

Data Analysis Strategy

D G Rossiter

Cornell University
New York State College of Agriculture & Life Sciences
Section of Soil & Crop Sciences

January 26, 2016

Copyright © 2016 Cornell University

All rights reserved. Reproduction and dissemination of the work as a whole (not parts) freely permitted if this original copyright notice is included. Sale or placement on a web site where payment must be made to access this document is strictly prohibited. To adapt or translate please contact the author (<http://www.css.cornell.edu/faculty/dgr2/>).

Definition

Data analysis:

Data **raw** information

OED: “Facts, esp. numerical facts, collected together for reference or information.”

Grammar: singular or plural? “The data show . . .” vs. “The data shows . . .”

Analysis extracting **interpreted** information

What does it mean? → directly useful for decision making
(helps, suggests) answers to **research questions**

Data and metadata

Data the raw materials to be analyzed

- directly measured
- observed and noted (interpretation of observer)
- transformed from some direct measurement(s)
- inferred by some model (e.g., pedotransfer function)

Metadata data about the data

- what is in the dataset?
- who created/edited/verified/distributed the dataset?
- how was the dataset created/edited/verified?
- for **geospatial** data:
 - * what are the spatial objects?
 - * what is the coordinate reference system?

Data items

- What are the **variables** and what do they represent?
- How were they **measured** in the “field”?
- What are the **units of measure**?
- What **type** of variables are these? (*next slide*)
- Which data items could be used to **stratify** the population into sub-populations for analysis?
- Which data items are intended as:
 - * **response** variables?
 - * **explanatory** or **predictor** variables?

Types of variables

- **Continuous** (with some **precision**)
 - * **ratio**: with a natural zero, ratios are meaningful
 - * **interval**: no natural zero
- **Classified**
 - * **ordinal**: ordered classes ($>$, $<$ meaningful)
 - * **nominal**: classes with no order (only $=$ meaningful)
 - * **binary**: yes/no, true/false, presence/absence ... vs. **multinomial**

Metadata

Metadata: structured information that describes an information resource:

- “data about data”
- **structure:** how it is organized, how to access
 - * including **geographic** reference (coördinate reference system)
- **content:** what it represents
- **lineage:** how it was created

Digital metadata: machine-readable (allows **data discovery**)

Geospatial metadata standards


Good introduction at <http://www.fgdc.gov/metadata>

CSDGM (USA) Content Standard for Digital Geospatial Metadata
(FGDC-STD-001-1998) <http://www.fgdc.gov/metadata/csdgm/>

ISO 19115-1:2014 (ISO)

“ISO 19115:2003 defines the schema required for describing geographic information and services. It provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data.”

Example of FGDC-STD-001-1998 metadata: dataset


Cornell University
Albert R. Mann Library

Search:

● CUGIR ● Cornell [more options](#)

CUGIR: Cornell University Geospatial Information Repository

Home About Browse Map browse Basket Help

Download now | Add to basket

Item	Value
Title	Tompkins County Land Use and Land Cover 2012
Mapsheet	Tompkins (data distributed for Tompkins Co. only)
Provider	Tompkins County ITS/GIS
Format	Shapefile .shp
Download	TClulc2012.109.shp.08162.zip Add to basket
Coordinate System	State Plane Coordinate System 1983 survey feet
Datum	North American Datum of 1983
Publication date	November 13, 2014
Date added or updated	November 13, 2014
Description	<p>Abstract</p> <p>The Tompkins County Planning Department produced a land use and land cover data set of Tompkins County, NY in 1995 and in 2009 updated the data using the 2007 natural color ortho-imagery. Further updates were made in 2012. The need for this data set has been identified by various county departments, local municipal agencies, and other not-for-profit organizations. High resolution natural color digital orthoimagery acquired from the NYS Office of Cyber Security and Critical Infrastructure Coordination (CSCIC), and a number of other secondary digital data sources (wetlands, hydrology, tax parcel and planimetric base data) were used to interpret and delineate land use and land cover directly on-screen. The land use and land cover classification system that was designed for the 1995 project to meet the needs of the users of these data was used for the 2007 data. In 1968, the Land Use and Natural Resource Inventory (LUNR) a state-wide land use and land cover mapping project, used aerial photographs to identify 130 land use and land cover classes. The Tompkins County land use and land cover classification system has been designed to be comparable with the LUNR classification system. Comparable classification systems will enable users to analyze changes in land use and land cover over the previous thirty year time period.</p> <p>Purpose</p> <p>Provides a county-wide data set of all land use and land cover classes in Tompkins County. Potential uses of these data include land use and land cover time change analysis, comprehensive planning and development suitability analysis.</p> <p>Supplemental Information</p> <p>This data set is derived from the interpretation and delineation of land use and land cover from high-resolution natural-color digital orthoimages.</p>
Topics	<i>environment, planning</i>
File size	10649 kilobytes
Use constraints	The following statement must be included with any products that use or are derived from this data set: "Data contained in this product was originally produced by the Tompkins County Planning Department and may not be reproduced or redistributed without the express written consent of the originator. The originator does not warrant the accuracy or completeness of the information portrayed by the data, as it is currently still in draft format."
Full metadata	HTML XML Text

