Geography 4203 / 5203

GIS Modeling

Class 13: Uncertainty in „Source Data“

Some Updates

Last Lecture

- We finished the **conceptual** part of uncertainty and spatial data quality
- You have seen some examples where uncertainty lead to a lack of the **fitness** of the data **for the intended use** (hydro-charts, bogs, forest in historical maps)
- We talked about general aspects of SDQ and we discussed some first **definitions** of **uncertainty/SDQ** together with some **examples**
- We started with looking at errror models for source data such as CSE, Perkal band

Today's Outline

- We will continue with error models and uncertainty assessment
- After looking at measurable errors in position (or ratio-scaled attributes) and **methodological** aspects how to assess these errors we will talk about **categorical/nominal** data that rather fit the perspective of **raster-based modeling** in a GIS
- We will go through the error table/confusion matrix and discuss some of the summary statistics available and where the limitations of using confusion matrices are
- You will see some examples of how to overcome these limitations

Learning Objectives

- You will understand the terms, concepts and meanings regarding **uncertainty** and **spatial data quality**
- You will be able to explain the differences between **error**, **vagueness**, **ambiguity** and what the **elements** of **SDQ** are
- You will know what the **SDTS** is and what stands behind the **famous five points**
- Finally you will be able to **explain** and **use** simple **error models** for **positional** and **attribute accuracy** (circular standard error, epsilon bands, confusion matrices)

Let’s look at some Error Models

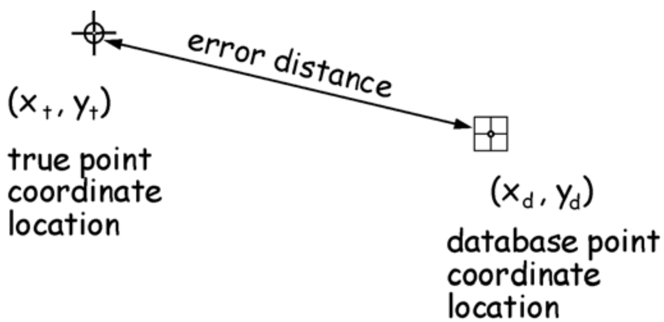
- **Fit for the intended use?** (we have seen 3 examples where they were not)
- Remember the **definitions** we have seen and the “**diversity**” of **conceptual** perspectives
- We will start with **error assessment** as the simplest set of methods available (“**truth**”?)
- **Interval/ratio values**: Positional & attribute accuracy (RMSE, CSE, Perkal)
- **Nominal/ordinal**: Attribute accuracy (Confusion matrices)

How Dependent and Systematic are my Errors?

- … for **positional** and **attribute** uncertainty
- Land cover map -> change in land cover type **moves boundary**
- Chloropleth map -> **Administrative** boundary (position) predetermined - boundary won’t change because of a change in an attribute
- For many classes, the class (attribute) is **predetermined** e.g. street names - class doesn’t change because of positional uncertainty
- Systematic errors follow a **pattern** (**constant** or **systematically varying**) and are easy to correct

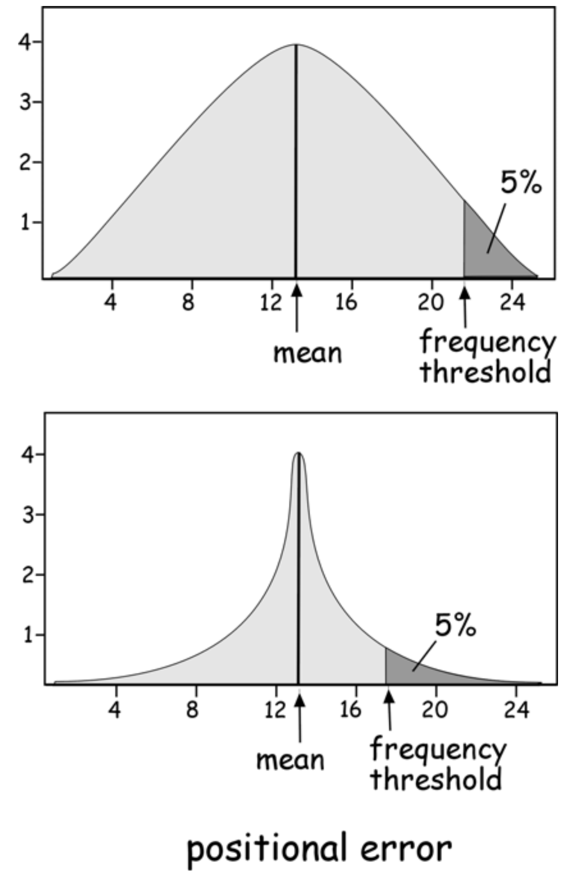
Random Errors in Points

For positions and attributes: RMSE: **Root** of the **Mean** of the **Squared Error**…!



$$\text{error distance} = \sqrt{(x_t - x_d)^2 + (y_t - y_d)^2}$$

$$RMSE = \sqrt{\frac{e_1^2 + e_2^2 + \ldots + e_n^2}{n}\partial}$$

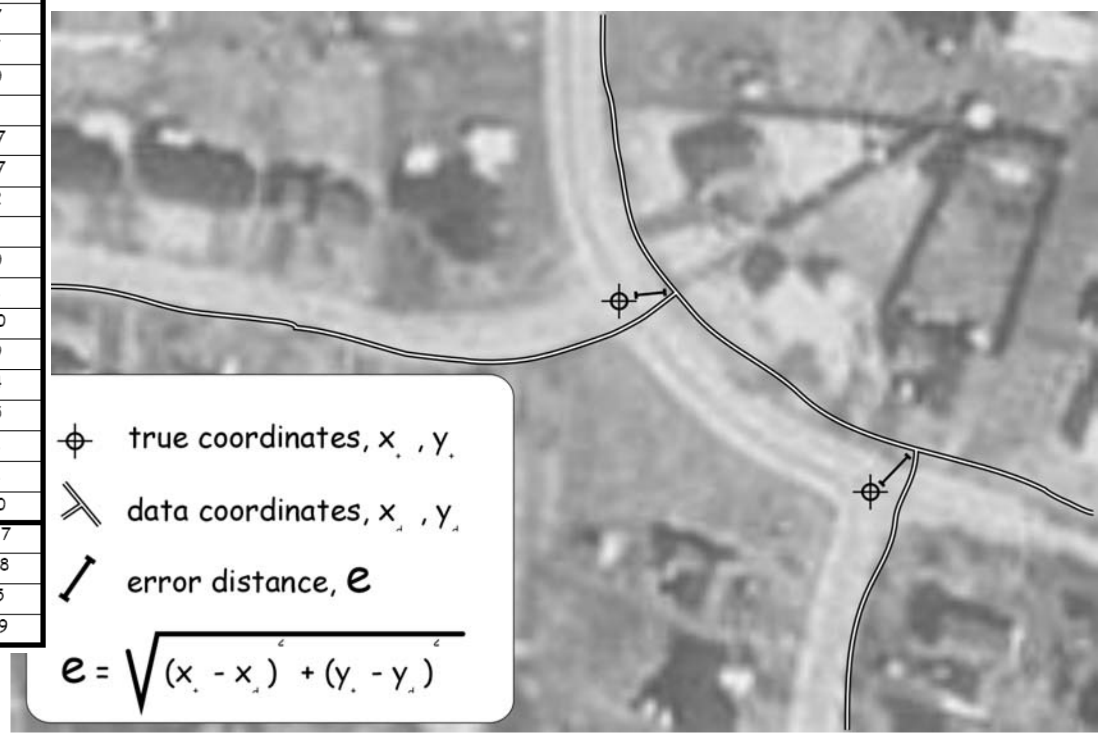


What is the difference between RMSE and standard deviation?

