# Uncertainty and data quality in spatial modelling

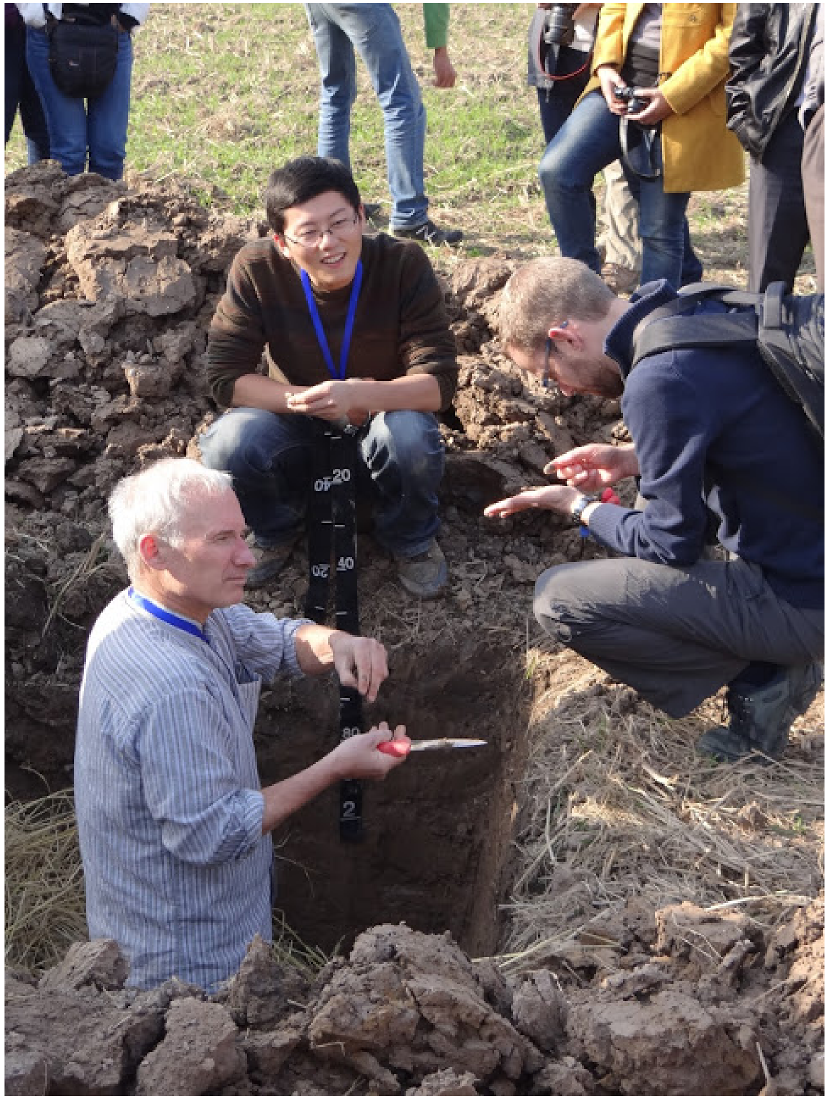
D G Rossiter
Section of Soil & Crop Sciences, Cornell University

April 27, 2020



Cornell University
College of Agriculture and Life Sciences

# Data



From observations/measurements of some kind. . .

Never “perfect” . . .

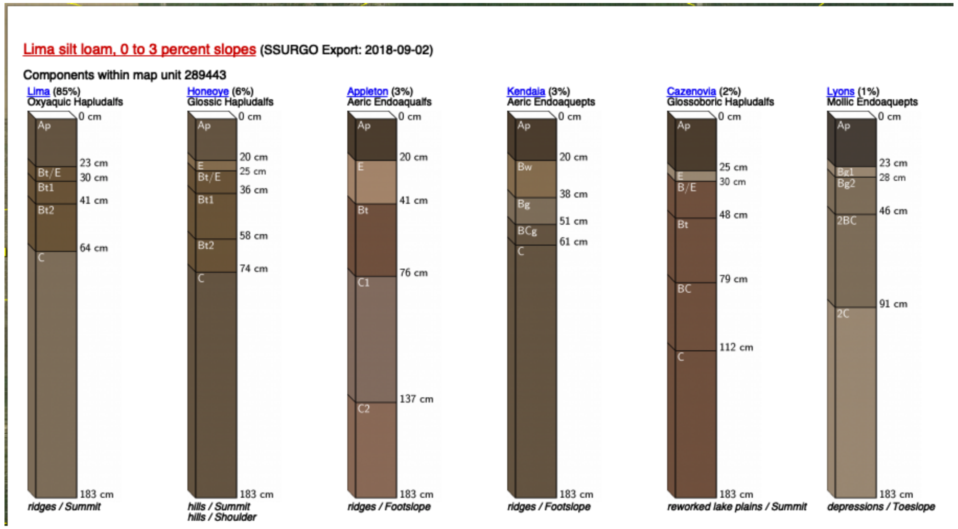
Natural variation, sampling error, observer bias . . .

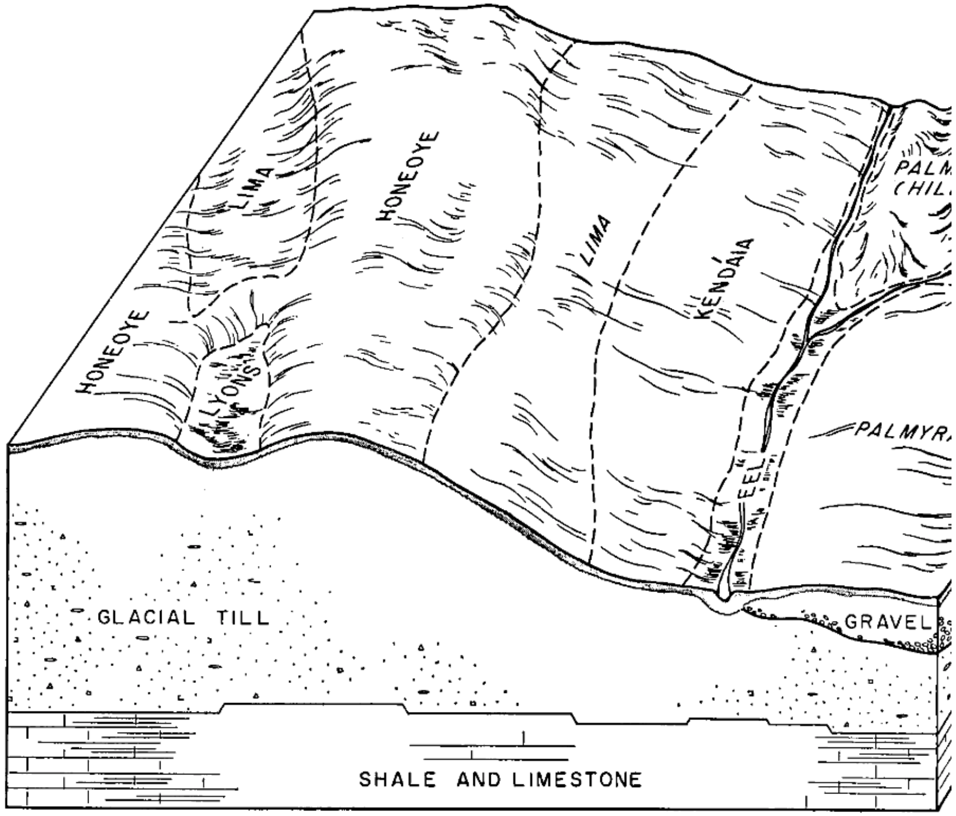
# Natural variability → Uncertainty

Natural variability in nature . . .

Where to describe the "representative" soil profile?

How to describe the variation?


# Map unit impurity

design scale 1:24 000;
minimum legible delination 2.3 ha
(0.4 $cm^2$ on map)
(Cornell experimental farm, Aurora NY)

# Uncertainty and data quality

Related concepts:

**Uncertainty** **lack of knowledge** about the "truth"

**Data quality** **fitness for use** of the data

So **uncertainty** is only one aspect of **data quality**

Uncertain data can be useful. . . but how "uncertain" is too much?



Cornell University
College of Agriculture and Life Sciences

# Topic: Data quality

- **External** quality is **"fitness for use"**, so depends on **intended uses**
  - EPA: "The totality of features and characteristics of data that bears on their ability to satisfy a given purpose"[1]
  - Emphasize: **"to satisfy a given purpose"**
    * Example: precision of georeference to find an area for further study vs. an area for direct intervention

- **Internal** quality is the consistency, completeness, documentation of a dataset
  - Explained by the **metadata** (see below)

[1] http://www.epa.gov/emap/html/pubs/docs/resdocs/mglossary.html

# Data Quality sources

- Glossary of terms from EPA's Environmental Sampling and Analytical Methods (ESAM) Program[2]
- Shi, W., Fisher, P., & Goodchild, M. F. (2003). *Spatial Data Quality*. CRC Press.
- Guptill, S. C., & Morrison, J. L. (2013). *Elements of Spatial Data Quality*. Elsevier (on behalf of International Cartographic Association)
- eBird. (2020). *The eBird review process*. Retrieved 27-April-2020, from `https://support.ebird.org/en/support/solutions/articles/48000795278-the-ebird-review-process`

[2] `https://www.epa.gov/esam/glossary`

# Data quality components

**Completeness** : degree to which the dataset represents the **population of interest**

- what is the population about which we want to make decisions or maps?

**Consistency** : degree to which different items in the dataset are **coherent**

- internal: among data items;
- external: with other sources of similar information

**Currency** : when was the data collected? To what **time period** is it relevant?

**Lineage** : how has the data arrived from original observations to its current state? how has it been "massaged"?

- Are the data as directly measured (how?) or manipulated? How and why?
- Were any observations ("outliers") adjusted or deleted? How and why?

. . .

...

**Accuracy** : **difference** between data and reality

- e.g., evaluation (“validation”) RMSE (average error), MAE (accuracy, bias)

**Precision** : **dispersion** of data around true value

- e.g., $\sigma^2$, IQR etc. of measurements

**Credibility** : reliability of **information source**

- is the data source **technically competent**?
- does the source have a **political or economic interest** in the data or its interpretation?
- is the data source explicit about its **funding sources** and possible biases?

**Subjectivity** : how much and what kind of **human interpretation** was used?

- e.g., automated vs. manual photointerpretation

# Topic: Metadata – documenting data quality

**"Data about the data"**; **document** and **communicate** all the above aspects of data quality

- **Formal**: according to a **standard**, in a machine-readable format (e.g., XML) can be **searched** by a program
- **Informal**: described in text or non-standard database
- It is a revealing exercise to create proper metadata – one rapidly discovers that one doesn't know as much about the dataset as one thought

For **geospatial** data: ISO 19115, (USA) Federal Geographic Data Committee (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM)[3]

Metadata **tools** built-into GIS or standalone[4]

[3] http://www.fgdc.gov/metadata/geospatial-metadata-standards
[4] http://www.fgdc.gov/metadata/geospatial-metadata-tools

# FGDC metadata sections

1. Identification Information
2. Data Quality Information
3. Spatial Data Organization Information
4. Spatial Reference Information
5. Entity and Attribute Information
6. Distribution Information
7. Metadata Reference Information
8. Citation Information
9. Time Period Information
10. Contact Information

# Metadata in plain language

1. What does the data set describe?
   (a) What is the title of the data set?
   (b) What **geographic area** does the data set cover?
   (c) Does the data set describe conditions during a particular **time period**?
   (d) Is this a digital map or remote-sensing image, or something different like tabular data?
   (e) How does the data set represent geographic features?
      i. How are geographic features stored in the data set?
      ii. What **coordinate reference system** is used to represent geographic features?
   (f) How does the data set describe geographic features?
      i. What are the types of features present?
      ii. For each feature, what **attributes** of these features are described?
      iii. What sort of values does each attribute hold?
      iv. For measured attributes, what are the units of measure, resolution of the measurements, frequency of the measurements in time, and estimated accuracy of the measurements?

...

. . .

2. **Who** produced the data set?

3. **Why** was the data set created?

4. **How** was the data set created?