URL: <http://cugir.mannlib.cornell.edu/bucketinfo.jsp?id=8162>

Example of FGDC-STD-001-1998 metadata: the metadata

Tompkins County Land Use and Land Cover 2012

- [Identification Information](#)
- [Data Quality Information](#)
- [Spatial Data Organization Information](#)
- [Spatial Reference Information](#)
- [Entity and Attribute Information](#)
- [Distribution Information](#)
- [Distribution Information](#)
- [Metadata Reference Information](#)

Identification Information:

Citation:

Citation Information:

Originator: Tompkins County Planning Department

Publication Date: 20141113

Title: Tompkins County Land Use and Land Cover 2012

Geospatial Data Presentation Form: vector digital data

Online Linkage: <http://cugir.mannlib.cornell.edu/bucketinfo.jsp?id=8162>

Description:

Abstract: The Tompkins County Planning Department produced a land use and land cover data set of Tompkins County, NY in 1995 and in 2009 updated the data using the 2007 natural color ortho-imagery. Further updates were made in 2012. The need for this data set has been identified by various county departments, local municipal agencies, and other not-for-profit organizations. High resolution natural color digital orthoimagery acquired from the NYS Office of Cyber Security and Critical Infrastructure Coordination (CSCIC), and a number of other secondary digital data sources (wetlands, hydrology, tax parcel and planimetric base data) were used to interpret and delineate land use and land cover directly on-screen. The land use and land cover classification system that was designed for the 1995 project to meet the needs of the users of these data was used for the 2007 data. In 1968, the Land Use and Natural Resource Inventory (LUNR) a state-wide land use and land cover mapping project, used aerial photographs to identify 130 land use and land cover classes. The Tompkins County land use and land cover classification system has been designed to be comparable with the LUNR classification system. Comparable classification systems will enable users to analyze changes in land use and land cover over the previous thirty year time period.

Purpose: Provides a county-wide data set of all land use and land cover classes in Tompkins County. Potential uses of these data include land use and land cover time change analysis, comprehensive planning and development suitability analysis.

Supplemental Information: This data set is derived from the interpretation and delineation of land use and land cover from high-resolution natural-color digital orthoimages.

Time Period of Content:

Time Period Information:

Single Date/Time:

Calendar Date: 2012

Currentness Reference: ground condition

Geospatial Metadata Tools

These allow creation, editing, and validation of metadata, according to various standards.

URL: <http://www.fgdc.gov/metadata/geospatial-metadata-tools>

stand-alone e.g., tkme and mp

on-line e.g., USGS Online Metadata Editor

embedded in a commercial program, e.g., ArcGIS 10

Repositories

- “Permanent” stores of **verified**, **documented**, and **cleaned** datasets
- Data is stored with its metadata
 - * so any researcher can directly use the data, without consulting its authors
- Required by some funding agencies

Example data repository

DATASETS
3TU.Datacentrum

RSS | FAQ | Contact | Login

Dataset: Limpopo National Park (Mozambique) Soil Organic Carbon study

Link/cite as [doi:10.4121/uuid:6cb98f84-f0de-47d4-8a2c-d6aaef5db08](https://doi.org/10.4121/uuid:6cb98f84-f0de-47d4-8a2c-d6aaef5db08) | [show link code](#) | [full citation](#)

▼ go to DATA section ▼

creator	?	Cambule, A. H. (Armindo)
creator	?	Rossiter, D. G. (David)
contributor	?	Stoorvogel, J. J. (Jetse)
date created	?	2013-06-29
date published	?	2013
description	?	410 field observations of topsoils in Limpopo National Park (Mozambique), 128 of which were analyzed by wet chemistry for ph, soil organic C, sand, silt, clay; all of which have predicted soil organic C concentration by lab. spectroscopy calibrated with lab. analysis
language	?	en
publisher	?	University of Twente
subject	?	soil organic Carbon
title	?	Limpopo National Park (Mozambique) Soil Organic Carbon study
▲ member of	?	Datasets of dissertations
spatial coverage	?	Limpopo National Park
map	?	<p>▶ Map of this location [kml]</p> <p>▶ Map including all data locations within <input type="text" value="500"/> km [<input checked="" type="checkbox"/>+circle] [kml]</p> <p><small>Note: the extra points shown may be related to completely different types of datasets.</small></p>
time coverage	?	months 2009-07 to 2009-11
related publication	?	http://www.itc.nl/library/papers_2013/phd/cambule.pdf

