagonal on the ith row/column;

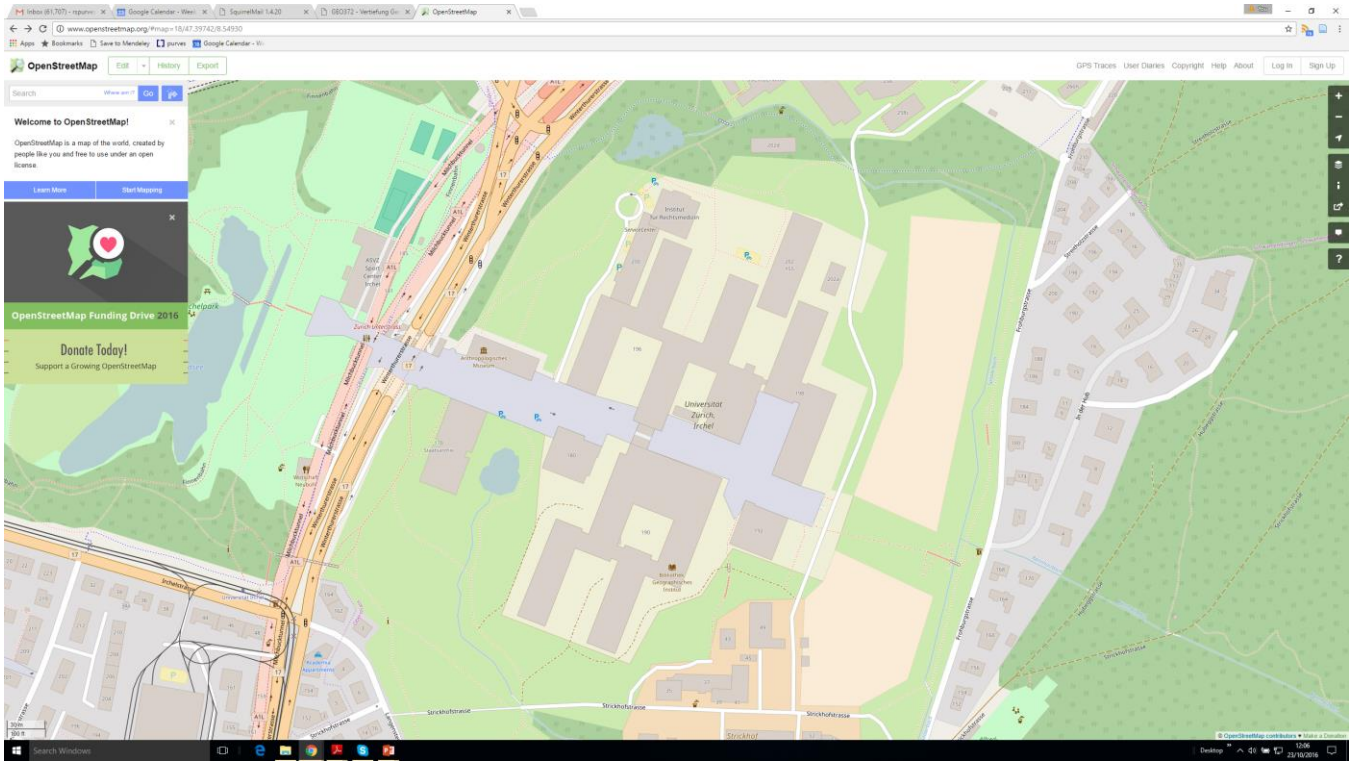
$c_{i.}$ is the sum of row i;

$c_{.i}$ is the sum of column i; and

$c_{..}$ is the overall sum.

Note that $c_{i.} c_{.i} / c_{..}$ is the probability of an entry being true due to chance

OpenStreetMap



- OpenStreetMap was **founded** by Steve Coast in **2004**
- The project's aim was to **create** and **provide free geographic data**
- OSM is *the* example of **Volunteered Geographic Information** – today we will use it to explore some issues of data quality

Some OSM basics

According to OSM (http://www.openstreetmap.org/about)

"Local Knowledge OpenStreetMap emphasizes local knowledge. Contributors use aerial imagery, GPS devices, and low-tech field maps to verify that OSM is accurate and up to date.