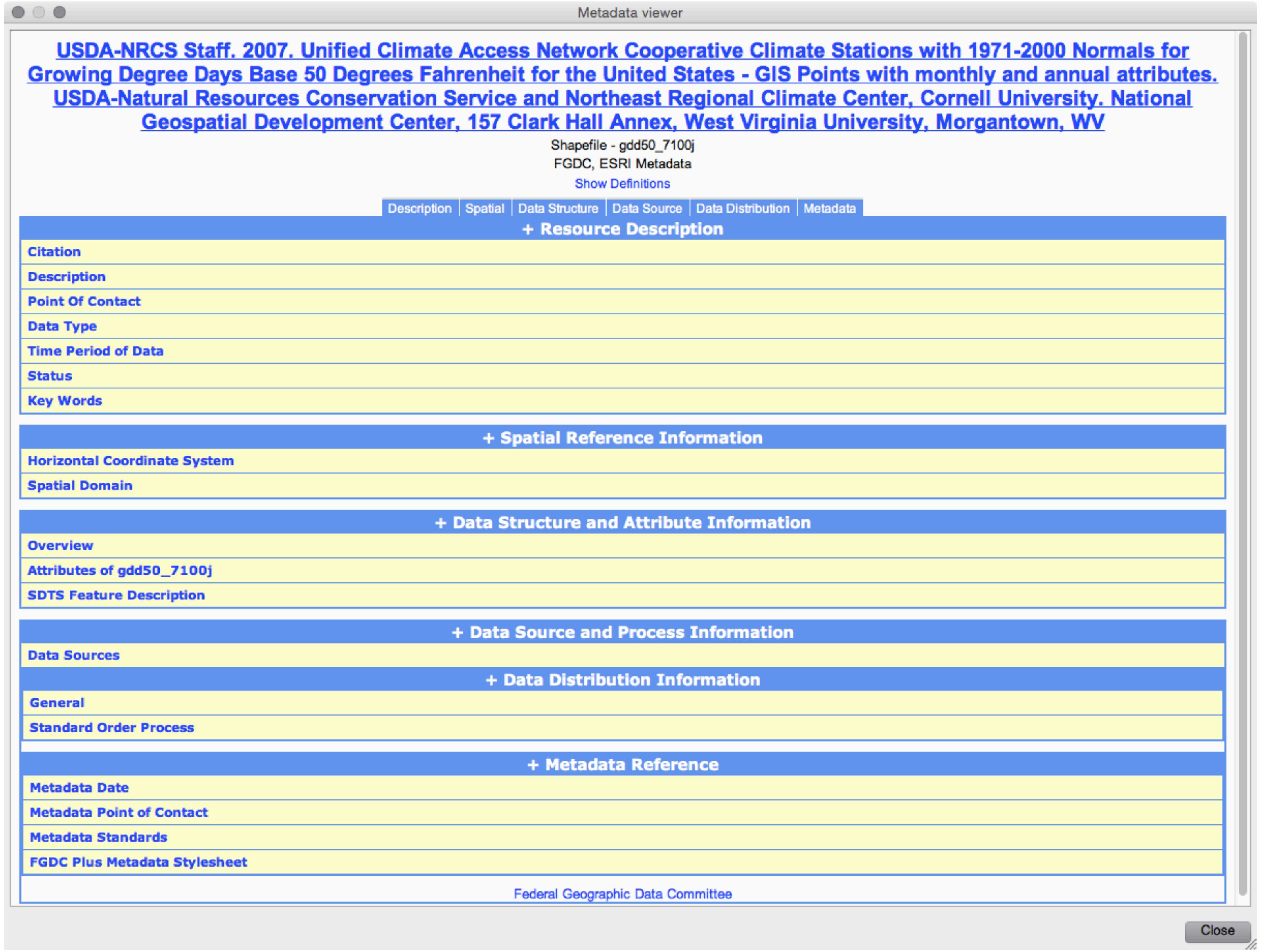
5. How **reliable** are the data; what problems remain in the data set?

6. How can someone **get a copy** of the data set?

7. Who wrote the metadata?

source: `http://geology.usgs.gov/tools/metadata/tools/doc/ctc/`

# Metadata template



# Example: Administrative units

## Download GADM data (version 3.6)

Country

Cambodia

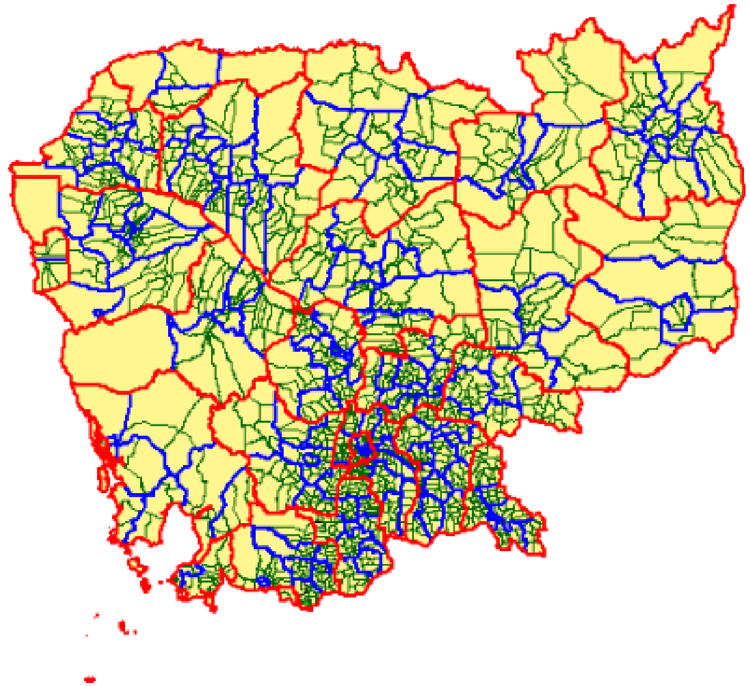
Geopackage

Shapefile

R (sp): level-0, level1, level2, level3, level4

R (sf): level-0, level1, level2, level3, level4

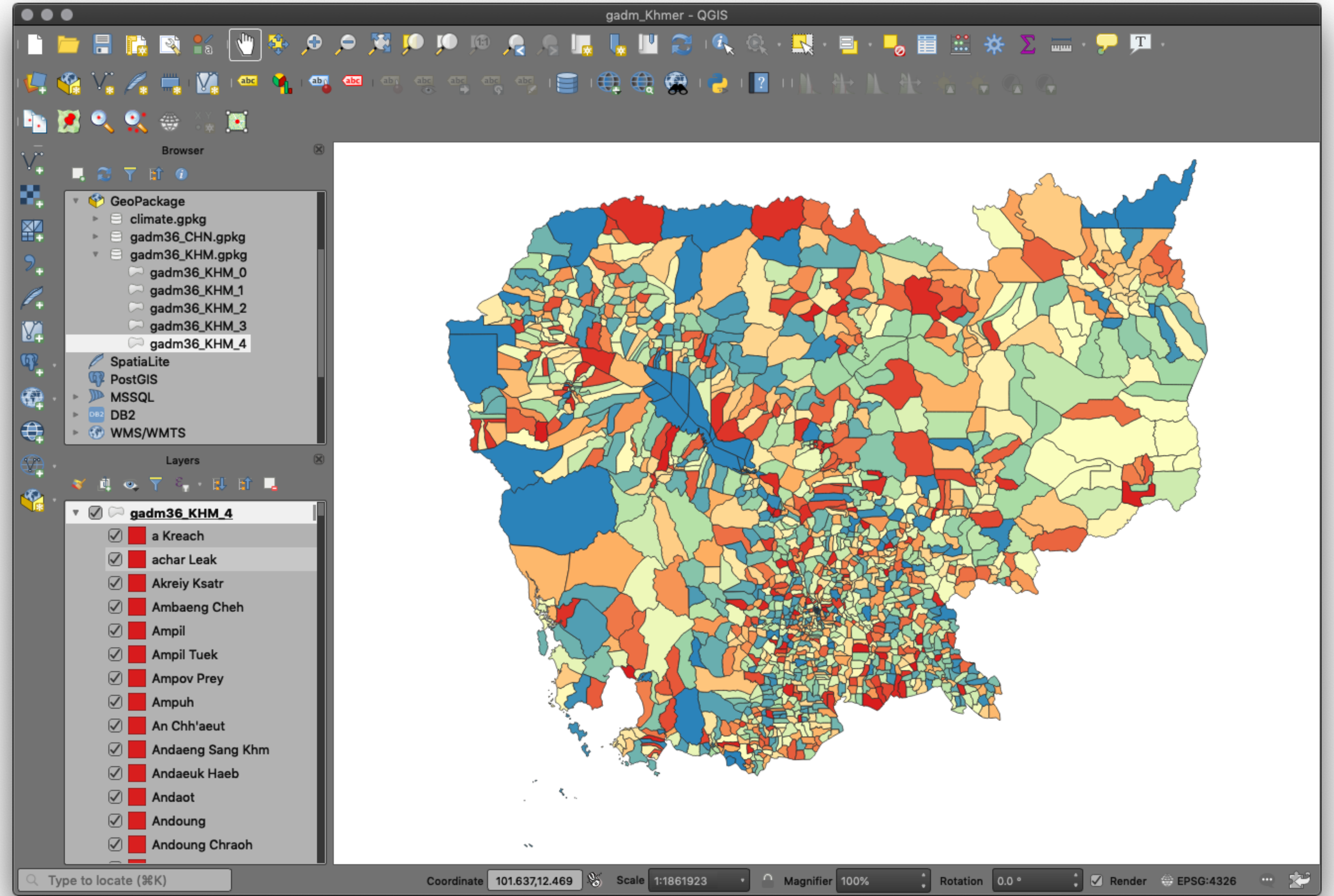
KMZ: level-0, level1, level2, level3, level4



The coordinate reference system is longitude/latitude and the WGS84 datum.
Description of file formats.

source: `http://gadm.org`

We know the political unit, file format, and CRS.




It opens in QGIS, with projection intact, good, but . . .

gadm36_KHM_4 :: Features Total: 1580, Filtered: 1580, Selected: 0

| | fid | GID_0 | NAME_0 | GID_1 | NAME_1 | GID_2 | NAME_2 | GID_3 | NAME_3 | GID_4 | NAME_4 | VARNAME_4 | TYPE_4 | ENGTYPE_4 | CC_4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 411 | KHM | Cambodia | KHM.6_1 | Kâmpóng Thum | KHM.6.1_1 | Baray | KHM.6.1.16_1 | Svay Phleung | KHM.6.1.16.1_1 | Khnay Tong | NULL | Phum | Village | NULL |
| 2 | 612 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.4_1 | Kaoh Thum Ka | KHM.8.3.4.1_1 | Chong Kaoh Thmei | NULL | Phum | Village | NULL |
| 3 | 613 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.5_1 | Kaoh Thum Kha | KHM.8.3.5.1_1 | Svay Ta Mekh | NULL | Phum | Village | NULL |
| 4 | 614 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.6_1 | Leuk Daek | KHM.8.3.6.1_1 | Khleang Lech | NULL | Phum | Village | NULL |
| 5 | 615 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.7_1 | Pouthi Ban | KHM.8.3.7.1_1 | Kampong Kor | NULL | Phum | Village | NULL |
| 6 | 608 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.2_1 | Kandal Stueng | KHM.8.2.23_1 | Trea | KHM.8.2.23.1_1 | Damrei Slab | NULL | Phum | Village | NULL |
| 7 | 609 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.1_1 | Chheu Khmau | KHM.8.3.1.1_1 | Traeuy Kaoh | NULL | Phum | Village | NULL |
| 8 | 610 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.2_1 | Chrouy Ta Kaev | KHM.8.3.2.1_1 | Lekh Bei | NULL | Phum | Village | NULL |
| 9 | 611 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.3_1 | Kampong Kong | KHM.8.3.3.1_1 | Preaek Ph'av | NULL | Phum | Village | NULL |
| 10 | 620 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.4_1 | Khsach Kandal | KHM.8.4.1_1 | Bak Dav | KHM.8.4.1.1_1 | bak Dav Kraom | NULL | Phum | Village | NULL |
| 11 | 621 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.4_1 | Khsach Kandal | KHM.8.4.2_1 | Chey Thum | KHM.8.4.2.1_1 | Chey Touch | NULL | Phum | Village | NULL |
| 12 | 622 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.4_1 | Khsach Kandal | KHM.8.4.3_1 | Kampong Chamlang | KHM.8.4.3.1_1 | Tboung Damrei | NULL | Phum | Village | NULL |
| 13 | 623 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.4_1 | Khsach Kandal | KHM.8.4.4_1 | Kaoh Chouram | KHM.8.4.4.1_1 | Kandal | NULL | Phum | Village | NULL |
| 14 | 616 | KHM | Cambodia | KHM.8_1 | Kândal | KHM.8.3_1 | Kaoh Thum | KHM.8.3.8_1 | Preaek Chrey | KHM.8.3.8.1_1 | Pak Nam | NULL | Phum | Village | NULL |

Show All Features

. . . What do these fields mean? (see next slide)
How current is the information? Or to what time period does it refer?
How precise are the boundaries?
Are these from field measurements, official gazette, a government map . . . ?
Are these legal or customary boundaries?
Any disputes?

