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 error 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???

![Table of data](image)

- **True coordinates, x, y**
- **Data coordinates, \( x', y' \)**
- **Error distance, \( e \)**

\[
e = \sqrt{(x - x')^2 + (y - y')^2}
\]
Positional Errors of Points

- **Circular Standard Error**
- Say: $x \pm \delta x$, $y \pm \delta y$
- 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 produce constant areas around the lines
Back to the rabbit-polygon example
### Categorical Data - Confusion Matrix

- **What can go wrong in a classification?**

<table>
<thead>
<tr>
<th></th>
<th>wheat</th>
<th>corn</th>
<th>soy</th>
<th>alfalfa</th>
<th>grass</th>
<th>fallow</th>
</tr>
</thead>
<tbody>
<tr>
<td>wheat</td>
<td>14</td>
<td>4</td>
<td></td>
<td></td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>corn</td>
<td>2</td>
<td>12</td>
<td>1</td>
<td>3</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>soy</td>
<td>1</td>
<td>18</td>
<td>2</td>
<td></td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>alfalfa</td>
<td>3</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td></td>
<td>23</td>
</tr>
<tr>
<td>grass</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>fallow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20</td>
</tr>
</tbody>
</table>

**overall accuracy** = \[ \frac{\text{sum of diagonal}}{\text{total number of samples}} \] = \[ \frac{92}{120} = 76.7\% \]
Summary Statistics

- **Overall accuracy**: Diagonal / Total
- **Error of omission (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 omitted)
- **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)

![Table of data]

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...

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

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

<table>
<thead>
<tr>
<th>Measure</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence</td>
<td>( \frac{a + c}{N} )</td>
</tr>
<tr>
<td>Overall diagnostic power</td>
<td>( \frac{b + d}{N} )</td>
</tr>
<tr>
<td>Correct classification rate</td>
<td>( \frac{a + d}{N} )</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>( \frac{a}{a + c} )</td>
</tr>
<tr>
<td>Specificity</td>
<td>( \frac{d}{b + d} )</td>
</tr>
<tr>
<td>False positive rate</td>
<td>( \frac{b}{b + d} )</td>
</tr>
<tr>
<td>False negative rate</td>
<td>( \frac{c}{a + c} )</td>
</tr>
<tr>
<td>Positive predictive power (PPP)</td>
<td>( \frac{a}{a + b} )</td>
</tr>
<tr>
<td>Negative predictive power (NPP)</td>
<td>( \frac{d}{c + d} )</td>
</tr>
<tr>
<td>Misclassification rate</td>
<td>( \frac{b + c}{N} )</td>
</tr>
<tr>
<td>Odds-ratio</td>
<td>( \frac{(ad)}{(cb)} )</td>
</tr>
<tr>
<td>Kappa</td>
<td>[ \frac{[(a + d) - (((a + c)(a + b) + (b + d))}{(c + d))/N]}{N - ((a + c)(a + b) + (b + d(c + d))/N]} ]</td>
</tr>
<tr>
<td>NMI n(s)</td>
<td>[ -a \cdot \text{ln}(a) - b \cdot \text{ln}(b) - c \cdot \text{ln}(c) - d \cdot \text{ln}(d) + (a + b) \cdot \text{ln}(a + b) + (c + d) \cdot \text{ln}(c + d) ]/[N \cdot \text{ln} N - ((a + c) \cdot \text{ln}(a + c) + (b + d) \cdot \text{ln}(b + d)) ]</td>
</tr>
</tbody>
</table>
Kappa Example

<table>
<thead>
<tr>
<th></th>
<th>Forest on ground</th>
<th>Water on ground</th>
<th>Row total (C_{i,})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forest in DB</td>
<td>1000</td>
<td>100</td>
<td>1100</td>
</tr>
<tr>
<td>Water in DB</td>
<td>200</td>
<td>700</td>
<td>900</td>
</tr>
<tr>
<td>Column total (C_{i,})</td>
<td>1200</td>
<td>800</td>
<td>2000</td>
</tr>
</tbody>
</table>

\[ k = \frac{\left( (1000 + 700) - \frac{((1200 \times 1100 / 2000) + (800 \times 900 / 2000))}{2000 - ((1200 \times 1100 / 2000) + (800 \times 900 / 2000))} \right)}{} \]

\[ = 0.69 \]

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)

<table>
<thead>
<tr>
<th></th>
<th>Reference map</th>
<th></th>
<th>Original map</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pontresina</td>
<td>St. Moritz</td>
<td>Pontresina</td>
</tr>
<tr>
<td>Forest area (correct)</td>
<td>8350</td>
<td>8223</td>
<td>5853</td>
</tr>
<tr>
<td>Non-forest area (correct)</td>
<td>6486</td>
<td>11496</td>
<td>5649</td>
</tr>
<tr>
<td>Misclassified proportion</td>
<td>–</td>
<td>–</td>
<td>3334</td>
</tr>
<tr>
<td>PCC</td>
<td>–</td>
<td>–</td>
<td>0.76</td>
</tr>
<tr>
<td>Kappa</td>
<td>–</td>
<td>–</td>
<td>0.55</td>
</tr>
<tr>
<td>NMI</td>
<td>–</td>
<td>–</td>
<td>0.26</td>
</tr>
</tbody>
</table>
First Example - Simple Accuracy Assessment

- Image extraction result to be evaluated against human inspection efforts
First Example - Simple Accuracy Assessment

<table>
<thead>
<tr>
<th>Hydro (Blue)</th>
<th>Elevation (Red)</th>
<th>Black Layer</th>
<th>Background (White)</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.76</td>
<td>0.91</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>Precision</td>
<td>0.80</td>
<td>0.92</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>ACC</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.96</td>
</tr>
<tr>
<td>Kappa</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.93</td>
</tr>
<tr>
<td>NMI</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.81</td>
</tr>
</tbody>
</table>
What is lacking with summary statistics?

<table>
<thead>
<tr>
<th>true value</th>
<th>wheat</th>
<th>corn</th>
<th>soy</th>
<th>alfalfa</th>
<th>grass</th>
<th>fallow</th>
</tr>
</thead>
<tbody>
<tr>
<td>wheat</td>
<td>14</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>corn</td>
<td>2</td>
<td>12</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>soy</td>
<td>1</td>
<td>18</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>alfalfa</td>
<td>3</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>grass</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>fallow</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Overall accuracy = \( \frac{\text{sum of diagonal}}{\text{total number of samples}} \) = \( \frac{92}{120} = 76.7\% \)
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

![Table and Diagram]

\[
\text{overall accuracy} = \frac{\text{sum of diagonal}}{\text{total number of samples}} = \frac{92}{120} = 76.7\% \]
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, exploration 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

\[ \text{var}_{\text{dep.}} = f(\text{var}_{\text{indep.}}) \]
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 $\rightarrow [0,1]$: uncertainty
- $Link(\text{Response}) = LinearPredictors_{comb}$
  
  $\log(\mu / (1 - \mu)) = \alpha + X^T\beta$

- Crosswise calibration and testing
Mapping local uncertainty or quality

Degree of local certainty (Kappa)

0  ...  1
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 polygons.
• We talked about the confusion matrix which represents the most prominent assessment approach for categorical/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

• … if you like endless reference lists: Leyk et al., 2005 in TGIS