Aseguramiento de Calidad de la Información Geoespacial

Guia didactica

22-27 septiembre 2019

INEGI, Aguascalientes, Mexico

Instructor: Nicholas Chrisman, PhD

Objectivo :

Al finalizar el curso, los participantes conocerán los conceptos, fundamentos y mejores prácticas sobre la calidad de la información geoespacial, para poder instrumentarla en su trabajo.

Temario

Aseguramiento de cualidad de la información geoespacial

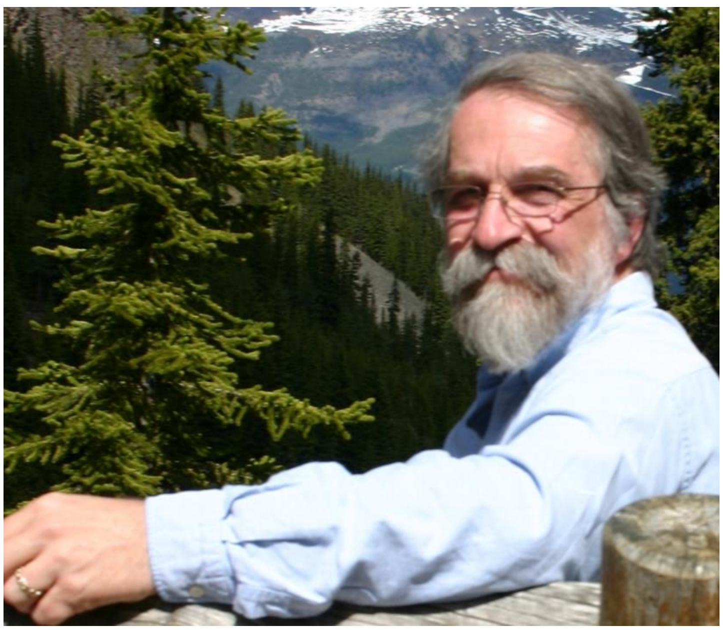
1. Normas internacionales de calidad geoespacial: indicadores y metricas
2. Precisión posicional;
3. Precisión de clasificación, exhaustividad;
4. Consistencia lógica; (Estructuras de datos topológicas)
5. Aspectos estadísticas y métodos de muestreo

Cada tema se pone sus materias y referencias bibliográficas, páginas siguientes (en ingles).

Each theme (and subtheme) is described below on a separate page with materials and references attached (details in English). There is a web resource page with access to online materials at http://www.nchrisman.fastmail.fm/Present/INEGI_DQ_2019.html

But first, who is Nick Chrisman? (Quien es el doctor Chrisman?)

Nicholas Chrisman is currently Editor of the journal Cartography and Geographic Information Science, based in Bellingham, Washington. For 2013-2014, he was Professor of Geospatial Science and Discipline Head at RMIT University in Melbourne, Australia. From 2005 to 2012, he was Professor of Geomatics Sciences at the Université Laval, Canada with his principal assignment as Scientific Director of the GEOIDE Network. His research has concentrated on time in GIS, data quality testing and the social and institutional aspects of Geographic Information Systems. From 1987-2004 he was Professor of Geography at University of Washington. From 1982-87 he was Assistant Professor in the Department of Landscape Architecture at the University of Wisconsin-Madison. During the 1980s, he was in charge of the Working Group on Spatial Data Quality of the National Committee for Cartographic Data Transfer Standards. From 1972-82 he was a programmer at the Harvard Lab for Computer Graphics. He participated in the design of prototype GIS software. His PhD is from the University of Bristol (UK) for research on error and statistics applied to categorical maps. For many decades, his writing has tried to connect the technical details of GIS to larger issues of philosophy and culture.



Ensuring spatial data quality

- What can be checked?
- Have you checked it?
- Have you fixed it? (and why not…)?

Elements of Data Quality
(date back to work for National Standards for US 1980s)
Now ISO / OGC

5 Basic components

- Lineage
- Positional Accuracy
- Attribute Accuracy
- Logical Consistency
- Completeness

(Temporal Accuracy, also an issue but often not separately testable).

Issues: metadata is a fine idea, but useless if the fields are all blank.
Resources required (time, expertise, etc.) to ensure quality.
Data quality varies spatially; to record data quality for a map, you need a map (a meta-map?)

