# Private, Public, Personal: Shifting Patterns in Geospatial Data Sources in Geographic Research

Gabriel Appiah, Mira Kaufman, Billy Cooney and Clio Andris
Georgia Institute of Technology

# Has there been a 'data shift' in peer-reviewed GIS research?

# Research Questions

(1) How has government geospatial data usage changed over time in GIS research?

(2) Which types of GIS subfields (spatial statistics, VGI, ABM) tend to use government data?

(3) Are funding sources listed in GIS analyses? Are data made available?

# Bibliometric Analysis

## -Six journals*

International Journal of Geographical Information Science (IJGIS);
Computers, Environment and Urban Systems (CEUS);
Transactions in GIS (TGIS);
Geographical Analysis (GA);
Environment and Planning B (EPB);
Annals of the American Association of Geographers (AAAG).

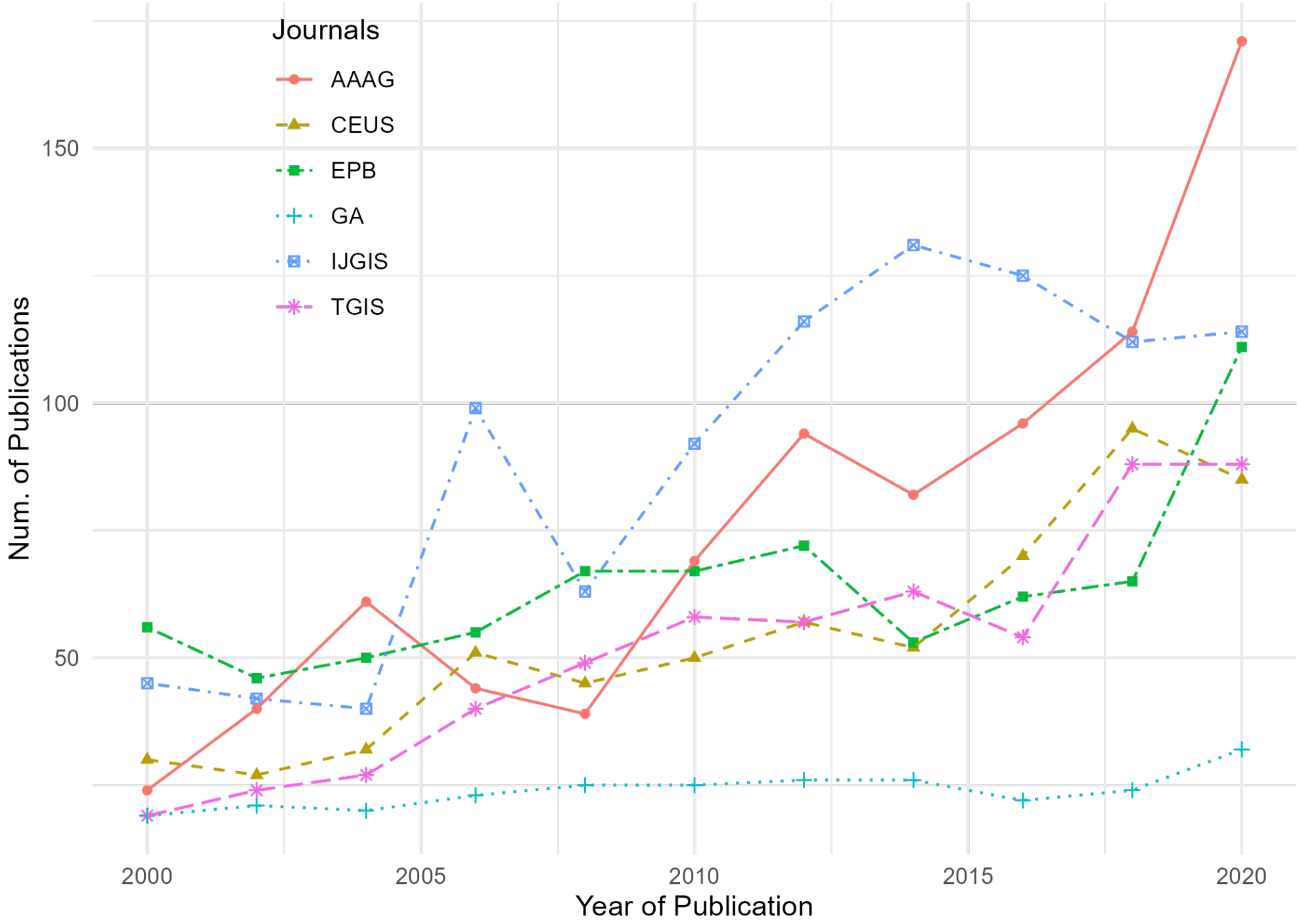
## -2,192 articles (2000 to 2020 every other year). In total, we reviewed 3,537 articles.

*"Please list the top six GIS/Geography/Urban Analytics flagship journals that publish articles that conduct geospatial data analysis."

# We classified data type as a private, public, or fieldwork source.

| Data Source Types | Description |
| --- | --- |
| Private | Data collected by private companies (e.g., Google, Weibo, Yelp, CitiBike, Facebook, Twitter) |
| Government | Data collected by governmental organisations (e.g., U.S. Census Bureau, EPA, USGS) |
| Field survey | Data collected on-location, through interviews, focus group discussions, field observations, etc. |
| Two or more sources | Articles using at least two different data sources |

