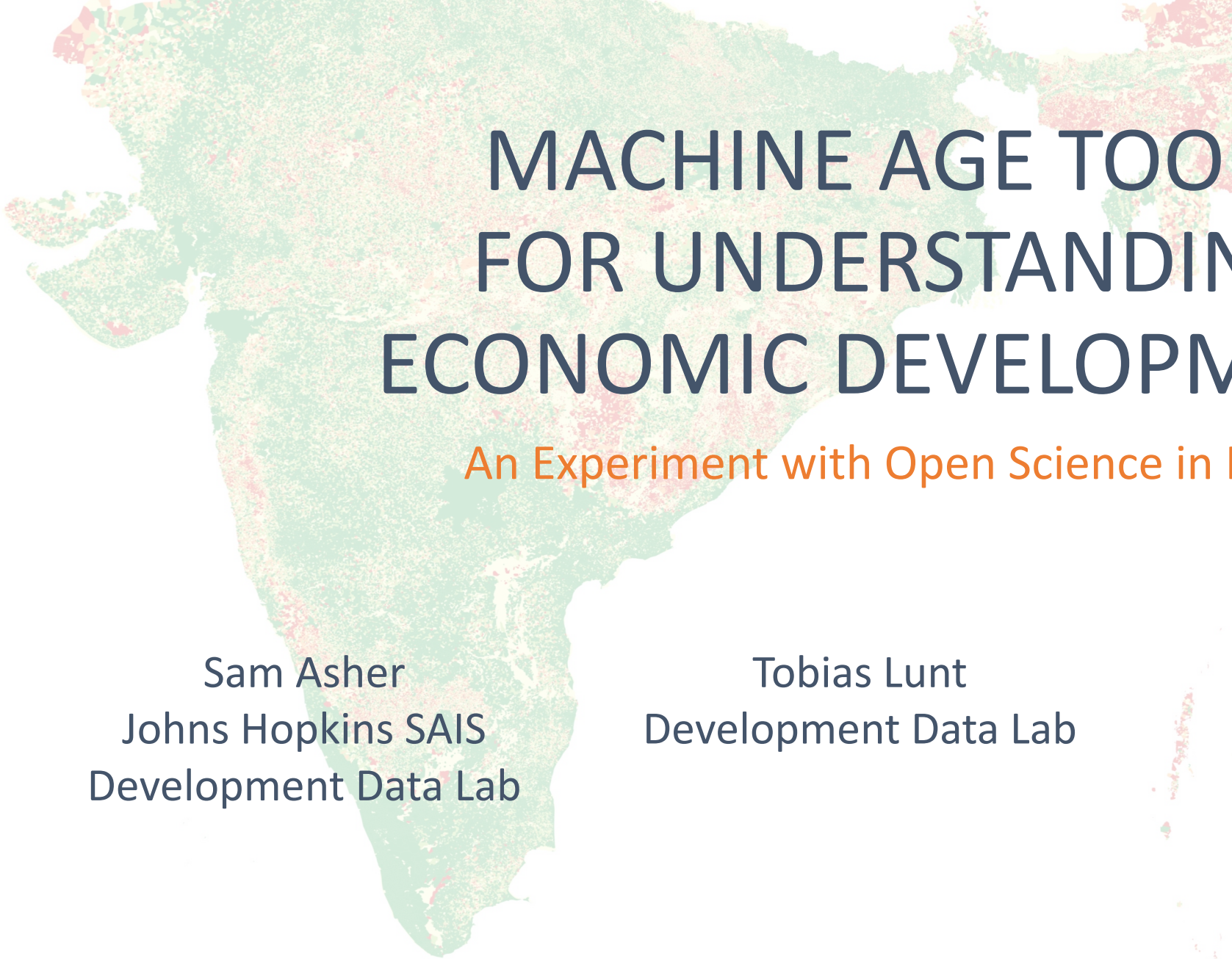


MACHINE AGE TOOLS FOR UNDERSTANDING ECONOMIC DEVELOPMENT

An Experiment with Open Science in India

Sam Asher
Johns Hopkins SAIS
Development Data Lab

Tobias Lunt
Development Data Lab

Paul Novosad
Dartmouth College
Development Data Lab

Presentation Preview

Tremendous unrealized scope for better data collaboration in the social sciences, policymaking, civil society, and the private sector

The SHRUG: A copyleft dataset and platform for research on India

Extending the open source software model to socioeconomic data

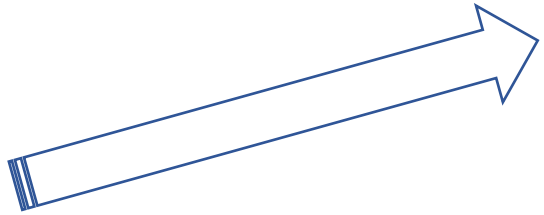
Open Science: In Theory

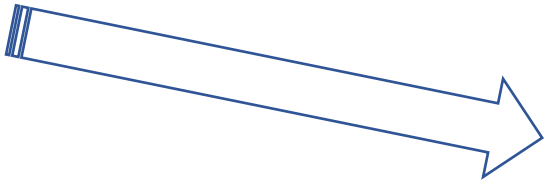
BUILD

Researchers spend years creating new data and publish results

SHARE

Data is posted for public use





REPLICATE

Other researchers can replicate/test results

RE-USE

Other researchers can use data for original analysis

Open Science: In Practice

Technical Barriers to Data Sharing

Creating usable public data ≠ just posting code and data

Institutional Barriers to Data Sharing

Keeping data private:

- Takes less work
- Lowers risk of failed replications
- Allows monopoly control of data for future projects
- Complies with restrictions on data use

Open Science: In Practice

The result: public data is often of limited use to future potential users

- Posted datasets are messy and undocumented
- Posted data are often limited to project samples, limiting wider usability
- Posted code is impenetrable, shows final steps but no construction

Journal policies focused on replication are not solving the usability problem

Administrative Data Raises the Returns to Openness

Socioeconomic research in developing countries is usually based on sample surveys

- India's NSS: equivalent to surveying 1500 households to study California
- Useful for aggregate statistics, but not for understanding local variation

Digital exhaust from government programs is barely used

- Universal digital multidimensional paper trail
- But: restricted access, limited documentation, unclear identifiers
- Research value scales with the number of datasets
  - Chicken / egg problem: isolated admin data is of limited use
- Other unconventional data sources (e.g. remote sensing) have similar features

Researchers in silos cannot mobilize these resources effectively

The SHRUG: The Socioeconomic High-resolution Rural Urban Geographic data platform for India

THE DATA BACKBONE

A comprehensive national socioeconomic dataset that is the best starting point for all research (on India)

EASY LINKING

Seamlessly links to all national datasets, so integration is almost costless

OPEN ARCHITECTURE

Lower both the technical and institutional barriers to sharing of data

- Citation: Data maintains original reference. Contribution -> citations
- Copyleft licensing: if you use, you must share what you link in a principled manner
- Cost reduction: Standardized data protocols lower cost of making data usable.

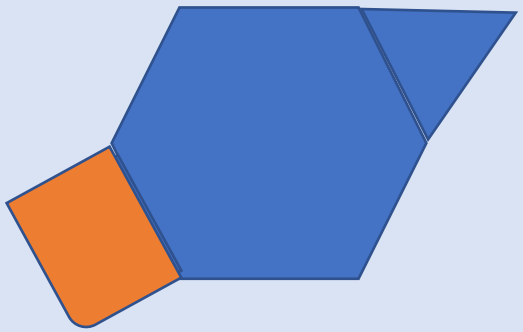
Acknowledgement: Every piece of this has an analog in the OSS movement

SHRUG: The Location Backbone

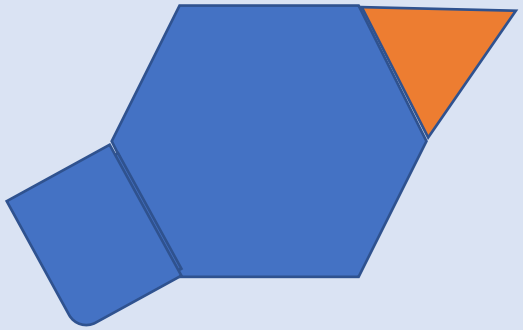
The backbone is a set of universal locations

- Indian Census locations have new (hard to link) identifiers every 10 years
- SHRUG has universal identifiers – time series analysis is a cinch
- We provide simple keys to link SHRUG to any major national Indian dataset
- Consistent industries, variables definitions, data structure, etc.