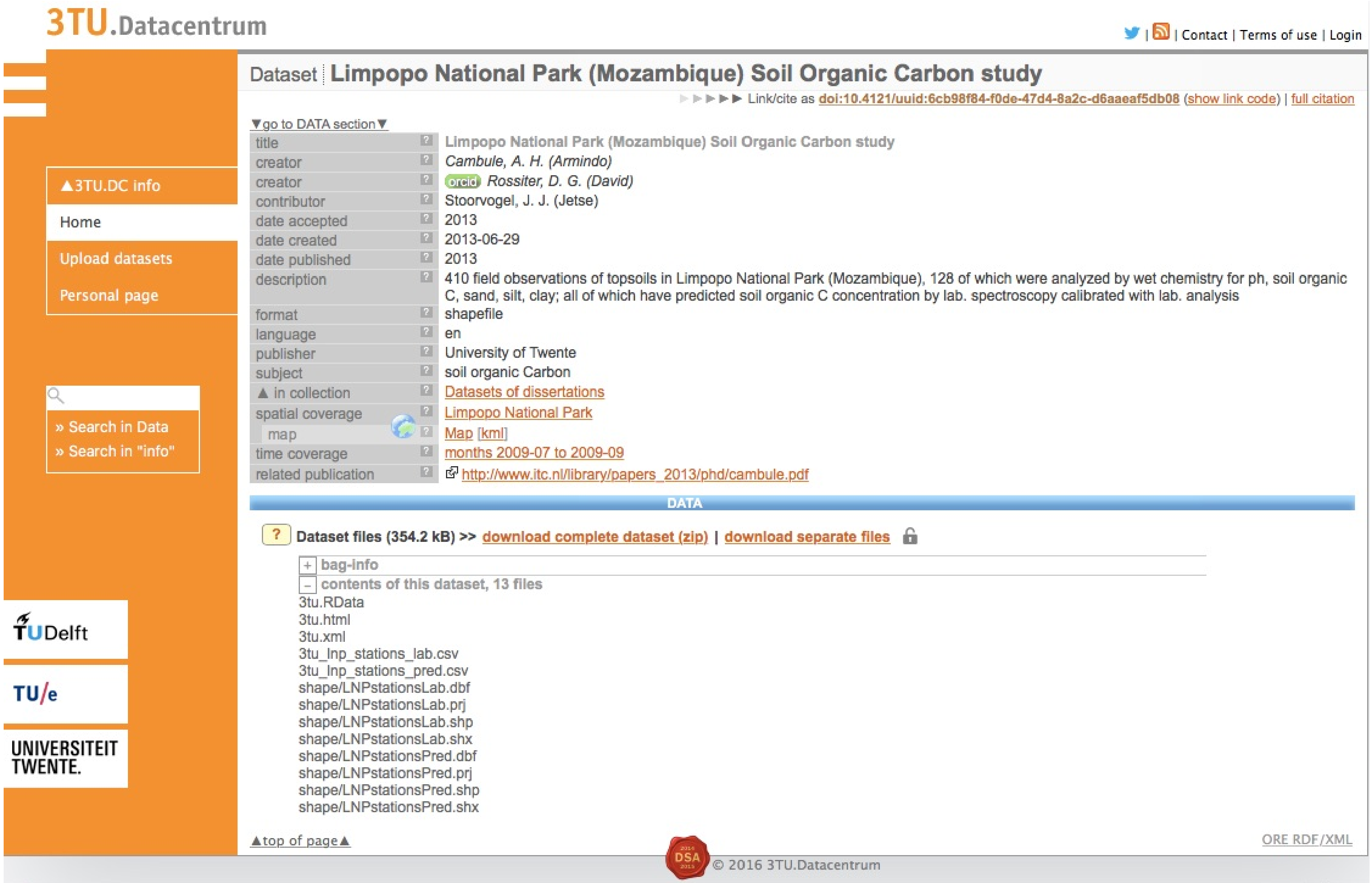
## Variable names for level 0 (country)

| Variable | Type | Description |
|---|---|---|
| UID | Integer | Unique ID across all geometries at the highest level of subdivisions |
| ID_0 | Integer | Unique numeric ID for level 0 (country) |
| GID_0 | String | Preferred unique ID for level 0 (see below). ISO 3166-1 alpha-3 country code when available |
| NAME_0 | String | Country Name in English |

## Variable names for level "i", where "i" can be 1, 2, 3, 4, or 5

| Variable | Type | Description |
|---|---|---|
| GID_i | String | Preferred unique ID at level i. See discussion below |
| ID_i | Integer | Alternative unique identifies at level 1. See discussion below |
| NAME_i | String | Official name in latin script |
| VARNAME_i | String | Variant name. Alternate names in usage for the place, separated by pipes \| |
| NL_NAME_i | String | Non-Latin name. Official name in a non-latin script (e.g. Arabic, Chinese, Russian, Korean) |
| HASC_i | String | HASC. A unique ID from Statoids |
| CC_i | String | Country code. Uniqe ID used within the country |
| TYPE_i | String | Administrative type in local language |
| ENGTYPE_i | String | Administrative type in English (following commonly used translations) |
| VALIDFR_i | String | Valiid From. Date from which data is known to have started. default: Unknown. Format is YYYY-MM-DD or YYYY-MM or YYYY |
| VALIDTO_i | String | Valid To. Date at which data is no longer valid. default: Present or Current. Format is YYYY-MM-DD or YYYY-MM or YYYY |
| REMARKS_i | String | Comments about edits, relevant to history. For example "This is a split from Matam region." |

# Reduced metadata standards



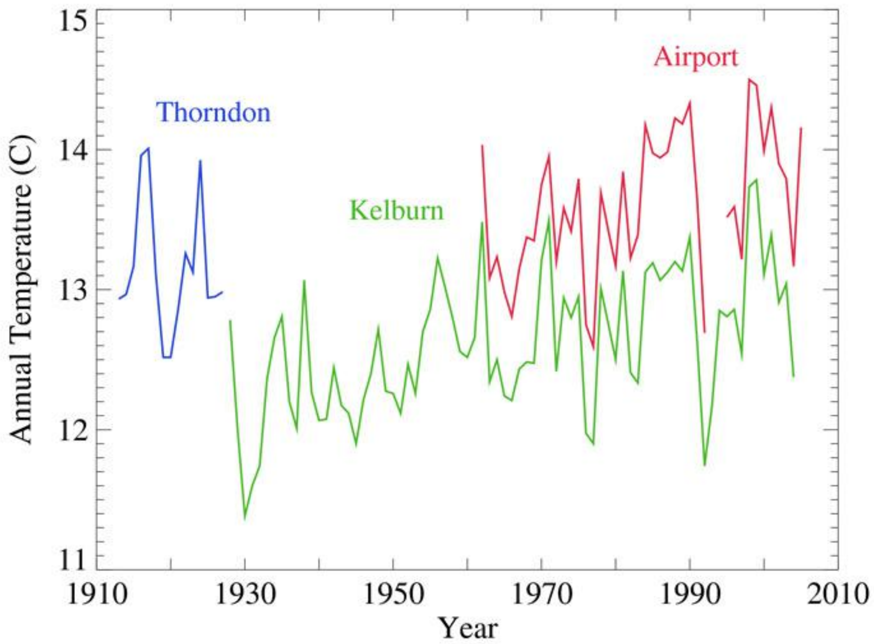

Refers to another document (here, a thesis) for further information.

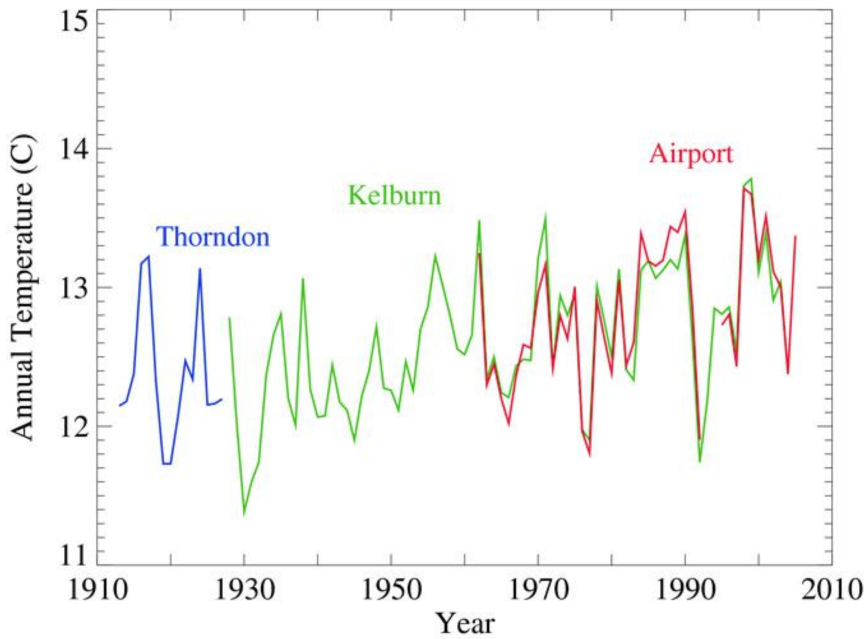
# Lineage

- Part of (2) "Data Quality Information"
- Shows how the product was **derived** from **original sources**
- Should **explain** the choices made
- Source(s) information + process step(s)
  - **Source information**: type of media, time period of content, source contribution
  - **Process step**: process description, process date; optional source used for process

# Lineage example: Raw and adjusted time series

Adjust T for a change in weather station location (Wellington, NZ):




"When we create a time series using adjusted data, we **retain all the original raw data**. It remains available on-line in the National Institute of Water & Atmospheric Research (NIWA) climate database so **others can conduct their own analysis**."[5]

[5] http://www.niwa.co.nz/our-science/climate/information-and-resources/nz-temp-record/why-climate-data-sometimes-needs-to-be-adjusted

# Lineage: Tompkins County (NY) Agricultural Districts

*Lineage:*
- *Source_Information:*
  - *Source_Citation:*
    - *Citation_Information:*
      - *Originator:* Tompkins County Planning
      - *Publication_Date:* unknown
      - *Title:* none
  - *Source_Scale_Denominator:* 24000
  - *Type_of_Source_Media:* Hard copy on Mylar, vellum or paper; digital on CD-ROM.
  - *Source_Time_Period_of_Content:*
    - *Time_Period_Information:*
      - *Multiple_Dates/Times:*
        - *Single_Date/Time:*
          - *Calendar_Date:* 20131010 (district #1)
        - *Single_Date/Time:*
          - *Calendar_Date:* 20090407 (district #2)
    - *Source_Currentness_Reference:* 8-year certification date
  - *Source_Citation_Abbreviation:* agTOMP
  - *Source_Contribution:* original district boundaries
- *Process_Step:*
  - *Process_Description:* 1) ORIGINAL SCAN PROJECT In 1996, the entire set of NYS Agricultural District maps in the collection of Cornell IRIS (originally CLEARS) was converted to digital format. This was done by shipping blueprint copies of the maps to the NYS DEC for scanning. Digital Line Graph files were returned, which were converted to ArcInfo Coverages. These coverages represented one map sheet apiece. Original maps with multiple sheets were represented by multiple coverages. Coverages were compared to the original maps and edited as necessary to create an accurate representation of the Ag District boundaries shown on the maps. After accuracy was confirmed, coverages representing multiple sheets were merged to create district coverages. Districts were then merged to create county coverages. Merged districts sometimes created slivers, which were eliminated, and gaps, which are flagged with district value of zero. Overlaps between districts also occurred in a few cases. These were flagged with district value "66". For each coverage, an attribute table was built to record the information shown on the Cornell IRIS title block of each Ag District hardcopy map. These tables are further described in the Entity and Attribute Information section of the metadata.
  - *Process_Date:* 19960100 through 20010131
  - *Process_Contact:*
    - *Contact_Information:*
      - *Contact_Organization_Primary:*
        - *Contact_Organization:* Cornell IRIS
      - *Contact_Address:*
        - *Address_Type:* mailing
        - *Address:* 1015 Bradfield Hall
        - *Address:* Cornell University
        - *City:* Ithaca
        - *State_or_Province:* New York
        - *Postal_Code:* 14853-1901

