Geospatial Project Data

Guidance and examples for building location (geospatial) information from administrative and project data to be used in a geospatial impact evaluation.

Topics include:
- Geospatial data basics
- Units of analysis
- Preparing project geospatial data
- Project metadata
Geospatial Data Basics
What is Geospatial Data?

• Any observation connected to specific location
  • E.g., at latitude Y and longitude X the temperature is 30 degrees

• Because all geospatial data have locations, you can always merge different datasets together (though it can be difficult).

• This locational merging process enables various types of analysis (e.g. targeting, monitoring, and evaluation) that would otherwise not be feasible
Examples of Geospatial Data

- GPS coordinates of household where a survey was conducted
- A description of villages that received improved crop seeds
- Series of coordinates forming a line which represents canal that was built
- Coordinates at 4 corners of a field where harvest data was collected
- A copy of a paper map showing fields growing specific crop types
- Government administrative boundaries of a district in which an agricultural education program was implemented
- Start and end points of a road project
- A construction firm’s drawing of site plans for an irrigation system
- Drone imagery of the site of a newly developed small earth dams
**Vector Data**

**POINTS**: Individual $x$, $y$ locations.
- ex: Center point of plot locations, tower locations, sampling locations.

**LINES**: Composed of many (at least 2) vertices, or points, that are connected.
- ex: Roads and streams.

**POLYGONS**: 3 or more vertices that are connected and **closed**.
- ex: Building boundaries and lakes.
Raster Data
Raster Data Resolution
Vector / Raster Transformation
Vector / Raster Integration

<table>
<thead>
<tr>
<th>Unit</th>
<th>NDVI 1990</th>
<th>NDVI 2005</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.15</td>
<td>0.10</td>
</tr>
<tr>
<td>B</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>C</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>D</td>
<td>0.20</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Geospatial Data Tools

```python
from pathlib import Path
import pandas as pd
import geopandas as gpd

base_path = Path('/home/user/Desktop')
geojson_path = base_path / 'SE2S_Ethiopia.geojson'
raw_gdf = gpd.read_file(geojson_path, driver='GeoJSON')

points_gdf = raw_gdf.loc[raw_gdf.type == 'Point']
points_gdf = points_gdf.set_crs(epsg=4326)

polygons_gdf = raw_gdf.loc[raw_gdf.type == 'Polygon']
polygons_gdf = polygons_gdf.set_crs(epsg=4326)
polygons_gdf = polygons_gdf[['geometry']].copy()

gdf = gpd.sjoin(polygons_gdf, points_gdf, op='contains')
gdf.drop(columns=['index_right'], inplace=True)

gdf = gdf.to_crs(epsg=32636)
gdf['buffer'] = 0

buffers = [5, 10, 15, 20, 25, 35]

bgdf_list = [gdf]
for bin in buffers:
    bgdf = gdf.copy()
    bgdf['geometry'] = bgdf['geometry'].buffer(bin)
    bgdf['buffer'] = bin
```
Geospatial Units of Analysis
Defining a Unit of Analysis

• In many non-spatial studies, the unit of analysis for an intervention is well defined:
  ✓ Giving a pill to individuals for a disease - the individual person is the unit of observation.
  ✓ Applying fertilizer to a number of fields - the individual field is the unit of observation.

• In geospatial analyses the unit of analysis usually needs to be theoretically defended.
  ✓ If you build a canal, over how many meters or kilometers from the canal do you expect that you should be able to observe an impact?
  ✓ If you conduct a training on sustainable farming techniques in a city center, what is the size of the ‘catchment area’ where you expect the benefits to accrue?

• Before any statistical analysis can be performed, the unit of analysis must be defined.
# Defining a Unit of Analysis

Examples of project data and corresponding units of analysis:

<table>
<thead>
<tr>
<th>Type of Project</th>
<th>Example Project</th>
<th>Unit of Analysis</th>
<th>Alternative  (less precise spatial analysis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Field Based</td>
<td>Field level irrigation improvements</td>
<td>Polygon outlines of fields</td>
<td>Point location of field</td>
</tr>
<tr>
<td>Household Based</td>
<td>Cash transfer provided to household to improve resilience)</td>
<td>GPS of household survey respondents</td>
<td>Village in which household is located</td>
</tr>
<tr>
<td>Area Based</td>
<td>Distributing improved seed or providing irrigation for entire village)</td>
<td>GPS of specific village households / fields benefitting from intervention or spatial feature defining precise extent of village</td>
<td>Buffer around approximate GPS centroid of village</td>
</tr>
<tr>
<td>Linear Features</td>
<td>Network of roads which were improved</td>
<td>Grid cells within buffer around improved roads</td>
<td>villages/districts in which roads were improved</td>
</tr>
<tr>
<td>Broad Scope Program</td>
<td>Agricultural subsidies allocated based on administrative unit</td>
<td>Administrative boundaries for district (or other units) at which interventions were allocated</td>
<td>-</td>
</tr>
</tbody>
</table>
Units of Analysis: Buffers

- Assign a specific distance (i.e., 1km, 5km, 10km) from the known intervention point, and test for impacts within that area.