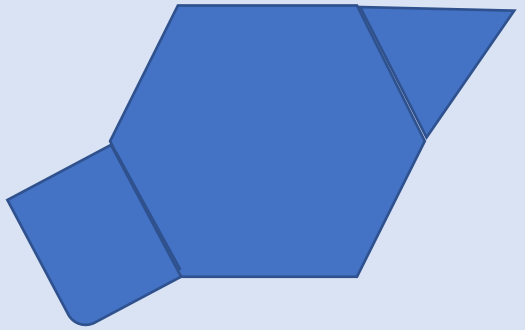
Locations are amalgamated to create the smallest consistent unit



2001 Census Boundaries



2011 Census Boundaries

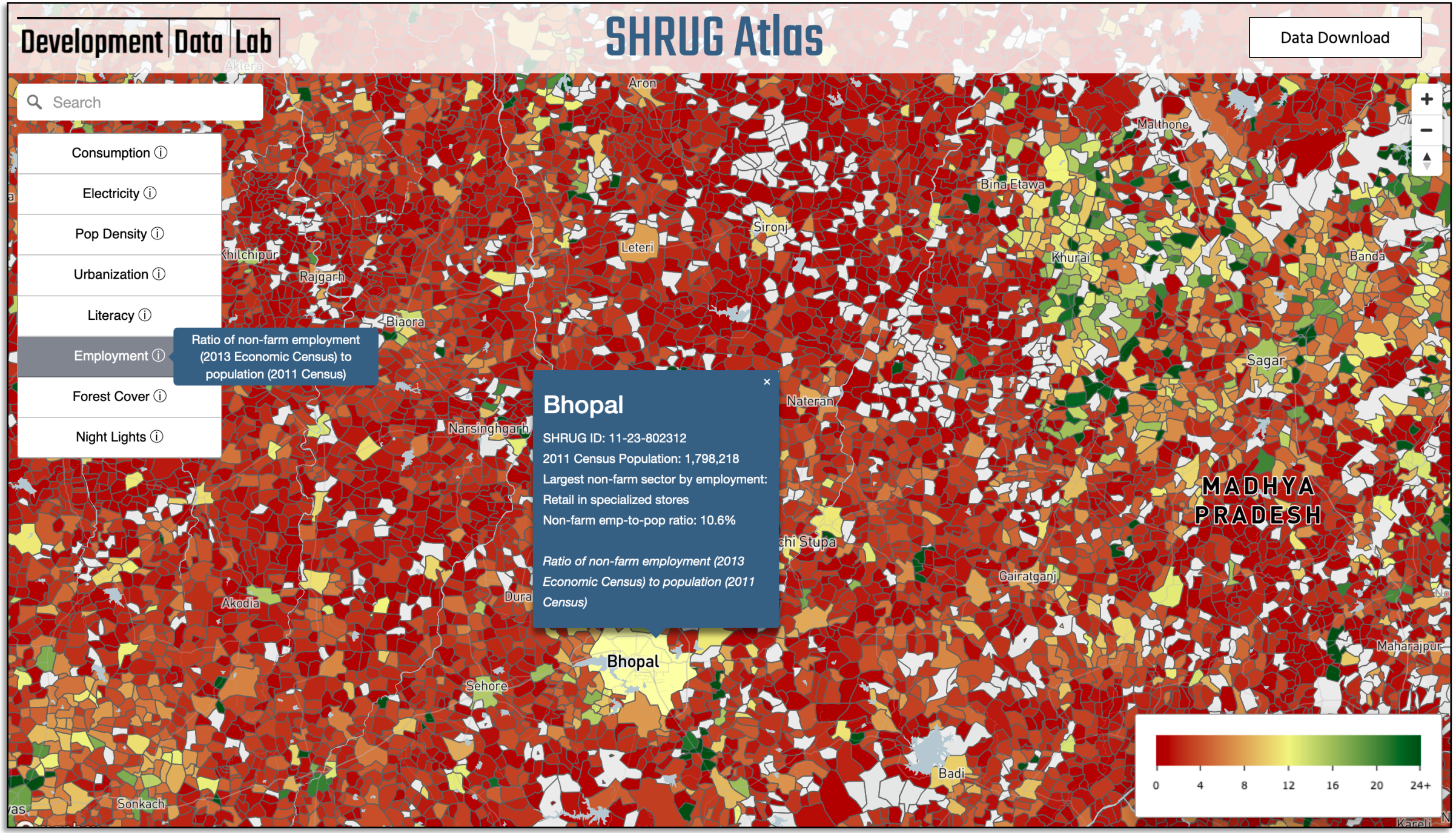


SHRUG Boundaries (all years)

SHRUG 1.4 (Samosa)

Coming soon:

- Rainfall & Temperature
- Parliamentary Constituencies and Panchayats
- Intergenerational Mobility
- City features, including segregation


Development Data Lab
SHRUG Atlas
Data Download
Search
Consumption ⓘ
Electricity ⓘ
Pop Density ⓘ
Urbanization ⓘ
Literacy ⓘ
Employment ⓘ
Forest Cover ⓘ
Night Lights ⓘ
Ratio of non-farm employment (2013 Economic Census) to population (2011 Census)
Bhopal
SHRUG ID: 11-23-802312
2011 Census Population: 1,798,218
Largest non-farm sector by employment: Retail in specialized stores
Non-farm emp-to-pop ratio: 10.6%
Ratio of non-farm employment (2013 Economic Census) to population (2011 Census)
MADHYA PRADESH
Aron
Malthone
Bina Etawa
Sironj
Leteri
Khurai
Banda
Rajgarh
Biaora
Sagar
Narsinghgarh
Nateran
Gairatganj
Akodia
Maharajpur
Bhopal
Sehore
Badi
Sonkach
0
4
8
12
16
20
24+

Use Cases for the SHRUG

Studying Local Development

Most variation in socioeconomic status and in policy is local.

Baseline Data for RCTs

Plug a village list into the SHRUG and get 30 years of multidimensional data.

Cities

SHRUG is the first broad dataset that identifies the full set of towns and cities.

Media / Civil Society

Journalists / citizens are hungry for data but lack resources to build it themselves.

Use Cases for the SHRUG

Targeting government policies

Maximizing impacts requires detailed multi-dimensional data at a high resolution

Guiding private sector investments

Data on local businesses and purchasing power can greatly reduce cost of market research.

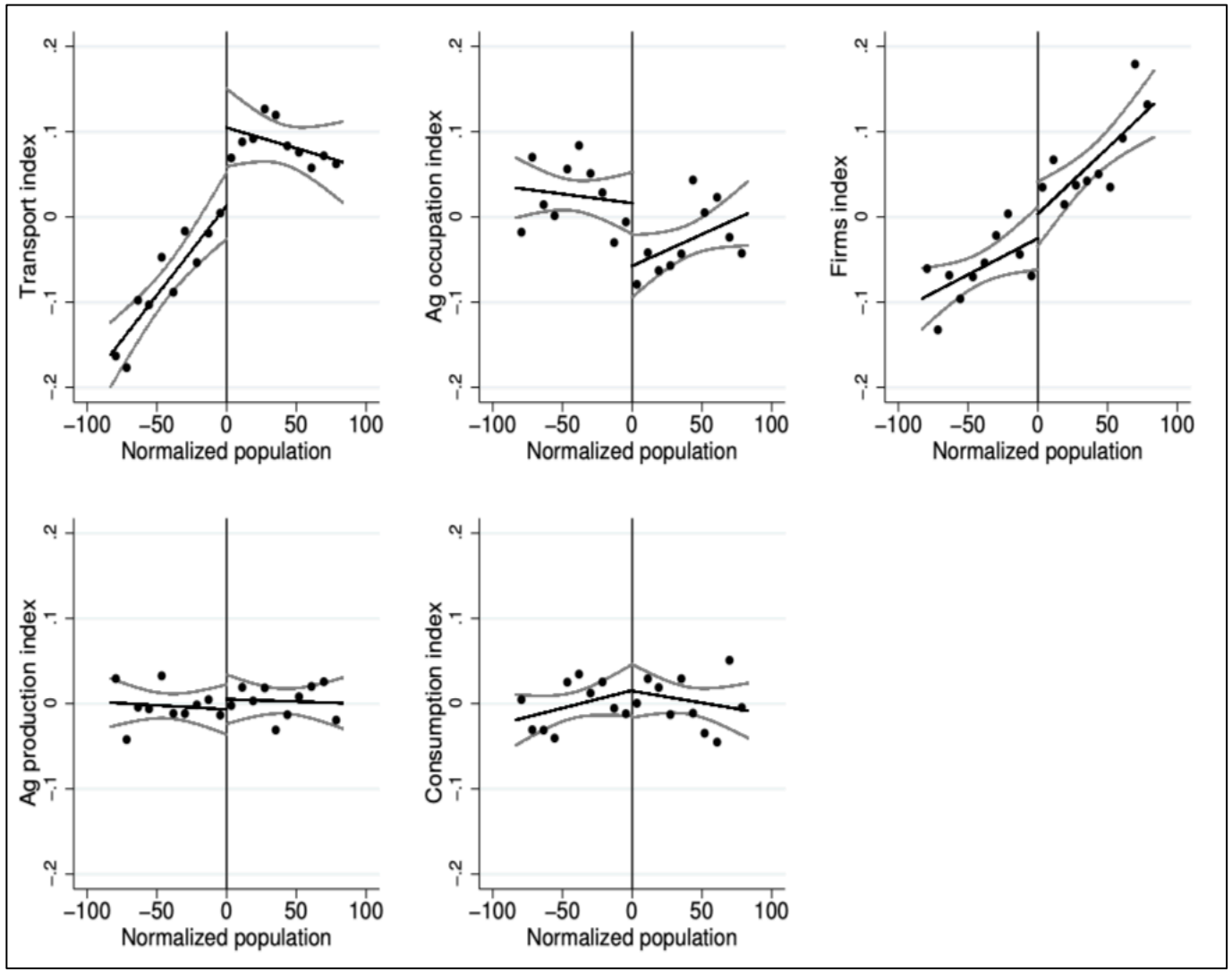
Example 1: Rural Roads

What are the impacts of India's large-scale rural roads program?