Community Driven OpenStreetMap's community is diverse, passionate, and growing every day. Our contributors include enthusiast mappers, GIS professionals, engineers running the OSM servers, humanitarians mapping disaster-affected areas, and many more. To learn more about the community, see the user diaries, community blogs, and the OSM Foundation website.

Open Data OpenStreetMap is *open data*: you are free to use it for any purpose as long as you credit OpenStreetMap and its contributors. If you alter or build upon the data in certain ways, you may distribute the result only under the same licence. See the Copyright and License page for details."

OSM and research

- OSM has been the subject of **lots of research** in GIScience
- Much of this has focussed on **data quality**
- Because of it's nature (collected by **volunteers** with **differing skills** and **motivations**) OSM has very **different properties to traditional data**
- Much data quality research has **compared it to traditional data** (e.g. as produced by SwissTopo, Ordnance Survey, etc.)
- Remember OSM is a **global product**, whose **data quality varies in space**

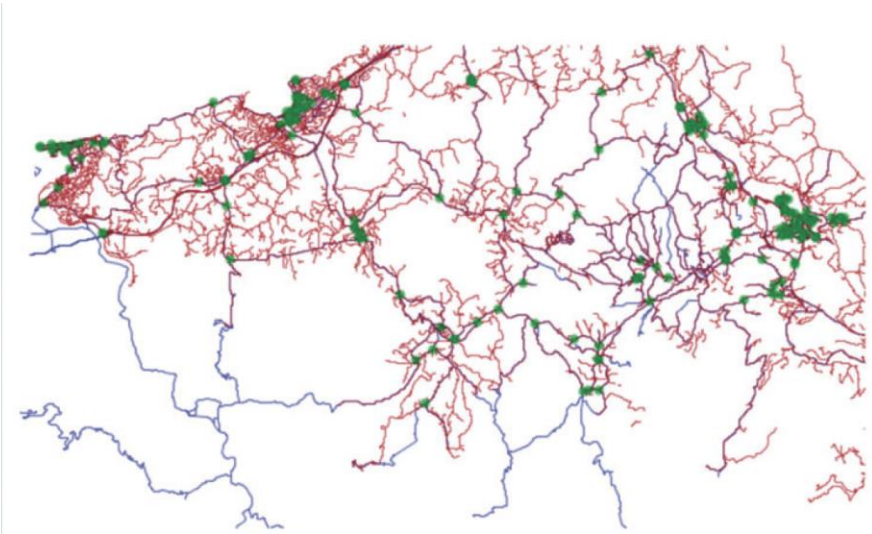
OSM and the SDTS

We're going to use OSM and look at each aspect with varying amounts of detail:

- **Positional accuracy** Comparing motorways around London and road junctions in France (Haklay 2010), (Girres & Touya 2010)
- **Attribute accuracy** Land cover in Portugal (Estima & Painho 2013)
- **Logical consistency** Road networks, Borders and coastline in France (Girres & Touya 2010)
- **Completeness** Road length in England (not UK) (Haklay 2010)
- **Lineage** Gender and OSM contributors (Stephens 2013)

Positional accuracy in OSM

- Girres & Touya **compared crossroads** (treated as point objects)
- They **matched 207 pairs** of points by hand (thus effectively ignoring blunders)
- RMSE in position of crossroads was **4.54m**



Source: Girres & Touya 2010

Note point pairs can only be compared where they can be matched – non-trivial if data are not sufficiently similar