Error Distributions

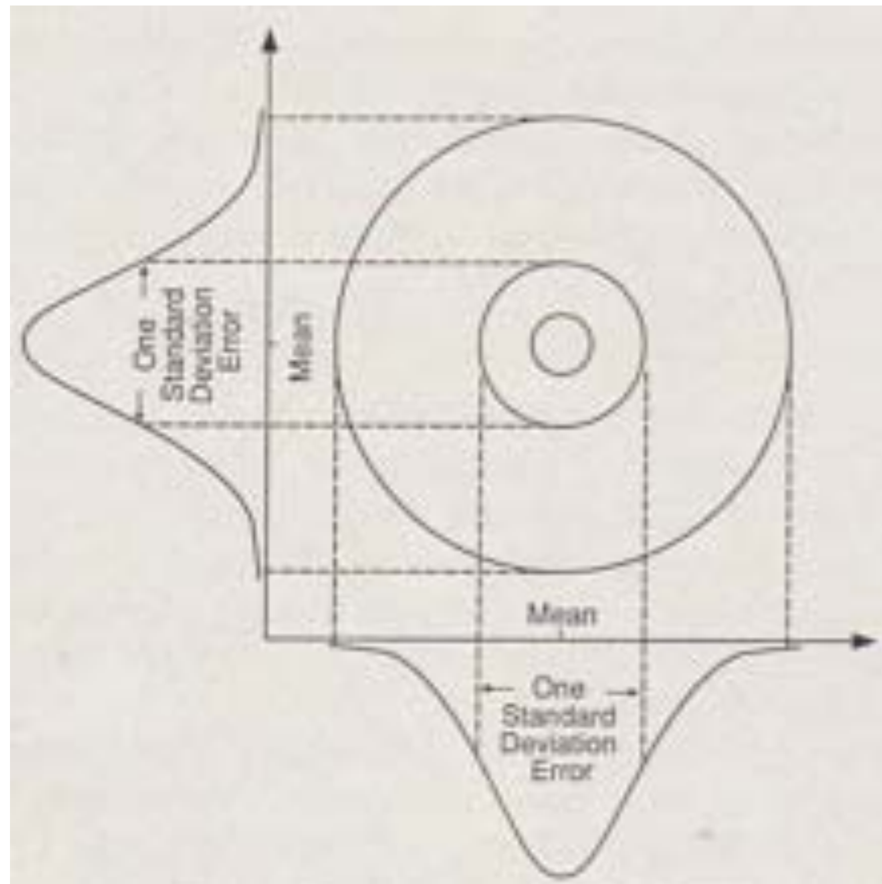
- No information on error **distribution** using RMSE (**Gaussian** often used because it is easy)
- Assumption that errors are **randomly** distributed… Why is this an **implication**???

ID	x (true)	x (data)	x difference	(x difference)^2	y (true)	y (data)	y difference	(y difference)^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



Positional Errors of Points

- **Circular Standard Error**
- Say: x ± δx, y ± δy
- Using assumptions of a **distribution** we can make **judgments** about the point set and its **accuracy**
- Guess if a rabbit's location can be assumed to be within a polygon…

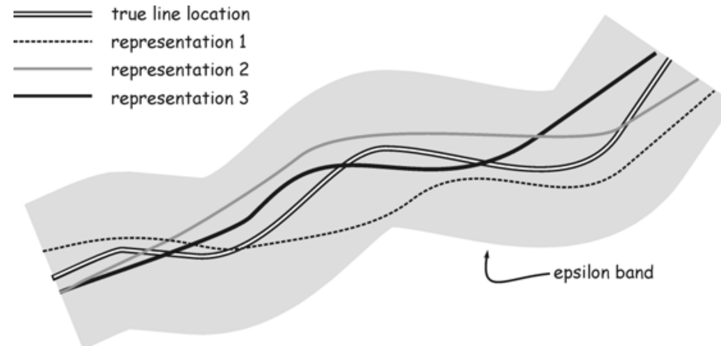


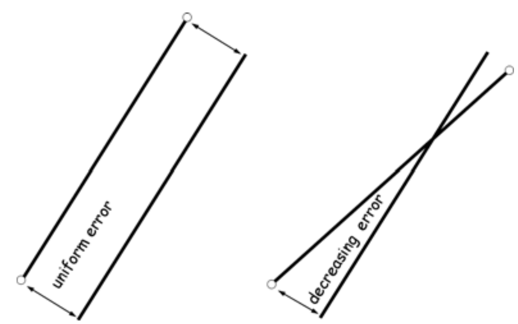
And what about Lines and Polygons?

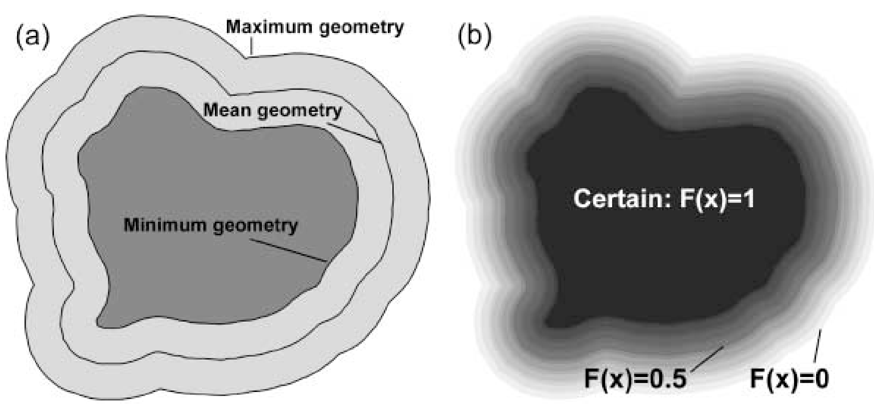
Epsilon or Perkal Bands

Extension of the CSE to lines (their vertices) to produe constant areas around the lines

Back to the rabbit-polygon example

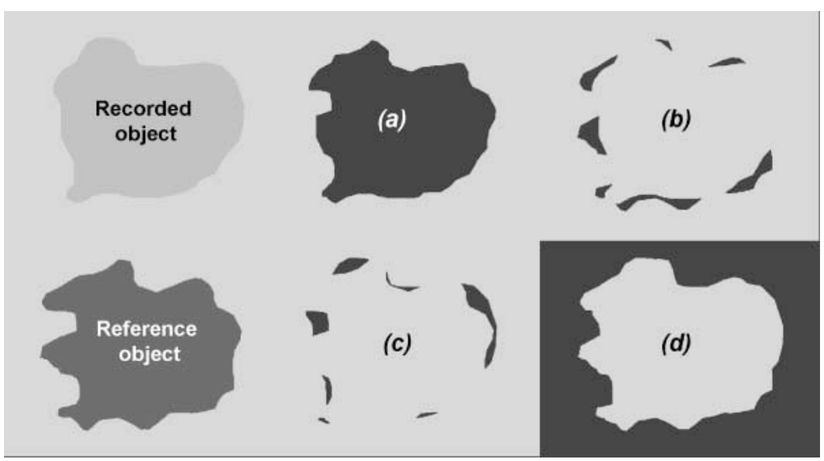






Categorical Data - Confusion Matrix

- What can go wrong in a classification?



true value

data layer attribute value

	wheat	corn	soy	alfalfa	grass	fallow	
wheat	14	4			4		22
corn	2	12		1	3		18
soy	1		18	2			21
alfalfa		3	2	16	1		23
grass	3	1		1	12		17
fallow						20	20
	20	20	20	20	20	20	92

$$\text{overall accuracy} = \frac{\text{sum of diagonal}}{\text{total number of samples}} = 92/120 = 76.7\%$$

Summary Statistics

- **Overall accuracy**: Diagonal / Total
- **Error of ommission (Producer's acc.)** : proportion of values in reality, which were interpreted as something else: Sum of column's non-diagonal elements / column total (e.g: corn 8/20 parcels were ommitted)
- **Error of commission (User's acc.)**: 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 corn 6/18 parcels were falsely assigned to another class

true value

data layer attribute value

	wheat	corn	soy	alfalfa	grass	fallow	
wheat	14	4			4		22
corn	2	12		1	3		18
soy	1		18	2			21
alfalfa		3	2	16	1		23
grass	3	1		1	12		17
fallow						20	20
	20	20	20	20	20	20	92

$$\text{overall accuracy} = \frac{\text{sum of diagonal}}{\text{total number of samples}} = 92/120 = 76.7\%$$

More Summary Statistics

- **PCC** does not take into account that a **random** classification will have an **accuracy > 0**
- **Cohen' Kappa** coefficient of agreement includes an **estimation** of **agreement** due to **chance**…

$$c_{i.}c_{.i} / c_{..}$$

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

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.