Resources/ bibliography
ISO standard
VMAP standard (coverages for data quality)
Chrisman AUTO-CARTO 6
Chrisman data quality chapter in Big Book

Report, FrontierSI (Australia) Upgrading the Spatial Accuracy of the Digital Cadastre – A Pilot Study (pdf)
USGS completeness levels for National Hydrological Network (NHN) (pdf)
Devillers, R, Stein, A, Bédard, Y, Chrisman, N, Fisher, P, Shi, W. 2010: 30 years of research on spatial data quality – Achievements, failure and opportunities. Transactions in GIS 14 (4) 387-400, doi: 10.1111/j.1467-9671.2010.01224.
Chrisman, N.R. 2005 : Chapter 1 Development in the treatment of spatial data quality, p. 21-30 in Devillers, R. & Jeansoulin, R. (eds.) Fundamentals of Spatial Data Quality, p. 21-30, Chapter.
Chrisman, N.R., 1991: The error component in spatial data, Chapter 12, p. 165-174 in Maguire, D.J., Goodchild M.F. and Rhind, D.W. (editors) Geographical Information Systems: Overview Principles and Applications, Longmans
Chrisman, N.R., 1983: The role of quality information in the long-term functioning of a geographic information system, Proceedings AUTO–CARTO 6, 2:303-321.
https://cartogis.org/docs/proceedings/archive/auto-carto-6/pdf/the-role-of-quality-information-in-the-long-term-functioning.pdf
.. /G460/Lec17.html
Web pages of Geog 458, (past course online) www.nchrisman.fastmail.fm/Geog_458.html

Basic concepts to understand data quality measurement (Day 1, Monday)

1. Confusing terminology…
    a. Uncertainty (at what level?)
    b. Error (in statistical sense, and in human performance)
    c. Accuracy/ precision; repeatability/ resolution
    d. 'relative' accuracy versus 'absolute' (how absolute is it?)
    e. Misunderstandings due to imperfect translations (from statistics and other foreign languages) – In Spanish: precision, exactitude, and other words. Let's use these words carefully.
2. Three kinds of error – with three different treatments
    a. *Systematic error* (bias) – treatment : calibration, remove the differences
    b. *Random error* (unavoidable variation) – treatment : predictive models (least squares)
    c. *Blunders* ('outliers', massive mistakes often human caused, far beyond the typical random bounds) – treatment : detections and removal (use robust techniques that do not get deflected by outliers)
    d. What happens if these three are not kept separate? Random variation is treated as much larger. Estimates may be biased.
    e. *Propagation of error:* each step in a process may introduce error, how to reduce the impact as it accumulates in the final product.
3. "Terrain nominale" (French term) very badly translated into some standards as 'abstract universe' (!!) – how to see the world through the lens of the established definitions and procedures
    a. practical example 1: to test the accuracy of position for a street intersection, you need to decide on the rule for center of the two roads and their angle of intersection. Oblique angles can lead to somewhat unexpected results.
    b. practical example 2: careful definition of objects. Does the parking area for an object (hospital/school/park…) belong to the object or to the street/transportation ?
4. Requirement to correct errors
    a. Policy choice of institution: how critical is the error?
    b. How do corrections get transmitted?
    c. Does the correction increase the chance for relative errors locally?

International standards for Data Quality in geospatial disciplines

- Map Accuracy Standards (dating back to early photogrammetry) USA 1947
- 1980s: National Committee for Digital Cartographic Data Standards, led to European committee CEN then ISO

Metadata reporting requirement: 5 components

1. Lineage
2. Positional accuracy
3. Attribute accuracy (or semantic accuracy)
4. Logical consistency
5. Completeness (exhaustivity)
6. (plus temporal accuracy)

Positional accuracy (Day 2, Tuesday)

- Positional Accuracy Handbook, 1999, Minnesota Land Management Information Center. pdf.
- National Standard for Spatial Data Accuracy 1998.
- Do not confuse (or conflate) : resolution and accuracy!

Major issue:

- Sampling: obtain unbiased estimate of average conditions; must apply to the same kind of object (well-defined points)
- Exhaustive inventory: Estimate accuracy by feature class (not all 'well-defined')

Field Work : How to apply the concept of « terrain nominale »

- Field trip (at least conceptually) (Testing various sources: Google, Bing, OSM, INEGI, and others)
- Consider how the specifications for a street (caretera) will change what point to test. Center-line? Edge of pavement?

Attribute accuracy (Classification) – Day 3, Wednesday

This aspect is best developed in the remote sensing field.
Classification accuracy is assessed by a square matrix comparing the product (rows) to 'ground truth' (columns).
Correctly classified pixels are in the diagonal.
Percent by row is 'producer's accuracy'. By column, 'user's accuracy'.

An example

For objects (not exhaustive rasters), there is an additional row and column - not found/nonexistent. Nothing can be in the diagonal cell for the misclassification matrix (clearly).

IGN France did an exhaustive sample in a limited area of the maps tested. ALL objects were assessed for their classification on the map and in the field (according to terrain nominale).

For example, road classes: highway, street, dirt road, footpath, NONEXISTENT in a square matrix. It showed the error rate where some highways were coded as 'street', but very few as 'dirt road'.