- Can be applied to point-based interventions (i.e., a training) or line-based interventions (i.e., a canal or road), or polygon (i.e., protected area).
Units of Analysis: Grid Cells

Advantages

✓ Many units of observation
✓ Can identify precise changes over study region
✓ Can mitigate challenge of not knowing the exact buffer radius you need to select

Disadvantages

x Both study region and grid cell size must be defended theoretically.

x Underlying data must support chosen resolutions.

x Statistical challenges are larger (independence of units can be scale dependent).
Advantages

✓ Works very well when the exact area an intervention would have impacted is known
✓ No or little spatial measurement imprecision

Disadvantages

✗ Many programs do not have a precisely defined spatial scope
✗ The costs to generate a precisely defined spatial scope can be significant
Advantages

✓ Helpful when an intervention is anticipated to affect an entire decision-making unit.
✓ Frequently the same units used for census activities.

Disadvantages

✗ Can change in unmeasured ways over time.
✗ Can be of variable size and have variable underlying measurement qualities.
Preparing Project Geospatial Data
Polygons to measure spatio-temporal rollout of indigenous land demarcation program in Brazil (1997-2008)
Impacts of Mine Clearance in Afghanistan

Project Data: Existing Geospatial Data

Source: Landsat 8 OLI/TM, 2015. Hazard data from MAPA Geo-Data database, admin. boundaries from GADM.
Map prepared by Christian Baeht, AeiData
Irrigation Projects in Mali
Project Data: Geocoding Project Records

Retrospective coding to cover records where explicit geospatial data was not collected

Source Material:
- Project documents
- Local government Aid Information Management Systems
- Donor information systems
- Other official, media, and third-party sources

Includes “geoparsing” and “geocoding”:
- **Geoparsing** - the initial process of identifying place names or other location information in documents (project reports, PDF files, etc.)
- **Geocoding** - translating location information into standardized, machine readable geospatial information (coordinates, administrative boundaries, etc.)
Multiple possible approaches to geocoding, with varying levels of precision and amount of time required.

**Point Based:** single point with codes to describe precision level of the location
- Ex: Coordinates in center of district and code which represents “district level”

**Precise:** uses lines or polygons to represent the true (or best approximation of) actual spatial features defining where investments went
- Lines defining infrastructure development such as roads, power lines, irrigation networks
- Polygons for land tenure, protected areas, etc.
Point based may be preferable at large scales where less precise location information is likely to be available in project records.

Point base coding schemes have inherently limited spatial detail (e.g., project in town recorded as coordinates of town center and code which identified it as a populated place).

Example: GEF historic portfolio of land degradation projects (202 projects, 1704 locations around world)
Precise geocoding requires more time and underlying information, but enabled far more types of spatial analysis and understanding of local impacts.

Example geocoding roads: point based vs precise. Left: recording centroids of districts road runs through vs true road path. Right: recording start and end point of road vs true road path.
Alternative to generating precise features is using data from OpenStreetMap (OSM)
Project Data: Geocoding Project Records

Currently being used to geocode AidData's Global Chinese Development Finance Dataset

Features from OSM representing Chinese financed projects
Project Data: Collecting New Data

Generating new geospatial data in the field using GPS to record borders, etc.

Example of farm plot data collected using two methods in Northern Ghana
All geospatial must have some unique ID for each feature that can be used to reference the feature during analysis and when joining with other data.
Project Metadata
Includes any additional information describing key characteristics of the project that can be tied to the geospatial unit of analysis. Examples:

- construction or project start/end dates, cost of the project / site, status of the project at a location,

The type of metadata collected depends on the particular project and evaluation needs, but in general you should address the three following steps.

1. Identify all relevant metadata which is available from project documentation and needs to be extracted and/or formatted.
2. Determine a methodology for standardizing and extracting the metadata consistently across all project records.
3. Include a unique identifier associated with the project so that metadata can be linked to geospatial data.
Joining existing data
Project Metadata: Overview

Extracting data from records

<table>
<thead>
<tr>
<th>Zanjeer Main Canal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Length Of Main Canal</td>
<td>5235 M</td>
</tr>
<tr>
<td>Length Of Branches</td>
<td>7110 M</td>
</tr>
<tr>
<td>Total Area</td>
<td>286 Ha</td>
</tr>
<tr>
<td>Cultivated Area</td>
<td>275 Ha</td>
</tr>
<tr>
<td>Non-Cultivated Area</td>
<td>11 Ha</td>
</tr>
</tbody>
</table>
Questions?

Module Contact Points:

- Seth Goodman (sgoodman@aiddata.org)
- Rachel Sayers (rsayers@aiddata.org)
- Ariel BenYishay (abenyishay@aiddata.org)