Geocoding GMD

Bringing Geospatial Insights to Household Surveys

Ben Brunckhorst & Minh Cong Nguyen

GeoPov webinar series

October 30, 2025

Overview

New spatial modules in the Global Monitoring Database

Image generated by Google Gemini

- Harmonized variables to describe the location and spatial context of interviewed households
- Locations in surveys are mapped to a global grid system (H3 hexagons)
- The H3 grid is a key to merge geospatial data from other sources
- Geocoded GMD data is currently available for 100+ surveys from 44 countries
- More than 3 million households mapped to 130,000+ unique locations

Today

1. Motivation
2. How were surveys geocoded?
3. How to access the data?
4. How to use the data?

Thanks to the extended team!

Daylan Alberto Gomez, Rostand Mbouendeu, Santiago Baquero, D4G, Climate (GSG5) & regional stats teams.

Motivation

- Many surveys capture the location of observations at a granular level
  - Global datasets like GMD have not harmonized this information
- Geocoded observations are useful to:
  - understand the relationship between welfare and local spatial correlates
  - assess the distributional impacts of shocks and policies across space

Location information in household surveys takes different forms!

1. Point locations (coordinates) of households or EAs, displaced to protect privacy
2. Administrative areas, whose names and boundaries may change over time

⚠ Existing approaches to merge household and geospatial data can be frustrating, time consuming and lack transparency

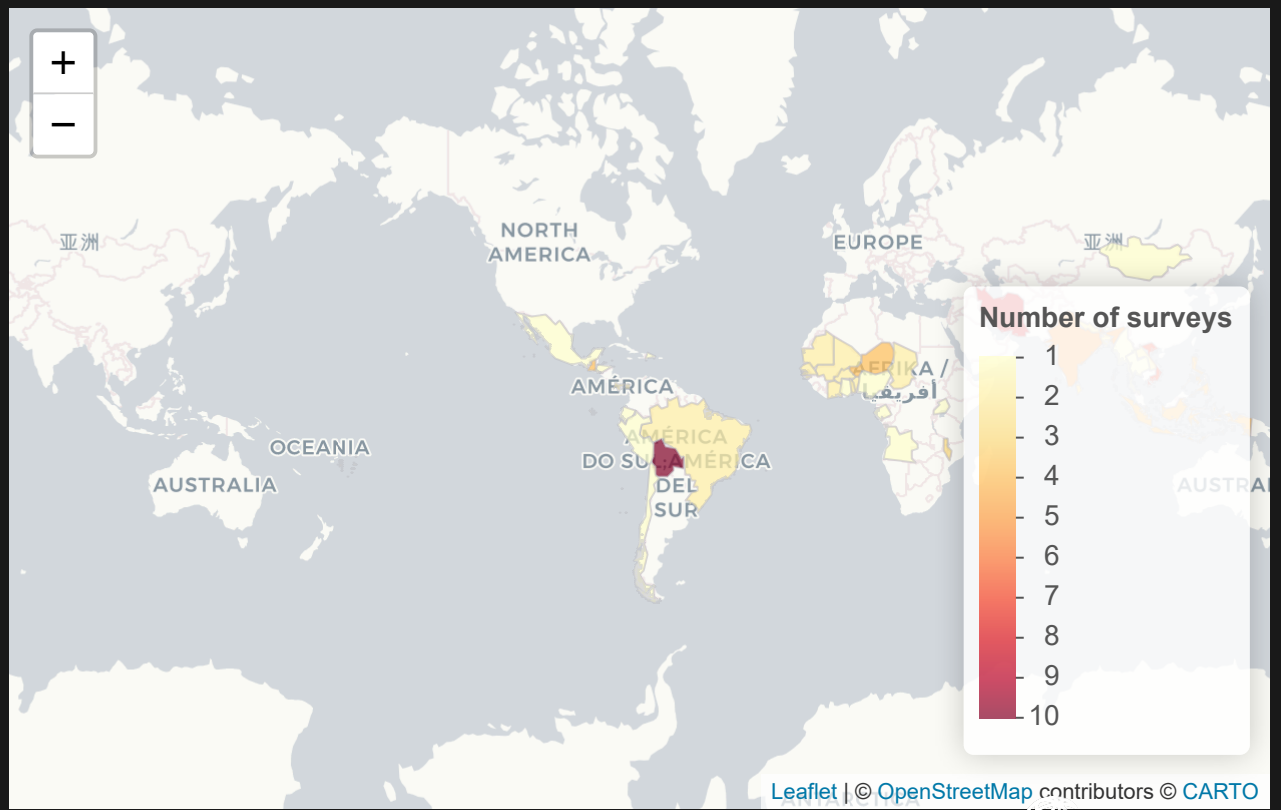
What's new?

