API-231 / GIS-PubPol
Meeting 16 (Building a GIS Research Project from Ground Up)

Yuri M. Zhukov
Visiting Associate Professor of Public Policy
Harvard Kennedy School

March 26, 2024

Overview

What steps are involved in a research project?

1. Topic selection: identify a motivating question/puzzle
2. Theory-building: advance one or more hypotheses/claims to test
3. Data collection: find the data needed to conduct this test
4. Pre-processing: prepare the data for analysis
5. Analysis: analyze the data using qualitative or quantitative methods
6. Discussion: draw tentative conclusions, identify next steps

Different steps will be more/less important for different types of projects

1. Theory-building and hypothesis-testing less central in descriptive projects ("what?" questions)
2. Analysis needs to be more sophisticated in explanatory projects ("why?" questions)

What will we do this week?

Illustrative example: a research project on Islamic State violence in Iraq and Syria

1. Review methods:
   - clipping data by extent of another layer
   - point-in-polygon analysis
   - line-in-polygon analysis
   - joins
   - raster-in-polygon analysis
   - polygon-in-polygon analysis
   - selection by expression
   - intersections
   - exporting processed data
2. Introduce new methods:
   - regular expressions
   - dissolve operations
   - regression analysis

QGIS step-by-step & R replication code on Canvas (no problem set!)

Theory

Research Question

Selecting (and refining) a research question

1. "What" versus "why"
   a. describe *what* happened:
      - goal: uncover general patterns of social behavior and attitudes
      - reasoning: (mainly) inductive (specific $\rightarrow$ general)
        make generalizations based on observations of data
   b. explain *why* something happened:
      - goal: make inferences about (potentially) causal relationships
      - reasoning: (mainly) deductive (general $\rightarrow$ specific)
        derive hypotheses from theory, test them with data
2. Descriptive $\neq$ atheoretical
   a. description is the first, exploratory step toward explanatory analysis
   b. deduction and induction can be mutually supportive

Illustrative example:
Rise and fall of the Islamic State

1. The Islamic State (ISIS/Daesh) is a Salafi-jihadist insurgency that captured large swaths of Iraq and Syria in 2014-2015
2. A U.S.-led air campaign culminated in 2018 with the liberation of the last major pockets of ISIS-held territory
3. ISIS still exists today, in diminished form

Research question:
Why was ISIS more active in some parts of Iraq and Syria than in others?

Figure 1: Capture the flag

Hypotheses

hypothesis

noun, plural hypotheses [www.dictionary.com/browse/hypothesis]

1. A proposition, or set of propositions, set forth as an explanation for the occurrence of some specified group of phenomena
   a. either asserted merely as a provisional conjecture to guide investigation (i.e. "working" hypothesis)
   b. or derived from first principles (e.g. game theory)
   c. or accepted as highly probable in light of established facts

What makes a "good" hypothesis?

1. clearly relates to your research question
2. specifies direction of relationship between X and Y (bad: "X matters"; good: "X has a positive effect on Y")
3. derived from some body of theoretical literature (not "just so")
4. concise (can summarize it in 1 sentence)
5. falsifiable (some chance it can be shown to be wrong)
6. policy relevant (can convincingly answer the "so what" question)

Hypothesis 1: state weakness

1. "insurgencies become entrenched in areas less accessible to government troops" (fewer roads $\rightarrow$ more ISIS sanctuaries)

Figure 2: A road to perdition

Hypothesis 2: demographics

2. "there will be more insurgent violence in more populous areas"
   (more people → more targets of attack)

Figure 3: A built-up area

Hypothesis 3: political economy

3. "insurgents will be more active where opportunity costs of rebellion are low" (fewer alternative sources of income → more ISIS recruits)

Figure 4: A steady paycheck

Hypothesis 4: key infrastructure

4. "there will be more insurgent violence in areas with critical infrastructure"
   (sites, whose capture would give ISIS significant leverage to shape physical or economic security, public health, safety)

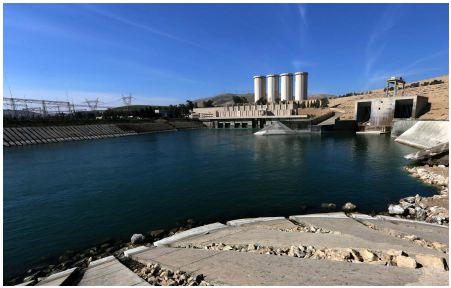

Figure 5: A critical object

Hypothesis 5: sectarian divisions

5. "there will be more insurgent violence in areas dominated by Sunni Arabs"
   (more Sunnis → more ISIS recruits)