- 100,000 new village roads were built from 2000-15
- District-level (old data) approach:
  - Districts building more roads were way better off
  - (But correlation ≠ causation, and effective districts built more roads)

- We use village variation and RD to measure causal impacts. Findings:
  - New roads did not affect consumption, entrepreneurship, investment, or agriculture
  - They did help people get jobs outside of villages

- Required national data on a broad set of village outcomes
  - Very hard to do without administrative data

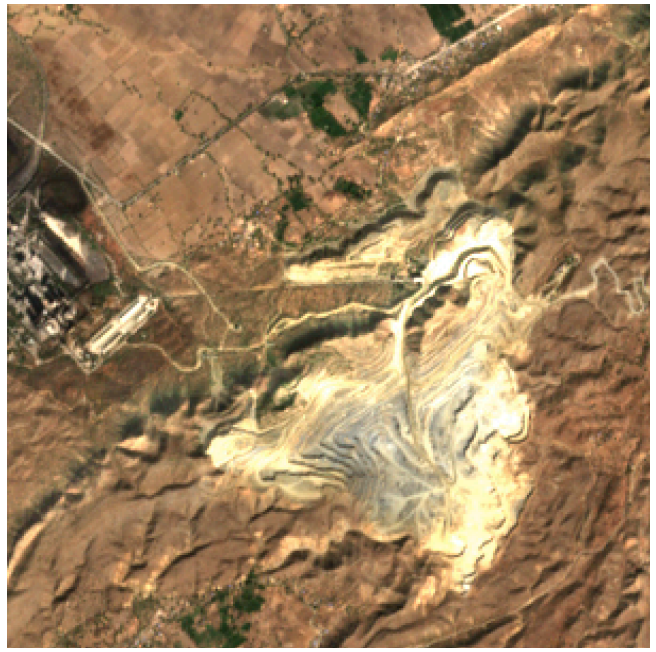


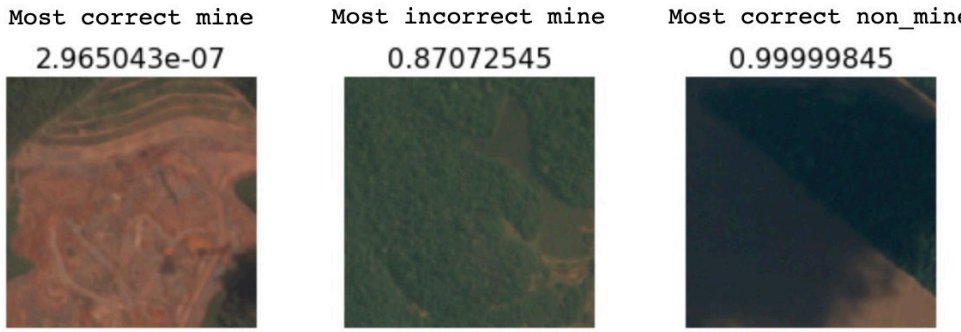
Asher and Novosad, "Rural Roads and Local Economic Development," AER (2020)

Example 2: Impacts of Mines

How does mineral extraction affect local opportunity?

- India has many mines but their impacts are highly local
- Few districts depend on mining: aggregate approach misses local effects
- We want to know:
  - How are the villages directly in the path of mining development affected?
  - Care about a wide range of outcomes: education, consumption, work, health, pollution
- Approach:
  - Computer vision and satellite data detect mine location and expansion
  - International prices generate exogenous variation in mine growth -> causality