More Summary Statistics

Measure	*Calculation*
Prevalence	(a + c)/N
Overall diagnostic power	(b + d)/N
Correct classification rate	(a + d)/N
Sensitivity	a/(a + c)
Specificity	d/(b + d)
False positive rate	b/(b + d)
False negative rate	c/(a + c)
Positive predictive power (PPP)	a/(a + b)
Negative predictive power (NPP)	d/(c + d)
Misclassification rate	(b + c)/N
Odds-ratio	(ad)/(cb)
Kappa	[(a + d) − (((a + c)(a + b) + (b + d)(c + d))/N)]/[N − (((a + c)(a + b) + (b + d)(c + d))/N)]
NMI n(s)	[−a.ln(a)−b.ln(b)−c.ln(c)−d.ln(d)+(a+b).ln(a+b)+(c+d).ln(c+d)]/[N.ln N −((a+c).ln (a+c) + (b+d).ln(b+d))]

		Actual	
		+	-
Predicted	+	a	b
	-	c	d

Kappa Example

	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**

```
κ = [(1000 + 700) -
      ((1200*1100/2000)+ (800 * 900/ 2000))]
    /[2000 -
      ((1200*1100/2000)+ (800 * 900/ 2000))]

= 0.69
```

For comparison: Overall Accuracy = 0.85

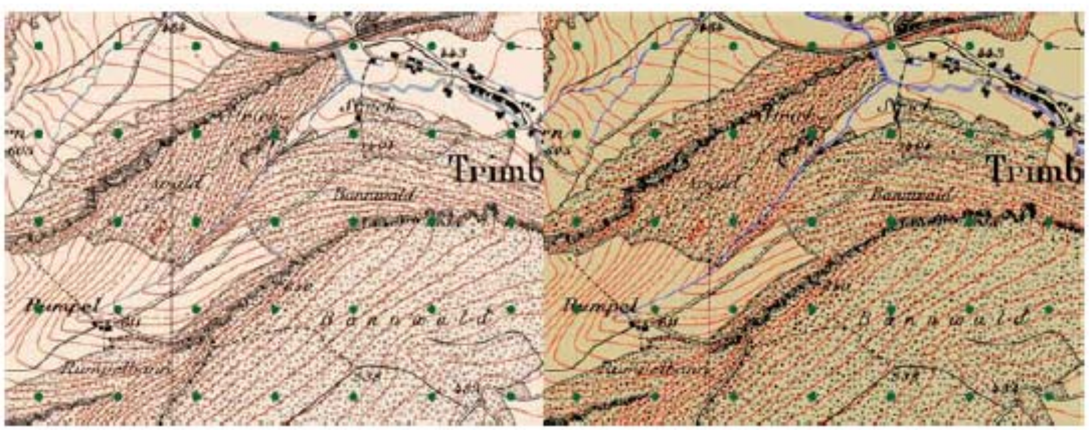
How Different look the Summary Statistics?

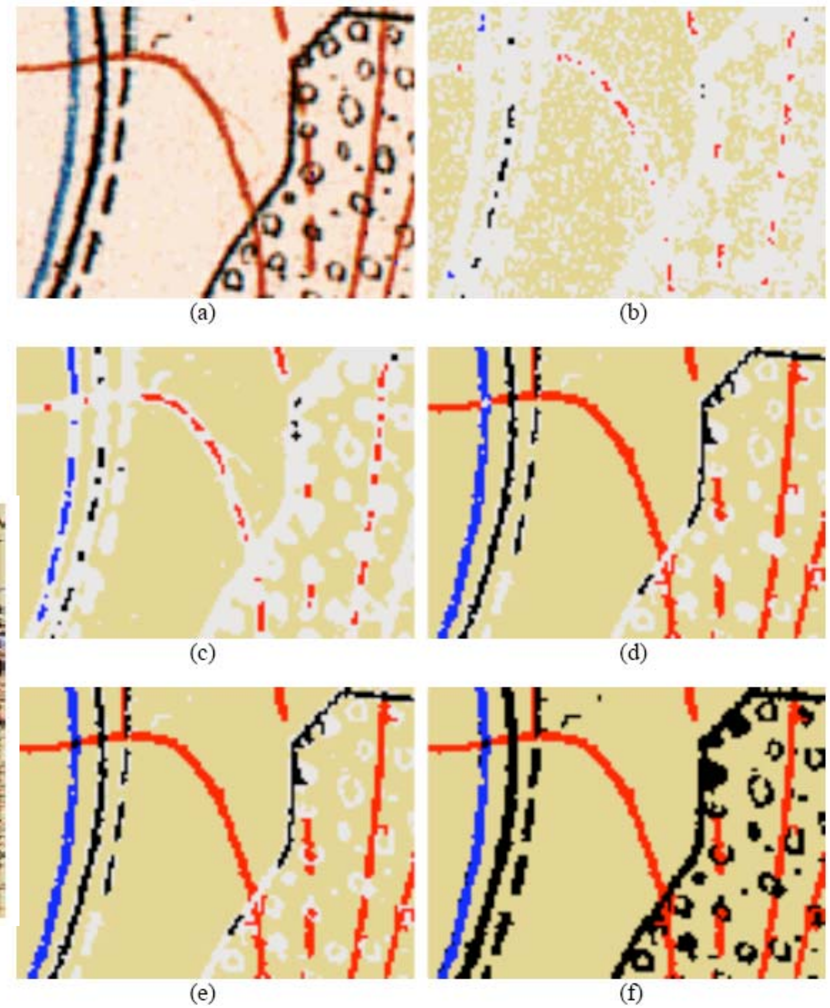
- How conservative?
- Chance agreement?
- Consideration of classes with low or high proportions (robustness)

	Reference map		Original map	
	Pontresina	St. Moritz	Pontresina	St. Moritz
Forest area (correct)	8350	8223	5853	6619
Non-forest area (correct)	6486	11496	5649	10016
Misclassified proportion	–	–	3334	3084
PCC	–	–	0.76	0.84
Kappa	–	–	0.55	0.67
NMI	–	–	0.26	0.36

