Early findings from a cross-domain need-finding study with users of geospatial data

Programming Systems Seminar // April 7, 2022
UC Berkeley EECS

Parker Ziegler // @parkie-doo
Acknowledgments

Sarah E. Chasins
Ph.D. Advisor, Principal Investigator

Rachel Leven
CDSS Media Communications Specialist

PLAIT Lab
Kind, thoughtful friends who let me rant about maps, KEXP, and the Red Sox

My Participants
Computing tools are built for those who participate!
A Note on Feedback

1. Put on your \{\text{CHI, UIST, OOPSLA, PLDI}\} reviewer hat!

2. Pay special attention to the Motivation section
What’s this talk about?

Sharing preliminary findings from a need-finding study with users of geospatial data.

25 participants
3 domains of interest
30 hrs 16 mins of video from observational interviews
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
What is geospatial data?

Tabular Data
What is geospatial data?

Data in which **location** is encoded alongside **attributes**

<table>
<thead>
<tr>
<th>GEOID</th>
<th>STATE</th>
<th>TOTAL_POP</th>
<th>NUM_SHOOTINGS</th>
<th>TOTAL_AREA</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>AK</td>
<td>733,891</td>
<td>13</td>
<td>1,477,953.4</td>
</tr>
<tr>
<td>40</td>
<td>CT</td>
<td>3,524,076</td>
<td>18</td>
<td>3,524,076.1</td>
</tr>
<tr>
<td>30</td>
<td>MI</td>
<td>1,527,979</td>
<td>6</td>
<td>1,527,979.6</td>
</tr>
<tr>
<td>15</td>
<td>KY</td>
<td>4,395,356</td>
<td>41</td>
<td>1,606,385.3</td>
</tr>
<tr>
<td>16</td>
<td>LA</td>
<td>4,517,562</td>
<td>63</td>
<td>4,517,562.4</td>
</tr>
</tbody>
</table>
What is geospatial data?

Data in which **location** is encoded alongside **attributes**

```
bbox: [
  [44.44, 40.49],
  [44.62, 40.45],
  [44.80, 40.49],
  [44.62, 40.53]
]
```
What is geospatial data?

**Location** is encoded in reference to a **coordinate system**

All coordinate systems trade-off between **size**, **shape**, and **distance**

Andy Woodruff’s projected heads, based off Charles Deetz’s 1921 book *Elements of Map Projection with Applications to Map and Chart Construction*
What is GIS?

Geographic Information System
What is GIS?

Geographic Information System

Software for **viewing**, **operating on**, and **managing** geospatial data.
What is GIS?

Zoomable, pan-able canvas for viewing geospatial data as layers

QGIS
What is GIS?

Processing tools for operating on geospatial data to produce new layers
What is GIS?

Attribute tables for connecting tabular data to geometry
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
Motivation

Why study the challenges and needs of geospatial data users?

1. There has been relatively little focus on tools for working with geospatial data in PL and HCI circles.

2. There is a growing gap between the amount of geospatial data and the experts capable of analyzing it.

3. Geospatial data is fundamental to understanding climate change, public health, election integrity, racial and economic inequity, and much more.
Motivation

Why study the challenges and needs of geospatial data users?

1. There has been relatively little focus on tools for working with geospatial data in PL and HCI circles.
Geospatial Data in HCI

“Why Are Geographic Information Systems Hard To Use?”
Traynor and Williams, CHI Short Papers, 1995

“End Users and GIS: A Demonstration Is Worth a Thousand Words”
Traynor and Williams, Your Wish is My Command: Programming By Example, 2001

“Human-computer interaction and geospatial technologies – context”
Mordechai Hackley, Interacting with Geospatial Technologies, 2010
Geospatial Data in HCI

The Many Skillset

Problem

Your Domain Knowledge

Geography
Cartography
Databases
Programming
Statistics
Geospatial Data in HCI

The Disconnected Toolbox Problem

Server Toolbox
Ready to Use Toolbox
Spatial Analyst Toolbox
Spatial Statistics Toolbox
... +35 More

ArcGIS

Bitwise Left Shift
Kriging
Raster Calculator
Iso Cluster Unsupervised
Fuzzy Overlay
Zonal Histogram
Darcy Flow
... +200 More
Geospatial Data in HCI

What does this work miss?

This work relies (mainly) on experts analyzing GIS software rather than directly observing GIS users.

This work hasn’t been revisited to address modern geospatial tools beyond desktop GIS software.

This work doesn’t consider domain-specific needs of geospatial data users.
Motivation

Why study the challenges and needs of geospatial data users?

2. There is a growing gap between the amount of geospatial data and the experts capable of analyzing it.
An Abundance of Geospatial Data

Rolf et al. estimate that global imagery data increases by **80TB per day.**

Motivation

Why study the challenges and needs of geospatial data users?