Positional accuracy in OSM

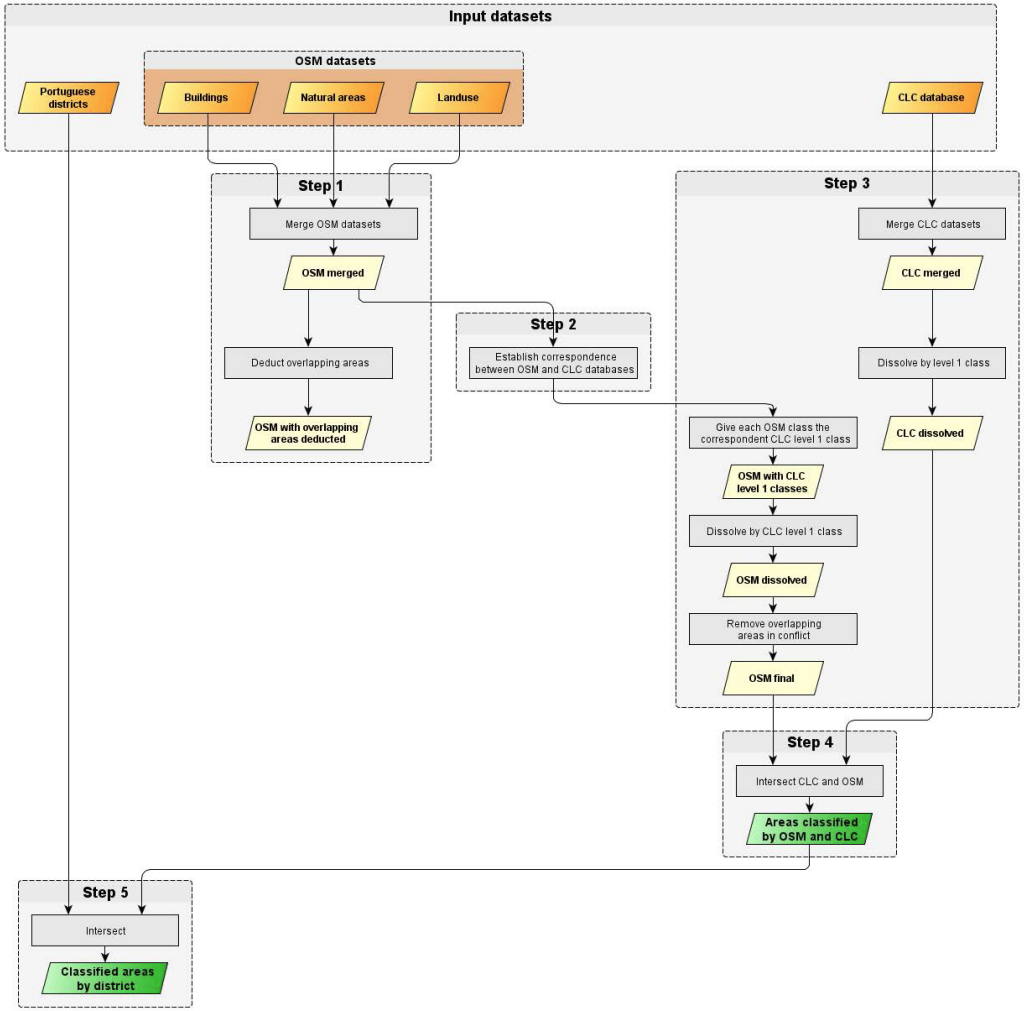
- Haklay compared **administrative data** for motorways (autobahn) around London because **motorways are significant**, and their mapping appeared to be **complete** in OSM
- He used **Epsilon bands**, modelled as 20m buffers for the administrative data and 1m (why?) for OSM and calculated **percentage overlap**

Motorway	Percentage
M1	87.36
M2	59.81
M3	71.40
M4	84.09
M4 Spur	88.77
M10	64.05
M11	84.38
M20	87.18
M23	88.78
M25	88.80
M26	83.37
M40	72.78
A1 (M)	85.70
A308 (M)	78.27
A329 (M)	72.11
A404	76.65

Source: Haklay 2010

Attribute accuracy and OSM

- Estima and Painho explored **attribute accuracy** in OSM
- They compared **land use/ land cover** as assigned to OSM polygons with a **European dataset** (CORINE)
- They **documented** the process nicely, though the results need to be **treated carefully**



Source: Estima and Painho (2013)

