DCN Planetears

Geospatial Data Curation:
an introduction
Module: Ethics and GIS Data

Module Objectives: Ethics and GIS Metadata

This module has two objectives. At the end of the module, learners should be able to:

1. Identify 3 examples where there would be ethical concerns with sharing geospatial data
2. Identify 3 recommendations for action when there are ethical concerns

The module has Lecture and Activity components to help reinforce new information.

Ethical checks

- Re-identification
- Sensitive data (beyond Protected Health Information)
- Permission to share

Re-identification

Examples of potential concern:

- Protected health information (PHI) tied to geographic areas with populations of less than 20,000
- Data about traffic accidents (combined with news articles and police reports)

Re-identification Prevention: Real Life Example

C
U
R
A
T
E
D

1) May 2023 a dataset that recorded traffic crashes, trauma center location, patient demographic metadata, and, alcohol and drug use, was submitted to a repository.

2) During the CURATE(D) Check step, the curator became concerned that patient demographic metadata and the crash and/or trauma center location data could lead to re-identification when paired with publicly available news reports or police reports of crashes.

3) The curator read through the final report, and realized that a number of demographic variables that the trauma centers collected were not analyzed by the researchers and were not needed to support research conclusions, and were not needed for reproducibility or replication.

4) Curator suggested researchers remove the variables **Year**; **Month**; **Race**; **Ethnicity (Hispanic)**; **Comorbidities**; and **Complications** to greatly reduce re-identification risk.
   - In states such as Iowa, the non-white population is a small percentage of the population, and the **Race** and **Ethnicity** variables could lead to easier to re-identify people, some who could become targets for discrimination or deportation.
   - Some medical **Comorbidities** or **Complications** are quite rare, and could lead to re-identification, and possible insurance or genetic discrimination.
   - The inclusion of **Month** and **Year** of crash could help to match event to news reports.
   - Taking all of these variables together, a re-identified driver or crash victim who was using prescription or non-prescription drugs or alcohol, or were or were not wearing a seatbelt or a helmet at the time of the crash could find themselves discriminated against for employment, health coverage, housing, and other social needs.

5) The curator also suggested that while the **Age** of the people involved in the crashes (as driver, passenger, collision victim, etc.) was binned to reduce re-identification, the age bin 18-20 was too small (3 years) while other bins were 10 to 20 years wide.

Recommendations when re-identification is a concern

1) Remove unnecessary attributes that are not part of the analysis
2) Aggregate data to geographic units that have populations of more than 20,000 people or remove geographic identifiers for non-compliant areas
3) Place data in embargoed state for sharing beneath restrictive agreements

Sensitive Data (beyond PHI)

Examples of potential concern:

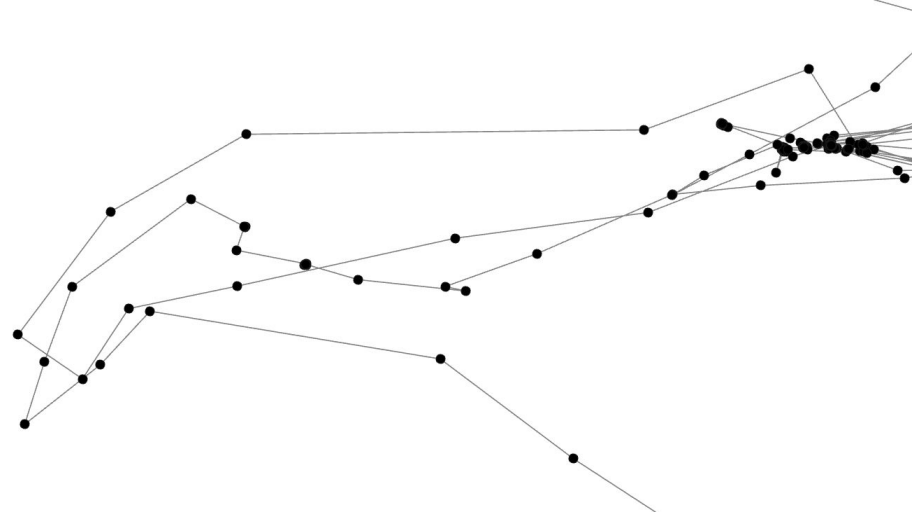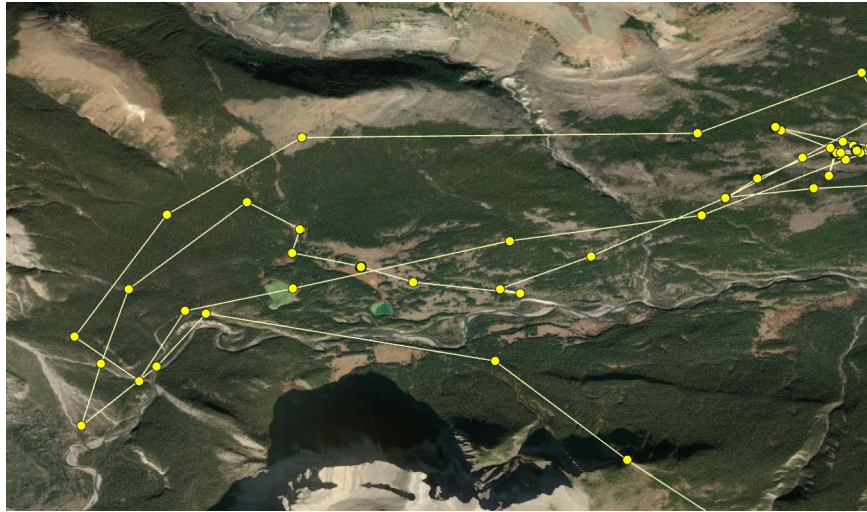- archeological sites
- movement of endangered species
- resources on lands of indigenous communities