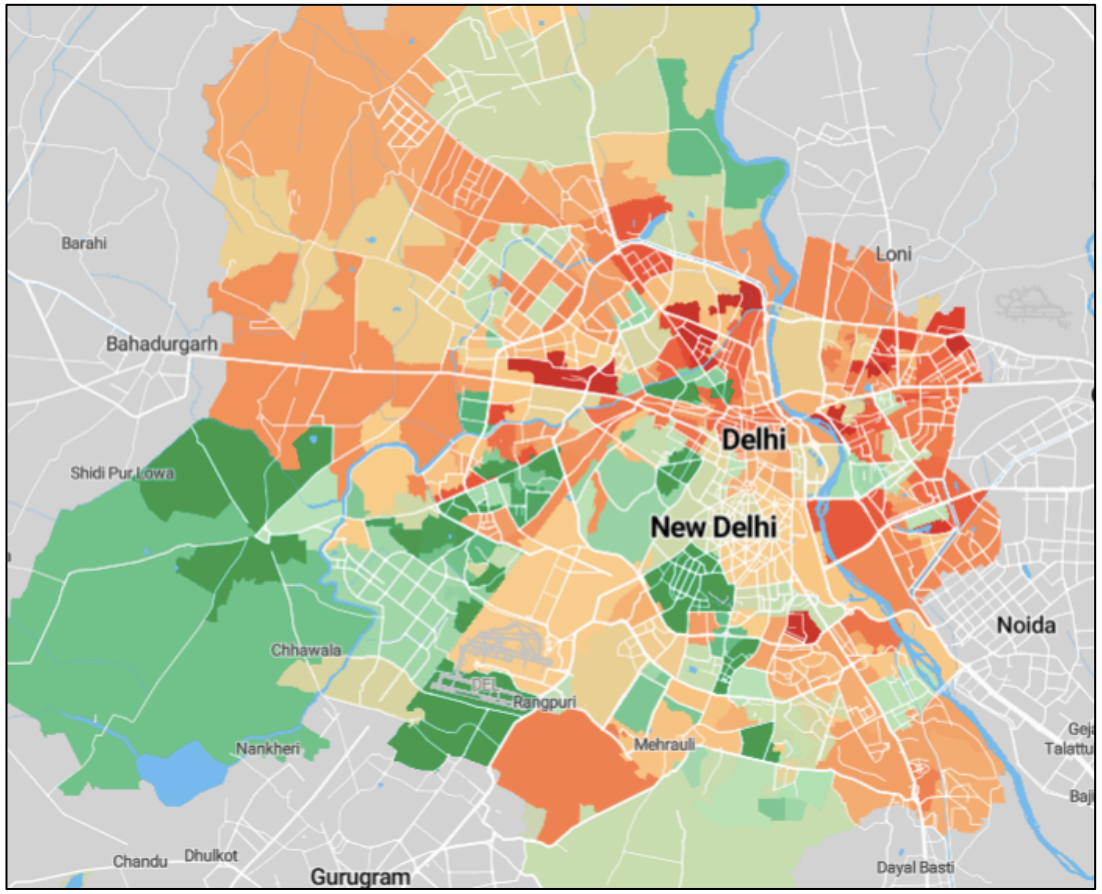


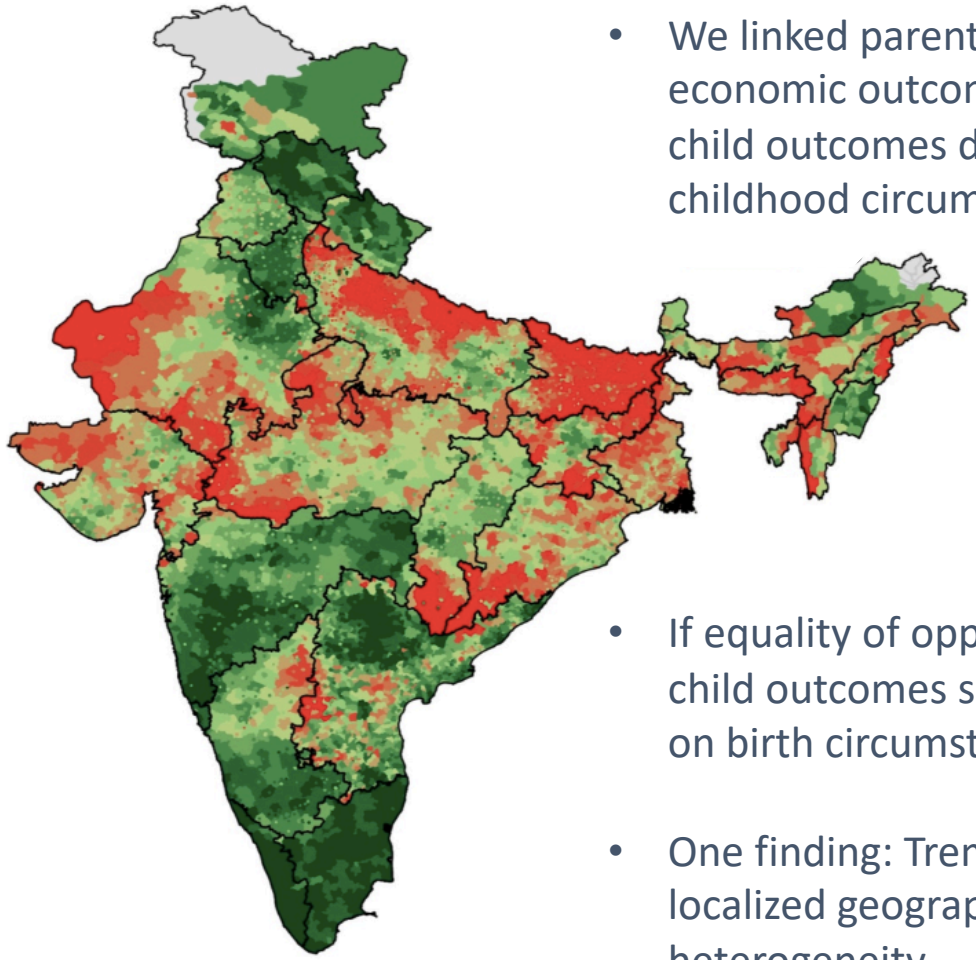


Asher, Lunt, and Novosad, "Digging for Development: Economic Impacts of Mining Booms"

Example 2: Equality of Opportunity

What is the geography of upward mobility in India?





- We linked parent to child economic outcomes to study how child outcomes depend on childhood circumstances.
- If equality of opportunity exists, child outcomes should not depend on birth circumstances.
- One finding: Tremendous highly localized geographic heterogeneity

Example 4: Social Determinants of Health

SHRUG is a unified framework for linking and analyzing social determinants of health

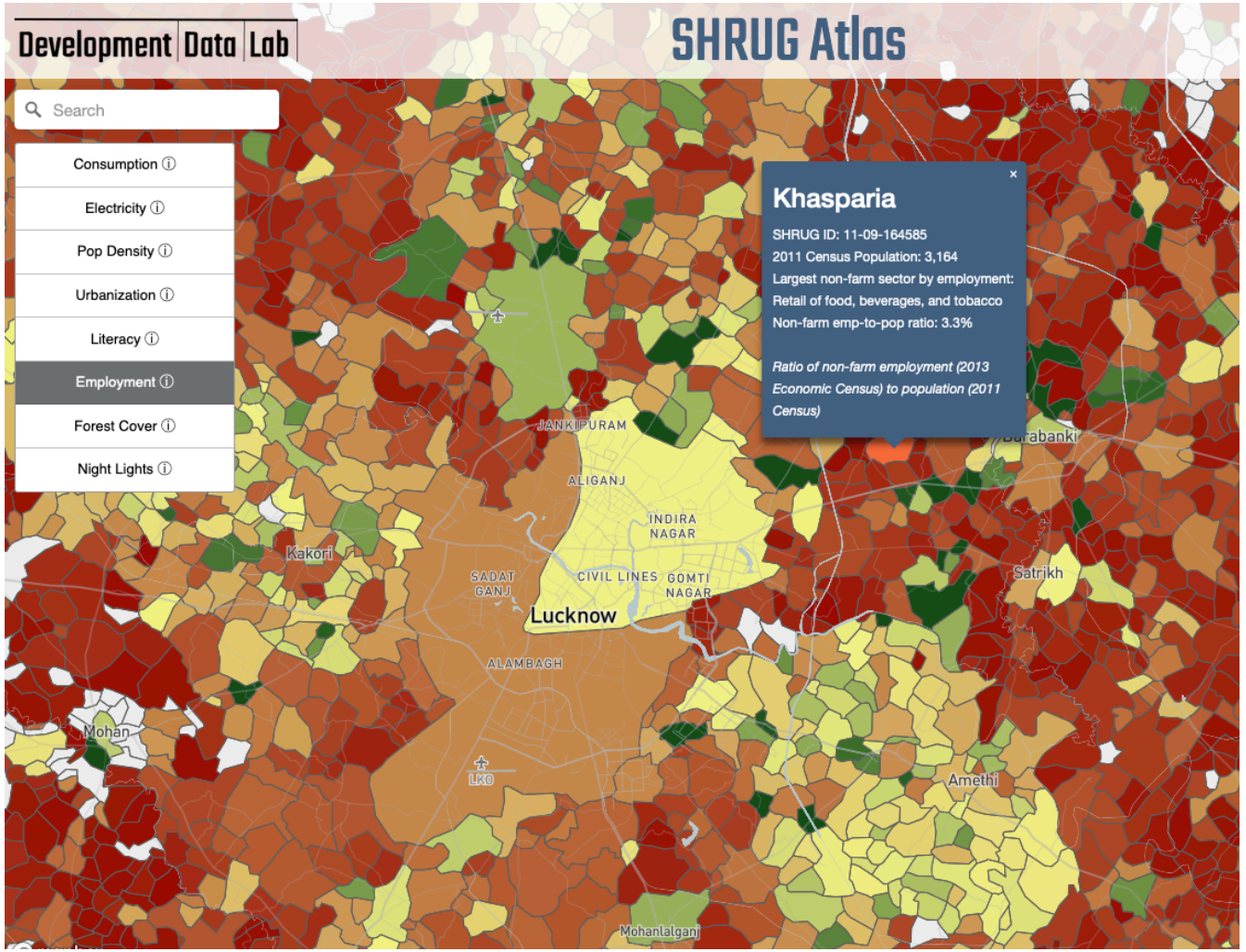
- Asset data for 1 billion individuals
- Employment/industry for 42 million Indian firms
- Demography, public goods, and features of the built environment for *all* towns/villages in India
- Very high spatial resolution data
- **Health data?**
  - DHS
    - Jittered geocodes: probabilistic linking
  - Epi monitoring
  - Claims data

US Department of Health and Human Services, "Healthy People 2020"

Example 5: Reducing the Costs of Market Research

SHRUG greatly lowers the cost of understanding potential markets

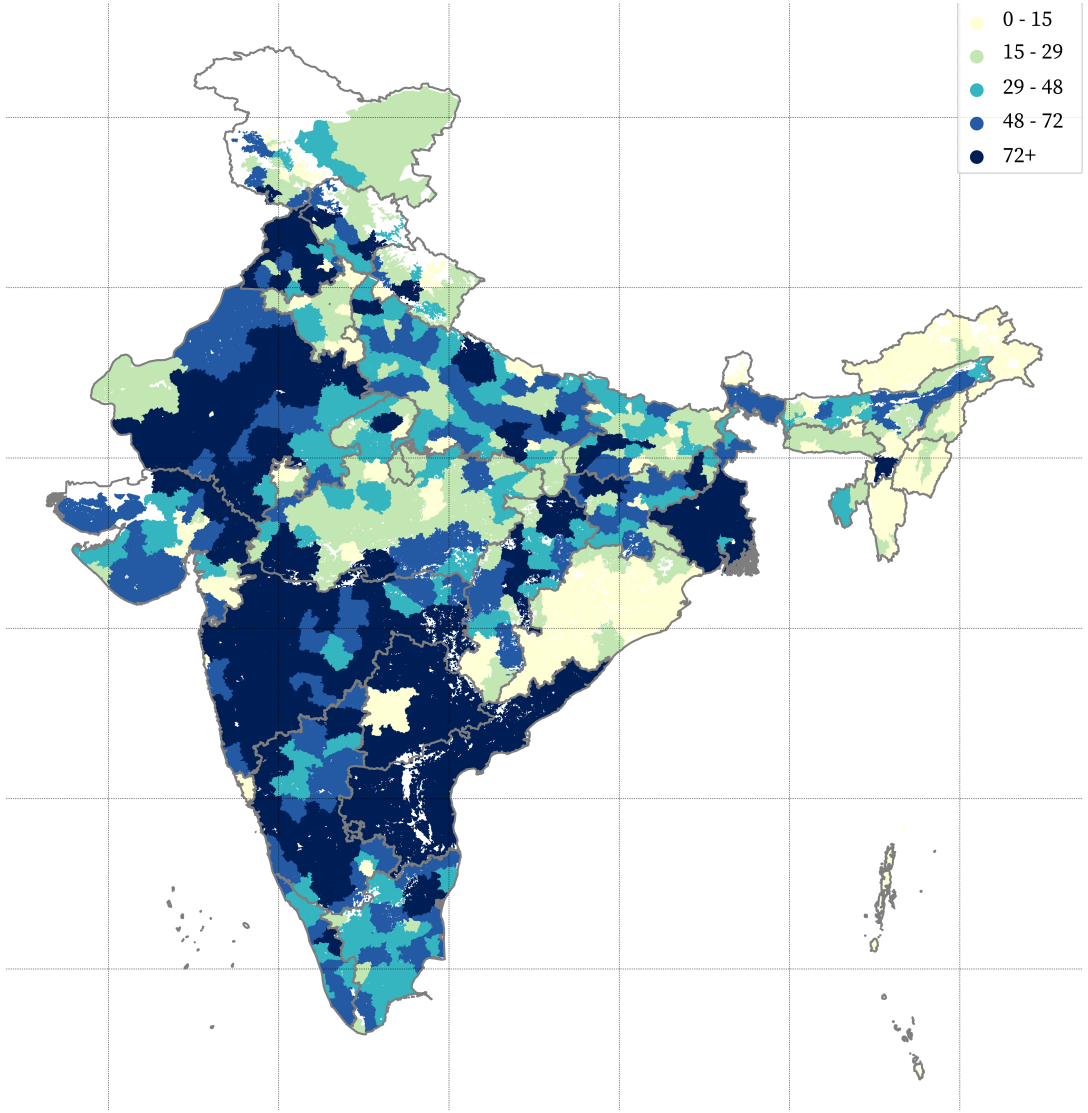
- Consumption data and other correlates of local economic activity can guide geographic expansion
- Encourages entrepreneurship by reducing information asymmetries between firms
- Examples:
  - Identify areas where families may be wealthy enough to support new private schools
  - Find locations with a large aging population lacking access to advanced medical care
  - Determine potential production sites combining educated workforce, low wages, related businesses, etc.