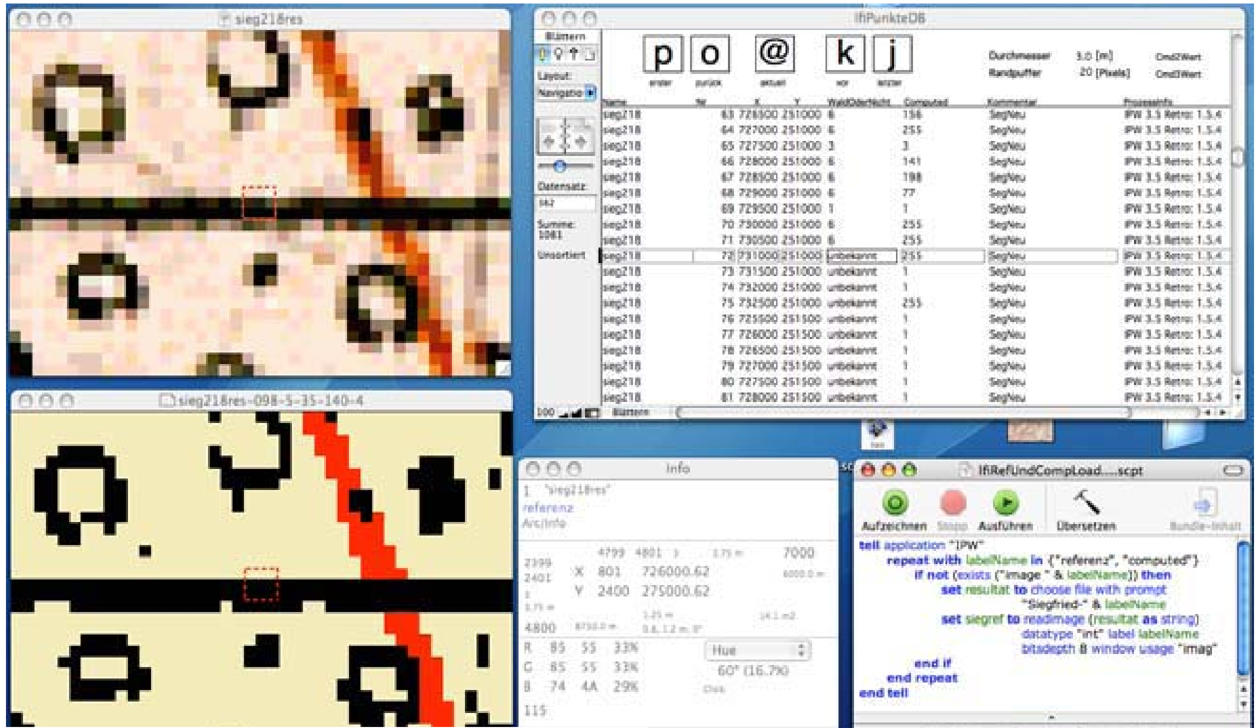
First Example - Simple Accuracy Assessment

- Image extraction result to be evaluated against human inspection efforts





First Example - Simple Accuracy Assessment



	Hydro (Blue)	Elevation (Red)	Black Layer	Background (White)	Global
Recall	0.76	0.91	0.97	0.97	-
Precision	0.80	0.92	0.93	0.99	-
ACC	-	-	-	-	**0.96**
Kappa	-	-	-	-	**0.93**
NMI	-	-	-	-	**0.81**

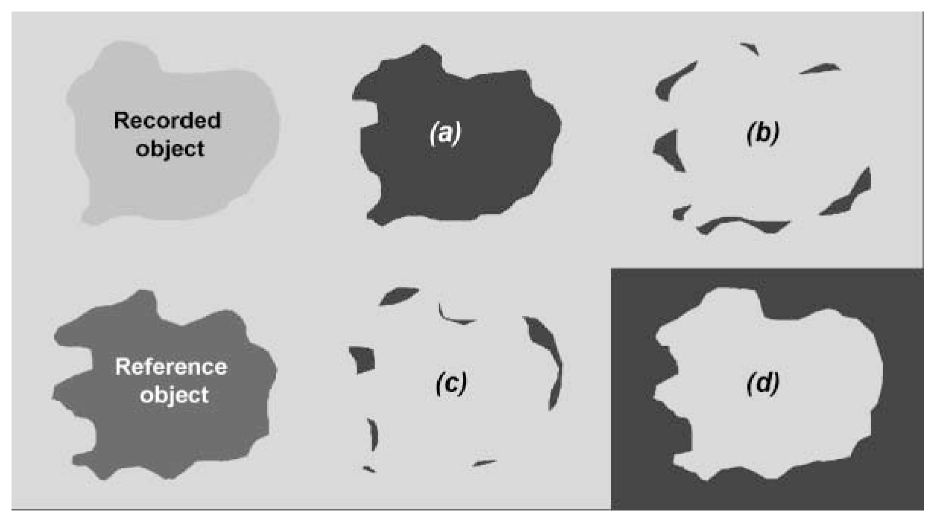
What is lacking with summary statistics?

true value

data layer attribute value

	wheat	corn	soy	alfalfa	grass	fallow	
wheat	14	4			4		22
corn	2	12		1	3		18
soy	1		18	2			21
alfalfa		3	2	16	1		23
grass	3	1		1	12		17
fallow						20	20
	20	20	20	20	20	20	92

$$\text{overall accuracy} = \frac{\text{sum of diagonal}}{\text{total number of samples}} = 92/120 = 76.7\%$$



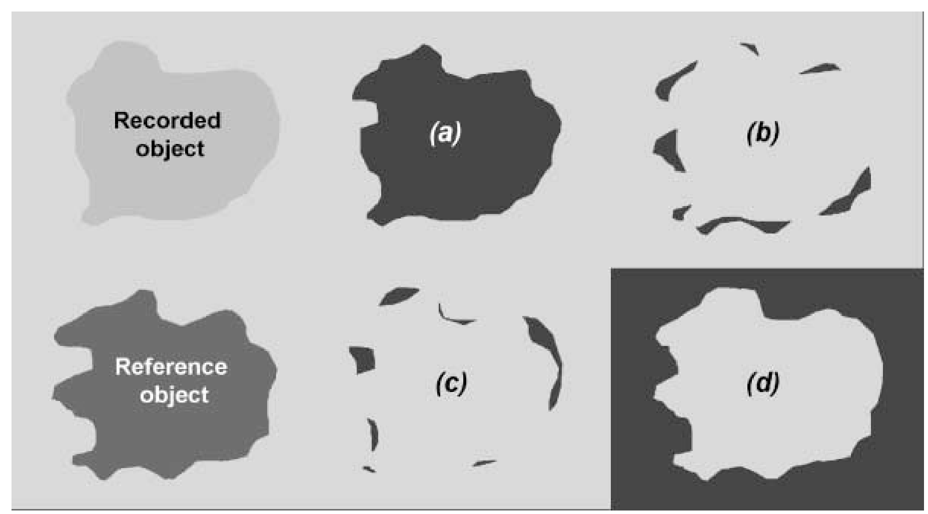
What is lacking with summary statistics?

- Spatial orientation?
- Judgments for the local unit/entity?
- Development of Geographical weighting, local summary statistics based on window operations

true value

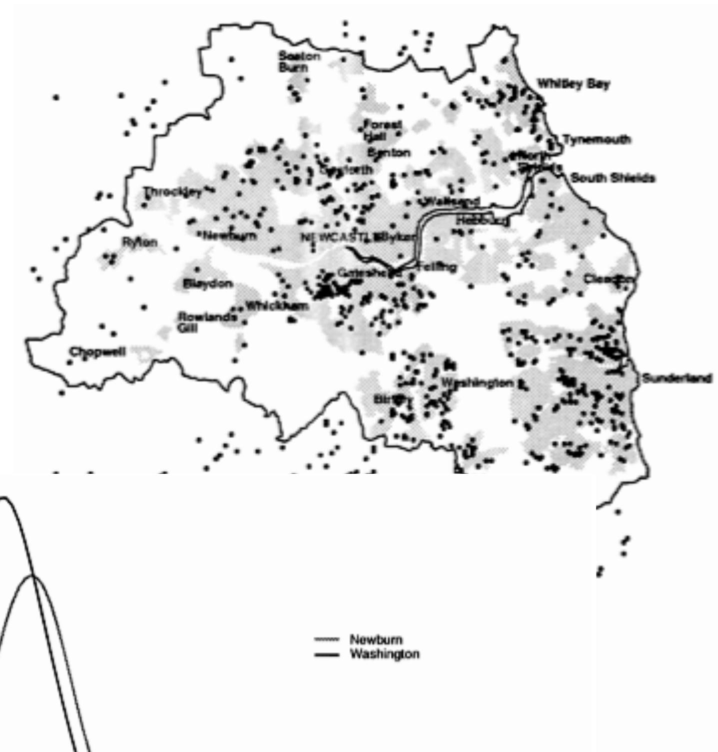
data layer attribute value	wheat	corn	soy	alfalfa	grass	fallow	
wheat	14	4			4		22
corn	2	12		1	3		18
soy	1		18	2			21
alfalfa		3	2	16	1		23
grass	3	1		1	12		17
fallow						20	20
	20	20	20	20	20	20	92