3. Geospatial data is fundamental to understanding climate change, public health, election integrity, racial and economic inequity, and much more.
The Importance of Geospatial Data

Participant 13 is a **data journalist** exploring whether Texas’ new electoral precincts “pack” or “crack” racial minority groups.
Participant 8 is a climate scientist building ML models to predict development of Cyanobacterial harmful algal blooms (CyanoHABs) from satellite imagery.
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
Research Questions

1. What are the challenges users face when attempting to gather, analyze, or visualize geospatial data?

2. Which of these challenges are shared between experts and non-experts alike? Which are unique to one group?

3. Which of these challenges are shared across our domains of interest? Which are unique to one group?
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
## Study Design

<table>
<thead>
<tr>
<th>Domain</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Earth and climate scientists</td>
<td>9</td>
</tr>
<tr>
<td>Data journalists</td>
<td>8</td>
</tr>
<tr>
<td>Social scientists</td>
<td>6</td>
</tr>
<tr>
<td>“Miscellaneous”</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tool Usage</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Environments</td>
<td>14</td>
</tr>
<tr>
<td>GUI-based Software</td>
<td>8</td>
</tr>
<tr>
<td>Both</td>
<td>3</td>
</tr>
</tbody>
</table>
Study Design

25 participants

Expertise

How long have you been working with geospatial data?

How would you assess your skill level in working with geospatial data?
Methods

Contextual Inquiry

“...go where the [user] works, observe the [user] as he or she works, and talk to the [user] about the work.”

Methods

**Thematic Analysis**

“...a method for identifying, analysing and reporting patterns (themes) within data. It minimally organizes and describes your data set in (rich) detail.”

Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
Roadmap

5. Findings and Discussion

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
Findings

⚠️ These findings are preliminary.

Creating informal program representations

Reasoning about geospatial operator behavior

Lack of spatial visibility in code-based tools
Findings

GIS software maintains only a minimal record of steps taken by a user in the interface, but users want detailed process information.

Creating informal program representations

GIS Software
Creating informal program representations

1. Load two shapefiles into a new QGIS project. **Change Layer Symbology** to use red outline with no fill.
Creating informal program representations

1. Creating informal program representations

2. Load two additional shapefiles into the QGIS project. **Toggle layer visibility** of one previous layer.
Creating informal program representations

1.

2.

3. Attempt a **Merge operation** on the three visible shapefiles. **Create a temporary layer** for the result.
Creating informal program representations

1. [Image of program representation 1]
2. [Image of program representation 2]
3. [Image of program representation 3]
4. Merge operation fails due to mismatching data types in attribute tables of source layers.
Creating informal program representations

5. Attempt Union operation of just two shapefiles.
Creating informal program representations

1.  
2.  
3.  
4.  
5.  
6. Union operation doesn’t error but produces an unexpected output.
Creating informal program representations

1.
2.
3.
4.
5.
6.
7. **Retry** **Union** **operation** but flip layer parameter order.
Creating informal program representations

1.

2.

3.

4.

5.

6.

7.

8. Union operation succeeds and gives expected output.
Creating informal program representations

1. [Image]
2. [Image]
3. [Image]
4. [Image]

5. [Image]
6. [Image]
7. [Image]
8. [Image]

Full trace of user actions
Creating informal program representations

1. Full trace of desired program

7.

8.
Creating informal program representations

What QGIS actually gives us
Creating informal program representations

Participants using GIS software persisted process information about their analysis outside of GIS software.

Participant 19 used macOS Notes.

Participant 4 used a hand-drawn diagram, with color distinguishing layers vs. operators.

Participant 3 used Google Sheets, adding annotations on data source quality and expressions to use in the Raster Calculator tool in QGIS.
Creating informal program representations

Participants using GIS software left hints to their analysis steps in layer names.

Participant 3 added the suffix “MERGE” to the layer generated by merging temperature and precipitation layers.

Participant 4 added the suffix “_no0” when filtering out Census block groups with no population from his dataset.
Creating informal program representations

Why do GIS software users create informal representations of their processes?

- Assists in **recall** of how the current program state was achieved
- Assists in **manual reproduction** of the full analysis at a future time
Creating informal program representations

What are the pitfalls of working from these informal program representations?

- Difficult to record a sufficient level of detail for precise reproduction
- Not independently verifiable by other researchers or journalists
“I don’t do any of my processing in [QGIS], mainly because I like that you can track what you did, the traceability of doing it in Python. Versus there’s like none of that if you do it in QGIS. It’s like you use a plugin or a function, but there’s no track record of it.”

— Participant 17
“If someone wants to go back and look at my code — ‘Oh, he got this shapefile from here and this shapefile from here and he's pushing them together.’ Whereas if you do that in Arc you can't really replicate that workflow in the same way.”

— Participant 15
Findings

Reasoning about geospatial operator behavior

Geospatial tools have hundreds of operators, and finding the right operator for an analysis context is difficult.
Reasoning about geospatial operator behavior

1. “I want to subset to just the points that are within the polygon.”
Reasoning about geospatial operator behavior

1. Copy .within example code from a blog post.
Reasoning about geospatial operator behavior

1. 

2. 

3. When that fails, check the geopandas API documentation for \texttt{intersection}.
Reasoning about geospatial operator behavior

1.  
2.  
3.  
4. Get confused by the geopandas API docs’ use of tables in lieu of geometries.
Reasoning about geospatial operator behavior

1. 
2. 
3. 
4. 