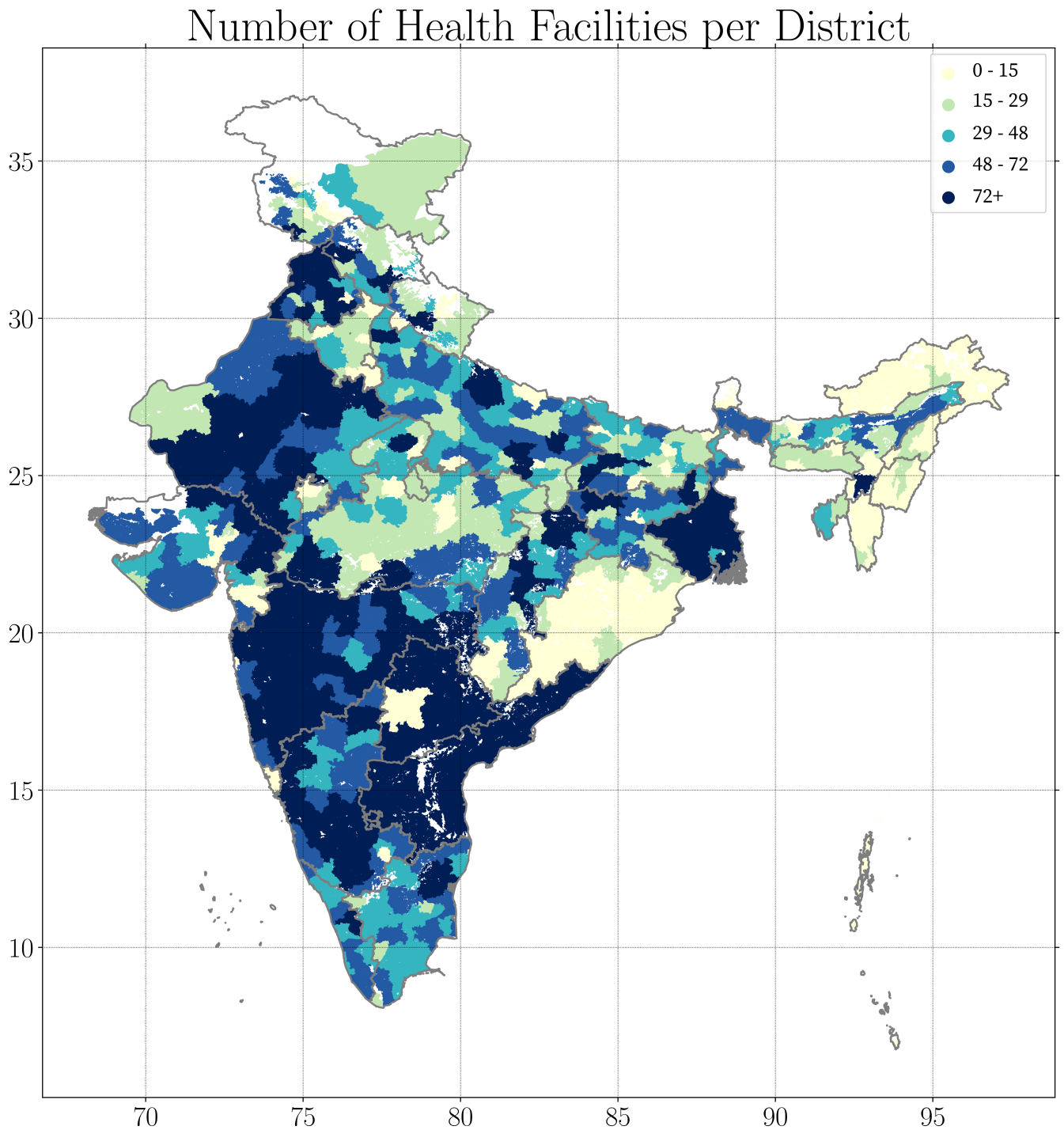


Example 6: Targeting Health Facilities

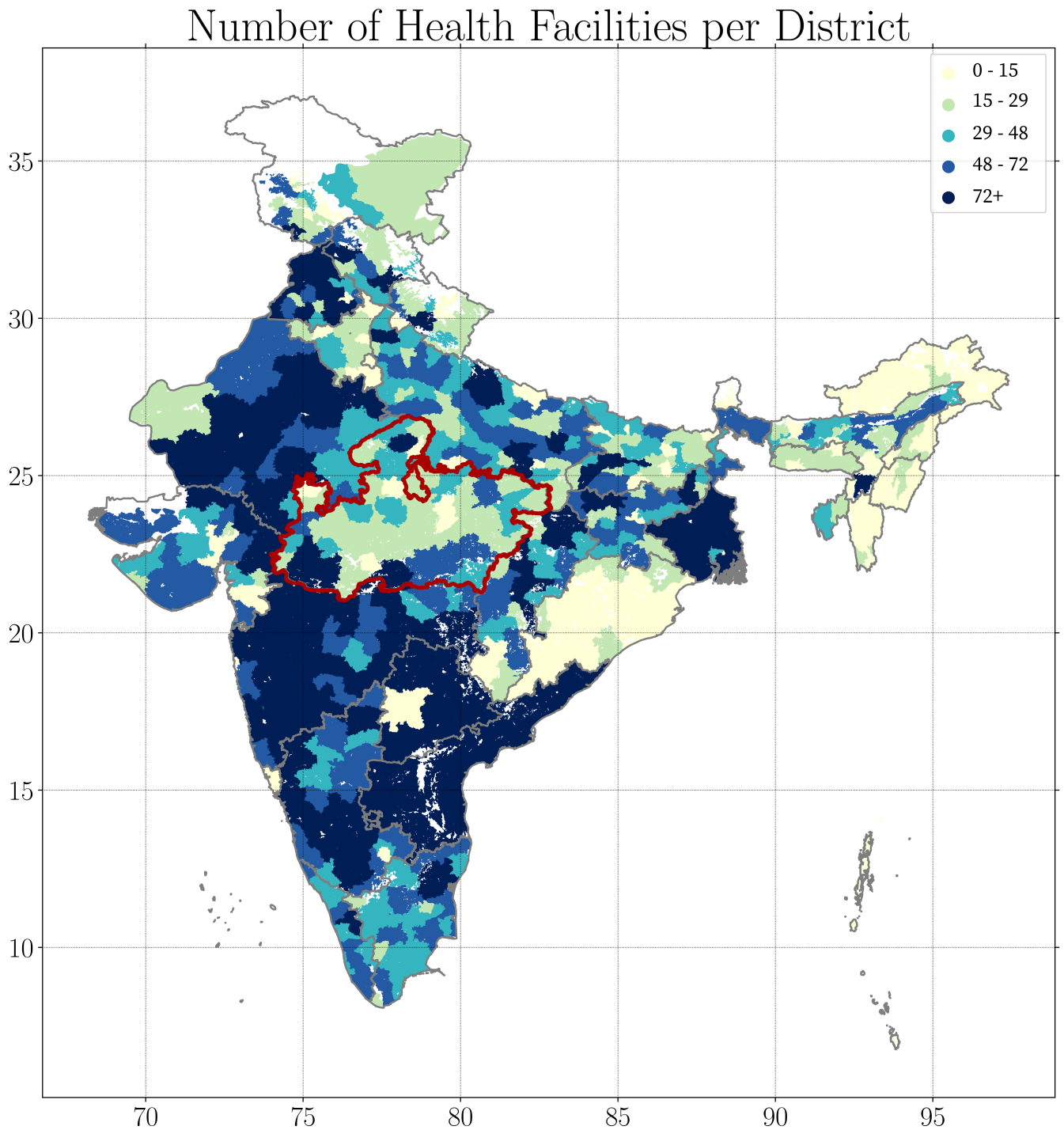
SHRUG gives policymakers the wide range of information they need to target goods and services