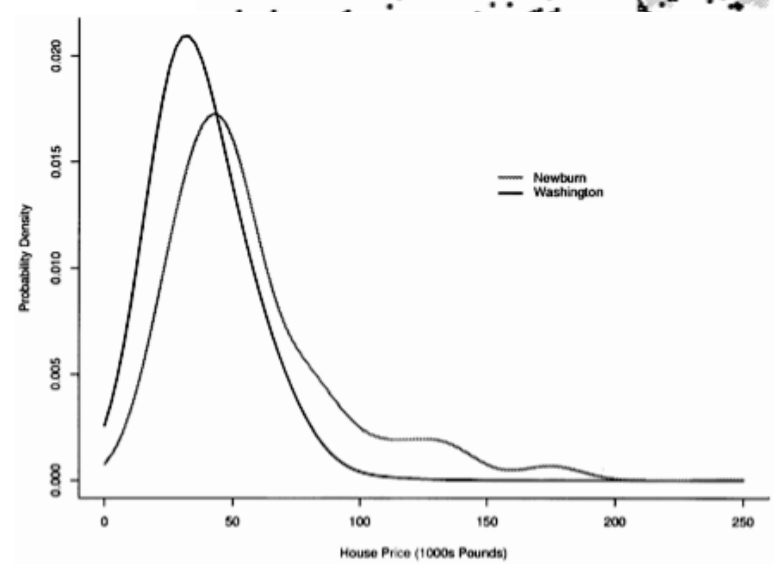
$$\text{overall accuracy} = \frac{\text{sum of diagonal}}{\text{total number of samples}} = 92/120 = 76.7\%$$



Example Geogr. Weighted Summary Statistics for House Prices

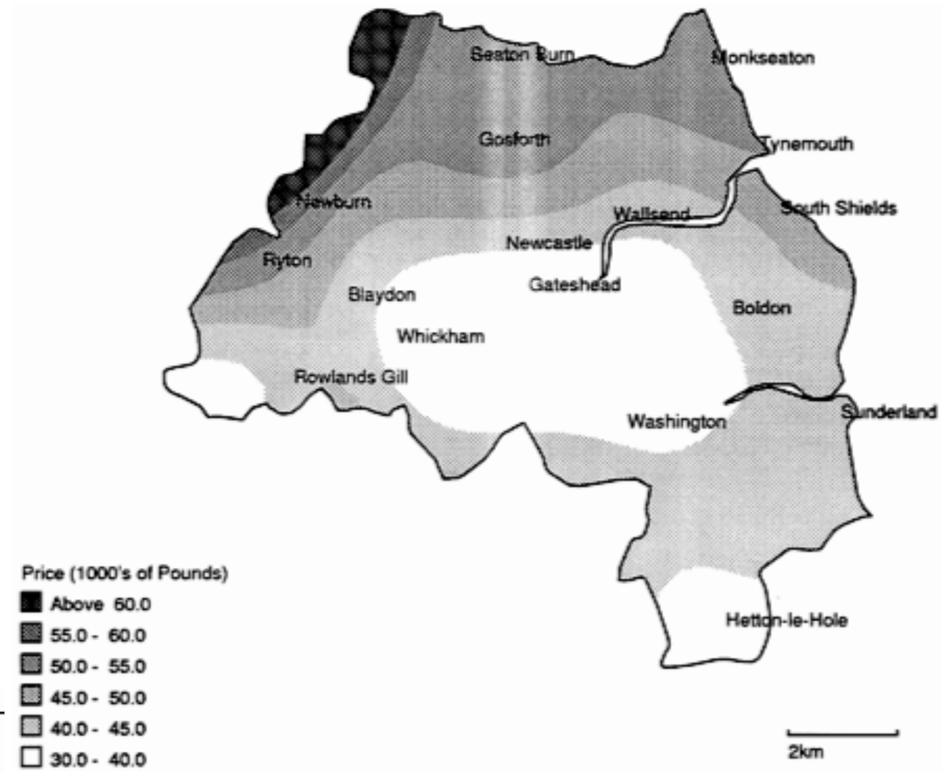
- Brundson & Fotheringham (2002)
- Two counties (Newburn and Washington) with characteristic “landscapes of housing prices”
- Of interest is how prices are different within the neighborhood and thus compared to aggregated data of prices