# Publication volume by journal



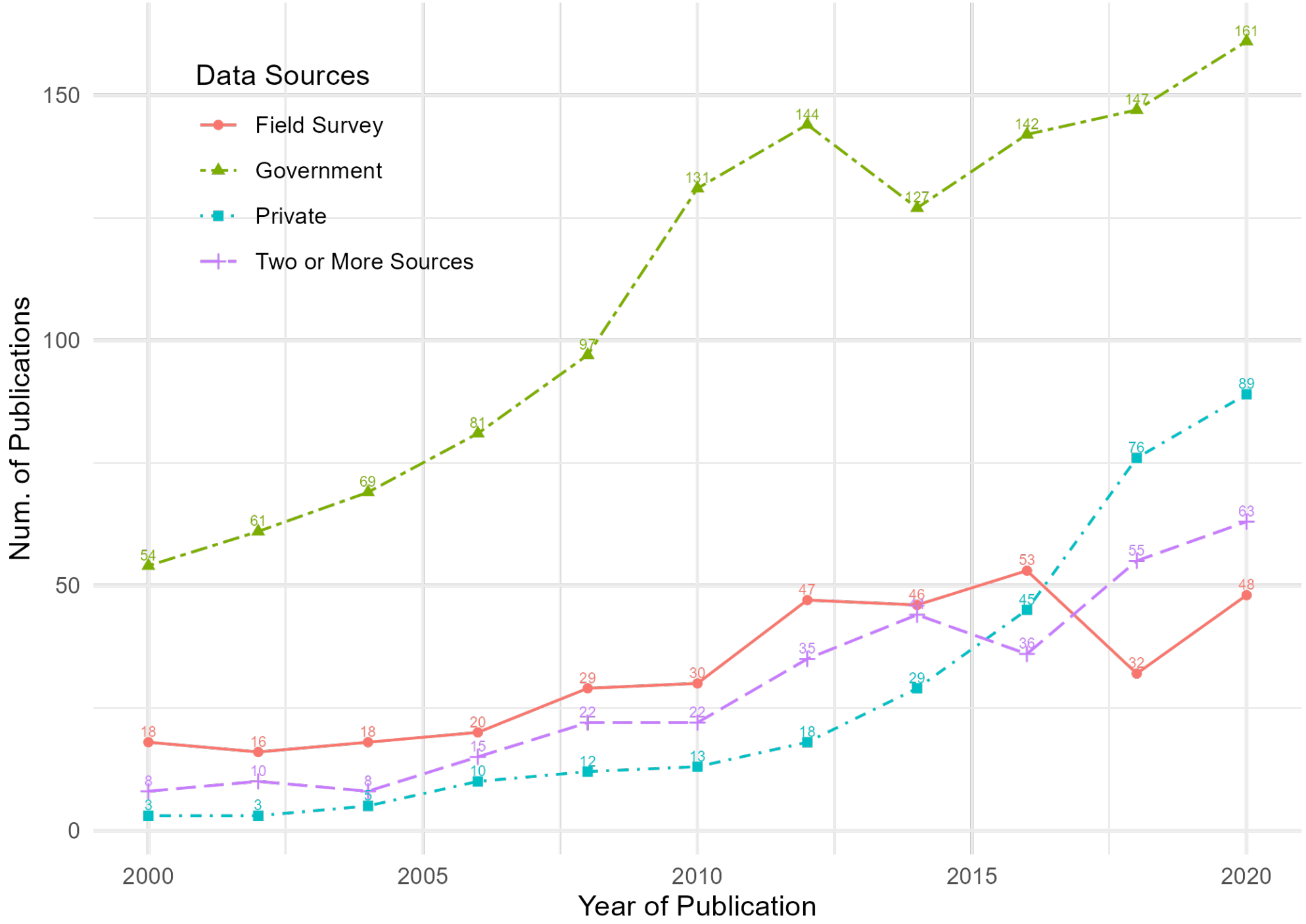
# Research Questions

(1) How has government geospatial data usage changed over time in GIS research?

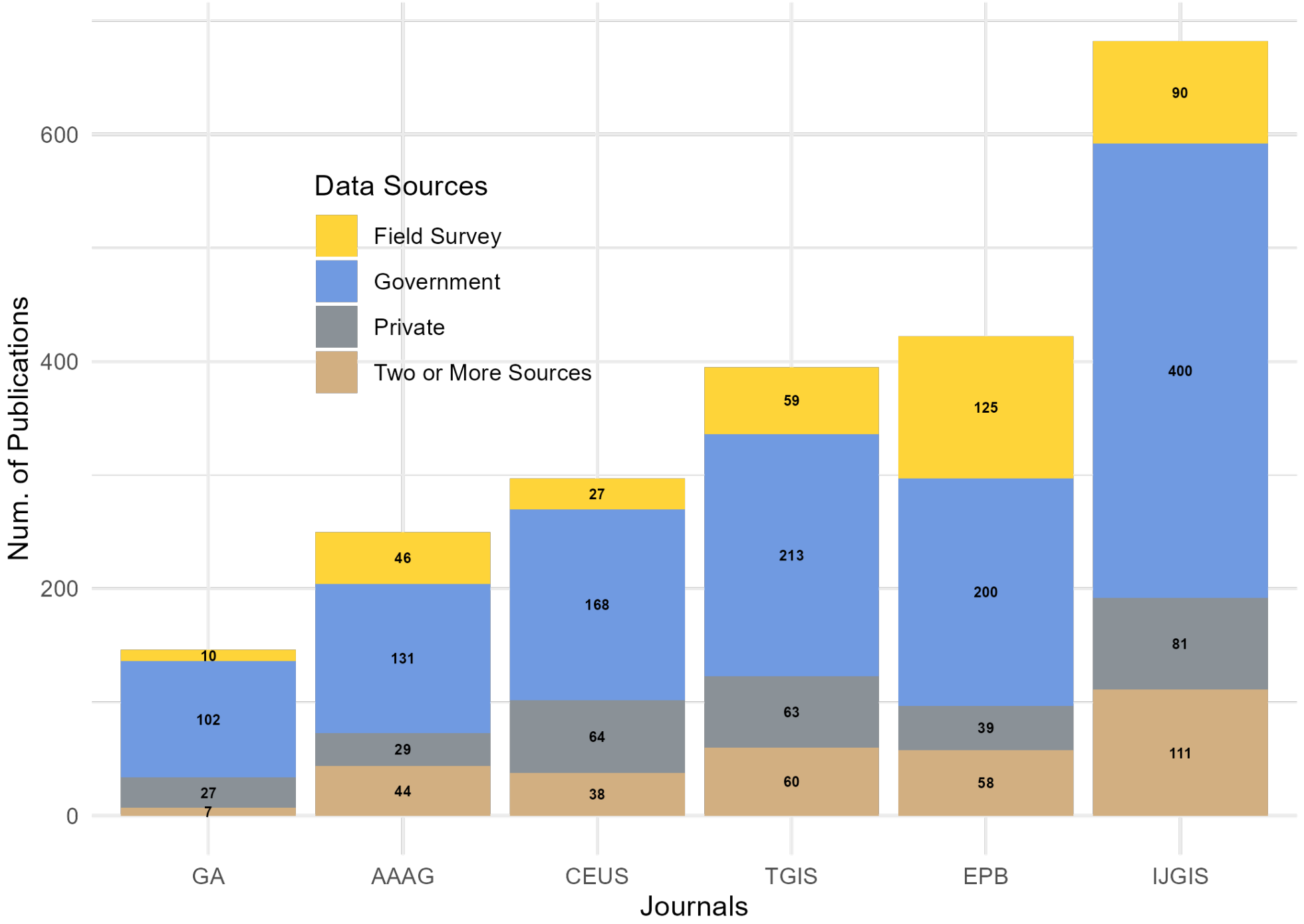
(2) Which types of GIS subfields (spatial statistics, VGI, ABM) tend to use government data?