- Much policymaking occurs at level of implementation
- Yet policymakers often have limited information over which they have power, or have to allocate many resources to get the necessary data
- Suppose you were tasked with targeting primary health centers in rural areas: **what do you need to know to target most effectively?**



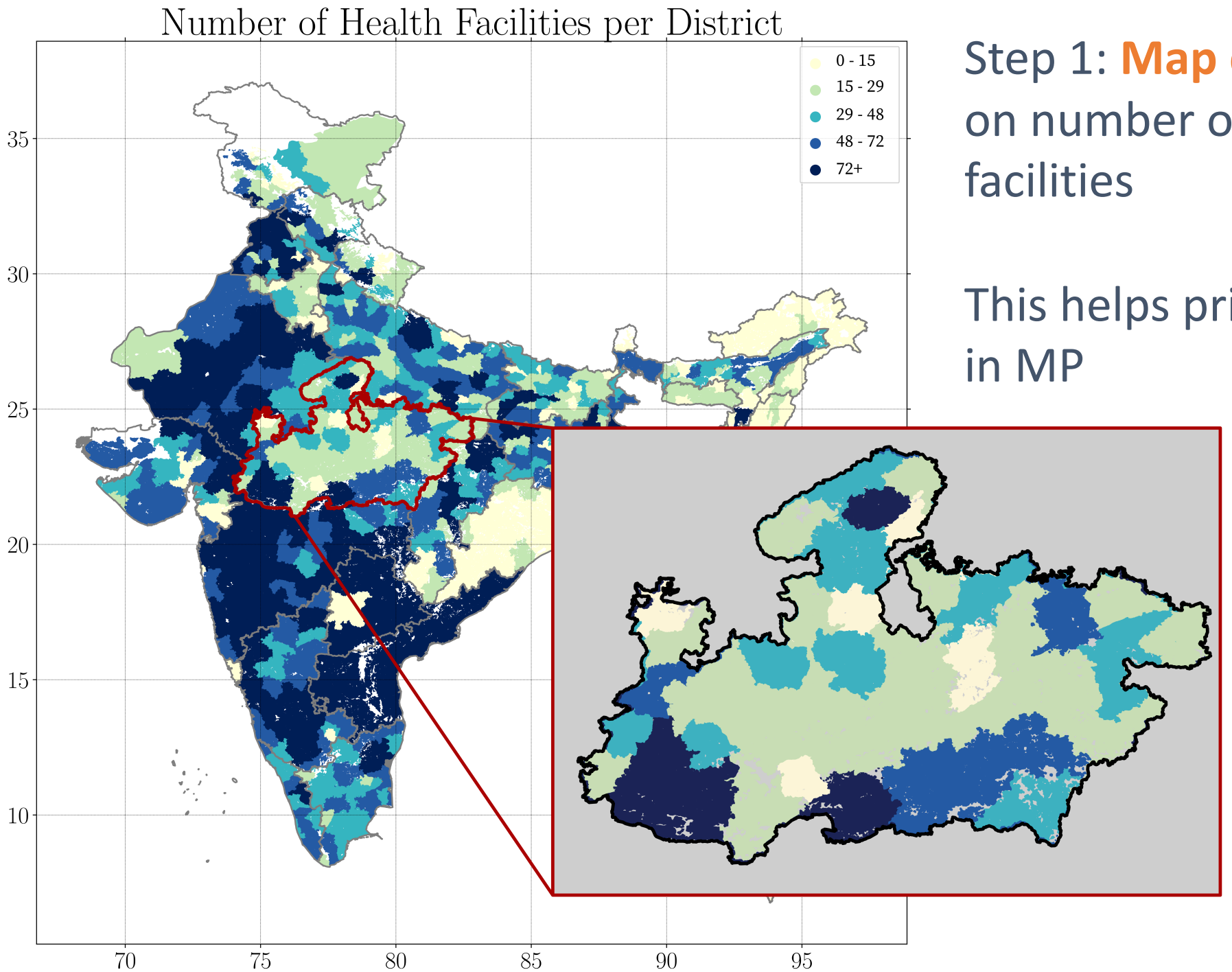


Step 1: **Map districts** based on number of public health facilities



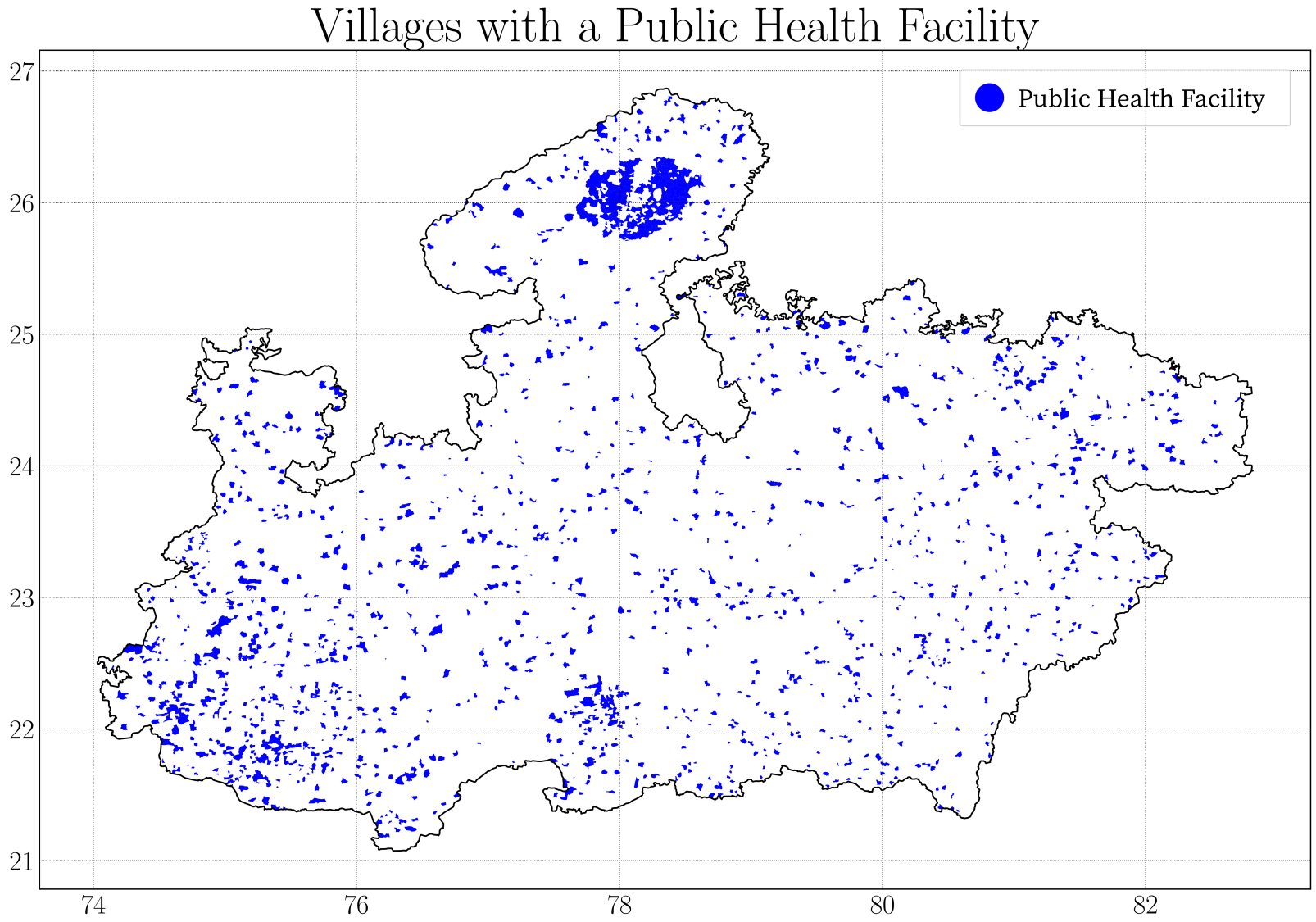
Step 1: **Map districts** based on number of public health facilities