5. Copy `.sjoin` example code from a different blog post.

```python
# Example use: get points in Los Angeles, CA.
pnt_LA = points[points.NAME=='Los Angeles']

# Plot map with points in LA in red
base = counties.boundary.plot(linewidth=1, edgecolor="black")
points.plot(ax=base, linewidth=1, color="blue", markersize=3)
pnt_LA.plot(ax=base, linewidth=1, color="red", markersize=3)
pnt.plot()
```
Reasoning about geospatial operator behavior

6. Adapt copied `.sjoin` for the notebook's context. Notice a runtime warning.
Reasoning about geospatial operator behavior

1. 
2. 
3. 
4. 

7. Change op to predicate to fix warning. The result is an empty GeoDataFrame 😱.

5. 
6. 

8. Go to StackOverflow post. “I feel really silly. I waste a lot of time doing things like this.”
Reasoning about geospatial operator behavior

The above sequence took **6 minutes**.

We continued for **another 12 minutes** before arriving at the desired output.

```python
so_2021_gdf[so_2021_gdf.geometry.within(Somalia_boundary.loc[0, 'geometry'])]
```
Reasoning about geospatial operator behavior

Participants understood an operator's behavior by running it and inspecting the output.

Participant 11 tried using geopandas’ `contains` API to see if it would work on point-within-closed-line operations.

Participant 3 executed a `Merge` on their temperature and precipitation rasters and learned that the operator, by default, combines all input rasters into a single band rather than preserving multiple bands.
Reasoning about geospatial operator behavior

Participants use specific features or pixels to validate operator semantics.

Participant 1 checked the pixel values of a lake in their normalized difference water index (NDWI) raster to validate the algorithm's results.

Participant 4 used color to validate that two features that should have been removed by Extract by Expression were indeed removed.
“I can never remember the vector operations. There’s like Union and Merge. Combine! I can never remember exactly what they do. I know exactly what the output should look like in the end, I’m just trying to figure out the tool that gets me that output.”

— Participant 15
Findings

Lack of spatial visibility in code-based tools

Code-based tools represent geospatial data as tables, **but users rely heavily on visual spatial reasoning during data exploration.**
Lack of spatial visibility in code-based tools

1. Load TIGER/Line shapefile of urban Census block groups in Imperial County into QGIS.
Lack of spatial visibility in code-based tools

1. Move to VisiData, a CLI tool for data wrangling, to Join 2010 Census data to block groups.
Lack of spatial visibility in code-based tools

1. 

2. 

3. Move back to QGIS and generate 2-mile buffers around the point locations of banks.
Lack of spatial visibility in code-based tools

1. Use the Overlap analysis operator to compute the area overlap between each block group and any buffer.

2. 

3. 

4.
Lack of spatial visibility in code-based tools

1. Use the Join Attributes By Field Value operator to join the areal overlay back to the 2010 Census data.
Lack of spatial visibility in code-based tools

1. [Image]
2. [Image]
3. [Image]
4. [Image]
5. [Image]
6. Move back to VisiData to multiply % overlay with bank buffers by total population for each block group.
Lack of spatial visibility in code-based tools

1. [Image 1]
2. [Image 2]
3. [Image 3]
4. [Image 4]
5. [Image 5]
6. [Image 6]
7. Use the macOS Calculator to compute the percentage of Imperial County residents living in a banking desert.
Lack of spatial visibility in code-based tools

1. [Image of code-based tool]
2. [Image of code-based tool]
3. [Image of code-based tool]
4. [Image of code-based tool]

5. [Image of code-based tool]
6. [Image of code-based tool]
7. [Image of code-based tool]
8. Move back to QGIS to repeat the analysis using 1-mile buffers.
Lack of spatial visibility in code-based tools

Participant 9 oscillates between a programming environment and GIS software, **moving to QGIS for all geometric operations.**
“I'm working in QGIS. I know that it’s slower than it would be to do it in PostGIS or maybe even geopandas and so I've considered switching to that. But I'm still sort of new enough that I need to kind of ‘see’ to make sure my projections are right and stuff like that.”

— Participant 9
Lack of spatial visibility in code-based tools

Participants rely heavily on visual overlays of layers to investigate spatial correlations, and recreate this functionality in programming environments.

Participant 16 compared an image of a plot against his Matlab figure to “eyeball” the spatial correlation between two variables of interest.

Participant 6 wrote an 8-line function in R to render three layers on a Leaflet map, allowing her to inspect the spatial overlap between them.
Lack of spatial visibility in code-based tools

“Now I *could* try to visualize it here with `matplotlib` and `geopandas`, but I know those things are shitty, shitty, shitty and *not interactive* and so I'm like, ‘I gotta take this to QGIS.’”

— Participant 18
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion
Roadmap

1. What is geospatial data? What is GIS?
2. Motivation
3. Research Questions
4. Study Design
5. Findings and Discussion

Highlights

1. The importance of informal program representations to the working process of GIS software users
2. The challenge of reasoning about the behavior of geospatial operators
3. The lack of spatial visibility in code-based tools