(3) Are funding sources listed in GIS analyses? Are data made available?

# Types of data used in journal articles over time



# Data type by journal



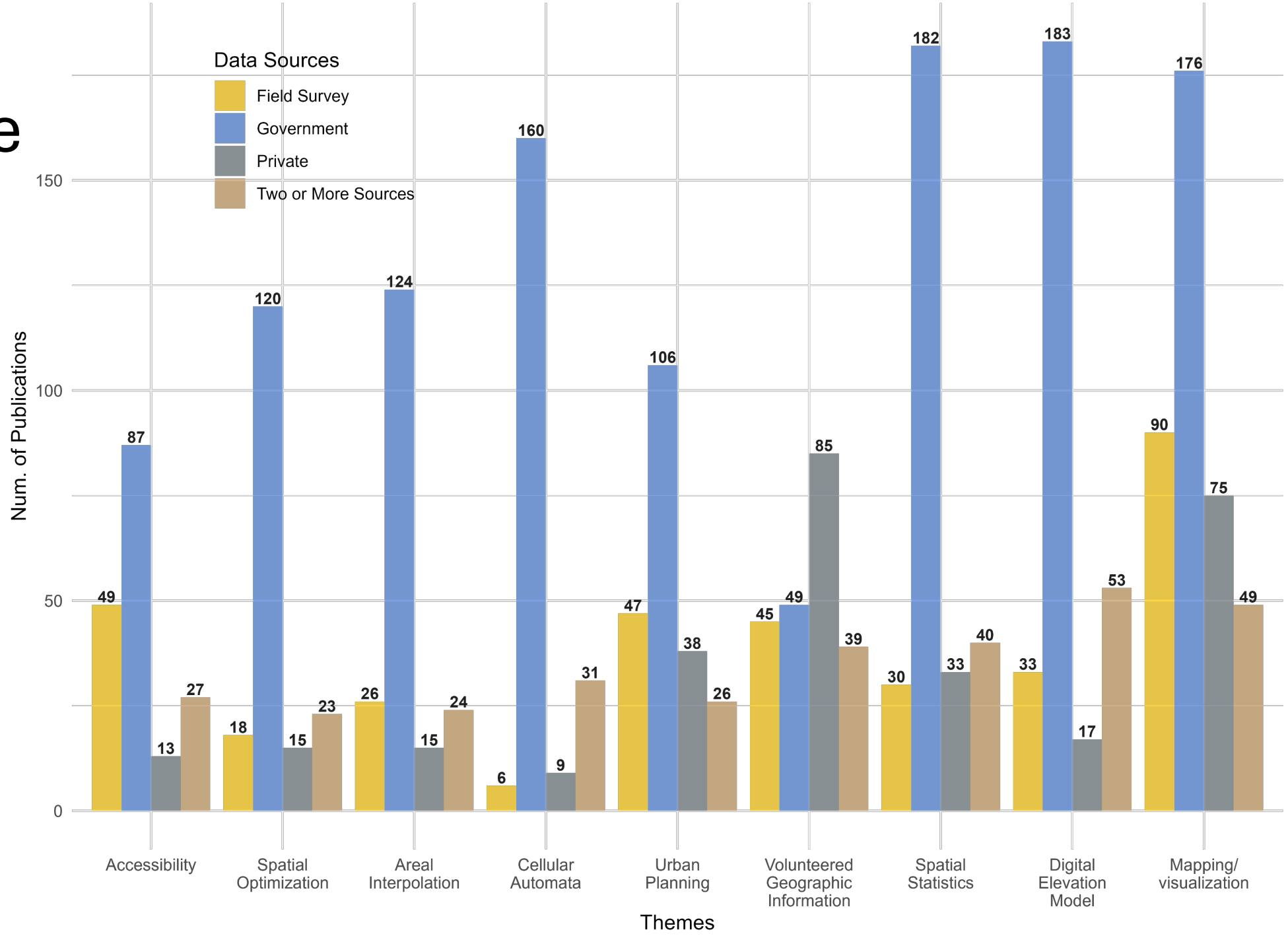
# Research Questions

(1) How has government geospatial data usage changed over time in GIS research?

(2) Which types of GIS subfields (spatial statistics, VGI, ABM) tend to use government data?

(3) Are funding sources listed in GIS analyses? Are data made available?

# Data type by article theme



# Unpublished Result

Did certain data types pair with analysis methods?

1 (basic)-Summary statistics
2 (medium) -Logistic regression
3 (advanced) -Agent based model or ML

Fieldwork tended to have more basic methods.
Govt and private sector tended to have more advanced methods.