Three spatial modules in GMD:

1. **LOC** - describes the location information available from the survey
2. **H3** - maps locations in the survey to a global grid (H3 hexagons)
3. **SPAT** - location (and time) matched variables extracted from spatial data

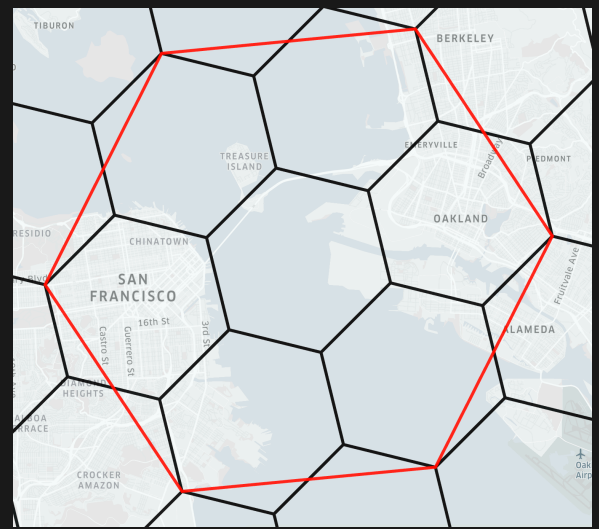
Coverage (July 2025):

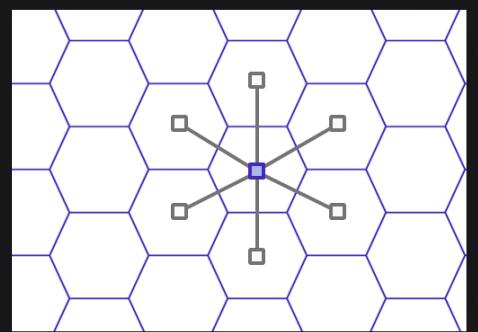
- 44 low- and middle-income countries
- 102 surveys
- 131,097 unique locations
- 3.2 million households



What is H3? Why hexagons?

- A geospatial indexing system provides a standardized, *time-invariant* key for joining disparate data sets by their location
- H3 is a hierarchical geospatial indexing system that partitions the world into hexagonal cells:
  - All neighboring cells are equidistant
  - Every hexagon has 7 children
  - 16 resolutions



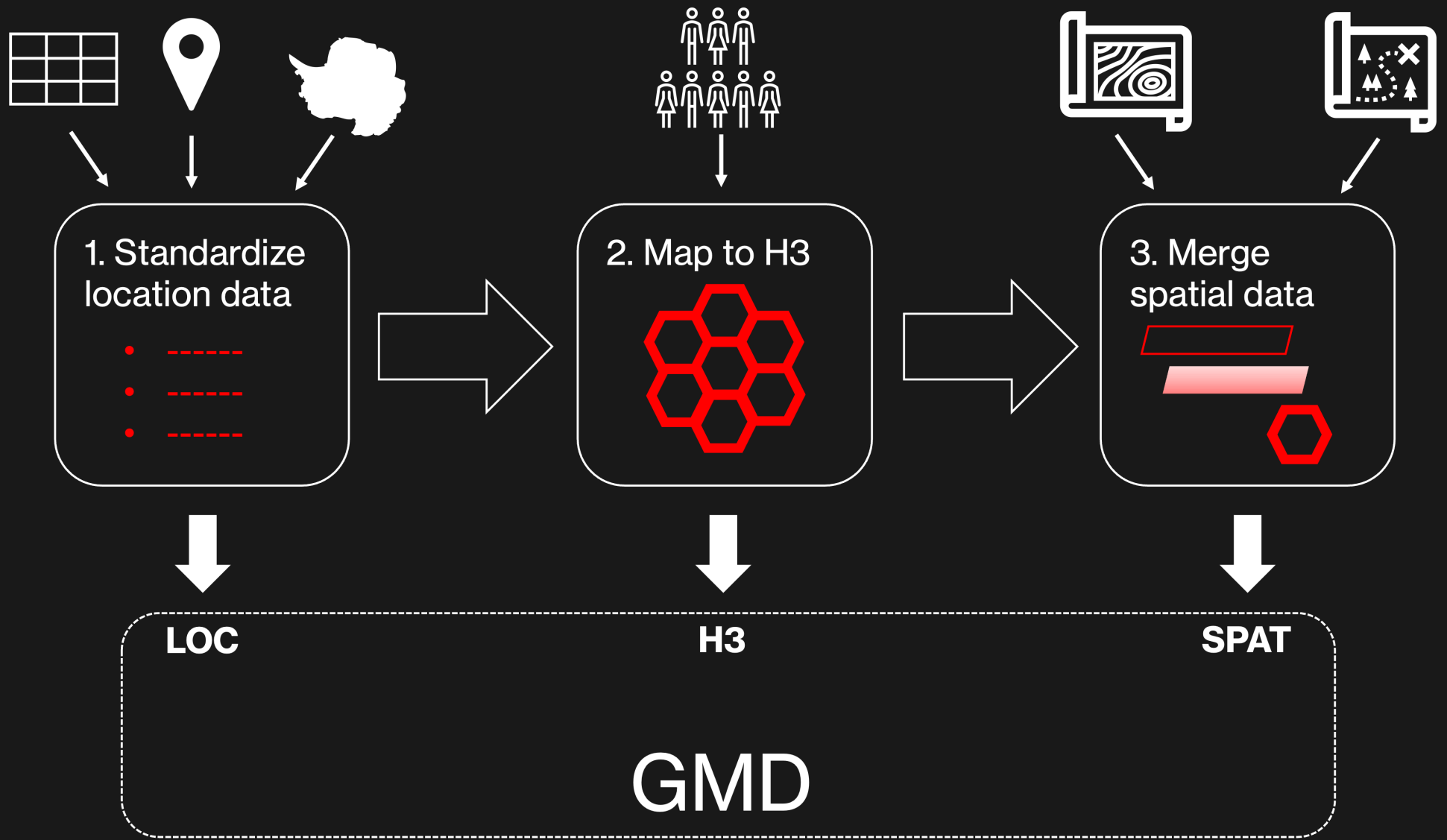


Other WBG projects using H3:

- Space2Stats - subnational geospatial aggregates
- WorldEx - data indexing and discovery
- geeLite - updating indicators derived from GEE

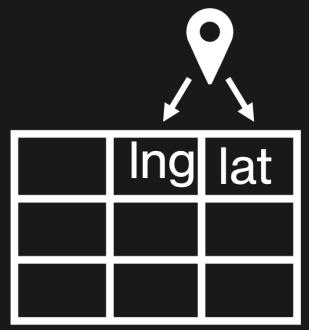
How were surveys geocoded?

Data pipeline



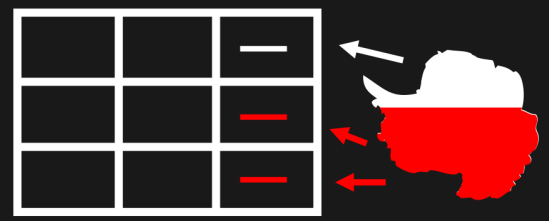
Step 1: Standardize location data

Inputs from POV economists, survey documentation, do-files, boundary research…



1. Mappable location data, either:
    - GPS coordinates (households or EAs), or
    - Boundary data (shapefile) with IDs corresponding to survey

- Info about the mappable location:
    - level of the GPS coordinates, boundary data source…



2. Household IDs matching GMD - or how to contruct `hhid`
    - Interview dates

LOC dataset

Describes location information available from a survey and used to map households

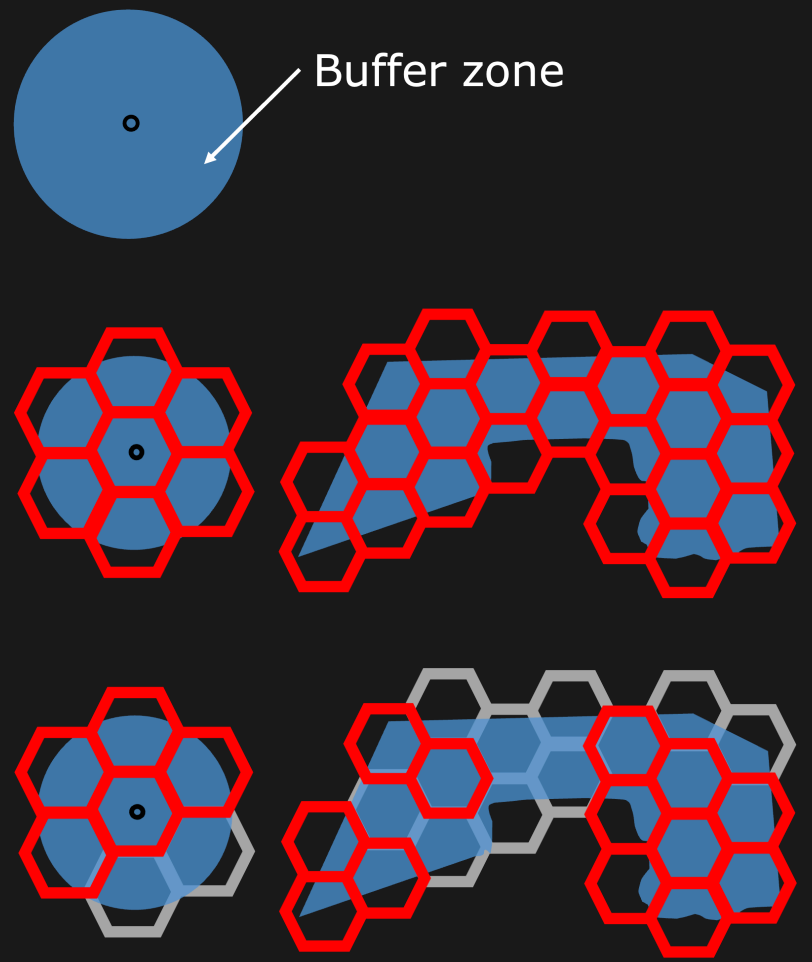
- household ID `hhid`
- interview date `int_year`, `int_month`, `int_day`
- location type `loc_type` “GPS” or “ADM”
- spatial unit ID `loc_id`
- mappable location `gps_lat`, `gps_lon` OR `adm_key` matching boundary data
- info about the mappable location `gps_level`, `adm_level`, `gps_mod`...

