Candidate: Matilde Bagnasco
Supervisor: Prof. Francesco Nicoli

Spatial Determinants of AI Data Center Location

A County-Level Econometric Analysis in the United States

Politecnico di Torino

Academic Year 2025–2026

1. Background & Research Motivation
2. Research Question & Hypothesis
3. Data & Methodology
4. Results, Implications & Limitations

Agenda

Politecnico di Torino

01.

Background & Research Motivation

Why Study Data Center Locations?

AI Expansion Drives Data Center Demand

- The launch of ChatGPT drove rapid AI adoption, reaching **100M users in 2 months**
- Global server spending hits **$229B in 2024** due to hyperscale growth

Energy & Water Consumption Concerns

- Data centers use **1–1.5% of global electricity**
- AI workloads consume **7–8x more power** than traditional tasks
- Each ChatGPT session may require **up to 500ml water**

Infrastructure Needs Intensify

- High-performance GPUs and advanced cooling systems are essential
- Microsoft, Meta, Google, Amazon rapidly expanding AI hardware clusters

Sustainability & Location Optimization

- Site selection impacts environmental footprint and operational costs
- Understanding determinants is key for **responsible AI infrastructure**

02.

Research Question & Hypothesis

Research Question & Hypothesis

RQ What factors drive the location of data centers in U.S. counties?

- Data centers are expected to locate in counties with **lower electricity prices.**
- Counties with **stronger infrastructure and economic activity** attract more data centers.
- Data centers may prefer **less populated areas** where land and resources are cheaper.

03.

Data & Methodology

Data Construction Pipeline

Stage 1: Data Sources

Collect ZIP-level electricity tariffs, water withdrawals, population, GDP, and data center locations from online directories.

Stage 2: Geographic Harmonization

Convert all datasets to county-level using ZIP–county crosswalks and FIPS codes for consistent spatial units.

Stage 3: Data Processing

Aggregate electricity prices, construct water availability proxies, harmonize GDP, and match city to county for data centers.

Stage 4: Variable Construction

Create key variables: electricity prices, water supply, population, GDP per capita, datacenter_presence, datacenter_count.

Stage 5: Final Dataset

Merge all variables into a unified county-level dataset for econometric analysis.

Electricity Prices Dataset Construction

1

DATA SOURCE: ZIP-Level Tariffs

Electricity pricing files by the **NREL** through the OpenEI platform (IOU and non-IOU providers)

2

GEOGRAPHIC HARMONIZATION

ZIP codes are matched to counties using the HUD ZIP–County Crosswalk.

3

DATA CLASSIFICATION

Records are classified as valid IOU, invalid IOU, or non-IOU. Priority rules select the tariff for each ZIP–county observation.

4

SPATIAL IMPUTATION

Missing tariffs are imputed using county averages. When no industrial tariffs exist in a county, prices are inferred from neighbouring counties.

5

COUNTY AGGREGATION

ZIP-level tariffs are aggregated to counties using TOT_RATIO weights.

Water Supply Dataset Construction

1 **DATA SOURCE**

County-level water withdrawal data from the **USGS Water Use database**

2 **VARIABLE DEFINITION**

Public water supply withdrawals are used as a proxy for water availability and potential cooling capacity for data centers.

3 **GEOGRAPHY CONSISTENCY**

Data are already reported at county level and are merged with the main dataset using county FIPS codes.

4 **HANDLING GEOGRAPHIC CHANGES**

Some inconsistencies arise due to administrative changes in U.S. county definitions: Connecticut Planning Regions and Alaska split of Valdez–Cordova Census Area into two new areas.

Dataset Construction: Population, GDP, Data Center Presence and Number

POPULATION ESTIMATES DATASET

- County-level population data from the U.S. Census Bureau.
- Used as a proxy for county size, and potential demand for digital services.
- Merged with the dataset using county FIPS codes.

GDP VARIABLE DATASET

- County-level GDP data from the U.S. Bureau of Economic Analysis (BEA).
- Special adjustments applied where administrative boundaries changed, including Connecticut Planning Regions, Hawaii and Combination Areas.