DATA

Dataset files (354.2 kB) >> [download complete dataset \(zip\)](#) | [download separate file\(s\)](#) [login required]

+ bag-info

+ contents of this dataset, 13 files

▲ about 3TU.DC




Home

Upload datasets

Personal page

Statistics

» Search In Data
» Search In "about"

URL: <http://www.datacentrum.3tu.nl/>

Steps in data analysis

1. Identifying the research questions
2. Identifying the sampling plan / (sub)populations
3. Examining data items
4. Data quality / data cleaning / transformations
5. Exploratory data analysis
 - (a) unusual observations and subpopulations
 - so-called “outliers”; may require return to “data cleaning” step
 - (b) possible lines of analysis
6. Selecting analytical methods
7. Modelling
8. Prediction
9. Answering the research questions

Research questions

- What **research questions** are supposed to be answered with the help of these data?
 - * \Rightarrow from the research design and **stated objectives**
- Are there **other** research questions that could be addressed?
 - * \Rightarrow motivated by the **data itself** (data exploration, data mining . . .)
 - * These might be discovered during analysis

Sampling plan

The data represent a **population** about which **inferences** should be made.

- What is the **target population**?
- Were there **subpopulations**? If so, how defined?
- What was the **sampling frame** (possible observations)?
- What was the **sampling plan** (selection out of frame)?
- What were the **sampling units** (individuals)?
- How were the selected units identified / located?
- How were the data collected in the “field”?

Data quality / data cleaning

- Purpose: check the reliability of the provided data.
- Problems:
 1. Faulty equipment / measurements
 2. Incorrect data entry / recording / transfer
 3. Inconsistent measurement methods (operators, calibrations ...)
- Methods:
 - * Chain of custody (lineage)
 - * Reality check (reasonableness vs. expected values)
 - * (Semi-)automatic identification of **outliers**

Data transformations

Many analytical methods are more **stable** when variables have certain **distributions**.

Most make **assumptions** about the data distribution, e.g.:

- Parametric (Spearman's) correlation assumes bivariate normal distribution of the two variables.
- Ordinary Kriging assumes a spatially-correlated Gaussian ("normal") random process

Almost always **extreme** values will have undue **influence** on inferences

Solution: **transform** variables to more-or-less **symmetric** distributions.

Then work with the transformed variables; inferences are about these.

Some transformations

- Variable left-limited (e.g. by 0) and right-skewed: **square root** transform
- Strongly-skewed: **logarithmic** transform
 - * If includes 0, shift by measurement precision
- Heteroscedastic (uneven variance): **Box-Cox** transform (variance-stabilizing)

$$\begin{aligned} z &= \frac{y^\zeta - 1}{\zeta}, \zeta \neq 0 \\ &= \ln(y), \zeta = 0 \end{aligned} \tag{1}$$

- Irregular
 1. **normal-score** transform
 2. Hermite polynomials → normal distribution
 3. continuous → categorical by slicing

Exploratory data analysis (EDA)

Purpose: know what is in the dataset

- **Summary graphs** appropriate to the data type
 - * Univariate: stem-and-leaf, histograms, boxplots ...
 - * Bivariate: scatterplots
 - * Multivariate: 3D scatterplots; classified boxplots ...
- **Univariate** descriptions: summary statistics
- **Bivariate** descriptions: cross-tabulations

Identifying unusual observations and subpopulations

Check whether all the data belong to the same **population**

- Are there **subpopulations**? How can these be identified?
- Are there **“outliers”**?
 - * better called “unusual/unexpected/inconsistent observations”
 - * Is there evidence that these are from a different **process**?
 - * Should they be **excluded**? On what basis?
 - * Should the definition of the **target population** be changed?
- If “outliers” are not errors and are part of the population, consider analyzing separately

Censored measurements

“Censored”: the value of a measurement is only partially known because of imprecision or limitations

- “at least” or “at most”
- e.g., depth of soil to bedrock, but only measure to 2 m (limitation of probe)

Approaches:

1. Impute values beyond measurement range by fitting a distribution
2. Analyze values at limit as binary (yes/no at limit) and others separately

Selecting analytical methods

1. Based on **research questions**
2. Based on **type** of variables
3. Based on **EDA**

Sources:

- **Textbooks** (discuss relative merits, applicability)
- **Journal papers** on the same research question

Example: selecting explanatory models

These for the typical “explanation-response” models

1. Based on the **explanatory** variables

- All continuous: **regression**
- All categorical: **ANOVA** (Analysis of Variance)
- Mixed: **Analysis of Covariance** (ANCOVA)

2. Based on the **response** variable:

- Continuous: **linear** models for regression, ANOVA or ANCOVA
- Proportions: **Multivariate logistic** regression
- Binary (yes-no): **Bivariate logistic** regression
- Counts: **Log-linear**
- Time to a given event: **Survival** analysis

Modelling

A **statistical model** summarizes the data for **explanation** or **prediction**.