*Process_Step:*
*Process_Description:* 2) CONVERSION FROM COVERAGES TO SHAPEFILES: Internal polygons were labeled zero; coverages were reprojected from UTMz18 NAD27 to UTMz18 NAD83; Coverages were converted to shapefiles; zero polygons were deleted; attribute table was modified: deleted fields are AREA, PERIMETER, FILE#, FILE_ID, AGDIST#, DOTQUADS. Modified fields are DISTCODE and DISTRICT. DISTCODE is 12 characters to accommodate changing abbreviations -- currently four characters and three digits to represent a key code for county name and district number. The field DISTRICT was enlarged to accommodate district numbers up to five digits. Also, DISTCODE, the key field, was moved to the end of the attribute table columns.
*Process_Date:* 20080101 through 20080331
*Process_Contact:*
*Contact_Information:*
*Contact_Organization_Primary:*
*Contact_Organization:* Cornell IRIS
*Contact_Address:*
*Address_Type:* mailing
*Address:* 1015 Bradfield Hall
*Address:* Cornell University
*City:* Ithaca
*State_or_Province:* New York
*Postal_Code:* 14853-1901
*Country:* USA
*Contact_Voice_Telephone:* 607-255-6520 or 607-255-6529
*Process_Step:*
*Process_Description:* 3) UPDATING COUNTY BOUNDARY DATA: County shapefiles are updated to reflect modifications that occurred during the eight-year review process. Boundaries are revised using one or more of the three methods: Tablet digitizing; On-screen digitizing; Copying boundaries from county-supplied shapefiles. All modifications are proofread against the original maps to confirm accuracy. Attributes are updated and checked against the information on the map title blocks, as well as information on file. If individual tax parcels are dissolved to form an aggregate boundary, slivers and gaps may be formed by drafting discrepancies. These are visually compared to the map and eliminated when they do not represent intended exclusions. Discrepancies between the title block information and file information are clarified by contacting the county and/or New York State Department of Agriculture and Markets.
*Process_Date:* 20130131 through 20140131
*Process_Contact:*
*Contact_Information:*
*Contact_Organization_Primary:*
*Contact_Organization:* Cornell IRIS
*Contact_Address:*
*Address_Type:* mailing
*Address:* 1015 Bradfield Hall
*Address:* Cornell University
*City:* Ithaca
*State_or_Province:* New York
*Postal_Code:* 14853-1901
*Country:* USA
*Contact_Voice_Telephone:* 607-255-6520 or 607-255-6529

Now we see exactly how the **delivered** product was **dervied** from the **original**.

# Topic: Uncertainty

Concepts related to uncertainty:

**Error** two uses of this word:

1. a **mistake**, incorrect measurement;
2. **lack of fit** of a statistical model (**residuals**).

**Uncertainty** **lack of knowledge** about reality, e.g.,:

- the true state of nature (**data** uncertainty)
- the true model form or model parameters (**model** uncertainty)
- the true location (**spatial** uncertainty)

**Risk** related uses of this word:

1. the **likelihood** of an **incorrect decision**
2. this, multiplied by the **consequences** of an incorrect decision
3. **hazard** (chance of something bad happening) times **vulnerability** to the event times **exposure** to the event (e.g., "earthquake risk")


Cornell University
College of Agriculture and Life Sciences

# Sources of uncertainty (0)

- "**uncertainty** uncertainty": not knowing the sources of uncertainty and how to assess them

"There are those who know, those who don't know, and then there are those who don't even suspect."
– standard English translation of a folk saying

# Sources of uncertainty (1)

- **measurement** uncertainty
  - instrument/operator **errors** (malfunction)
  - instrument/operator **precision** (signal vs. noise)
  - instrument/operator **accuracy** (systematic **bias**)

- **observation** uncertainty
  - classification uncertainty (compare complicated vs. simple legends)
  - observer bias (e.g., soil classification)

- **scale** uncertainty
  - attribute space: precision; categorization/classification
  - geographic space: location precision vs. support

...

# Sources of uncertainty (2)

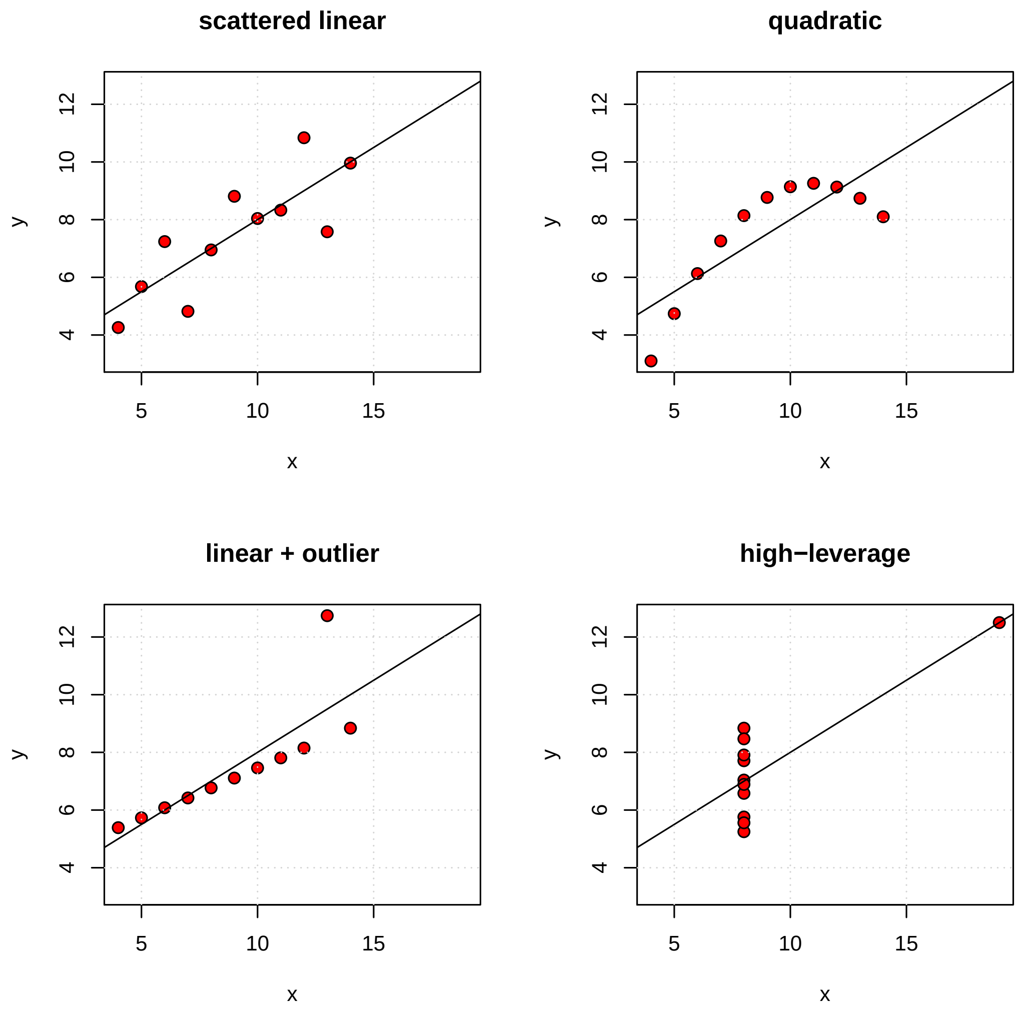
. . .

- **sampling** uncertainty: we do not see the whole population
  - object/location selection uncertainty (probability sampling vs. purposive sampling)
  - if probability, can be quantified by e.g., the sampling error
- **algorithm** uncertainty
  - e.g., supervised classification, any machine learning algorithm: representativeness of the target population
- **model form** uncertainty: does the **model form** accurately represent the underlying **process** that produced the observations?

. . .

# Model form uncertainty

Four uses of a **linear** model – in which cases is it justified?




(Use regression diagnostics to detect non-linearity)

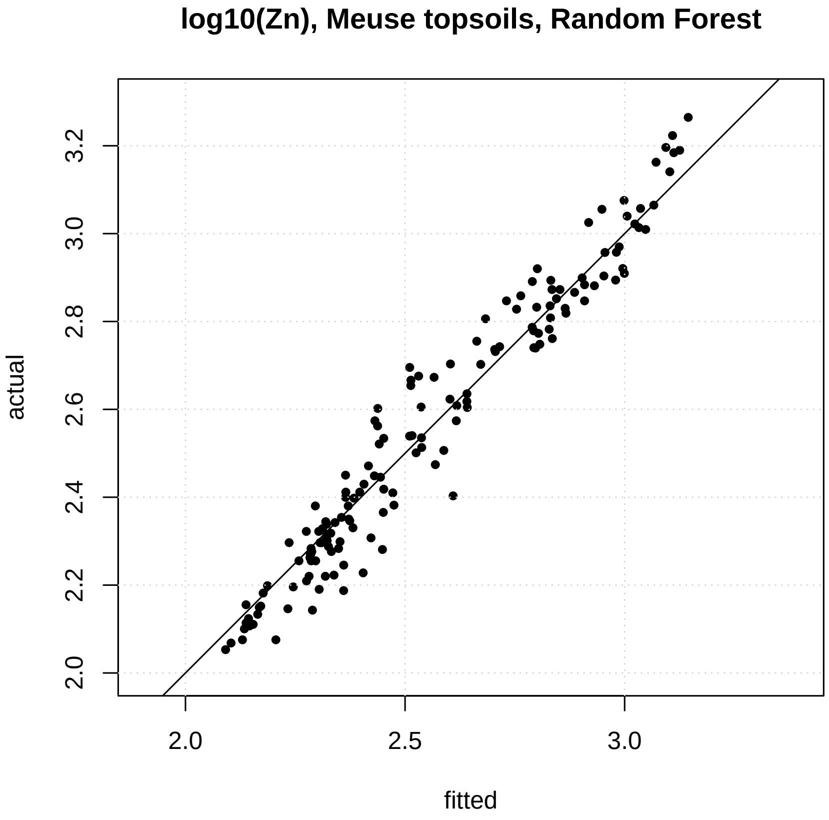
# Sources of uncertainty (3)

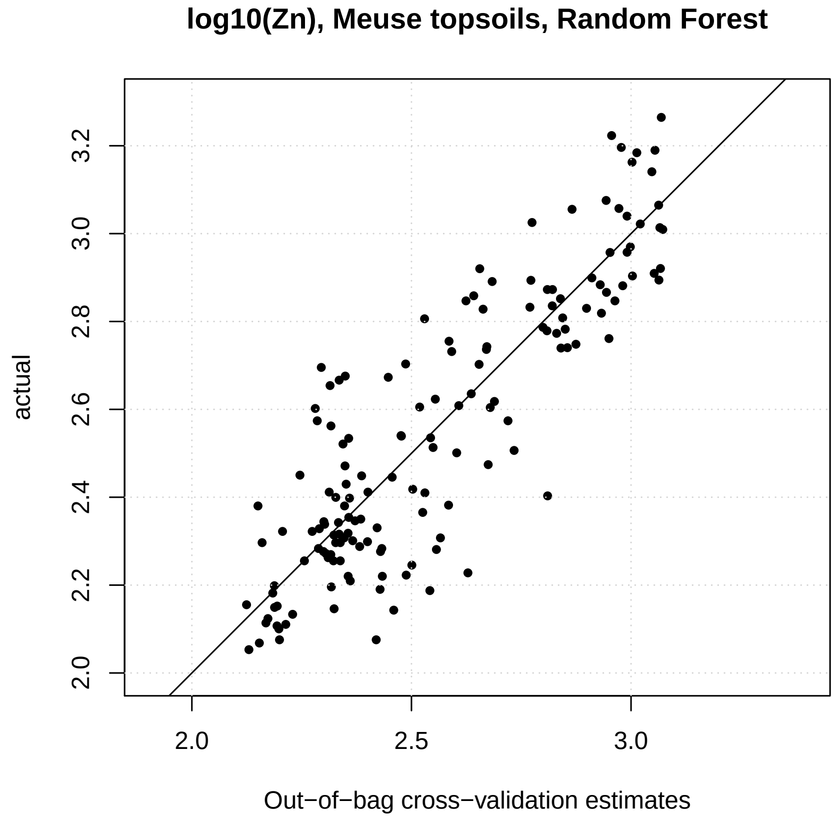
...