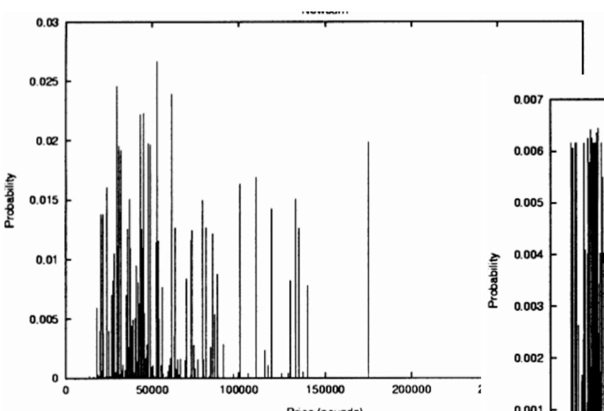


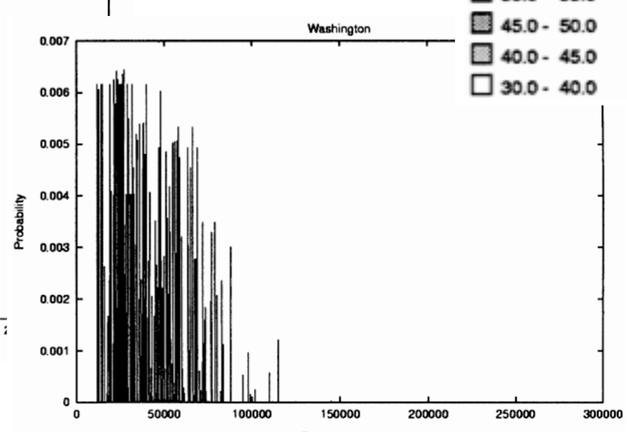


Local Summary Statistics

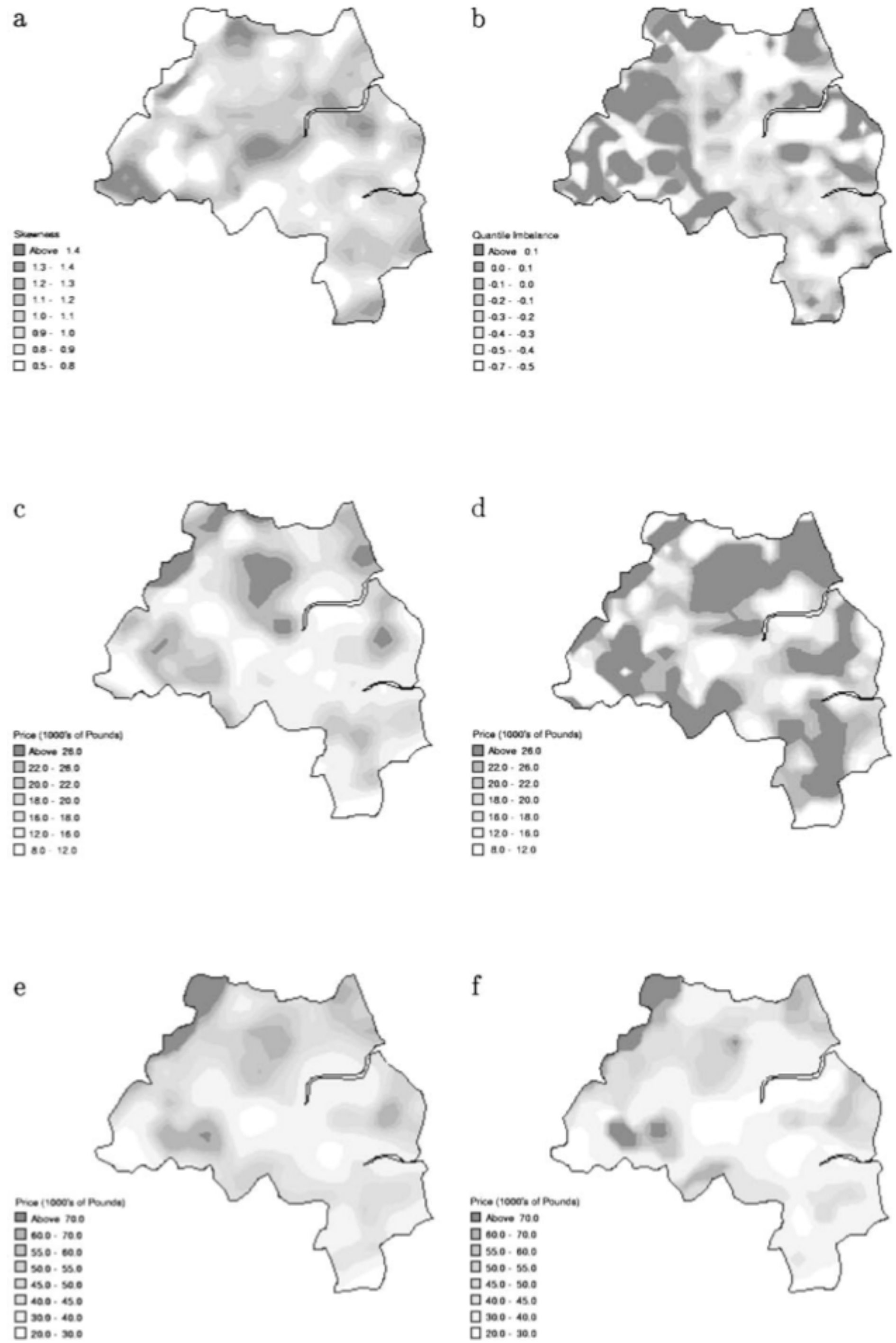
- Rough estimates for housing price trends using large Kernels to assess local statistics
- How about local variation?





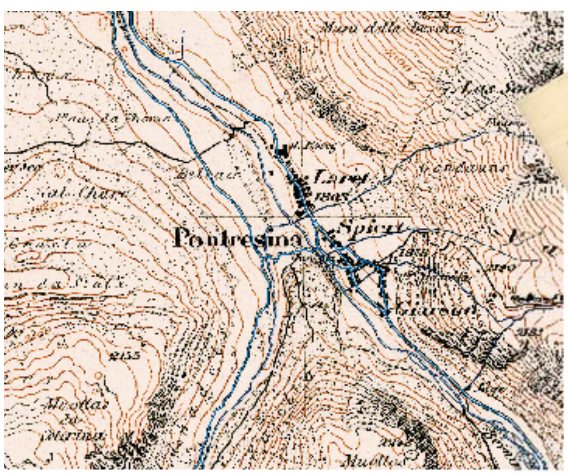


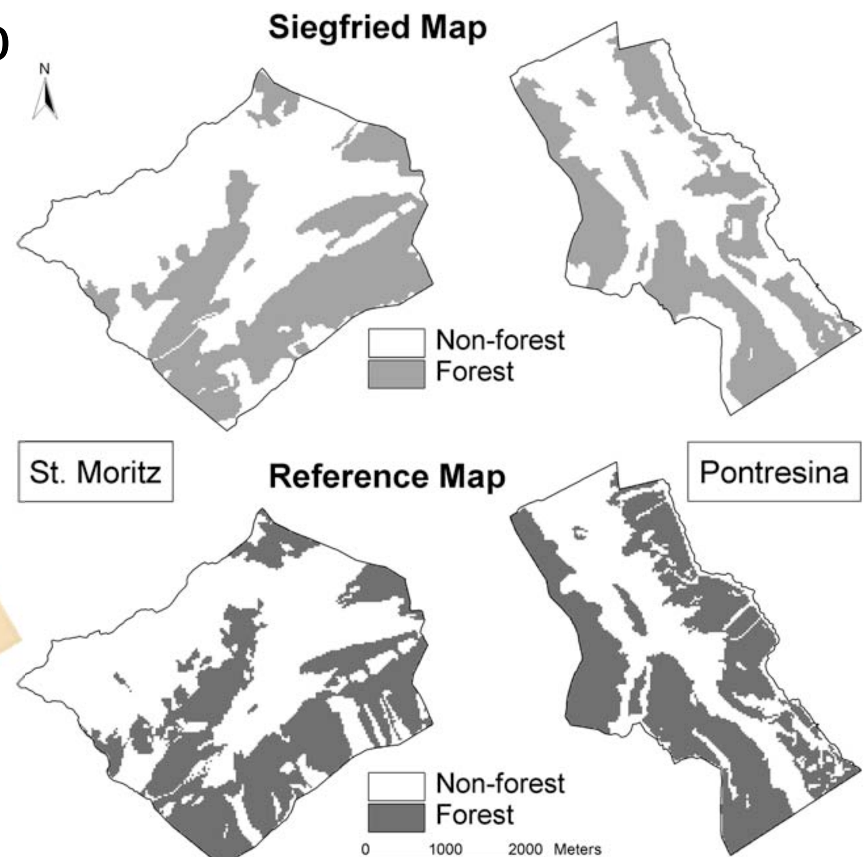
- Local summary statistics can be compared with actual point data for house prices
- Contrast of price with neighborhood using error tables
- Price ranges identifiable for geographical entities,…



Example Uncertainty Modeling

- Global accuracy measures?
- Like to apply knowledge to other regions?



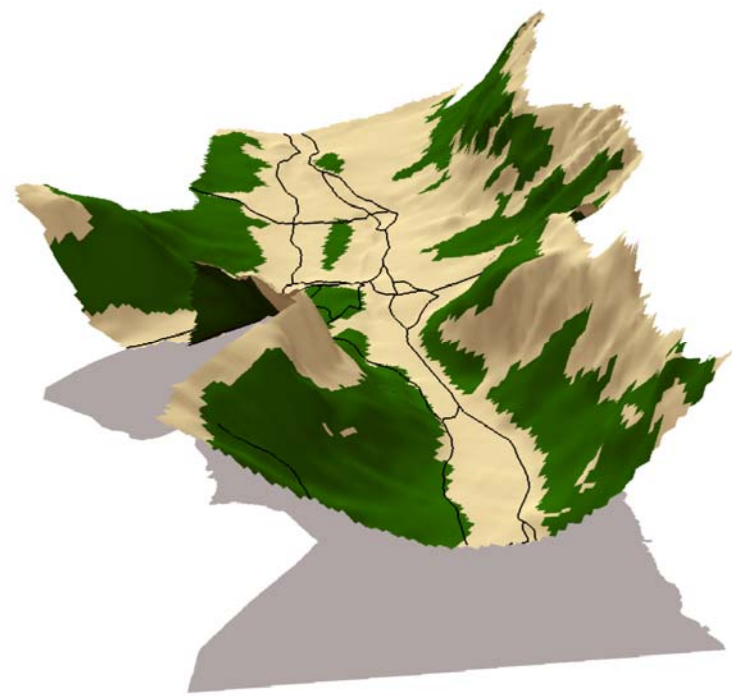


Trying to explain uncertainty and how it is caused

- What are influences to think of
- Survey, access, exploraqtion of the region
- Mountainous area, elevation, steepness