Suggest that the researcher obscure or randomize spatial information

1. Offset geospatial locations
2. Aggregate data to a larger area
3. Introduce noise or randomly distribute points within a larger area

Offset geospatial locations

Image credit: Melinda Kernik

Recommendations when sensitive locations are concern

Aggregate data to a larger area

Recommendations when sensitive locations are concern

Introduce noise or randomly distribute points within a larger area
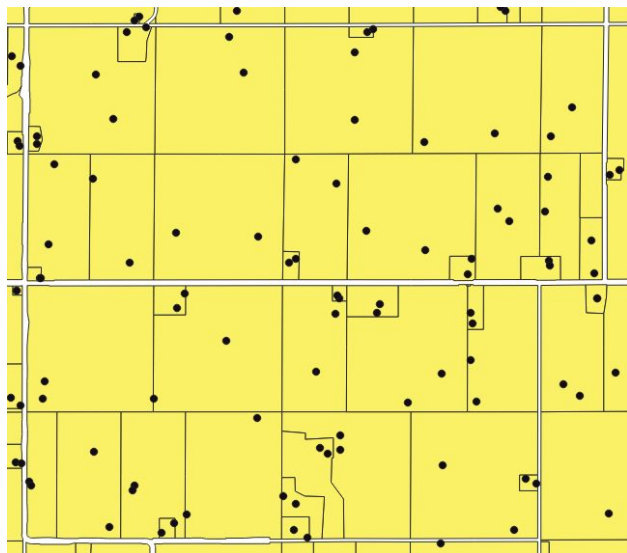
Original locations
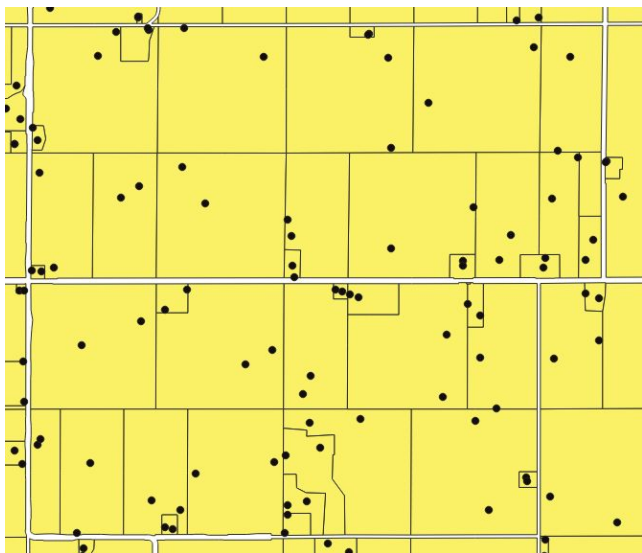
Moved randomly within the analysis polygons

Image credit: Melinda Kernik

1) Suggest that the researcher obscure or randomize spatial information
2) Suggest that the data be submitted to a repository with restricted access (or split out sensitive content as a separate deposit)
3) Request evidence that the researchers consulted with communities represented or affected by the data (signed permission, governance mechanisms, etc.)

Permission to share

Examples of potential concern:

- Multiple data layers from outside sources combined together to identify the best location for conservation actions
- Columns appended to a spreadsheet of information collected by another graduate student in a lab

1) Check that the authors have permission to share the data
   - Data use agreements
   - Licenses
2) Remove data for which permission has not been granted and replace it with information about how to acquire it
3) Add additional authors and attribution as needed

Resources for later

- [De-identification of Protected Health Information in Accordance with HIPAA](#)
- [Human Subjects Data Essentials Data Curation Primer](#)
- [CARE Data Principles, Indigenous data, Data related to Indigenous Peoples and Interest](#)

Ethics Discussion

1. Have you encountered any situations working with sensitive GIS data?
2. What were your ethical concerns and what actions did you end up taking?

Ethical. Reusable. Better.   **DATA CURATION NETWORK**   datacurationnetwork.org