- **model fit** uncertainty: lack of fit of the model to the observations; "noise"
- **prediction** uncertainty: making statements about (some individuals in) the population that have not been observed
  - spatial: unobserved locations
  - temporal: unobserved times (future; past, e.g., gap filling)

  "Det er svært at spå, især om fremtiden", i.e.,
  "Prediction is very difficult, especially if it's about the future"
  – Niels Bohr, quoting Robert Storm Petersen, Danish cartoonist

# Model fit vs. prediction uncertainty




Uncertain fit, more uncertain predictions

# Example of prediction uncertainty

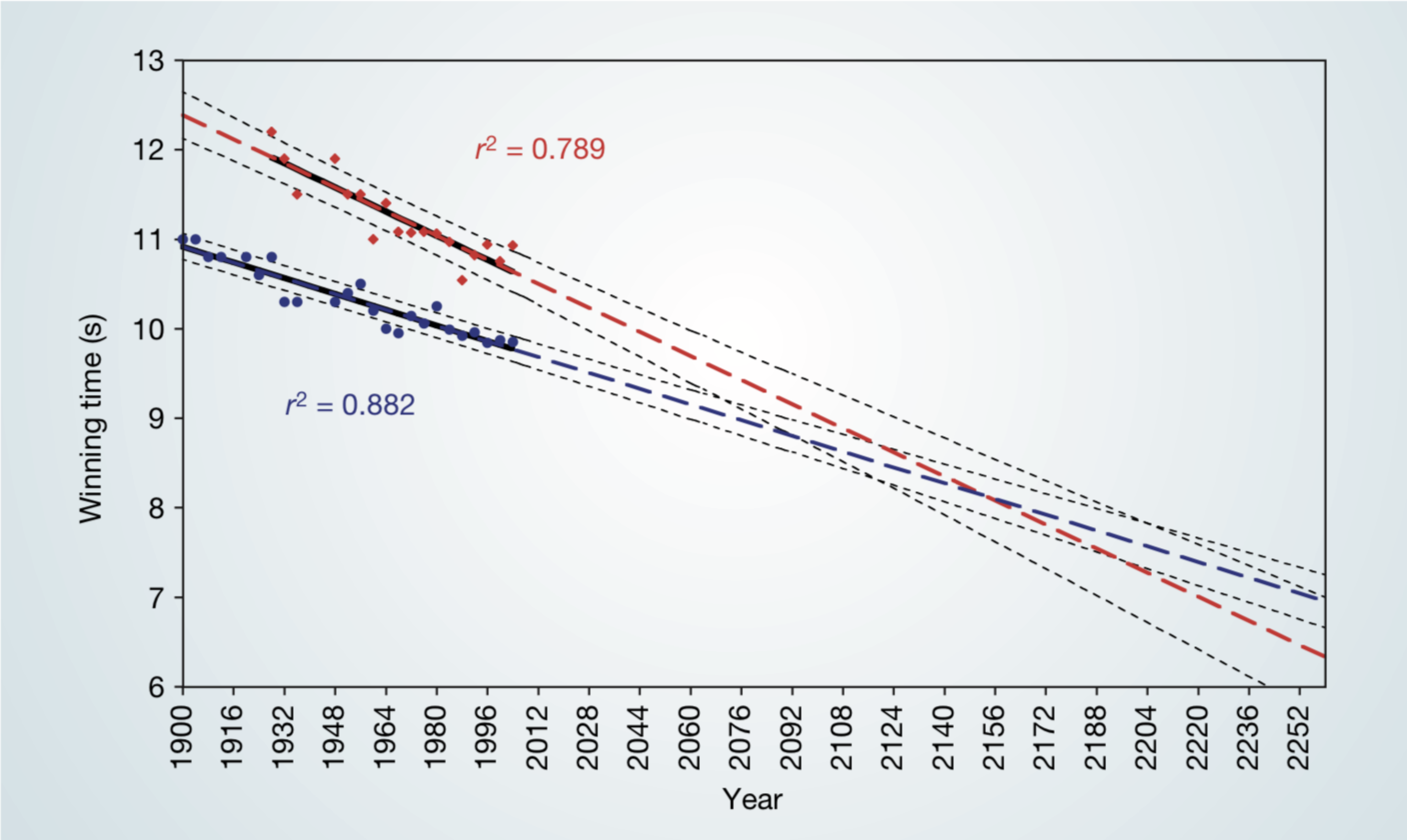
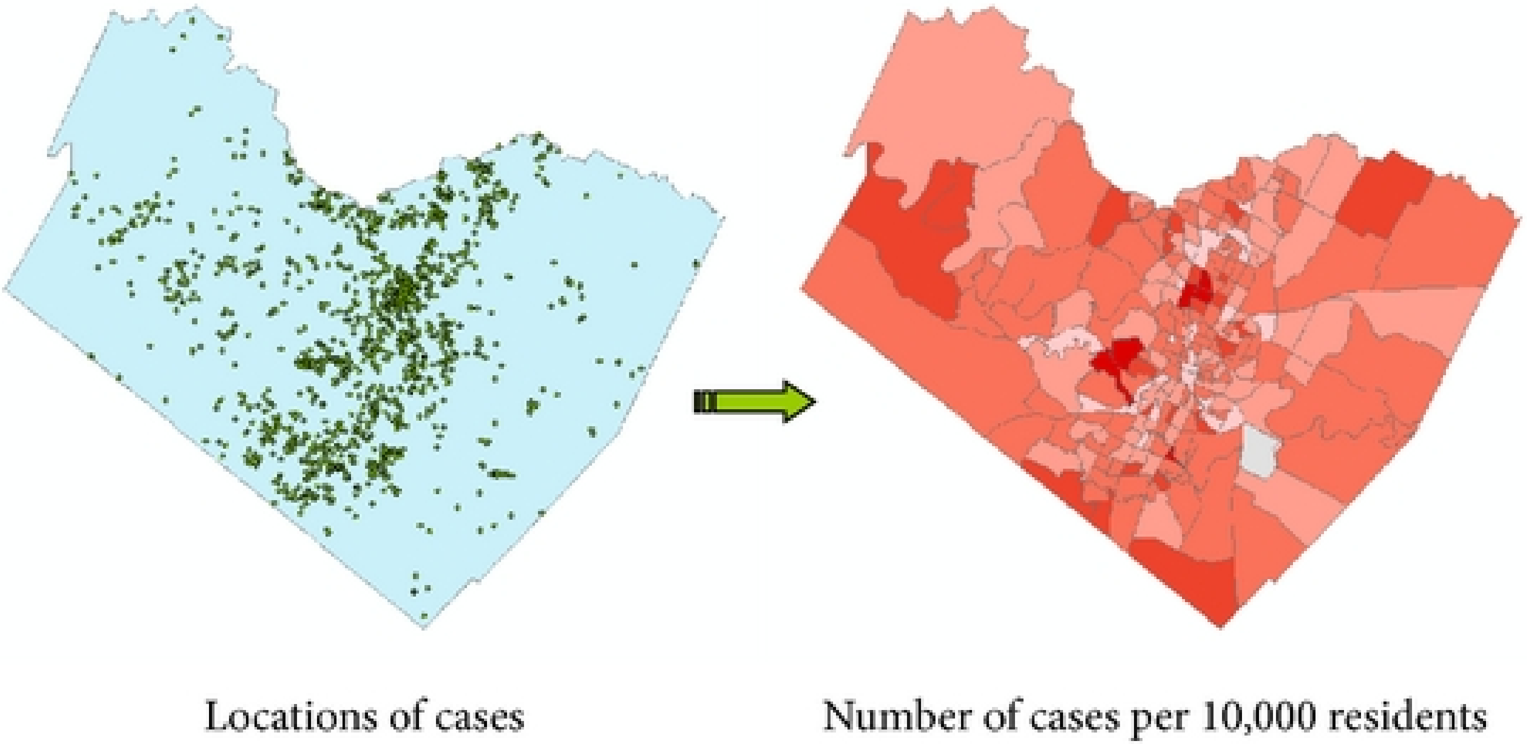



**Figure 1** The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

Source: Nature, 431, 525.

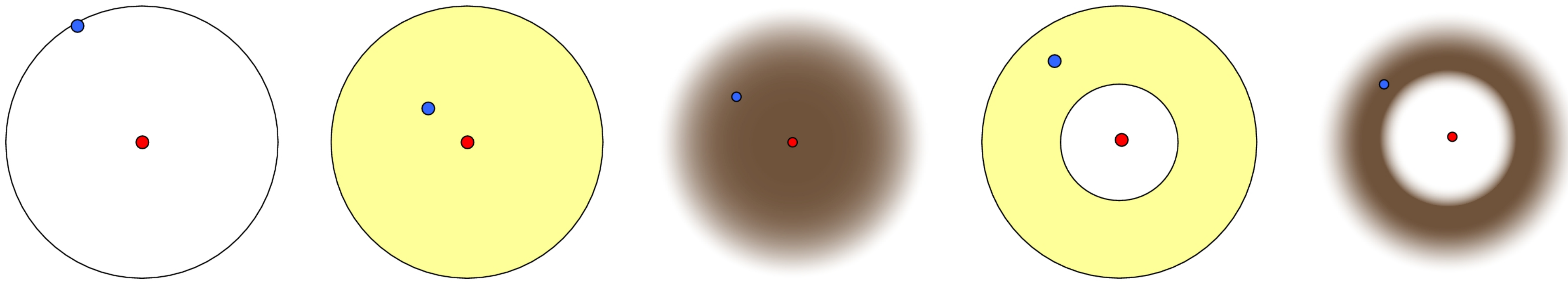
# Sources of uncertainty (4)

- **purposive** uncertainty, e.g., to ensure confidentiality




source: Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data *Advances in Medicine*, e567049. `http://doi.org/10.1155/2014/567049`

# Some techniques for anonymizing points



direction | direction & distance | Gaussian | donut | bimodal Gaussian

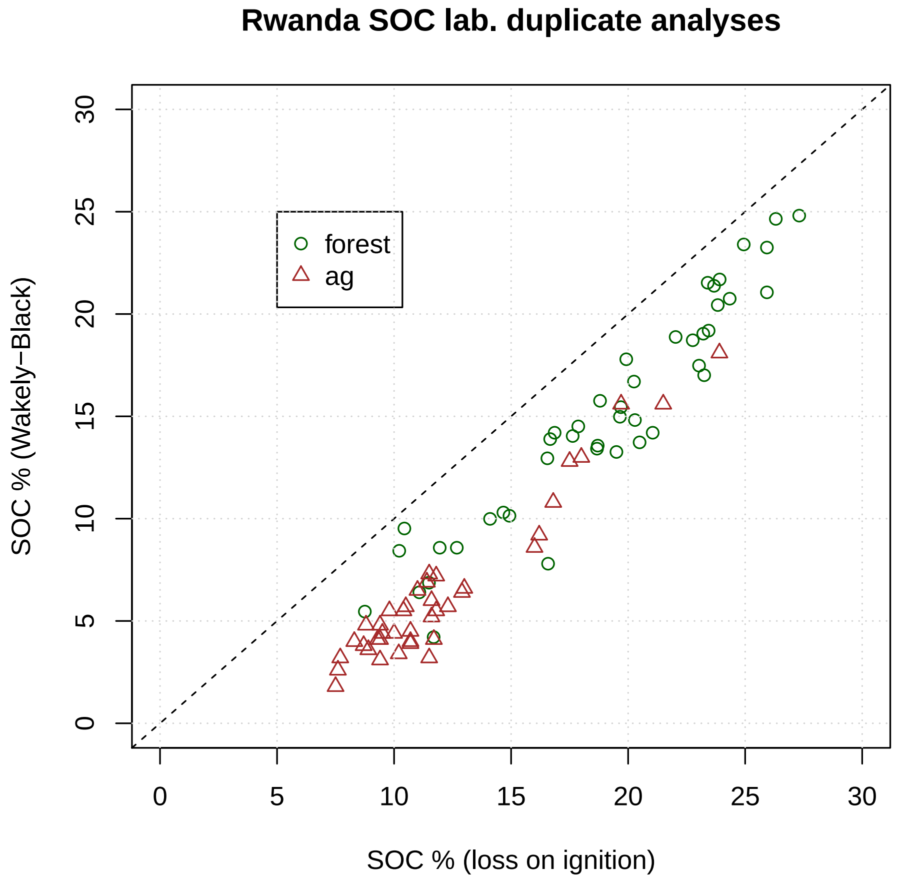
This uncertainty is known from the algorithm used and should be explained in the "lineage" section of the metadata.