Household level data - one row per HH

- Merges 1:m with individual level GPWG/ALL modules on `hhid`
- Typically many households will be mapped to a spatial unit (`loc_id`)
  - not household specific locations

Step 2: Map to H3

1. Draw a buffer zone around *point locations*
   - Rural 5 km, Urban 2 km
   - Coordinates are typically modified
2. Cover the buffer zone *or* the survey region mapped by boundary data with H3 hexagons
   - Resolution 7 (area ~ 5 km^2)
3. Exclude hexagons with zero population or those across national borders
   - Households not mapped to a specific hexagon



H3 dataset

Maps locations in the survey to a global grid (H3 hexagons)

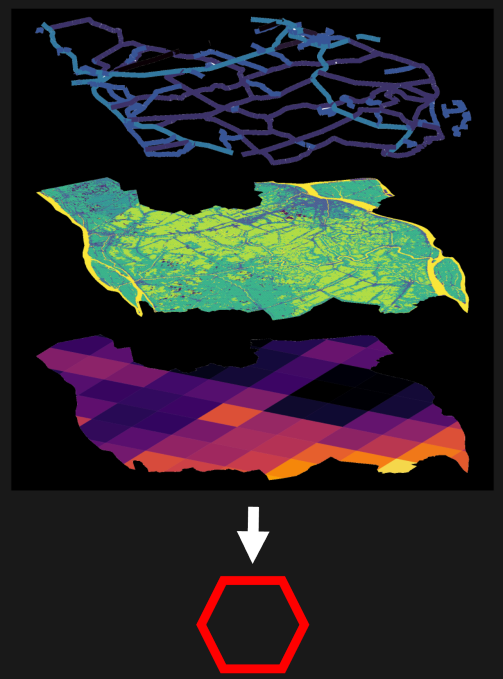
- spatial unit ID `loc_id`
- geocoded H3 resolution 7 cell ID `h3_7`
- parent H3 resolution 6 cell ID `h3_6`
- 2020 population extracted from gridded data (GHS-POP) `pop_2020`
  - generate population weighted zonal stats
  - estimate the probability a HH is located in a given H3 cell
  - check spatial anonymity

Location level data

- Each survey spatial unit (*multiple households*) is mapped to at least one hexagon
- merges m:m with LOC / SPAT on `loc_id`

Step 3: Merge spatial data

1. Index spatial data to H3 grid
    - DEC Space2Stats - already indexed to H3 level 6
    - Raster (gridded) data - e.g. temperature at hexagon centroid
    - Vector data - e.g. name of region containing hexagon, count of buildings in hexagon



2. Merge H3 module with H3 indexed spatial data - `h3` level
3. Aggregate H3 level data to survey spatial units - `loc_id` level
4. Merge with LOC module - `hhid` level

SPAT dataset

- Includes a set of standardized geospatial indicators
- Spatial variables are defined at the level of the geocoded location
- Some variables are defined *relative to the interview date*
  - these can vary within the same location if interview dates differ

Image generated by Google Gemini

Household level data - one row per household

- Merges 1:m with individual level GPWG/ALL modules on `hhid`

Geospatial variables in SPAT

To be expanded…

Population, Urbanization

- Area of spatial units
- Population
- Degree of urbanization

Divisions

- World Bank Official Boundaries (2025) - admin-1 and admin-2 IDs
- GAUL 2024 - level 1 and level 2 IDs
- GADM 4.1 - level 1 and level 2 IDs
- GHSL Urban Centre IDs

Economy, Infrastructure, Accessibility

- Nightlights
- Travel time
- Road density

Weather and climate

- 18 indicators for heat, cold, wet & dry
  - Mean/min/max, anomaly, days > X
- 3 reference periods
  - survey year, month before survey, climate reference period (1991-2020)

Practical considerations

- LOC, H3, and SPAT modules will merge with the *same version* of GMD data
  - `hhid` can change between versions
- Typically, *many* households mapped to *one* spatial unit represented by *many* hexagons

- we do not know exact household locations and need to protect privacy
- survey spatial units cover all possible locations of interviewed households

- Geocoded locations do not have a fixed area (number of hexagons can vary)
  - area and population are included in the SPAT module

Practical considerations

- Locations with different `loc_id` can:
  - include hexagons that overlap
  - include the exact same set of hexagons!

- `loc_id` identify unique locations in the *survey data* (LOC module)
  - GPS coordinates close together will be mapped to the same hexagons
- Possible to group `loc_id` based on the overlap / proximity of H3 hexagons
  - define panel locations across survey years

How to access the data?

Use the datalibweb API via Stata or R

- Install Stata or R package - if not already installed
- Agree to disclaimer and generate a token at datalibweb/
  - new token needed every 30 days

Get the SPAT module for Burkina Faso 2018 and merge with the ALL module:

Stata | R

```
dlw, token(your token)

dlw, country(bfa) year(2018) type(gmd) mod(spat)

tempfile data
save `data'

dlw, country(bfa) year(2018) type(gmd) mod(all)
merge m:1 year hhid using `data'
```