Let us focus on Madhya Pradesh



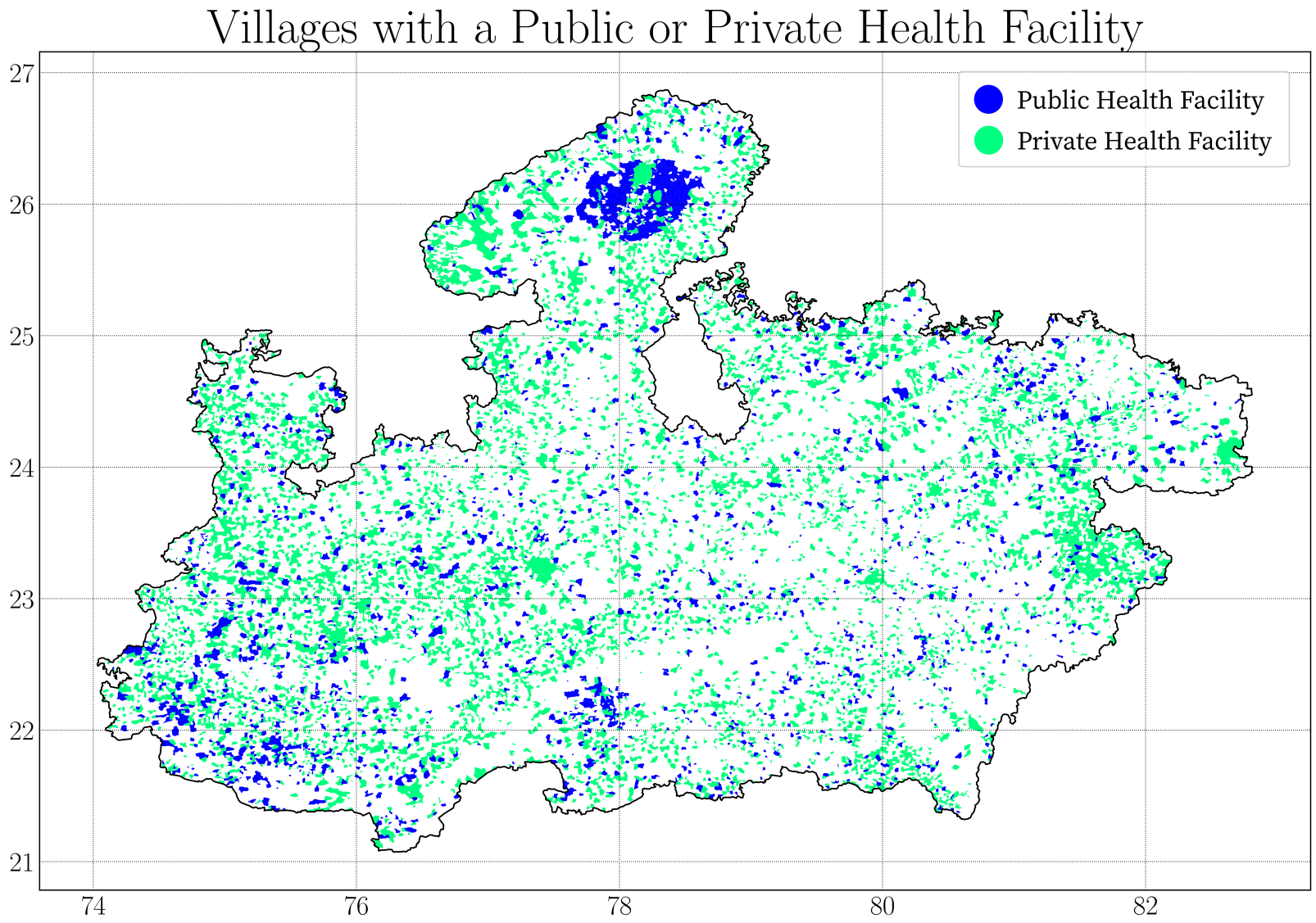
Step 1: **Map districts** based on number of public health facilities

This helps prioritize districts in MP



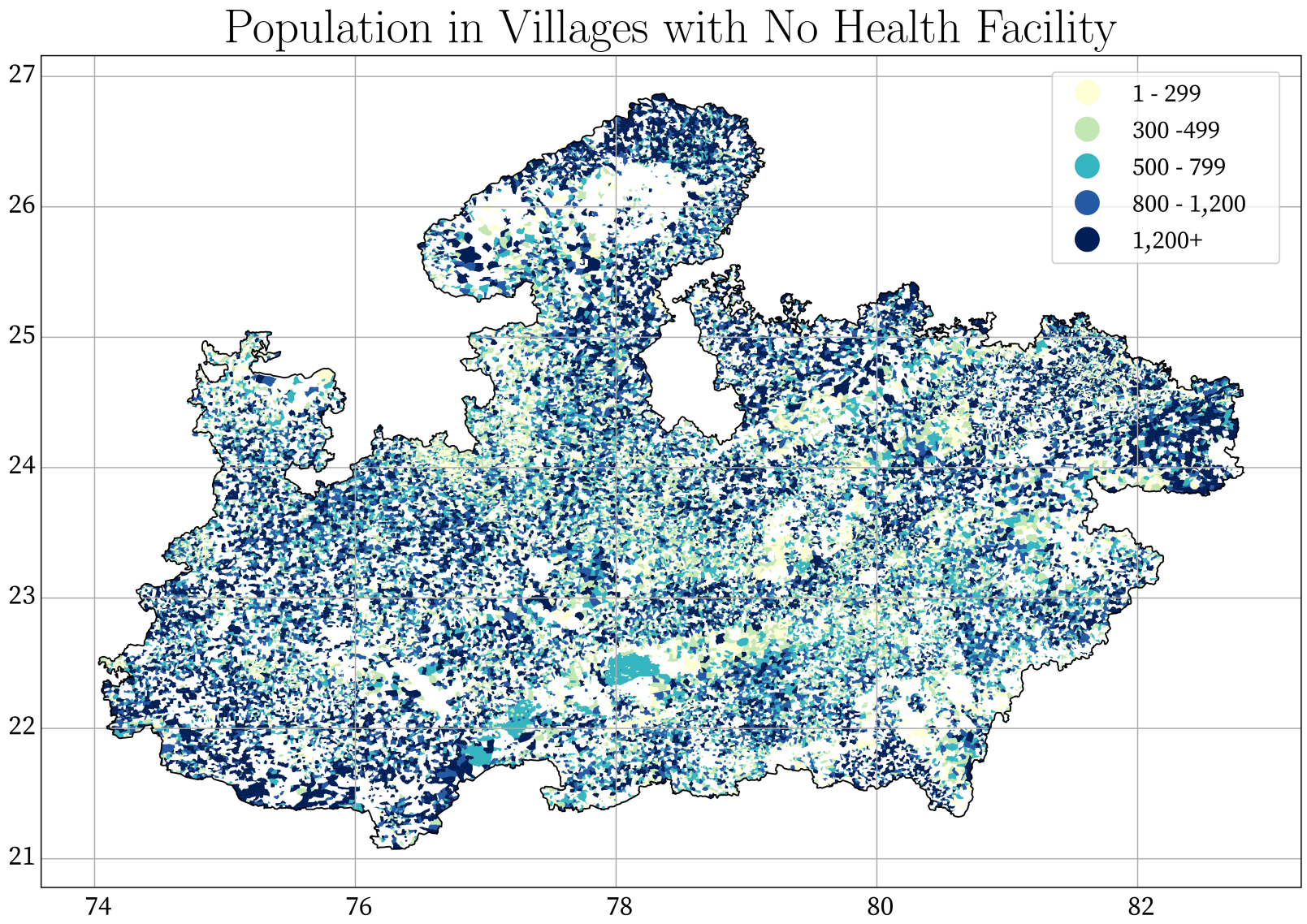
But district-level data is insufficient for targeting.