# Dealing with measurement uncertainty

- **Best practices** in field, lab., transcription, data processing
- Instrument **calibration** / check against **standards**
  - quality control / quality assurance procedures
- **Exploratory data analysis** for **unusual values** ("outliers")
  - Non-spatial, non-temporal: unusual values overall
  - Spatial: unusual values in spatial context
  - Temporal: unusual values in temporal context (e.g., quality control in a process; sensor drift)
- Automated detection of unusual values by a **rule set**
  - "unusual" just means to examine the cause; it may not be an error

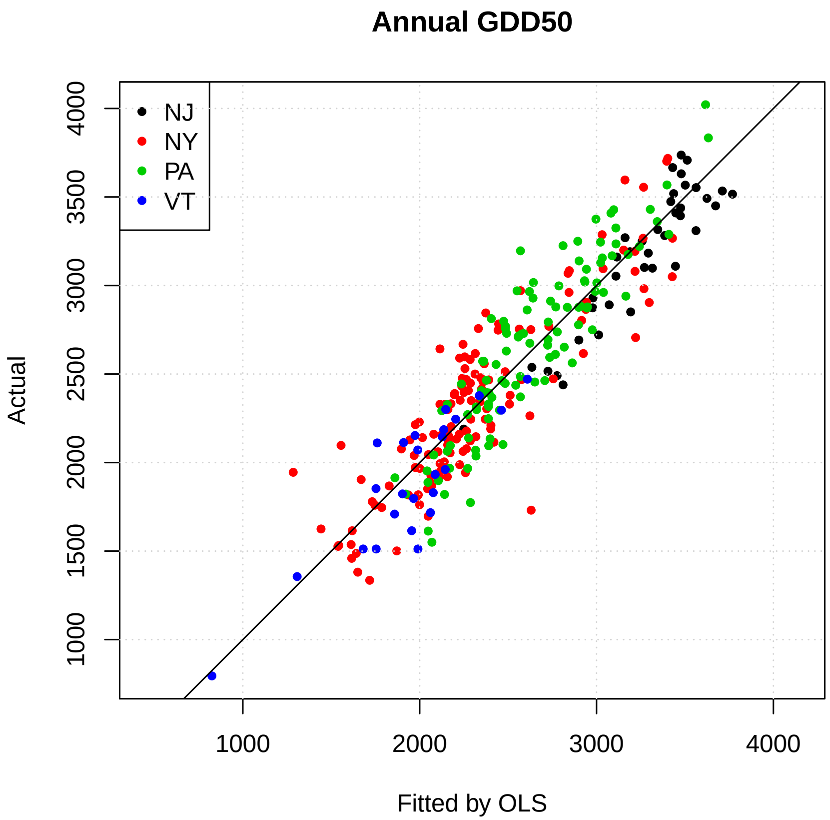
# Example of EDA

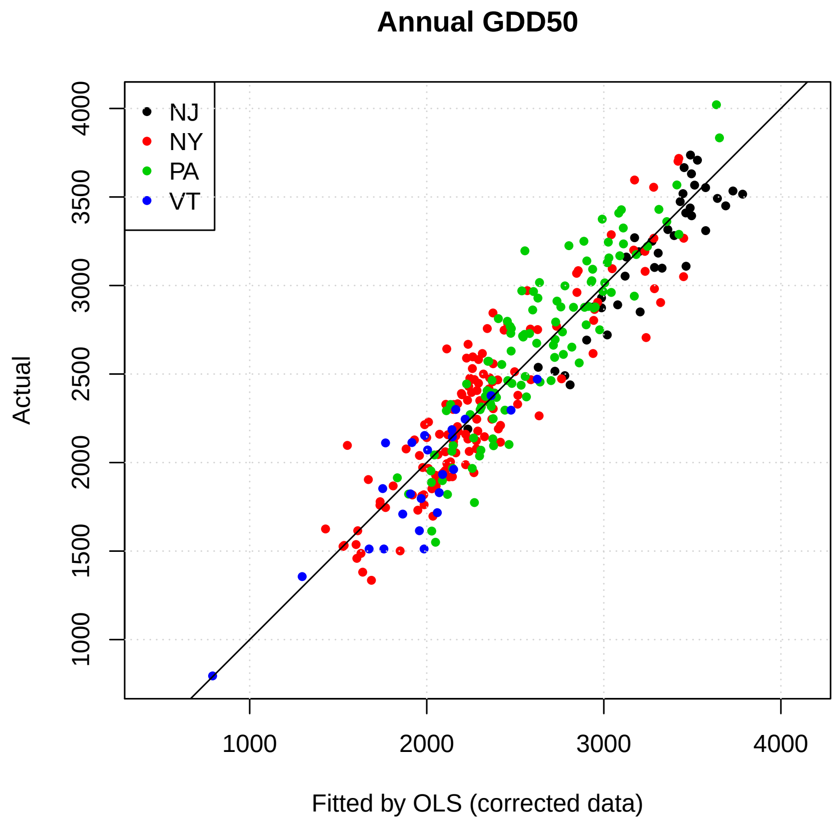
Check if two lab. methods / sample sets are consistent;
Develop transfer functions between them.




Note “forest” points at (LOI = 12, WB = 4), (LOI = 16.5, WB = 7.5)

# Unusual model residuals can reveal data problems




Note original points at ($\approx 1250$ fit, $\approx 2000$ observed), underfit, and ($\approx 2600$ fit, $\approx 1700$ observed), overfit.

We have a **well-fit model for almost all observations**; the worst fits may be good data but with some unusual circumstance; but they may be **incorrect data**

# Dealing with observation uncertainty

- Operator training / consistency checks
- Document methods, make sure they are achievable (simplify?)
- Allow **fuzzy classification** – observer records *degree of agreement* with *all* classes
  - Gopal, S., & Woodcock, C. (1994). *Theory and methods for accuracy assessment of thematic maps using fuzzy sets*. Photogrammetric Engineering & Remote Sensing, 60(2), 181–188.
  - Woodcock, C. E., & Gopal, S. (2000). *Fuzzy set theory and thematic maps: Accuracy assessment and area estimation*. International Journal of Geographical Information Science, 14(2), 153–172.
  - Laba, M., *et al.* (2002). *Conventional and fuzzy accuracy assessment of the New York Gap Analysis Project land cover map*. Remote Sensing of Environment, 81(2-3), 443–455.
- Report statistics at different levels of certainty.

# Dealing with sampling uncertainty

- If a **probability** sample, easily quantified
  - e.g., $\sigma_e^2 \approx \sigma^2 / \sqrt{n}$
- Compute **required sample size** to achieve a desired **statistical power** or **confidence interval**
  - power analysis; programs such as G*Power:`http://www.gpower.hhu.de/en.html`; also in R
  - depends on variance of the target variable
  - depends on the target parameter

# Dealing with model form uncertainty

- Check that **model assumptions** are met
  - e.g., linear models: independent and normally-distributed residuals; no dependence of residuals on fits; no spatial or temporal correlation of residuals; no excessively influential (high-leverage) residuals . . .
  - e.g. Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression. New York: Chapman and Hall.

- Attempt to reduce models to their most **parsimonious** form: the *fewest* predictors and *simplest* form to give a reasonable fit/prediction.
  - variable selection by principal components, removing colinearity with variance inflation factors, stepwise models . . .

# Dealing with model fit uncertainty

- Quantify model fit to the **calibration** (“training”) dataset
  - Amount of Variance Explained (AVE $\approx R^2$)
  - Root of Mean Squared Error of fit (RMSE): precision
  - Mean Error (ME): bias, systematic fitting error
  - Linn’s concordance coefficient, etc. (composite measures)

# Dealing with prediction uncertainty

- Quantify fit to an **evaluation** ("validation") dataset
  - Requires **independent dataset** from the **target population** to be predicted
  - Requires observations of a **probability sample** from this dataset
    * some **cross-validation** techniques – but the training dataset *must* represent the target populatuion
  - Amount of Variance Explained (AVE $\approx R^2$) against 1:1 line predicted:actual
  - Root of Mean Squared Error of fit (RMSE): precision
  - Mean Error (ME): bias, systematic fitting error

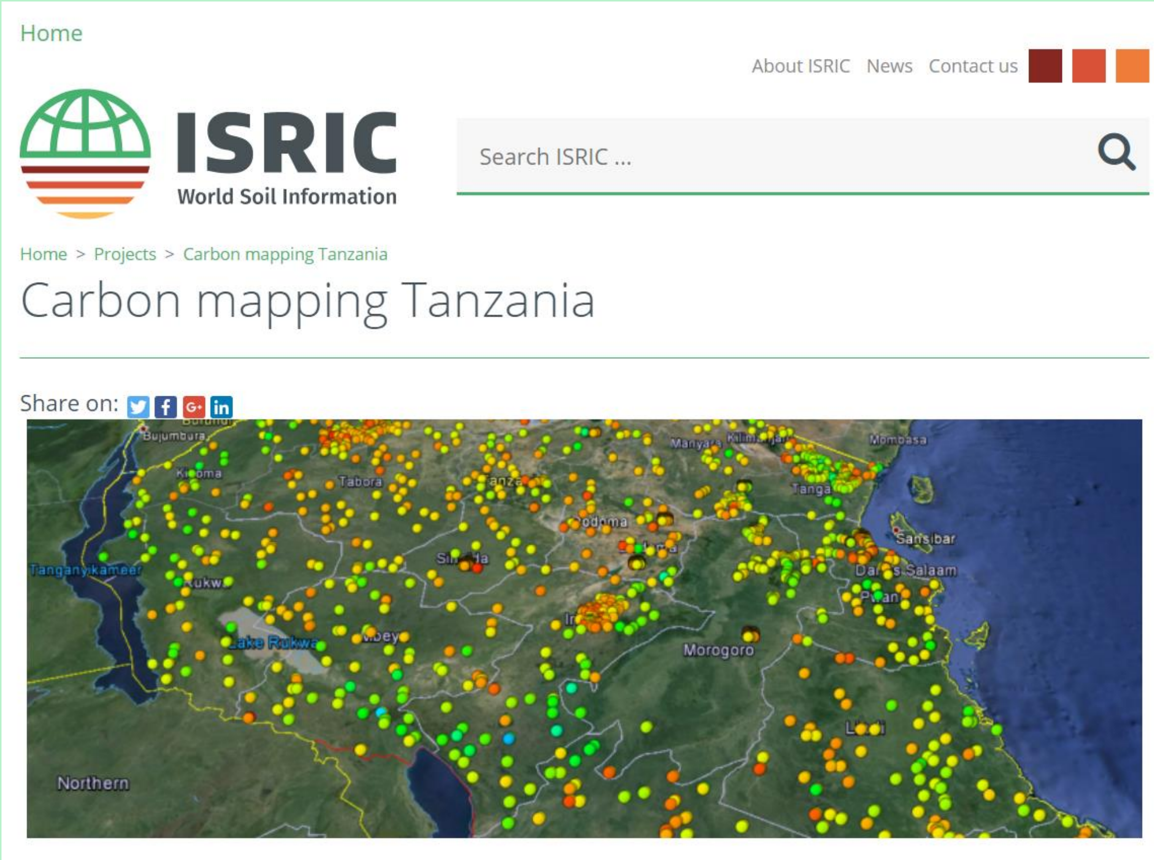
# Uncertainty in spatial models

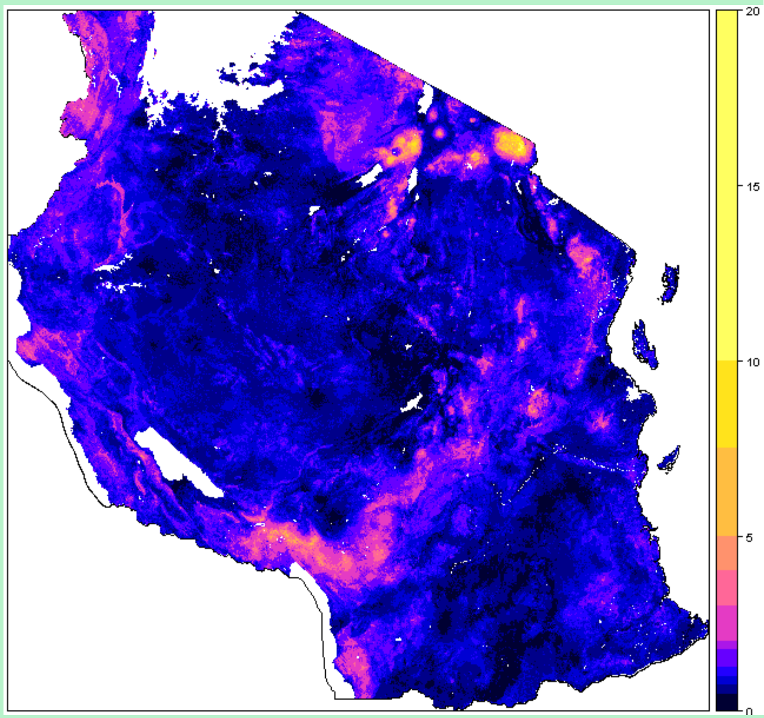
Components:

1. **Structured, non-spatial**; explainable in attribute space
   - linear, non-linear, GAM, regression tree, random forest . . .
2. **Structured, spatial**; explainable by spatial covariables (including coördinates)
   - SAR, GLS trend surfaces . . .
3. **Stochastic, spatial**; partially explainable by models of spatial autocorrelation
   - OK, CoK; with previous GLS, RK, KED . . .
   - “partially”: decreasing spatial correlation with separation
4. **Stochastic, non-spatial**: unexplainable
5. **Stochastic, spatial**: partially unexplainable
   - these two combined in the *nugget variance* of a variogram model

# Mapping uncertainty due to spatial uncertainty

## Example: topsoil organic carbon mapping Tanzania

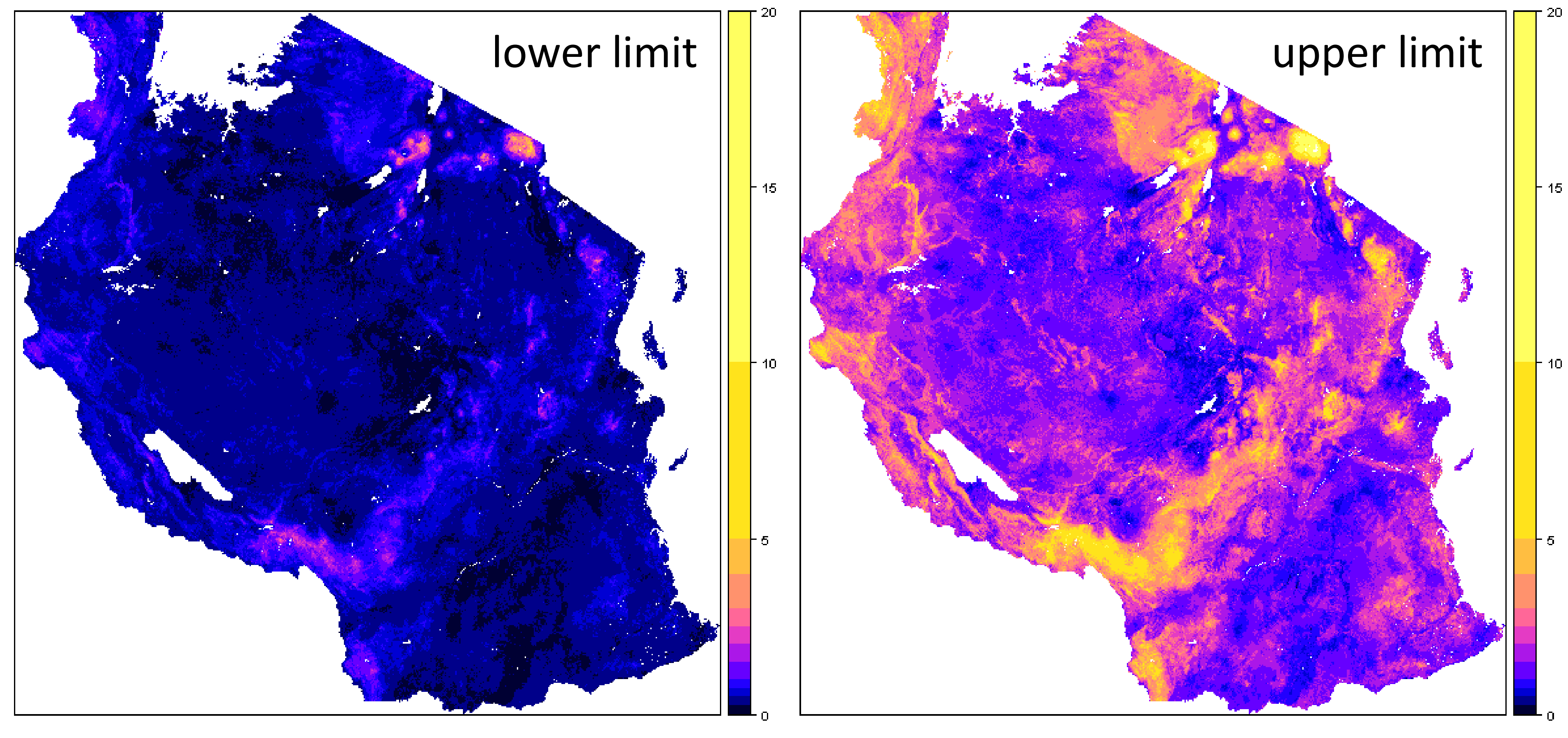



point observations

predictions by regression kriging

# Map quality quantified by lower and upper limits of a 90% prediction interval




Show both the **prediction** and its **uncertainty** (here, the kriging prediction variance).

# How much uncertainty is "too much"?

- A problem in **decision theory**
  - correct **representation** of the uncertainty
    - e.g., probability distribution of some parameter
  - **Sensitivity** of decision to the uncertainty
  - Expected **loss** due to incorrect decision due to uncertainty

- For **monitoring** or **change detection**: how much is the parameter expected to change? Is our measurement sensitive enough to detect this?

# Uncertainty propagation

Data → data manipulation → models → predictions

Heuvelink, G. B. M. (1998). *Error propagation in environmental modelling with GIS*. London: Taylor & Francis.

- Closed-form solutions are sometimes not possible; often not realistic
- Solution: Monte Carlo simulation through the entire chain, summarize results

# Example

- correct **representation** of the uncertainty
  - e.g., kriged map of probability of exceeding a defined threshold
  - e.g., kriged map of pollutant concentration; map of kriging prediction variances; combine to upper confidence level
  - e.g., statistical summary of a design-based sample of whole area, tested against $H_0 : \bar{x} > x_t$; decide based on probability of a Type 1 error

- **Sensitivity** of decision to the uncertainty
  - how far above the threshold is the prediction?

- Expected **loss** due to incorrect decision due to uncertainty
  - How expensive to clean up? How expensive if houses later have to be destroyed and residents treated?
  - e.g., famous case in Lekkerkerk (Zuid Holland)[6]

[6] https://nl.wikipedia.org/wiki/Gifschandaal_Lekkerkerk

# Topic: Assessing the effect of uncertainty

- Question: how to know if uncertainty affects decisions?

- Answer: **simulate** possible (uncertain) values and make the decision on this basis

1. Must assume the **univariate probability distribution** of the uncertain value of each model input
2. If several (partially) correlated inputs, must assume the **multivariate** probability distribution
3. Then, **sample** from this (univariate, multivariate) distribution
4. Collect the model outputs and summarize as **risk** of incorrect decisions

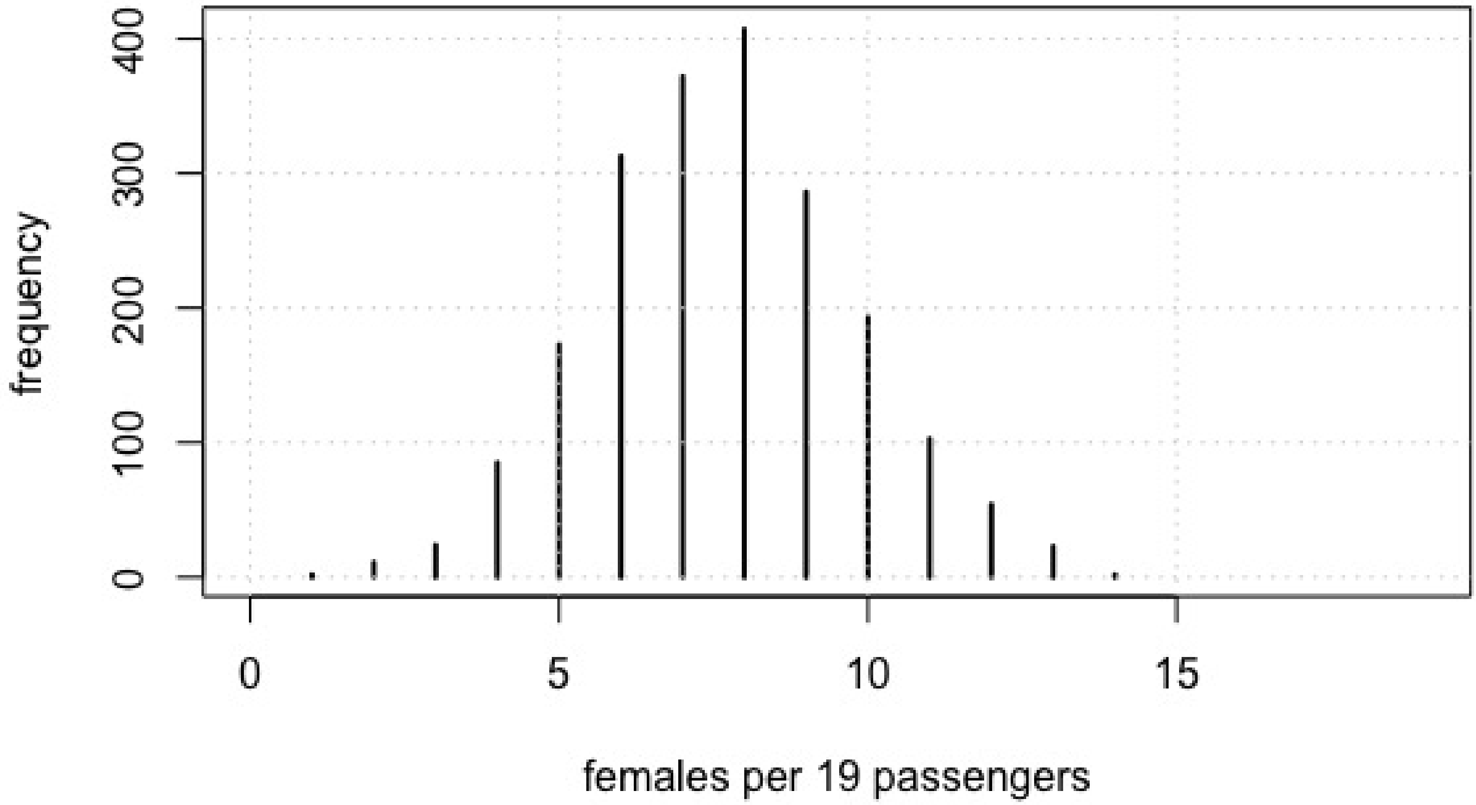
# Example: non-spatial

- Risk of an overweight airplane on 19-seat plane
- Passengers weights assumed to follow a **normal** distribution
  - Estimate mean and standard deviation from measurements from the **target population**
    - separate distributions for males/females
  - Estimate proportion of female passengers (**binomial**, estimate $\theta$)
- Random sample of 19 passengers
- Binomial proportion of females/males
- Simulate each individual's weight; sum all 19
- Compare to maximum allowable weight; find proportion overweight