Data structure

LOC | H3 | SPAT

code	year	survname	hhid	int_year	int_month	int_day	urban	loc_id	loc_type	gps_lat	gps_lon	gps_
BFA	2018	EHCVM	349197	2018	11	17	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349115	2018	11	16	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349073	2018	11	16	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349038	2018	11	18	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349045	2018	11	16	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349170	2018	11	16	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349178	2018	11	17	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349030	2018	11	16	1	1	GPS	9.874635	-2.910476	EA cent

code	year	survname	hhid	int_year	int_month	int_day	urban	loc_id	loc_type	gps_lat	gps_lon	gps_
BFA	2018	EHCVM	349094	2018	11	17	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349189	2018	11	17	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349227	2018	11	17	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	349075	2018	11	16	1	1	GPS	9.874635	-2.910476	EA cent
BFA	2018	EHCVM	385097	2018	11	21	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385018	2018	11	20	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385050	2018	11	22	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385114	2018	11	19	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385112	2018	11	21	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385057	2018	11	19	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385122	2018	11	18	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385105	2018	11	19	0	2	GPS	9.891972	-4.316297	EA cent

code	year	survname	hhid	int_year	int_month	int_day	urban	loc_id	loc_type	gps_lat	gps_lon	gps_
BFA	2018	EHCVM	385118	2018	11	21	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385103	2018	11	20	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385045	2018	11	18	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	385022	2018	11	20	0	2	GPS	9.891972	-4.316297	EA cent
BFA	2018	EHCVM	351046	2019	5	2	1	3	GPS	9.905053	-2.936254	EA cent

How to use the data?

“Imagine how the new geocoded GMD data could help achieve the World Bank’s mission of ending poverty and boosting shared prosperity on a livable planet”

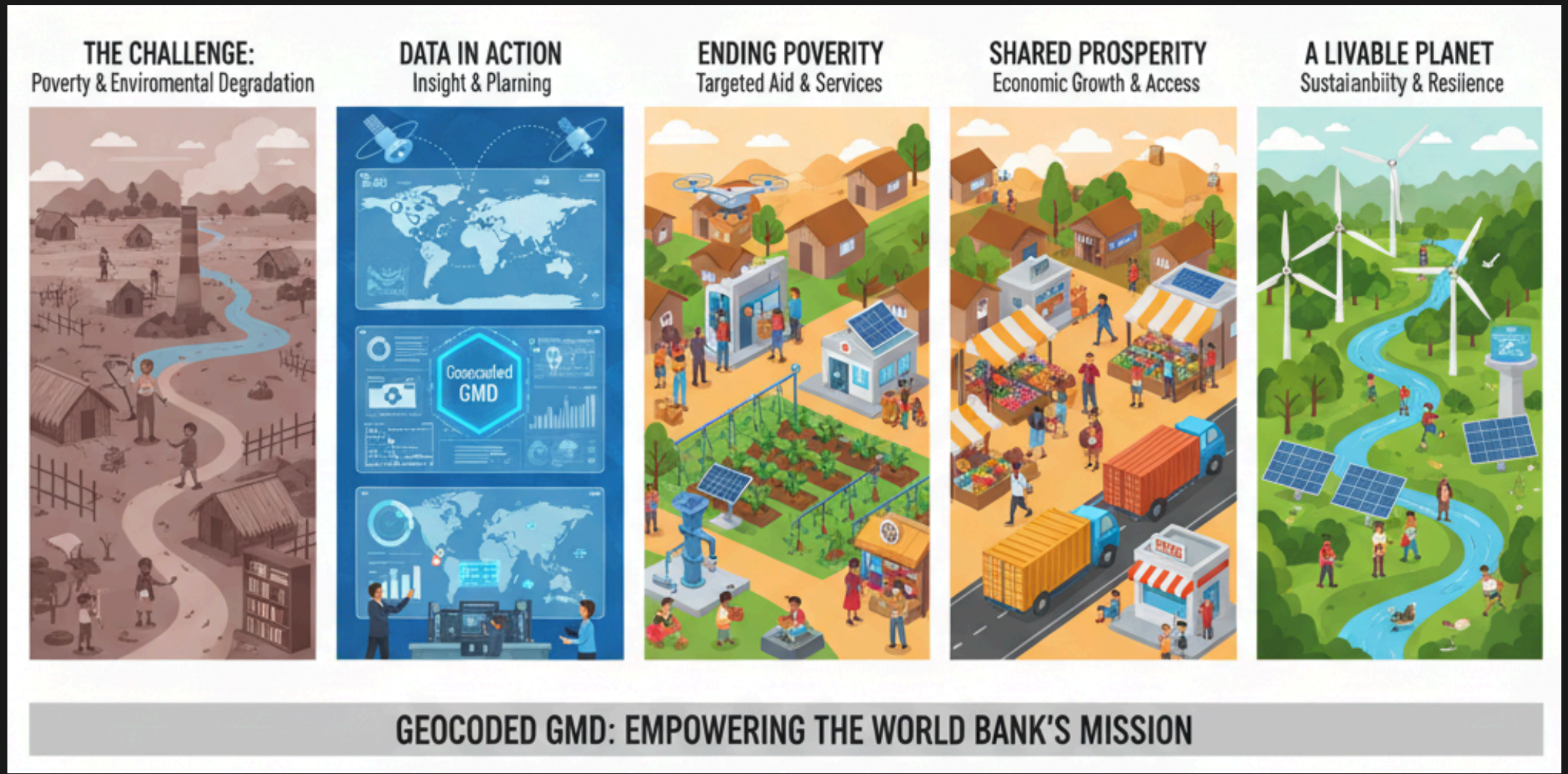
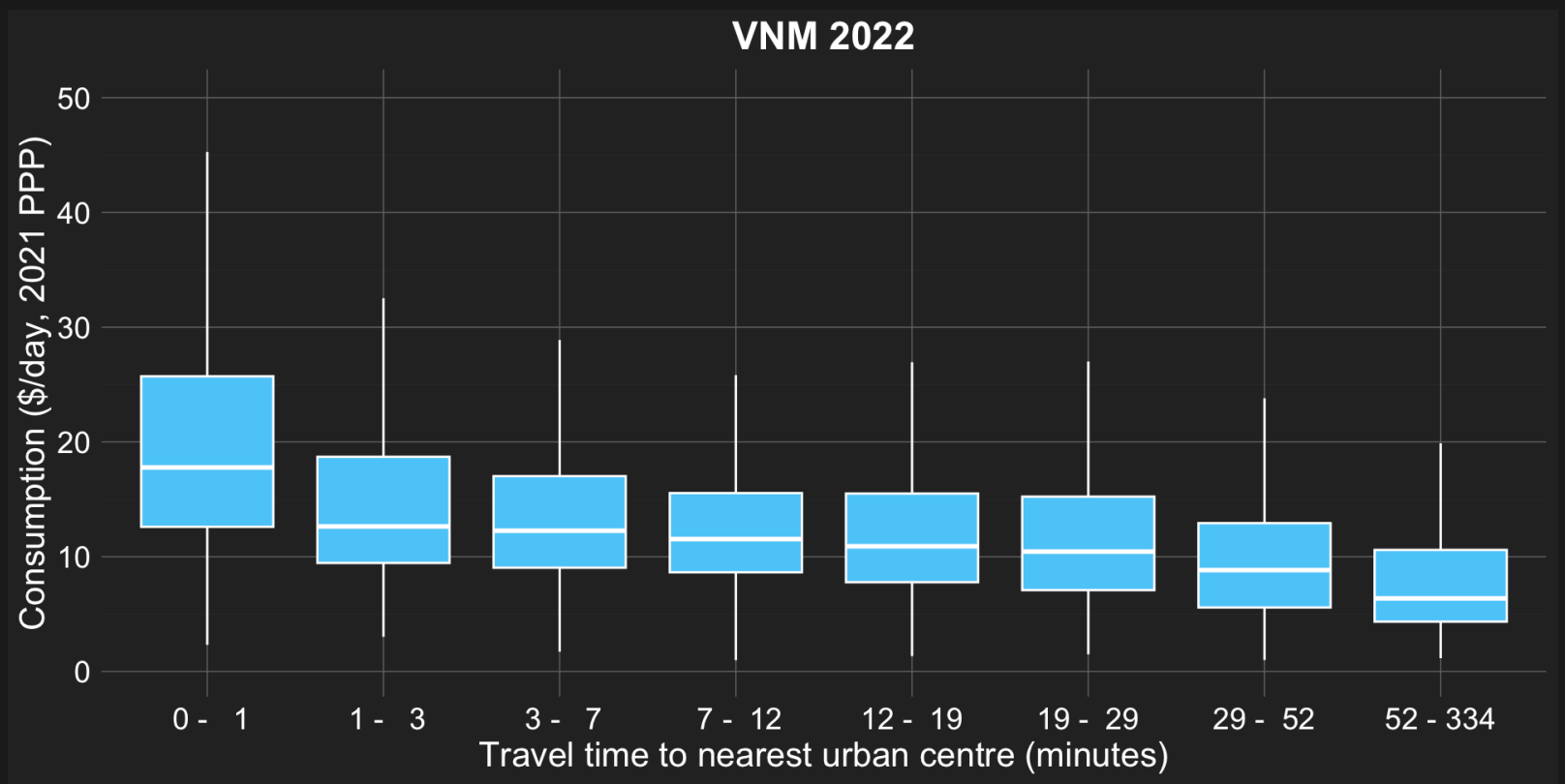