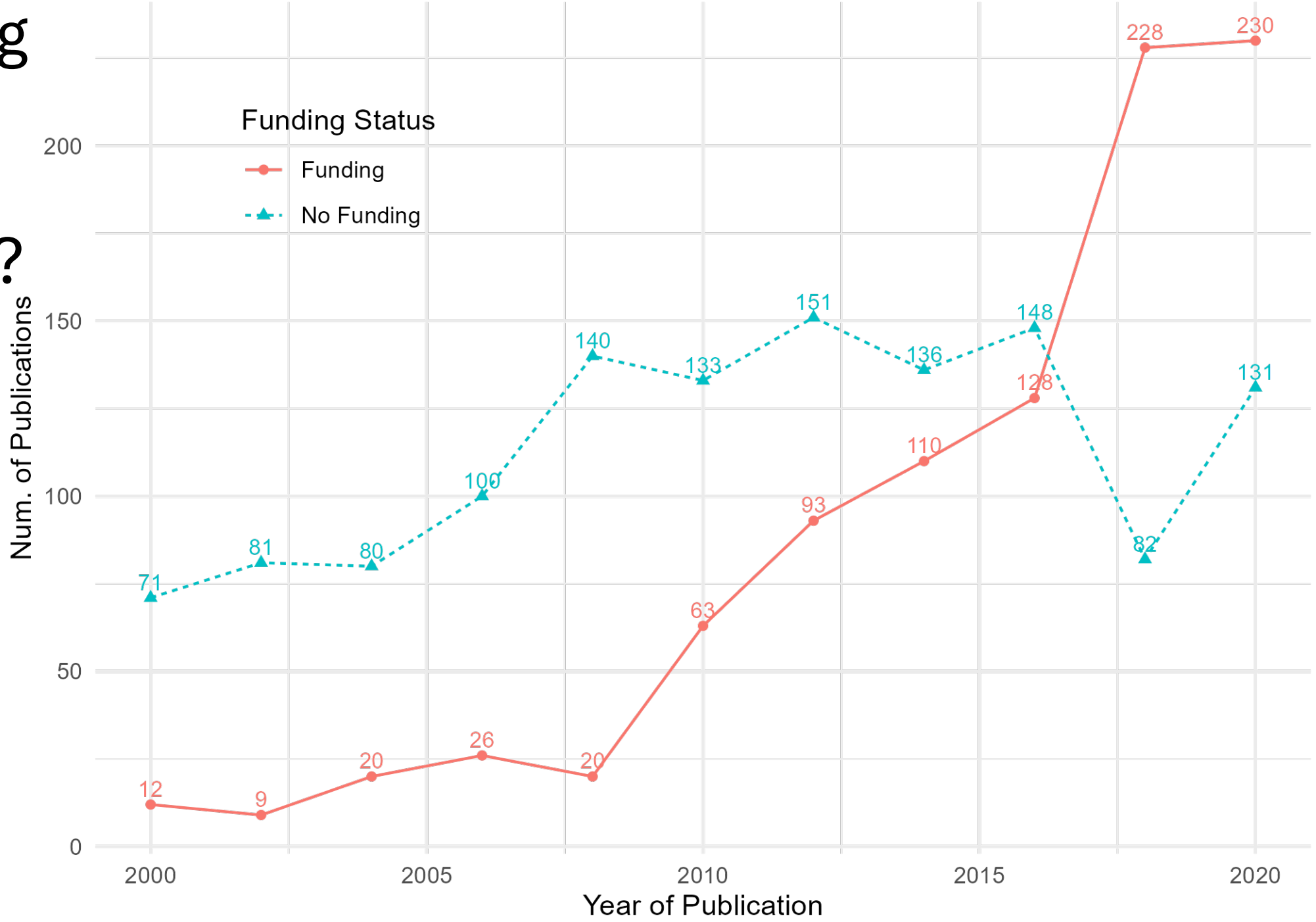
# Research Questions

(1) How has government geospatial data usage changed over time in GIS research?

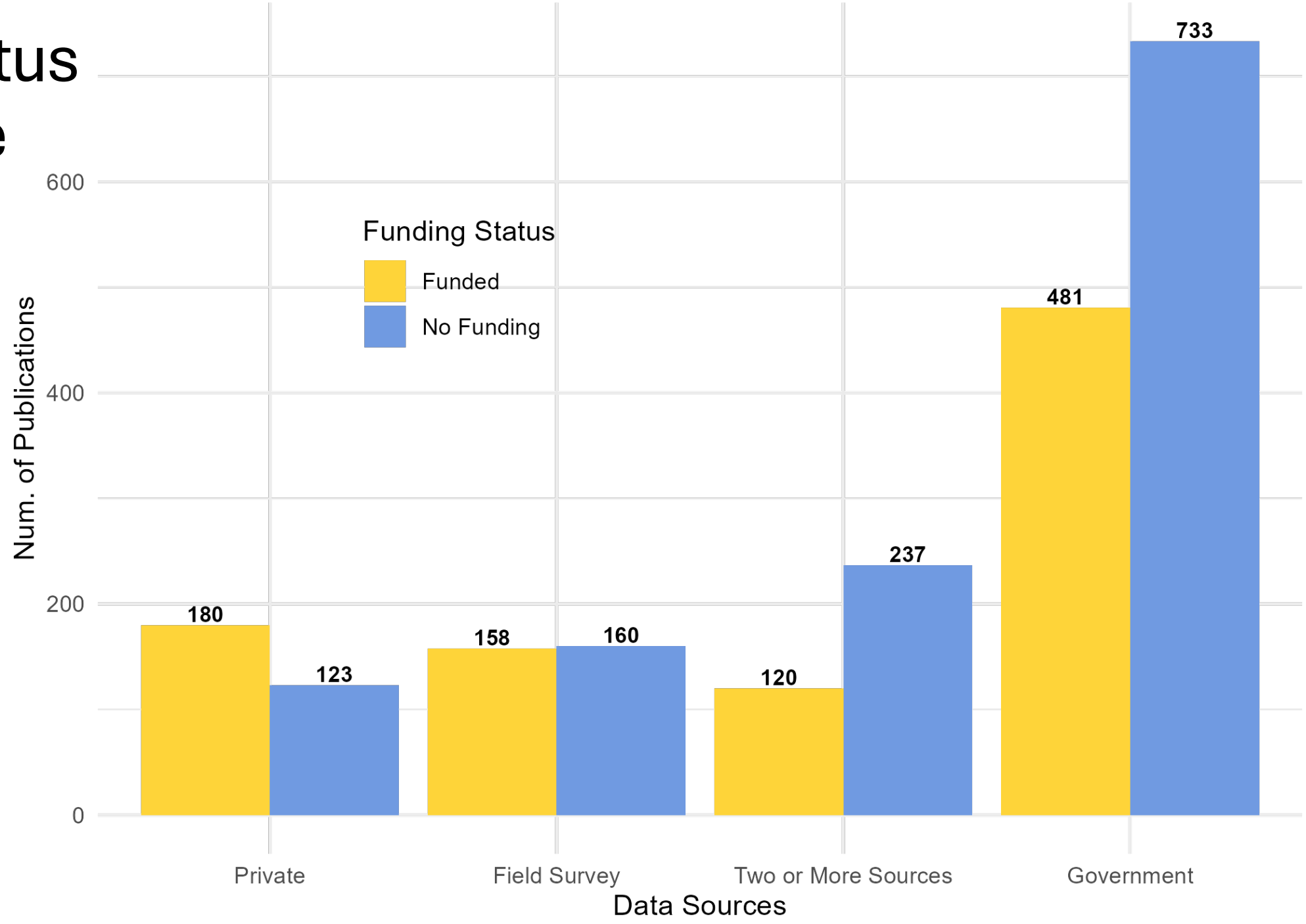
(2) Which types of GIS subfields (spatial statistics, VGI, ABM) tend to use government data?