Models depend on the research questions. Some possibilities:

- **Univariate** tests of **fits** to theoretical distributions
- **Bivariate** relations between variables (correlation)
- **Multivariate** relations between variables: partial correlation, principal components, factor analysis . . .
- Analysis of Variance (ANOVA) on predictive factors (confirms **subpopulations**)

Models are just ... models!

George E.P. Box & Norman R. Draper, *Empirical Model-Building and Response Surfaces*, Wiley (1987)

p. 74: “Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

p. 424: “Essentially, **all models are wrong, but some are useful.**”

So, how do we decide if the model is not “too wrong” and is “useful”?

- Match model form with known or hypothesized **process**
- Internal evidence from **model fit** and **diagnostics**

Competing models

1. the **null** model: everything is modelled by the overall mean
2. the **minimal adequate** model: explains “enough” of the variability
3. the **current** model
4. the **maximal** model: using all levels of all predictors
5. the **saturated** model: as many parameters as data values!

Choosing among competing models

Einstein: “A model should be as simple as possible. But no simpler.”

- **Theoretical** form of model; based on known or hypothesized **process**
- Greatest **predictive** power
 - * **Calibration** (goodness-of-fit) vs. **evaluation** (often called **validation**)
 - with an **independent** dataset that represents the target population
 - * **cross-validation**, “bootstrapping” from the calibration dataset
- Most **parsimonious** model (evaluate with AIC, adjusted R^2)
- “Best” **goodness-of-fit** to data
 - * Analysis of variance of **hierarchical** models (increasingly complex)
 - * **stepwise** procedures (supervised!)

Parsimony

- Prefer **simpler** explanations
- But, add **complexity** if it adds significantly to the explanatory (or predictive) power
- Conversely, simplify complex models until too much explanatory (or predictive) power is removed
- Prefer **simpler** model forms
 - * e.g., **linear** models are easy to fit and interpret; try to **linearize**
- Prefer models with **fewer assumptions**, and **test** the reasonableness of these
 - * e.g. for **linear** models: residuals are IID normal; no trend of residuals with data; no difference in variance with subpopulations or by data value . . .

Prediction

An objective of the study may be to make **predictions** based on a developed **model**:

- **Response variables** for unsampled individuals in the population
- Predictions about the **future**
- Predictions about **related populations**
 - * **Interpolation** vs. **Extrapolation**

How **accurate** are the predictions? Is this adequate for the purposes of the study?

- Evaluate with **prediction intervals**
 - * e.g., in linear models or Ordinary Kriging prediction variance
- Evaluate with **independent dataset**

Answering the research questions

This is the reason we **analyze** data!

- To what extent do the data and methods applied **answer** the research question?
- Are **new** research questions suggested by the analysis?
- Are **more data** needed? If so, how many and where?
- Are **further analyses** indicated? Which?
- Would **other methods be appropriate**? Why or why not? If yes, which?

And of course . . .

- what **new questions** do the results suggest?
- what ideas for **further research**?

Geospatial data

If the **location** of observations in some **space** is known, this is a **spatial** data set.

What is special about spatial data?

1. All data sets from a given area are **implicitly related** by their coordinates
2. Values at sample points are often **dependent**
3. That is, there may be a **spatial structure** to the data
 - Classical statistics assumes independence, at least within sampling strata
 - Major implications for sampling design and statistical inference
4. Data values may be related to their coordinates → **spatial trend**

Exploratory spatial data analysis

If the data were collected at known points in geographic space, we should visualise them in that space.

- **Postplots**: where are which values?
- **Geographic postplots**: with images, landuse maps etc. as background: do there appear to be any explanation for the distribution of values?
- Spatial **structure**: range, direction, strength . . .
- Is there **anisotropy**? In what direction(s)?
- Populations: do there appear to be geographically-compact **zones** with different processes (e.g. form of spatial dependence)?

Spatial modelling

If the data were collected at known points in geographic space, it may be possible to model this. Again, this depends on the research questions.

- Model the spatial structure
 - * **Local** models (spatial dependence)
 - * **Global** models (geographic trends, feature space predictors)
 - * **Mixed** models
- **Point-patterns** (spatial distribution of observations)
- **Directional** analysis

Spatial Prediction

An objective of the study may be to make spatial predictions, e.g. a **map** over the population area.

- Values at **points or blocks**
- **Summary** values (e.g. regional averages)
- **Uncertainty** of predictions