Completeness (Exhaustivity)

Completeness is measured by the 'extra' row and column. Did all objects appear in the product?

The cartographic scale of a product dictates some issues of completeness. Certain classes of features are not expected (no dirt roads at 1:100,000?).

A direct test of completeness requires an external database of all features that should be on the map. In some applications, there is such a list (eg. parcels in a tax register, inventory of roads for maintenance), but the identifiers need to match or some geocoding is required.

Logical Consistency (Day 4)

Not a test against external sources, but using internal structure of the data to test itself.

Simple level: legal codes

Are all feature codes in the specification? (Remove illegal codes).

Next level: consistency with other elements

- Are 'mountain' categories found at low elevation? (High mountain vegetation code at 5m elevation? Very unlikely…)
- Navigational buoys on land? Stop lights in the water?
- Is the river coded as outside the floodplain?

Geometric test: topological structure

The most powerful test of logical consistency involves the geometry of a map using the topological data structure. This test was first implemented by the US Census Bureau prior to the 1970 Census. Back then the geometric databases were crude, and incomplete. Adopting a topological data structure permitted them to test for completeness and geometric integrity.

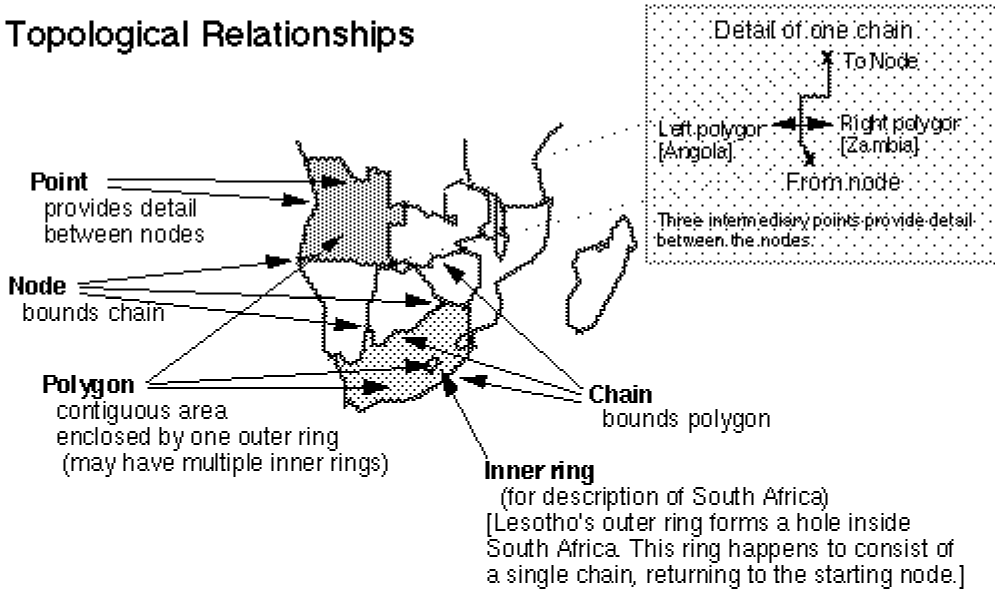
- Do all lines intersect only at established nodes?
- Do all polygons close properly?
- Are any lines missing?
- Are all lines labelled correctly for polygons to left and right?

A bit of history of the DIME and TIGER developments at US Census Bureau:

- James Corbett and the Euler network
- Don Cooke and Bill Maxfield publish paper about DIME: Census Use Study, New Haven
- DIME: hand coded from maps, key punched (Geographic Data Technology)
- ARITHMICON software (Corbett and Marvin White)
- TIGER – use USGS topographic maps (1:100,000), cooperation Census and mapping agency
- Conflation: adding attributes from one source to better geometry from another

Topological Data Structures

A longer issue in GIS and cartography- capturing the fundamental spatial relationships
Connectivity- connectedness
Containment - Boundary (recursively)



Resources:

Peucker, T.K. and Chrisman, N.R., 1975: Cartographic data structures, American Cartographer, 2:55-69.

Vector Product Format (a standard for geospatial data) – NATO/US Military (developed for Digital Chart of the World project) MIL-STD-2407

Experience of US Census Bureau in development of topological data structures: DIME & TIGER

.. /G458/lec15.html

.. /G460/Lec06.html

Day 5: Statistical issues and Sampling methods (Conclusion)

A question of intent and audience:

- is the data quality for internal uses?
- or for the users?

Questions of sampling bias – classic methods exclude the less-well defined points that may be the real test for some users (particularly for environmental analysis).

Conclusion, summary, further explorations

Gaps in the standards: still elements to evolve...

Entering metadata is very slow and incomplete unless it is directly tied to the data entry procedures.