(3) Are funding sources listed in GIS analyses? Are data made available?

# Is a funding source listed with the article?

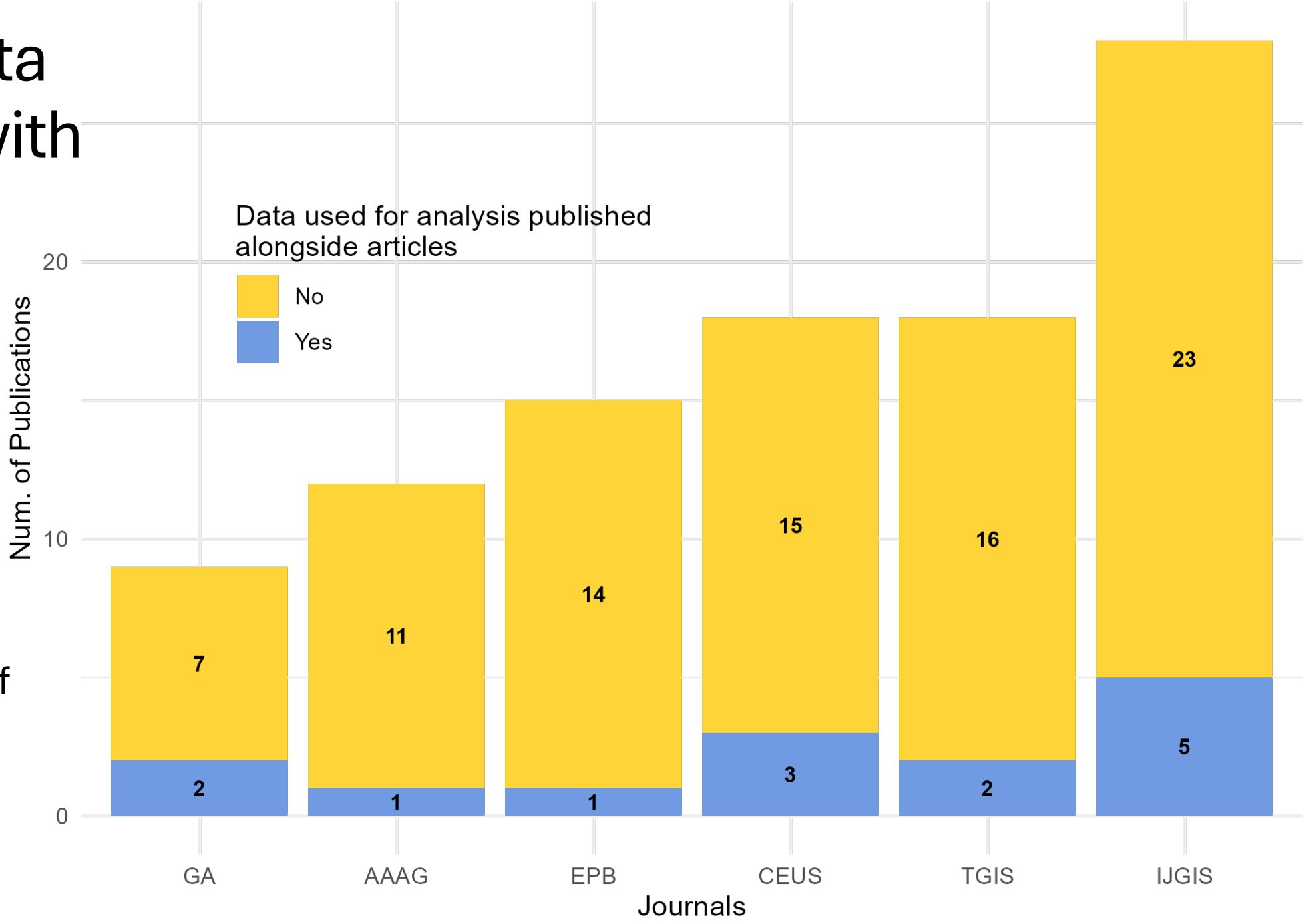


# Funding status by data type



# Is article data published with articles / available?

Random sample of 100 articles. 14% of articles using private data published their underlying data.



# Limitations

-None.

# Limitations

-Journal choices and time choices

-Human error

-Line between government and private sector is blurry

-Unclear the role of each data type in the paper

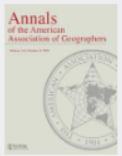
# Takeaways

1) Private sector data in peer reviewed research has increased but researchers still rely on government-collected data.

2) We should have access to a wide variety of sources (for replicating studies and helping less-funded researchers).

3) Private sector data should have metadata standards so researchers can be aware of the context of data.

4) We have a really nice data corpus and are looking for co-authors.