Image generated by Google Gemini

Example 1

What is the relationship between welfare and accessibility in Vietnam?

- Stata
- R



Example 2

What spatial variables determine poverty and welfare *within* subnational regions?

- Pooling geocoded GMD surveys from Vietnam
  - 2012, 2014, 2016, 2018, 2020, 2022

- : welfare of household in location in year
- : vector of household characteristics
- : vector of area characteristics
- : region fixed effects
- : year fixed effects

▶ Stata code

▶ R code

	Poor ($8.30)	Log welfare
Household size	0.056***	−0.104***
Urban	−0.100*	0.223**
Travel time to city (mins)	0.002**	−0.003***
Nighttime lights brightness	−0.075***	0.142***
Days >35C last month vs historical	0.000*	−0.000**
Heavy precipitation days last month vs historical	0.000*	−0.000*
Year fixed effects	X	X
Region fixed effects	X	X
Observations	209133	209133
R^2	0.265	0.404
* p < 0.1, ** p < 0.05, *** p < 0.01		
s.e. clustered at region level		

Many potential applications

- Revealing spatial inequality and exclusion
- Spatial determinants of job opportunities and employment outcomes
- Geographic targeting of social assistance and public investment
- Monitoring welfare impacts of shocks and policy

WISE-APP tool

1. Model the relationship between weather and welfare
2. Simulate welfare conditional on weather
3. Assess alternate policy scenarios

Your ideas!