Attribute accuracy and OSM

Key steps

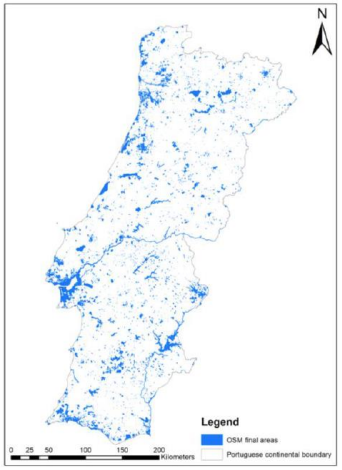
- Extracting land cover/ land use data from OSM (only covers 3.24% of Portugal) – **no planar enforcement!**
- Linking **OSM classes** to **top-level CORINE classes** (artificial surfaces, agricultural areas, forest and semi-natural areas, wetlands, water bodies)
- **Confusion matrix** created (and stratified spatially)

Important results

Confusion matrix lets us explore issues – overall accuracy (76.7%) not very interesting

		OSM classes					
		1	**2**	**3**	**4**	**5**	**Total**
CLC classes	**1**	**44160.56**	1059.00	4086.69	0.00	663.20	52369.87
	2	12934.72	**31884.28**	10716.09	4.94	12088.20	68459.87
	3	5182.27	1214.07	**83362.66**	0.07	6322.15	99843.05
	4	42.27	114.77	238.65	**59.57**	4402.91	4870.53
	5	87.66	37.81	132.53	0.00	**59145.14**	59433.67
	Total	62407.48	34309.93	98536.62	64.59	82621.61	284976.99

Data are very far from complete and very unevenly distributed across classes



Source: Estima and Painho (2013)

Completeness and OSM

- Haklay explored **completeness** using road **length**
- He compared two datasets for the whole of England, and calculated **length of the road network** at a resolution of 1km
- He hypothesised that because **administrative data** were generalised, that the **length of OSM roads** would be **greater**
- Black areas indicate cells where this is the case (good coverage), grey (poor coverage) and white coverage the same (any comments on these areas?)
- In this study, **24.5%** of the total area had **good coverage**

Source: Haklay 2010

Logical consistency and OSM

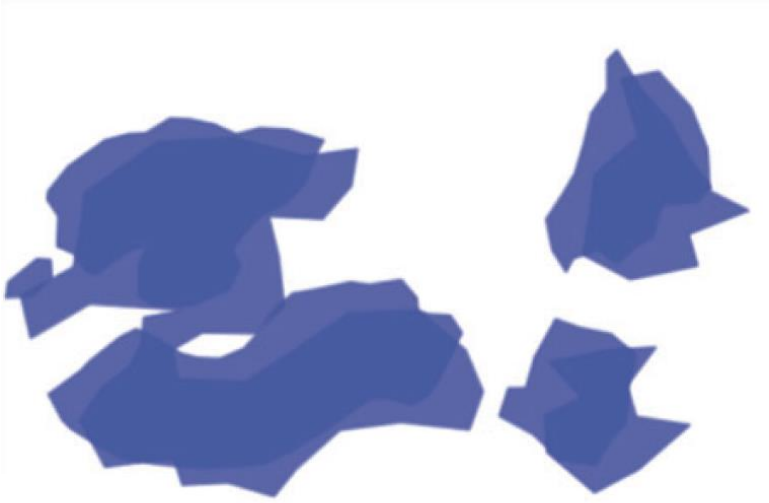
- Logical consistency is **challenging** to **quantify**
- Girres and Touya gave some examples including:
  - **Broken networks**
  - Duplicated objects at the same location
  - Inconsistency between features (recall sliver polygons)

Roads not connected at intersection

Source: Girres & Touya 2010

Logical consistency and OSM

- Logical consistency is **challenging** to **quantify**
- Girres and Touya gave some examples including:
  - Broken networks
  - **Duplicated objects at the same location**
  - Inconsistency between features (recall sliver polygons)