Siegfried

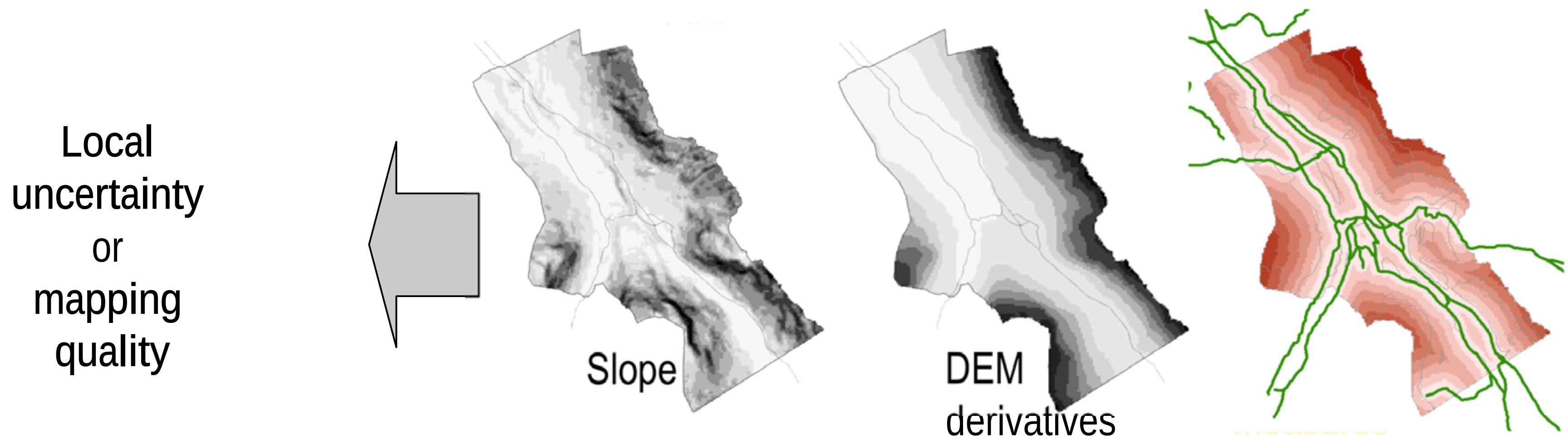
Few References



Modeling based on local summary statistics

Explanation of local uncertainty based on independent "explanatory" variables

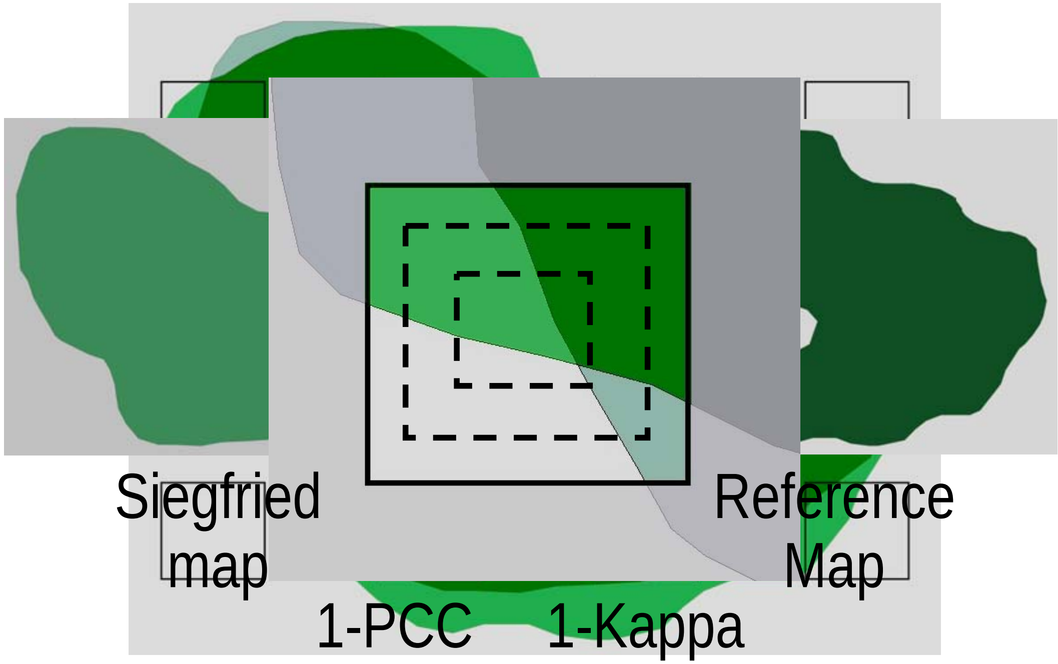
$$\mathbf{var}_{\mathrm{dep.}} = f(\mathbf{var}_{\mathrm{indep.}})$$



Modeling based on local summary statistics

The Dependent Variable

- Spatially oriented local uncertainty
- Map comparison: local disagreement
- Bounded error rate (0=perfect fit; 1=no agreement at all)