Figure 6: A potential base of support

Empirics

Data Collection and Pre-Processing

Data collection: categories of geospatial data

1. "Off-the-shelf" geospatial data (ready to use)
   (e.g. 500+ sites with free GIS data: `freegisdata.rtwilson.com`)
   a. *vectors*: points, polylines, polygons
      examples: events, roads, administrative borders
      (stored as shapefiles, GeoJSON, Geodatabase, KML/KMZ)
   b. *rasters*: continuous fields of regular grid cells
      examples: weather, elevation, land cover, population
      (stored as ASC, GeoTIFF, IMG, DEM,DTED)
2. Raw geospatial data
   a. map sheets, satellite imagery
      (need to be digitized, georeferenced, vectorized)
   b. lists of locations/addresses
      (need to be geocoded)
   c. tabular data organized by geographic units
      (need to be joined/merged to spatial data)

Illustrative example

What is our geographic unit of analysis?

Administrative level 0	Administrative level 1	Administrative level 2	Other
e.g. country	e.g. province/state	e.g. district/county ✓	e.g. grid cell

Illustrative example

What outcome are we trying to explain?

Hypotheses	Dependent variable	Data needed	Format
1-5	Number of ISIS attacks per district	event locations	vector

Illustrative example

What data on explanatory variables do we need to test our hypotheses?

Hypothesis	Explanatory variable	Data needed	Format
1. Power projection	road density	roads	vector
2. Demographics	local population size	population	raster
3. Political economy	% of land used for agriculture	land cover/use	raster
4. Infrastructure	proximity to dams	dam locations	vector
5. Sectarian division	local presence of Sunni Arabs	ethnic settlement	vector

Pre-processing

Having the data in the hand does not mean you're ready for analysis

Common *pre-processing tasks*:

1. Merging datasets (e.g. join 2 tables by a common field/variable)
2. Queries and subsets
   a. select (extract subset of data by attribute or expression)
   b. clip (extract subset of data by location)
   c. intersection (extract overlapping features)
3. Overlay operations
   a. point-in-polygon (e.g. calculate number of airstrikes per district)
   b. line-in-polygon (e.g. calculate length of roads per district)
   c. polygon-in-polygon (e.g. calculate share of overlapping territory)
   d. raster-in-polygon (e.g. calculate average grid cell values per district)
4. Simplification and generalization
   a. dissolve (combine multiple features into one)

Analysis and Discussion

What kind of evidence is needed to confirm/reject a hypotheses?

Compare empirical observations to theoretical expectations

- are observed patterns in the data consistent with what we would expect if our hypothesis was true?

Expectation	Observation	
X is positively associated with Y	positive correlation between X and Y	✓
X is positively associated with Y	negative correlation between X and Y	✗

Methods for hypothesis testing:

1. Visual inspection of maps
   a. disadvantage: observed pattern could be due to chance
2. Descriptive statistics (e.g. difference in means)
   a. disadvantage: most tests are for analyzing 1-2 variables at a time
3. Statistical graphics (e.g. box plots, bar plots, histograms)
   a. disadvantage: same... (most analyze 1-2 variables at a time)
4. Statistical modeling (e.g. multivariate regression)
   a. disadvantage: results sensitive to modeling assumptions, specification

There is no silver bullet! Best practice is to use multiple methods.

Illustrative example

Use regression analysis to test all 5 hypotheses at once

$$\begin{aligned}\text{violence}_i =& \beta_1 \text{road density}_i + \beta_2 \text{population}_i + \beta_3 \text{cropland}_i \\ &+ \beta_4 \text{dams}_i + \beta_5 \text{Sunni presence}_i + \epsilon_i\end{aligned}$$

where

- violence_i is the observed number of ISIS attacks in district i
- road density_i, ..., Sunni presence_i are explanatory variables
- ϵ_i are errors (deviation of observed values from model's predictions)
- β are coefficient estimates corresponding to each Hypothesis

Hypothesis	Expectation	Observation
1. Power projection	$\beta_1 < 0$	?
2. Demographics	$\beta_2 > 0$	?
3. Political economy	$\beta_3 < 0$	?
4. Key infrastructure	$\beta_4 > 0$	?
5. Sectarian divisions	$\beta_5 > 0$	?

Discussion

What are the results of the analysis?

1. Which hypotheses did you confirm?
2. Which hypotheses were you unable to confirm?
3. What were some of the limitations of your analysis?
4. Are there any alternative explanations for your findings?

What are the broader implications of these results?

1. Can your analysis lead to any specific policy prescriptions/lessons?
2. How do your findings advance the academic or historical debate?
3. Whose mind have you changed about what?
4. what additional research is needed on this topic?

Let's go! (switch to lab)