# Simulation R code

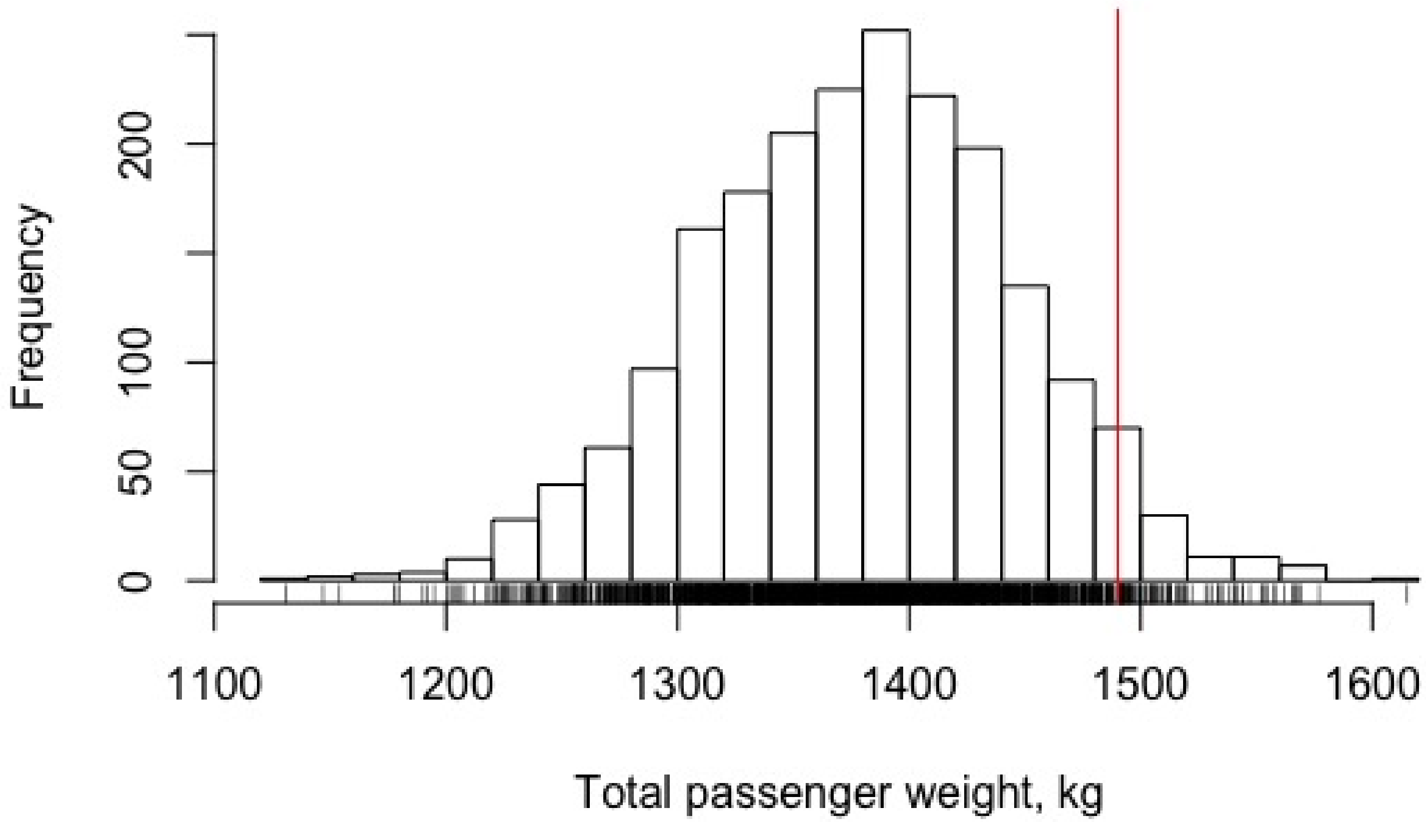
```
# parameters: mean, s.d. of fe/male weights, kg
mu.m <- 80; sd.m <- 14; mu.f <- 65; sd.f <- 12
# parameter: mean proportion of female passengers
prop.f.mu <- 0.35
# Fairchild Metro II: empty 3380 kg, max takeoff 5670kg
load.wt <- (5670-3380); pilots.wt <- 200; fuel.wt <- 600
n <- 19 # number of passengers

nsim <- 2048 # number of simulations
n.females <- vector(mode="integer", length=nsim)
wt.sum <- vector(mode="integer", length=nsim)
for (run in 1:nsim) {
  num.f <- rbinom(n=1, size=n, prob=prop.f.mu)
  num.m <- n - num.f
  wts.f <- rnorm(num.f, mean=mu.f, sd=sd.f)
  wts.m <- rnorm(num.m, mean=mu.m, sd=sd.m)
  n.females[run] <- num.f
  wt.sum[run] <- ceiling(sum(wts.f) + sum(wts.m))
  }
(n.overweight <- sum(wt.sum > (load.wt-pilots.wt-fuel.wt)))
(prob.overweight <- round(n.overweight/nsim,3))
```

# 2048 simulations; number of females




Per 19 passengers; $\theta = 0.35$.

# 2048 simulations; proportion of flights overweight 4.5%



# Key concepts

- Simulate reality: "what if?"
- Inputs are **probabilistic**
- So we need reliable **probability distributions**
- More runs → more accurate results, especially "long tails"

# Example: spatial

- Aim: see how much **positional uncertainty** in species occurrence records affects a **model** of species distribution (≈ habitat suitability)[7]

- Distribution is modelled by comparing **species occurence locations** with **spatially-distributed covariables**
  - e.g., elevation, slope, land cover, distance to ocean ...

- Occurence locations are not precise, so **randomly perturb** recorded locations $E_i$: $E_i^* = E_i + \varepsilon_{E_i}$, same for $N_i$
  - example: $\varepsilon \sim \mathcal{N}(0, 5000)$: no positional bias, standard deviation 5 km

- Then run models and compare maps – how much do they differ? in which areas?

[7] Naimi, B. *et al.* (2011). *Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling*. **Journal of Biogeography**, 38(8), 1497–1509. https://doi.org/10.1111/j.1365-2699.2011.02523.x

# Simulating the effect of spatial uncertainty

*Journal of Biogeography* **38**, 1497–1509 **1501**


B. Naimi *et al.*

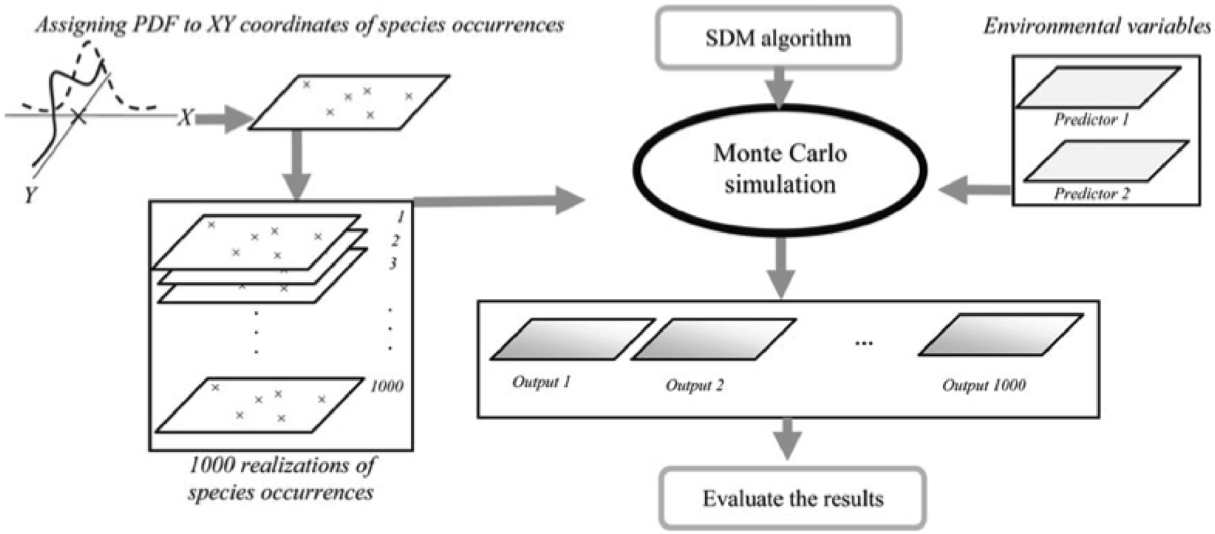



**Figure 3** Conceptual framework of species positional error propagation analysis. PDF, probability density function; SDM, species distribution model.

Repeated with different assumptions about the degree of spatial correlation

# Topic: Representing /communicating uncertainty

1. Blanket statement of accuracy and/or precision
2. Statistical reports
3. Cartographic techniques to visualize degree and type of uncertainty

Requires understanding the **psychology** of the intended reader/viewer – different cultural, educational, professional contexts and assumptions.

There are, however, universal psychological/perceptual facts.

# Example of accuracy statement

NMAS, National Map Accuracy Standards. Created in 1941, revised in 1947.

Scale dependent, 90% confidence intervals.

**Horizontal** accuracy:

> “For maps on publication scales larger than 1:20,000, not more that 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale; for maps on publication scales of 1:20,000 or smaller, 1/50 inch.”

**Vertical** accuracy:

> “. . . not more than 10 percent of the elevations tested shall be in error more than one-half the contour interval.”

# Example of statistical reports

NSSDA, National Standard for Spatial Data Accuracy, 1998

Reports positional accuracy at ground scale, and **does not set thresholds**. Users can evaluate if these are sufficent for their purposes.

> "Accuracy is reported in ground distances at the 95% confidence level. Accuracy reported at the 95% confidence level means that 95% of the positions in the dataset will have an error with respect to true ground position that is equal to or smaller than the reported accuracy value. The reported accuracy value reflects all uncertainties, including those introduced by geodetic control coordinates, compilation, and final computation of ground coordinate values in the product."

Problem: How to determine this over a whole map?

# Cartographic methods

- **Geometric simplification** (e.g., remove intermediate points in lines/boundaries)
  - Scale reduction: area → line (road, river), area → point (city)
  - Map readers understand this simplification – everyone knows a city is not a point
  - Experiment at `https://bost.ocks.org/mike/simplify/`

- **Attribute simplification**: grouping into more general categories or fewer classes
  - Example: low-accuracy detailed land cover map from remote sensing, generalize classes, should have higher accuracy

- **Visualization**: visual display of classification or continuous uncertainty

# Example: visualizing classification uncertainty

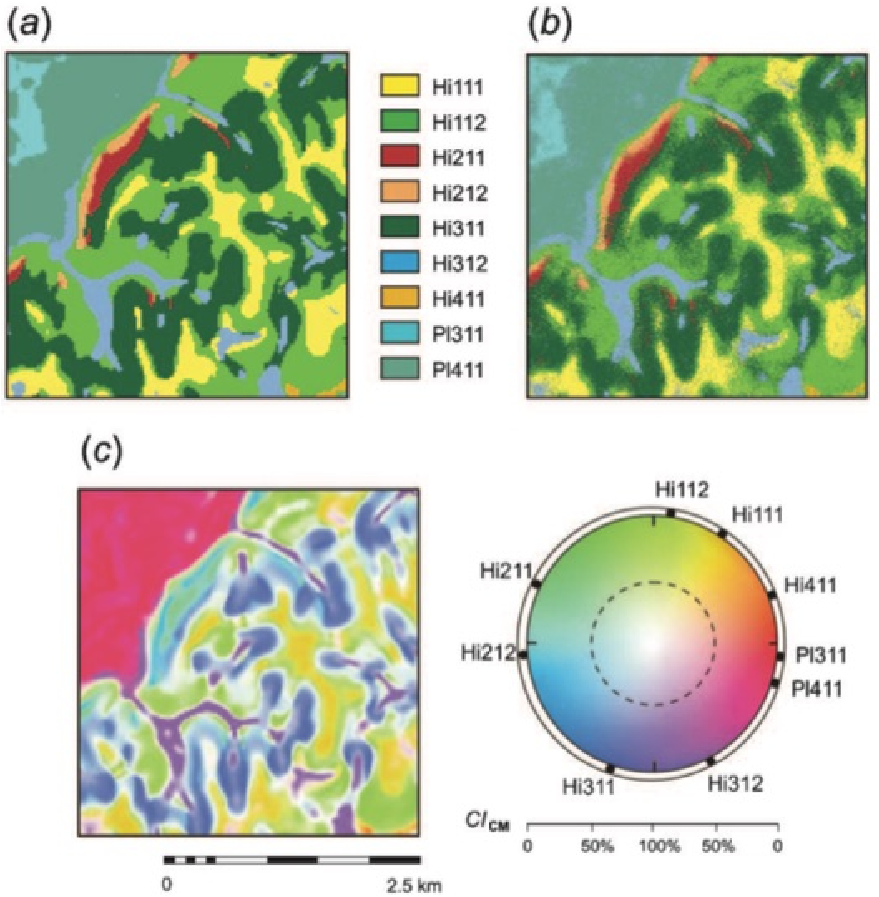



Figure 5. Comparison of different cartographic techniques: (*a*) defuzzification; (*b*) pixel mixture; (*c*) colour mixture with the circular fuzzy-metric legend.

source: Hengl, T., Walvoort, D. J. J., Brown, A., & Rossiter, D. G. (2004). A double continuous approach to visualization and analysis of categorical maps. International *Journal of Geographic Information Science*, 18(2), 183–202. http://doi.org/10.1080/13658810310001620924

# Conclusion


Uncertain world,
uncertain observations,
uncertain models . . .

Uncertain inferences,
uncertain decisions.

(Madras Crocodile Bank Trust and Centre for Herpetology)