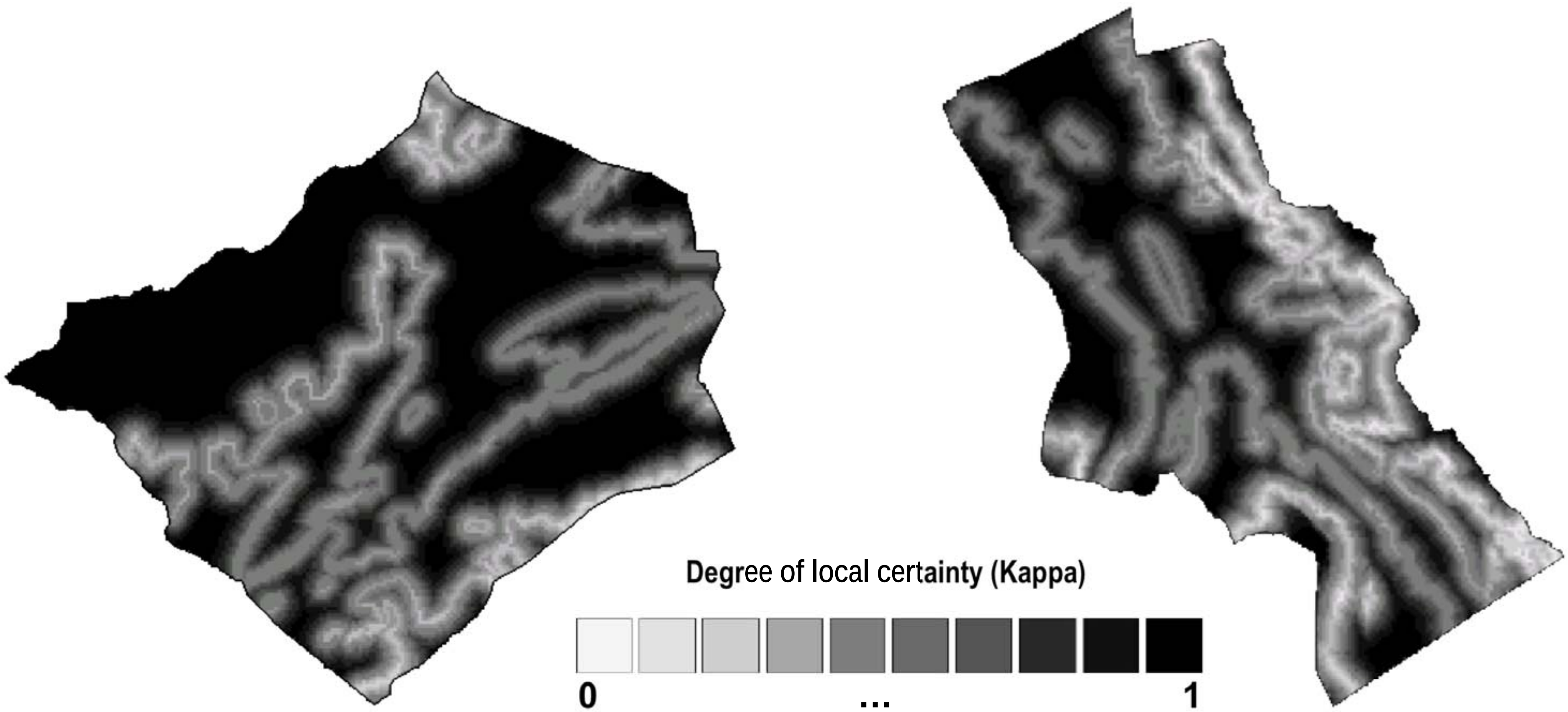


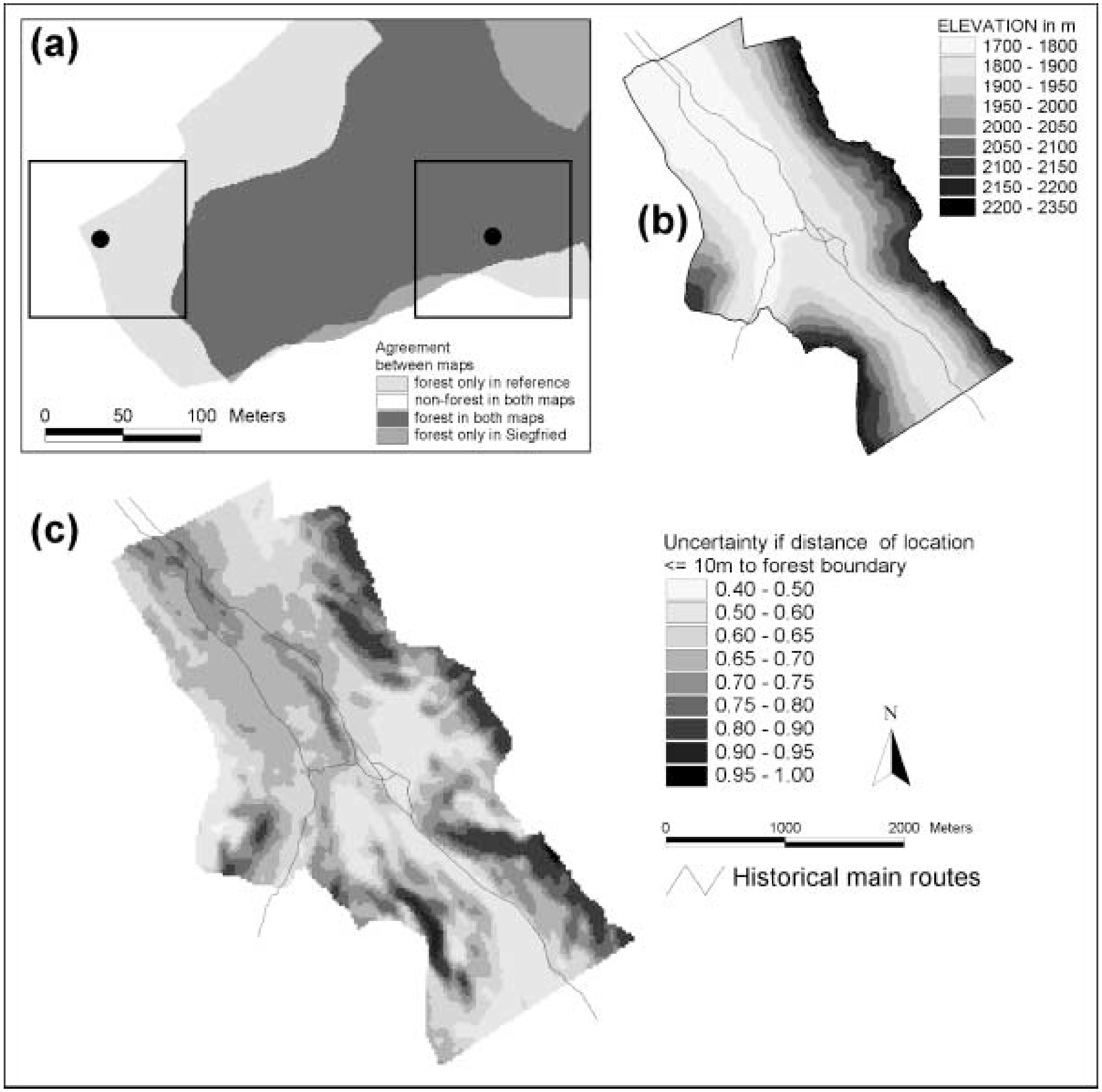
Local uncertainty and the Statistical Model

- Generalized Linear Models (GLM)
- Response $\longrightarrow$ [0,1]: uncertainty
- *Link(Response) = LinearPredictors*$_{comb}$

  $log(\boldsymbol{\mu} \,/\, (1 - \boldsymbol{\mu})) = \boldsymbol{\alpha} + X^T\boldsymbol{\beta}$
- Crosswise calibration and testing

Mapping local uncertainty or quality




(a)
Agreement
between maps
forest only in reference
non-forest in both maps
forest in both maps
forest only in Siegfried
0
50
100
Meters
(b)
ELEVATION in m
1700 - 1800
1800 - 1900
1900 - 1950
1950 - 2000
2000 - 2050
2050 - 2100
2100 - 2150
2150 - 2200
2200 - 2350
(c)
Uncertainty if distance of location
<= 10m to forest boundary
0.40 - 0.50
0.50 - 0.60
0.60 - 0.65
0.65 - 0.70
0.70 - 0.75
0.75 - 0.80
0.80 - 0.90
0.90 - 0.95
0.95 - 1.00
N
0
1000
2000
Meters
Historical main routes

Summary

- The assessment of uncertainty of our **source data** is one of the basic requirements we should be aware of
- You have seen different error models such as **CSE**, **Perkal** or **epsilon bands** and their application for **positional error** assessment for points, lines and polygones
- We talked about the **confusion matrix** which represents the most prominent assessment approach for **categorial/ nominal** data in a **classification** process
- You have seen some examples how to use and how to overcome **limitations** of the **summary statistics** derived

References

- Burrough, P.A. and McDonnell, R.A. (1998):Principles of Geographical Information Systems. Second Edition. Oxford University Press.
- Jones, C.B. (1997): Geographical Information Systems and Computer Cartography. Longman.
- Longley et al. 2001. Geographic Information Systems and Science. Wiley.
- Fisher P 1999 Models of uncertainty in spatial data. In Longley P, Goodchild M F, Maguire D J, and Rhind D W (eds) Geographical Information Systems: Principles, Techniques, Management and Applications (Volume 1). New York, John Wiley and Sons: 191–205
- Fisher P 2003 Data quality and uncertainty: Ships passing in the night! In Shi W, Goodchild M F, and Fisher P (eds) Proceedings of the Second International Symposium on Spatial Data Quality. Hong Kong, Hong Kong Polytechnic University: 17–22
- Guptill S C and C Morrison J L (eds) 1995 Elements of Spatial Data Quality. Oxford, Pergamon
- … if you like endless reference lists: Leyk et al., 2005 in TGIS