Resources

- This presentation!
- Data Dictionary and Survey List
- GMD documentation website - under development
- GMD GitHub repo

GeoPov Data & Methods repository

- examples and code to work with the GMD data
- coming FY26

Thanks

Annex

Variable definitions

LOC | H3 | SPAT

Variable name	Variable label	Note
code	Country code	NA
year	Starting year of survey	NA
survname	Survey acronym	NA
hhid	Household ID	Match with GMD hhid
hhid_orig	Household ID	Household ID in source, optional
int_year	Interview year	NA
int_month	Interview month	[1-12]
int_day	Interview day	[1-31]
urban	Urban	Household defined as urban in survey (0/1)
loc_id	Spatial unit ID	For loc_type = “ADM”, loc_id = adm_key. For loc_type = “GPS”, loc_id defines a unique combination of gps_lat and gps_lon.
loc_type	Location type	“GPS” or “ADM”
gps_lat	Latitude (decimal degrees)	If loc_type = “GPS”

Variable name	Variable label	Note
gps_lon	Longitude (decimal degrees)	If loc_type = “GPS”
gps_level	Coordinates level	If loc_type = “GPS”. Location identified by coordinates, e.g., “HH”, “EA centerpoint” or “Other”
gps_mod	Coordinates modified	If loc_type = “GPS”. Are the coordinates modified from true point location? “Yes” or “No”
gps_priv	Coordinates confidential	If loc_type = “GPS”. Coordinates data privacy: “Confidential”, “Official Use Only” or “Public”
adm_key	Spatial unit ID for shapefile	If loc_type = “ADM”. Corresponding with “adm_key” in shapefile
adm_name	Region name	If loc_type = “ADM”. Optional
adm_src	Map source	If loc_type = “ADM”. Source of boundary data, e.g., “NSO”, “GOV”, “WB”, “GADM”…
adm_year	Map year	If loc_type = “ADM”. Year of boundaries mapped
adm_level	Map level	If loc_type = “ADM”. Level of admin area/region mapped, e.g., “admin3”, “admin4”, “PSU”…

Survey list

Last updated July 31, 2025

code	year	survname
AGO	2018	IDREA
BEN	2018	EHCVM
BEN	2021	EHCVM
BFA	2018	EHCVM
BFA	2021	EHCVM
BGD	2016	HIES
BGD	2022	HIES
BOL	2012	EH
BOL	2013	EH
BOL	2014	EH
BOL	2015	EH
BOL	2016	EH
BOL	2017	EH
BOL	2018	EH
BOL	2019	EH
BOL	2020	EH

code	year	survname
BOL	2021	EH
BRA	2020	PNADC-E5
BRA	2021	PNADC-E5
BTN	2022	BLSS
CHL	2022	CASEN
CIV	2018	EHCVM
CIV	2021	EHCVM
DOM	2022	ECNFT-Q03
ECU	2022	ENEMDU
FJI	2013	HIES
FJI	2019	HIES
GAB	2017	EGEP
GHA	2016	GLSS-VII
GMB	2015	IHS
GMB	2020	IHS
GNB	2018	EHCVM
GNB	2021	EHCVM
GTM	2000	ENCOVI
GTM	2006	ENCOVI
GTM	2014	ENCOVI
GTM	2023	ENCOVI

code	year	survname
HND	2019	EPHPM
IDN	2005	SUSENAS
IDN	2010	SUSENAS
IDN	2015	SUSENAS
IDN	2020	SUSENAS
IND	2009	NSS-SCH1
IND	2011	NSS-SCH2
IND	2019	CPHS
IND	2020	CPHS
IND	2021	CPHS
IRN	2013	HEIS
IRN	2014	HEIS
IRN	2015	HEIS
IRN	2016	HEIS
IRN	2017	HEIS
IRN	2018	HEIS
IRN	2019	HEIS
IRN	2020	HEIS
LAO	2012	LECS
LAO	2018	LECS
LKA	2012	HIES