Annals of the American Association of Geographers ›
Latest Articles





Articles

# Private, Public, Personal: Shifting Patterns in Geospatial Data Sources in Geographic Research

Gabriel Appiah, Mira Kaufman, Billy Cooney & Clio Andris

Received 31 Jan 2023, Accepted 02 Jun 2024, Published online: 25 Sep 2024

Cite this article | https://doi.org/10.1080/24694452.2024.2394078 | Check for updates



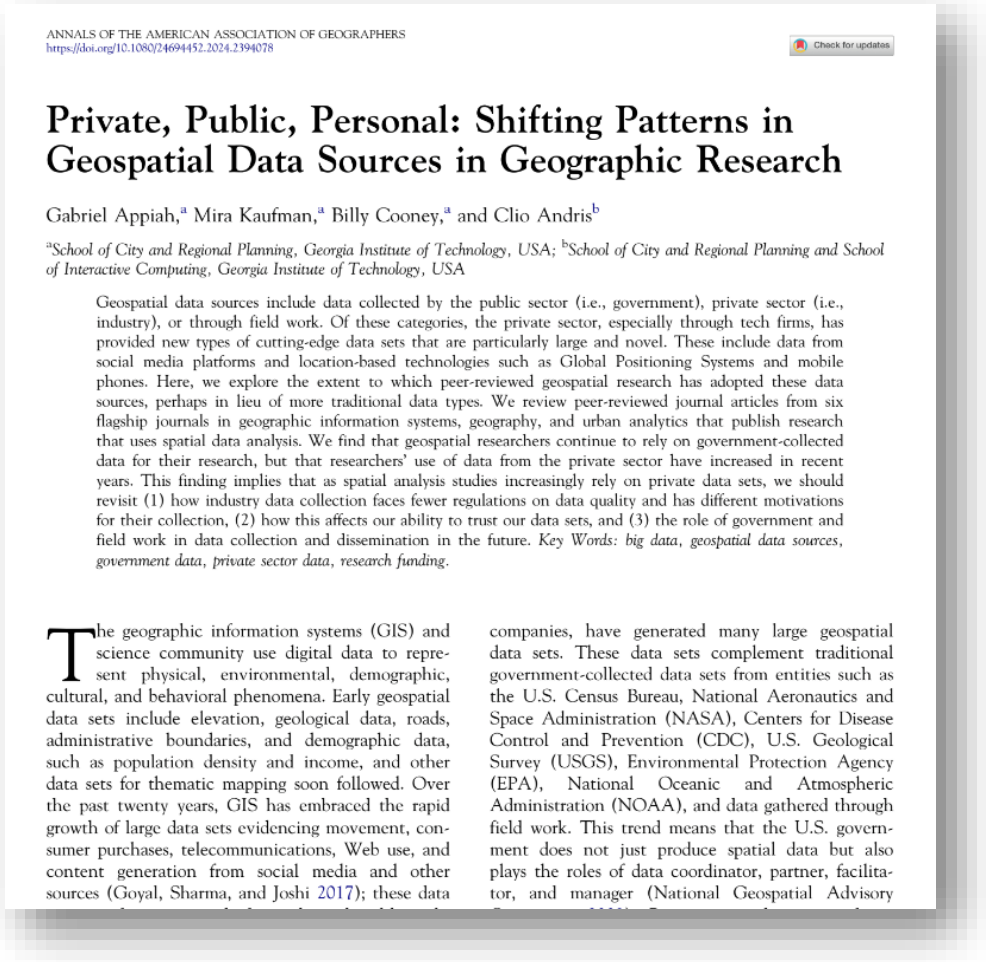
## Abstract

Geospatial data sources include data collected by the public sector (i.e., government), private sector (i.e., industry), or through field work. Of these categories, the private sector, especially through tech firms, has provided new types of cutting-edge data sets that are particularly large and novel. These include data from social media platforms and location-based technologies such as Global Positioning Systems and mobile phones. Here, we explore the extent to which peer-reviewed geospatial research has adopted these data sources, perhaps in lieu of more traditional data types. We review peer-reviewed journal articles from six flagship journals in geographic information systems, geography, and urban analytics that publish research that uses spatial data analysis. We find that geospatial researchers continue to rely on government-collected data for their research, but that researchers' use of data from the private sector have increased in recent years. This finding implies that as spatial analysis studies increasingly rely on private data sets, we should revisit (1) how industry data collection faces fewer regulations on data quality and has different motivations for their collection, (2) how this affects our ability to trust our data sets, and (3) the role of government and field work in data collection and dissemination in the future.

Key Words: big data | geospatial data sources | government data | private sector data | research funding

# Source:

Appiah, G, Kaufman, M, Cooney, B and Andris, C. Private, Public, Personal: Shifting Patterns in Geospatial Data Sources in Geographic Research. Annals of the American Association of Geographers (2024): 1-19. doi.org/10.1080/24694452.2024.2394078



ANNALS OF THE AMERICAN ASSOCIATION OF GEOGRAPHERS
https://doi.org/10.1080/24694452.2024.2394078

Check for updates

## Private, Public, Personal: Shifting Patterns in Geospatial Data Sources in Geographic Research

Gabriel Appiah,[a] Mira Kaufman,[a] Billy Cooney,[a] and Clio Andris[b]

[a]School of City and Regional Planning, Georgia Institute of Technology, USA; [b]School of City and Regional Planning and School of Interactive Computing, Georgia Institute of Technology, USA

Geospatial data sources include data collected by the public sector (i.e., government), private sector (i.e., industry), or through field work. Of these categories, the private sector, especially through tech firms, has provided new types of cutting-edge data sets that are particularly large and novel. These include data from social media platforms and location-based technologies such as Global Positioning Systems and mobile phones. Here, we explore the extent to which peer-reviewed geospatial research has adopted these data sources, perhaps in lieu of more traditional data types. We review peer-reviewed journal articles from six flagship journals in geographic information systems, geography, and urban analytics that publish research that uses spatial data analysis. We find that geospatial researchers continue to rely on government-collected data for their research, but that researchers' use of data from the private sector have increased in recent years. This finding implies that as spatial analysis studies increasingly rely on private data sets, we should revisit (1) how industry data collection faces fewer regulations on data quality and has different motivations for their collection, (2) how this affects our ability to trust our data sets, and (3) the role of government and field work in data collection and dissemination in the future. *Key Words: big data, geospatial data sources, government data, private sector data, research funding.*