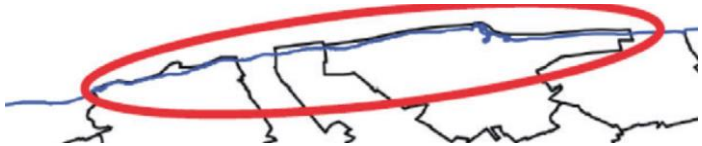


Lakes added to database multiple times but not treated as the same object

Source: Girres & Touya 2010

Logical consistency and OSM

- Logical consistency is **challenging** to **quantify**
- Girres and Touya gave some examples including:
  - Broken networks
  - Duplicated objects at the same location
  - **Inconsistency between features (recall sliver polygons)**



Note how the coastline and administrative borders don't overlap, although they are the same features

Source: Girres & Touya 2010

Logical consistency and OSM

- These problems are **not specific to OSM**
- However, because of the methods of data collection they were (are) **sometimes more likely**
- I**f detectable** (e.g. our road intersections) they may be **easy to solve** through post-processing
- OSM data has a **simple structure** which makes adding data easy, but increases the likelihood of some problems (e.g. the **lack of planar enforcement**)

OSM and lineage

- Stephens (2013) carried out a survey on OSM
- 23.5% of females and 61.6% of males had heard of OSM
- **20.8%** of females had contributed and **40.1%** of males
- She argued that **gender** therefore **influences the contribution** of information to OSM
- She illustrates this by exploring discussions about **amenities proposed as OSM features** (classes represented)
- She argues that features which might be characterised as **more male** (e.g. linked to prostitution) were subject to much **more debate and differentiation** than those characterised as **female** (e.g. childcare facilities)

Fitness for use

- Reporting on uncertainty alone is not enough…
- …we have to think about the **implications** for our particular application and decide what level of uncertainty is acceptable
  - Data used in **car navigation systems** must have very high **positional and attribute accuracy, be logically consistent and complete**
  - If we wish to measure **change over time**, **completeness** is very important
- In some application domains uncertainty is deliberately added:
  - **Census data** is used to identify trends and explore patterns –often data are processed to make it impossible to identify individuals
  - **Positional data** locating your telephone may be degraded (obfuscation)

Summary

- We've looked at different sources and types of error
- We've introduced simple error models for both numerical and nominal data
- I've illustrated these models with some examples – next week we will look at how these can influence products of spatial data
- We've illustrated all aspects of spatial data quality using OSM

Key references

Burrough, P.A. et al., (2015): *Principles of Geographical Information Systems*. Third Edition. Oxford University Press. Jones, C.B. (1997): *Geographical Information Systems and Computer Cartography*. Longman. (Chapter 7)

Longley, PA. et al. (2015): *Geographic Information Systems and Science*. Second Edition. (Chapter 5)

Girres, J. F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, *14*(4), 435-459.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets.*Environment and planning B: Planning and design*, *37*(4), 682-703.

Stephens, M. (2013). Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, *78*(6), 981-996.

Estima, J., & Painho, M. (2013, November). Exploratory analysis of OpenStreetMap for land use classification. In *Proceedings of the second ACM SIGSPATIAL international workshop on crowdsourced and volunteered geographic information* (pp. 39-46). ACM.

Example of the Kappa Index

	Forest on ground	Water on ground	Row total ($C_{i.}$)
Forest in DB	1000	100	*1100*
Water in DB	200	700	*900*
Column total ($C_{.i}$)	*1200*	*800*	**2000**

$$\kappa = [(1000 + 700) - ((1200*1100/2000)+ (800 * 900/ 2000))] / [2000 - ((1200*1100/2000)+ (800 * 900/ 2000))]$$

$$= 0.69$$

For comparison: Overall Accuracy = 0.85

Exercise

- For the grid given, **classify** each cell as either:
  - Housing
  - Forest
  - Educational
  - Transport
- Use the **dominance** or **central point principal**
- Then, with a colleague, create a confusion matrix

Source: Google

Confusion matrix

	Housing	Forest	Educational	Transport	Total
Housing					
Forest					
Educational					
Transport					
Total					20

Calculate the measures we discussed…