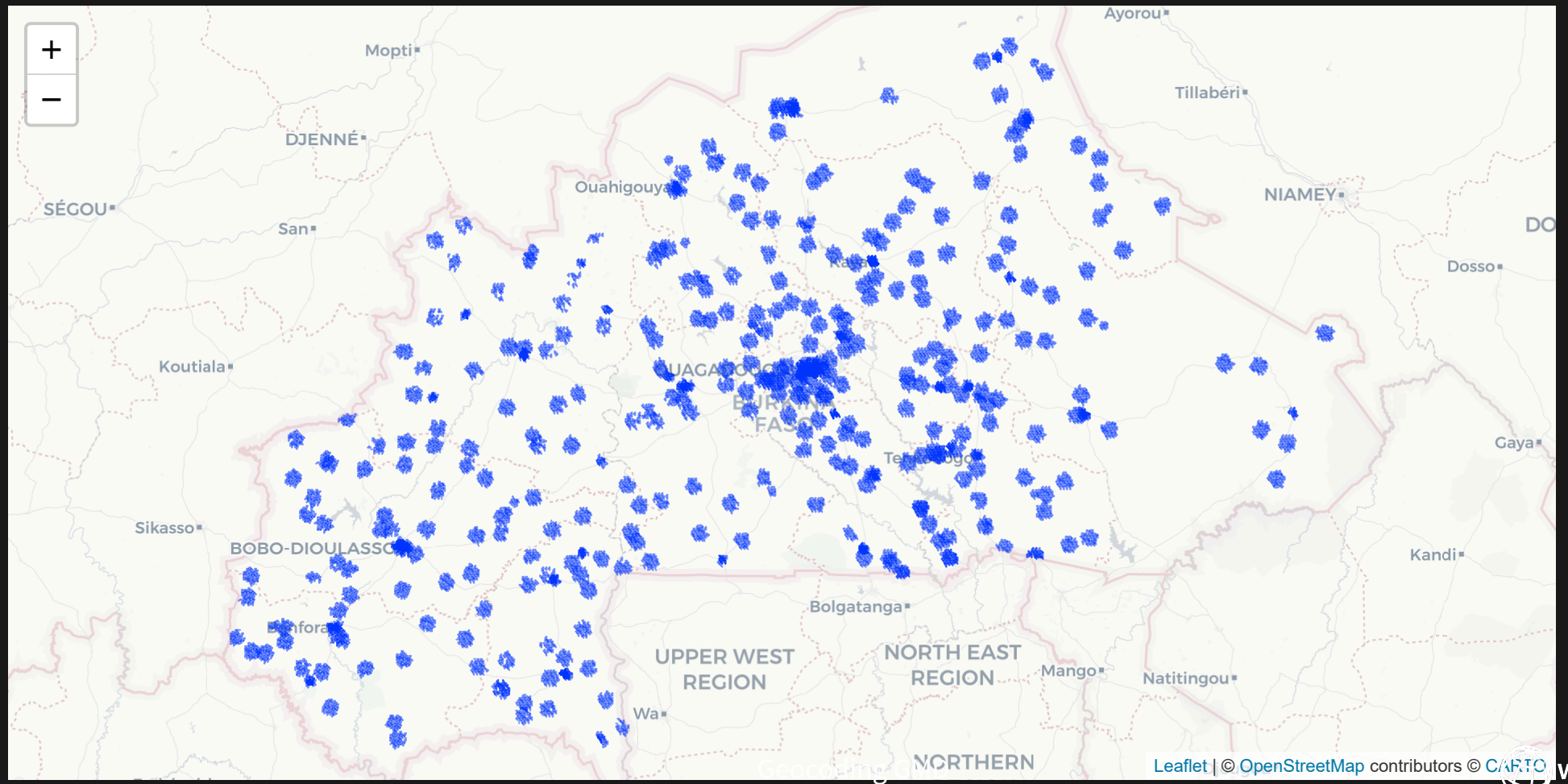
code	year	survname
LKA	2016	HIES
LKA	2019	HIES
MEX	2022	ENIGHNS
MLI	2018	EHCVM
MLI	2021	EHCVM
MMR	2017	MLCS
MNG	2022	HSES
MRT	2014	EPCV
MRT	2019	EPCV
MWI	2010	IHS-III
MWI	2016	IHS-IV
MWI	2019	IHS-V
NER	2011	ECVMA
NER	2014	ECVMA
NER	2018	EHCVM
NER	2021	EHCVM
NGA	2018	LSS
NPL	2010	LSS-III
NPL	2022	LSS-IV
PAN	2021	EH
PAN	2023	EH

code	year	survname
PER	2022	ENAHO
PHL	2006	FIES
PHL	2009	FIES
PHL	2012	FIES
PHL	2015	FIES
SEN	2018	EHCVM
SEN	2021	EHCVM
SLV	2022	EHPM
TCD	2018	EHCVM
TCD	2022	EHCVM
TGO	2018	EHCVM
TGO	2021	EHCVM
THA	2009	SES
TLS	2014	TLSLS
TON	2021	HIES
UGA	2019	UNHS
VNM	2010	VHLSS
VNM	2012	VHLSS
VNM	2014	VHLSS
VNM	2016	VHLSS
VNM	2018	VHLSS

Mapping survey locations in R

H3 hexagons | Point locations

▶ Show code



Merging spatial data with the H3 module in R

H3 indexed data | Raster/gridded data | Vector data

```r
# get GMD H3 data
library(dlw)
h3 <- dlw_get_gmd("BFA", 2018, "H3")

# get h3 indexed data (e.g. built up area from space2stats)
library(httr2)
library(jsonlite)

base_url <- "https://space2stats.ds.io"
hex_ids <- unique(h3$h3_6)
request_payload <- list(hex_ids = hex_ids,
                        fields = list("sum_built_area_m_2020"))
req <- request(base_url) |>
  req_url_path_append("summary_by_hexids") |>
  req_body_json(request_payload)
s2s <- req |> req_perform() |> resp_body_string() |> fromJSON(flatten = TRUE)

# merge on h3_6
library(dplyr)
h3_s2s <- left_join(h3, s2s, join_by(h3_6 == hex_id))
head(h3_s2s)

# aggregate to loc_id level (e.g., sum of built area)
loc_s2s <- h3 |>
```