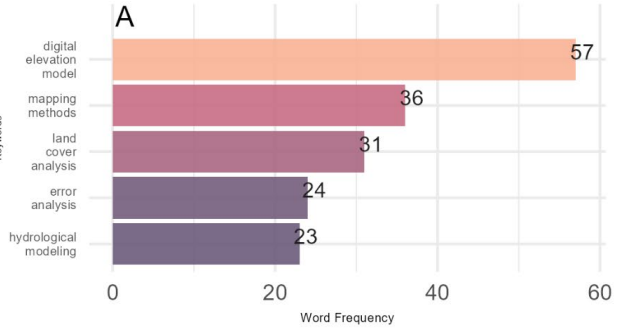
The geographic information systems (GIS) and science community use digital data to represent physical, environmental, demographic, cultural, and behavioral phenomena. Early geospatial data sets include elevation, geological data, roads, administrative boundaries, and demographic data, such as population density and income, and other data sets for thematic mapping soon followed. Over the past twenty years, GIS has embraced the rapid growth of large data sets evidencing movement, consumer purchases, telecommunications, Web use, and content generation from social media and other sources (Goyal, Sharma, and Joshi 2017); these data companies, have generated many large geospatial data sets. These data sets complement traditional government-collected data sets from entities such as the U.S. Census Bureau, National Aeronautics and Space Administration (NASA), Centers for Disease Control and Prevention (CDC), U.S. Geological Survey (USGS), Environmental Protection Agency (EPA), National Oceanic and Atmospheric Administration (NOAA), and data gathered through field work. This trend means that the U.S. government does not just produce spatial data but also plays the roles of data coordinator, partner, facilitator, and manager (National Geospatial Advisory

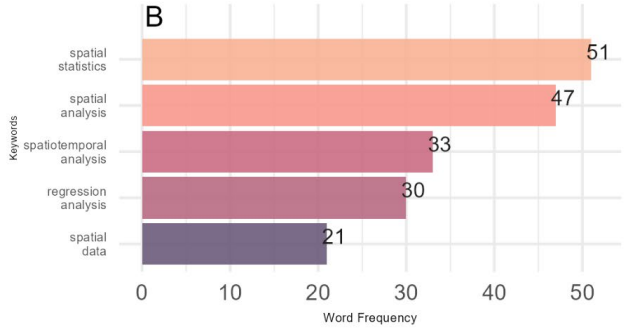
QR CODE PDF (choose Open In Browser)

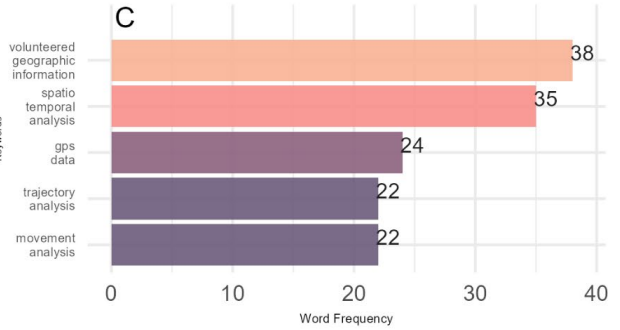
Slides available for sharing.

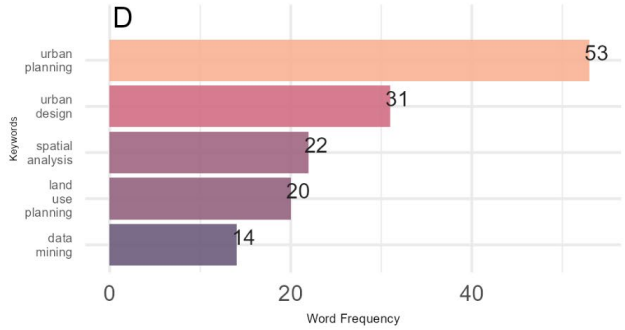
# Appendix: Creating GIS themes

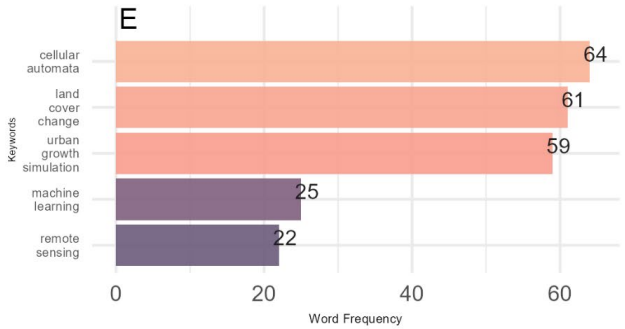
We used VOSviewer software. We generated an article network with 2,143 papers and over 69,000 edges where two papers had similar bibliographies, and we used the community detection algorithm in VOSviewer to create 9 classes.