Step 2: **Map villages** with a public health facility



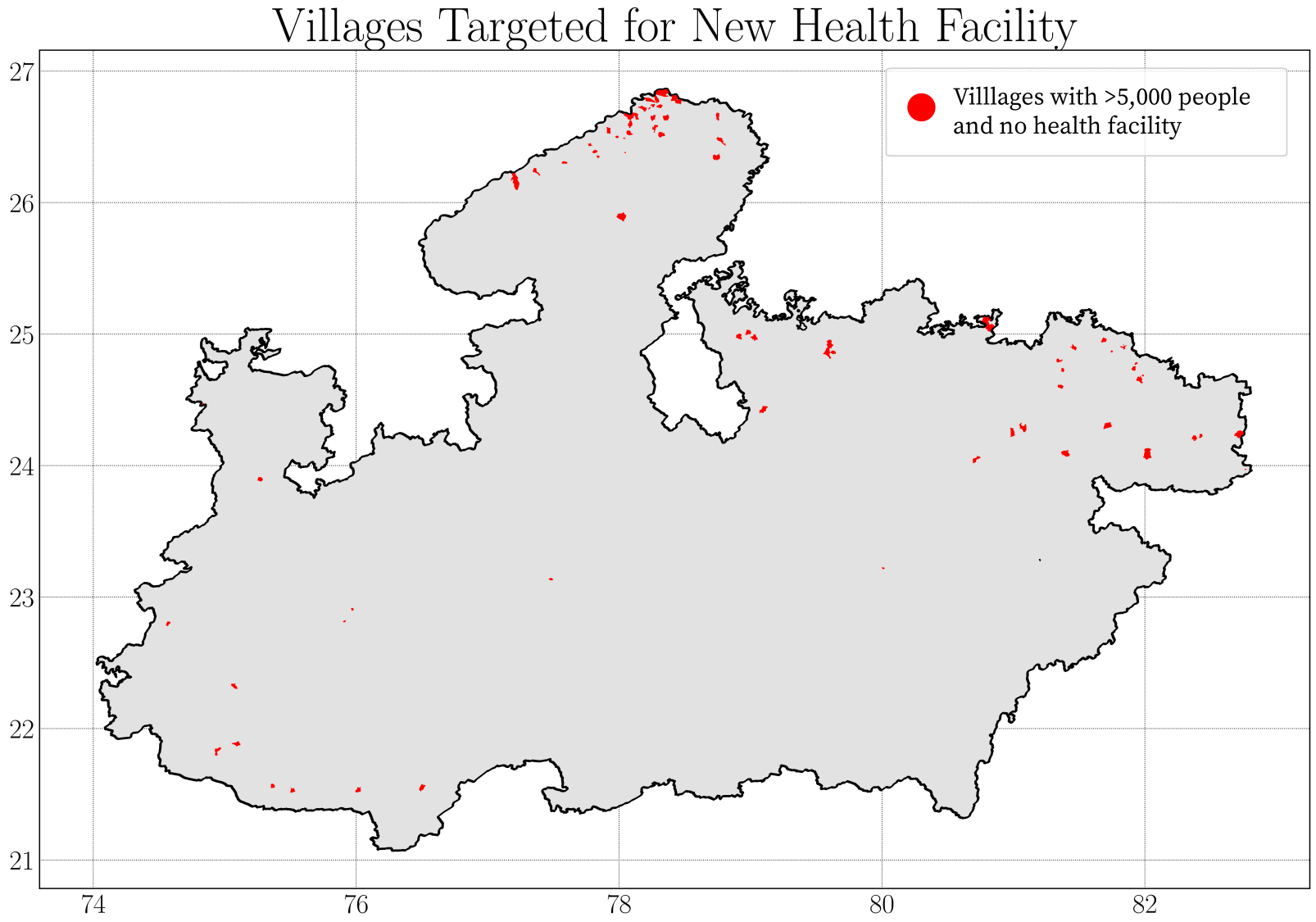
But what about private health facilities?

Step 3: Map villages with **public and private** health facilities



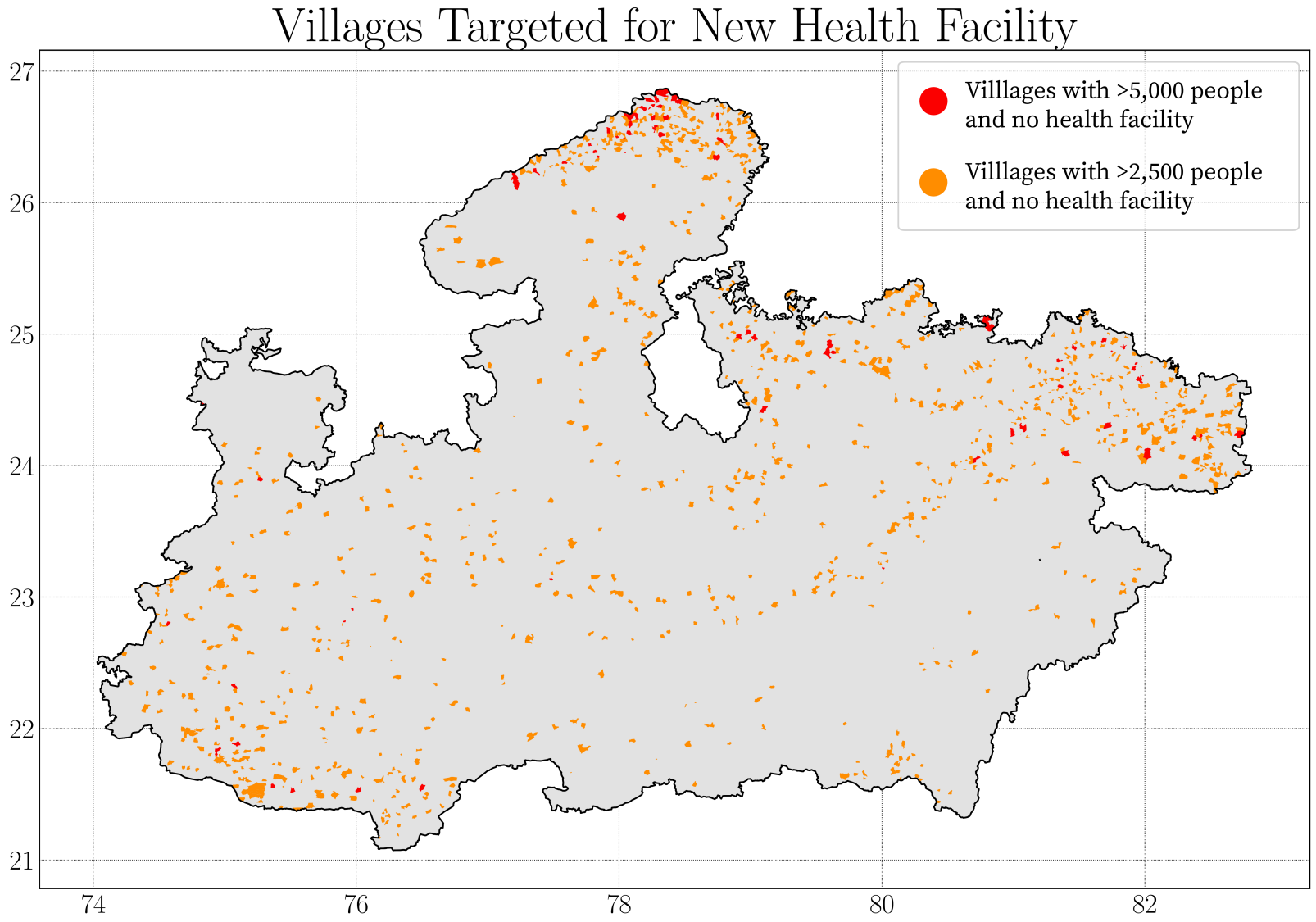
You need population data to know which villages most in need

Step 4: **Map villages with no facility by population size**



Step 5: Identify first priority villages:

74 villages in MP with no health facility and a population of over 5000



Step 6: Identify next priority villages:

1032 villages in MP with no health facility and a population of over 2500

SHRUG: Returns to Scale

The SHRUG is fully open

We are continuing to build this data platform

But to achieve maximal scale, we need to mobilize the crowd

1. Rewards for Contributors

Contributing to SHRUG helps research get found

Posting own data is great, but very slow to find, evaluate, and link.

SHRUG-connected data has a high-quality standard and is immediately linked to dozens of other data sources.

SHRUG is structured to maintain attribution

- Use three components → cite three papers
- Downloads automatically generate citation files
- Repeated nudges to eliminate accidental omissions

2. Copyleft Licensing

If you use SHRUG, you publish your data with SHRUG standards

ODbL-based license requires derivative products to be released with same license at time of publication.

Modeled on the Gnu Public License, a copyleft license for software that undergirds the open source software movement.

Like a time-limited patent, the license trades off the scientist's interest in not getting scooped and the public interest of having open data.

3. We Will Help You

Releasing highly usable data takes a lot of additional work.

Research teams aren’t rewarded for this work and may not have the capacity to release highly usable products.

We work with teams generating high value national data to help them normalize and integrate it with SHRUG.

Committing funds to this phase ex ante could further improve quantity and quality of data sharing in equilibrium.

A Vision for Data Collaboration

A health researcher is working with state-wide medical claims data.

By linking the data to SHRUG, she can study the highly local social determinants of health.

When she publishes, she also publishes village-level aggregates describing health outcomes with SHRUG identifiers.

Health module is now available to future users of the SHRUG, enabling dozens of additional studies.

Scale this process by all the researchers working with high resolution data in India

The Limits to Scale

There is a popular idea that assembling and linking datasets can be fully automated.

Our experience is that creating maximally usable products from messy administrative data requires significant high-skill labor input.

We are setting standards and building tools to make those human inputs as efficient as possible.

Next Steps

New data

Low-tech platform accessibility

Working with governments

Making contribution seamless

Development Data Lab

Building on a decade of research on the micro-determinants of growth and poverty alleviation in India

Founded in 2019 to make the world's data available for development policymaking, research, and private use

Current projects include:

- **Making the SHRUG the Wikipedia of dev data**
- **Building database of urban policies around India**
- **Research projects:**
  - **Urbanization: segregation, governance**
  - **Temporary migration**
  - **Intergenerational mobility**

Conclusion

Better data collaboration will unlock tremendous social value in policymaking, research, civil society, and the private sector.

The SHRUG framework mitigates the technical and institutional barriers to sharing.
This model is highly replicable in other contexts.

Many are energized about an open source model for the sciences.
We are building tools and institutions to harness that amazing energy.