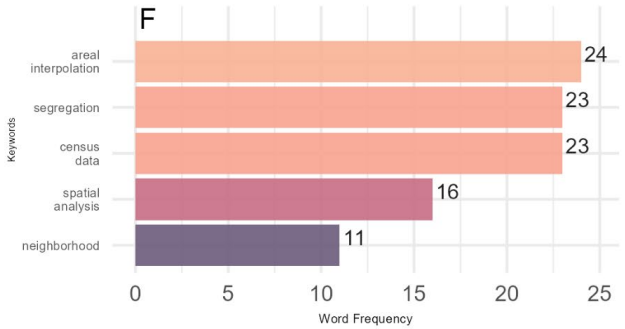


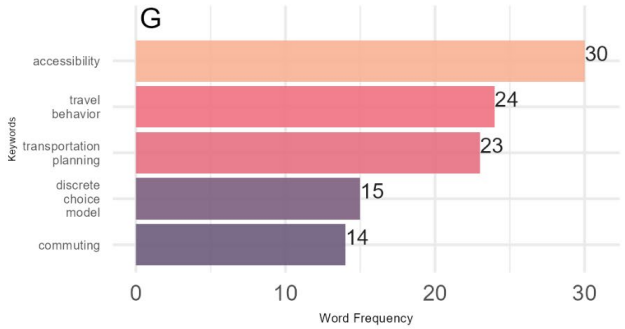



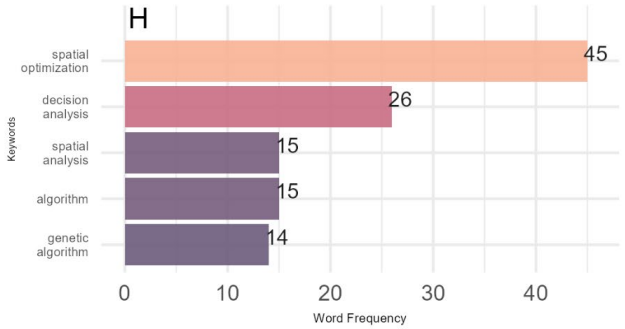



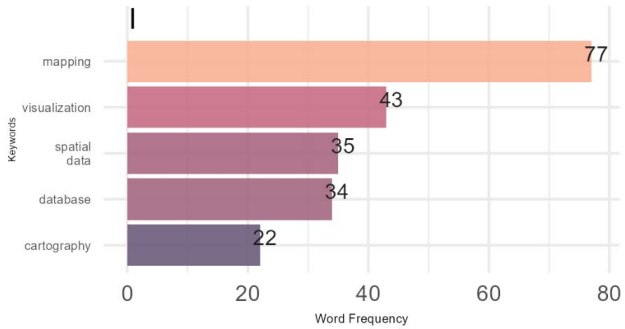



**Table B.1.** Data type, data sources, and methods used by selected articles

| Data type | Example of data | Data sources | Methods | Reference |
|---|---|---|---|---|
| Social media | Tweets, check-in data, points of interest, Flickr users | Twitter, Weibo, Flickr, Web crawling, Yelp | Shannon entropy measure, K-means cluster algorithm, space–time multiple regression model | Longley and Adnan (2016), Lwin, Sugiura, and Zettsu (2016) |
| | | | Labeled Dirichlet allocation, radius of gyration. Monte Carlo test for spatial independence | Xin and MacEachren(2020) |
| | | | Support vector model (SVM), weighted most frequently visited, kernel density estimation | Church and Li (2016), Lin and Cromley (2018), Ristea et al. (2020) |
| Transportation | Road network, public transport system data, vehicle trajectory data set, bike sharing trip data, routing information, digital tracing of taxi, GPS tracing data | ESRI, OpenStreetMap (OSM), TomTom, Movebank, Mapillary, Google | Jam flow algorithm, cold scan algorithm, network analysis and principal component analysis, radius of gyration, power law function | Kohan and Ale (2020), Derudder and Taylor (2020), Juhász and Hochmair (2016) |
| | | | Time semi-Markov process for vehicular mobility, shortest path algorithm, nonnegative matrix factorization, hierarchical clustering, frequent pattern mining | Song et al. (2016), Mayhew and Hyman (2000), Turdukulov et al. (2014) |
| | Street network, GTFS, TIGER lines, travel survey | U.S. Census Bureau, Transit Agencies, Statistics Canada, National Mapping Agency of Lower Saxony | Simulation, network routing Dijkstra algorithm, accessibility analysis | Keon et al. (2014), El-Geneidy et al. (2016) |
| Remote sensing data | LiDAR points, aerial images | Info Terra, Google Earth, Horizons Inc. | K-means algorithm, rough set theory, random forest regression analysis | Wan, Lei, and Chou. (2012), Redo, Aide, and Clark (2012) |
| | Landsat images, DEMs, LiDAR, National Land Cover data sets, building footprint | U.S. Geological Survey and Multi-Resolution Land Characteristics (MRLC) Consortium | Random forest regression, multilinear regression, model, gradient boost model, land-use classification, interpolation, hydrology modeling | Yin et al. (2020), Zhao et al. (2016), Kaučič and Žalik (2004), Brovelli, Cannata, and Longoni (2004) |
| Demography and socioeconomic | Population, income, employment status, percentage of immigrants, high-resolution global population data set, etc. | U.S. Census Bureau, Statistics Canada, Dubai Statistical Survey Department, Oak Ridge National Laboratory | Area interpolation techniques, spatial dynamic model, generic algorithm, Monte Carlo approach, simulated annealing and iterative proportional fitting | K. Li and Lam (2018), Durán-Heras, García-Gutiérrez, and Castilla-Alcalá (2018) |
| | | | Regression analysis, functional principal component analysis, structural equation model, agent-based modeling, logistic modeling | Parry et al. (2018), Ewing, Hamidi, and Grace (2016), Jepsen et al. (2006) |
| Environmental data | Natural hazard damage data, location of water bodies, slope, vegetation cover, soils, meteorological data, flood plans, bathymetric survey data | NOAA National Climate Data Center, U.S. Geological Survey, European Commissions Soil Geographical Database, Czech Meteorological Institute, German Federal Meteorological Authorities | Area interpolation techniques, genetic algorithm, spatial dynamic model, Monte Carlo approach, multicriterial least cost path analysis | K. Li and Lam (2018), Hanke, Lambert, and Smith (2014) |
| Energy | Location of gas reservoirs, pipelines, and gas wells | China National Petroleum Corporation, Louisiana Department of Natural Resources Oracle database | Backpropagation artificial neural network, geological empirical evaluation methods, qualitative analysis | Chen, Wang, and Li (2016), Hill (2002) |
| Phone location data | Mobile phone activity data, cellular phone activities | Seoul Institute and S.K. Telecom, Kokusai Denshin Denwa Inc. | Space–time multiple regression model, functional principal component analysis | Lwin, Sugiura, and Zettsu (2016), Kim (2020) |
| Housing data | Parcel data, housing data, appraisals of residential houses, property prices | OSM, Zillow Inc., UniCredit Bank Austria, private real estate company | Delaunay triangulation, Gaussian function, distance decay effect, hierarchical classification, univariate kriging variants and multivariate extensions | Bruhns et al. (2000), Kuntz and Helbich (2014) |
| | | | Hedonic pricing model, space syntax, regression-kriging, multivariate regression, fuzzy set approach | Lai et al. (2006), Morales et al. (2020), Oh and Jeong (2002) |
| Paper maps | Topographic maps, cadastral maps, land-use maps, vegetation maps, swamp maps, and glacier | U.S. Geological Survey, National Mapping Agency of Lower Saxony, Ordnance Survey | Deep convolutional neural-network-based framework, semantic similarity | Saeedimoghaddam and Stepinski (2020), Al-Bakri and Fairbairn (2012) |
| | | | Qualitative analysis, hierarchical model, Bayesian probability | Williams et al. (2006), Winter et al. (2008), Ran et al. (2012) |

*Note:* GPS = Global Positioning System; GTFS = general transit feed specification; DEMs = digital elevation models; NOAA = National Oceanic and Atmospheric Administration.