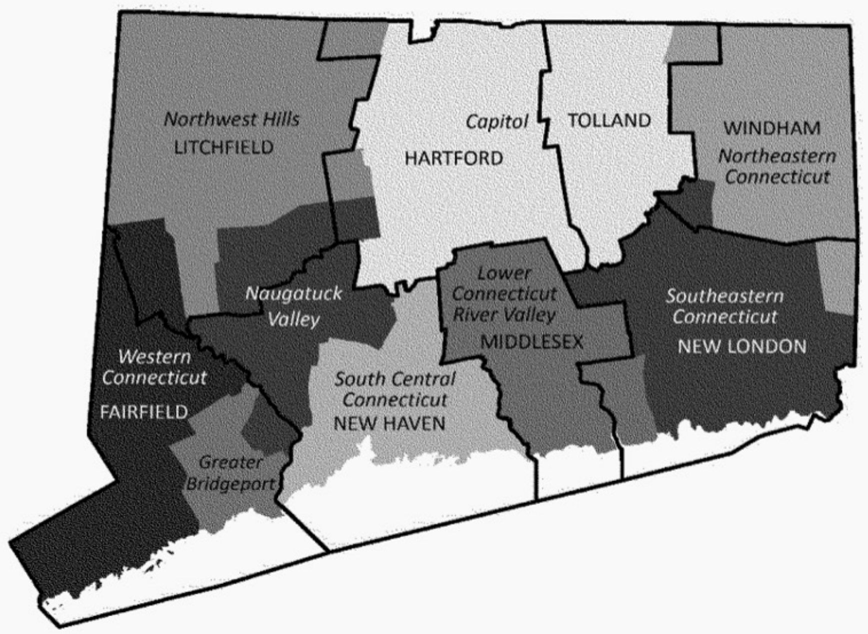
DATA CENTER PRESENCE AND NUMBER DATASET

- Data center locations collected from datacentermap.com.
- City-level locations matched to counties using the SimpleMaps U.S. Cities database.
- Construction of two key variables: datacenter_presence datacenter_number

SPECIAL GEOGRAPHIC CASES

Connecticut: Counties to Planning Regions

- In 2022–2023, Connecticut replaced counties with Planning Regions as county-equivalent units.
- Water withdrawals manually reassigned to Planning Regions using official maps and geographic correspondence tables.



Alaska: Valdez–Cordova Census Area Split

- In 2019, Valdez–Cordova split into Chugach and Copper River Census Areas.

Hawaii: Maui and Kalawao Counties and Other Combined Geographic Areas

- In the BEA dataset, some GDP values are reported for aggregated geographic units (e.g., Maui and Kalawao Counties in Hawaii).

Water withdrawals and GDP redistributed using population-weighted allocation to preserve total values.

$$Value_i = Value_{total} \times Population_i \,/\, \Sigma\, Population_j$$

Value = water withdrawals or GDP
i = counties within the aggregated geographic unit

04.

Results, Implications & Limitations

Empirical Strategy and Model Specification

Research Objective

- Analyze the **determinants of data center location** across U.S. counties.
- Identify how economic, demographic, and infrastructure characteristics influence the probability of hosting a data center.

Extensive Margin Model

- Dependent variable: **datacenter_presence (binary).**
- Tests **probability of hosting at least one data center.**
- Key explanatory variables: population, GDP per capita, water supply, electricity prices (commercial and industrial rates)

Intensive Margin Analysis

- Additional **analysis of data center clustering** within counties.
- Outcome variable: **datacenter_number**.
- Used to explore the distribution and concentration of infrastructure.

Robustness & Validation

- **Robustness checks** excluding counties with high electricity price imputation, neighbor-imputed industrial tariffs.
- **Interaction effects** between electricity prices, population, and GDP per capita and **correlation**.

Extensive Margin Results: Probability of Hosting a Data Center

$$Pr(DataCenter\_c = 1) = \Lambda(\alpha + \beta1 \ln(Population\_c) + \beta2 \ln(GDPpc\_c) + \beta3 \ln(Water\_c) + \beta4\ Electricity\_c + X\_c'\gamma)$$

Variable	Logit (Commercial)	Logit (Industrial)	AME (Commercial)	AME (Industrial)
Population	1.066***	1.055***	0.063***	0.062***
GDP per capita	1.157***	1.149***	0.068***	0.068***
Electricity price	-2.288	-1.037	-0.134	-0.061
Water supply	0.224***	0.224***	0.013***	0.013***

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$, Observations: 3,144, Pseudo R^2= 0.40

Key Takeaways

Data center presence primarily driven by **population size, economic development, and infrastructure availability.** Electricity prices do not significantly affect initial location decisions.

Interaction Effects & Robustness Checks

Interaction Effects - Key Takeaways

- **No statistically significant interaction effects**
- Structural factors remain the main drivers of data center location

Interaction tested	Coefficient	p-value
Population x Comm Electricity	-0.840	0.484
Population x Ind Electricity	-0.404	0.723
GDP per capita x Comm Electricity	2.168	0.355
GDP per capita x Ind Electricity	0.388	0.932
Population x Water Supply	-0.006	0.895
Water supply x Comm Electricity	-0.979	0.387
Water supply x Ind Electricity	-0.764	0.439

Robustness tests performed

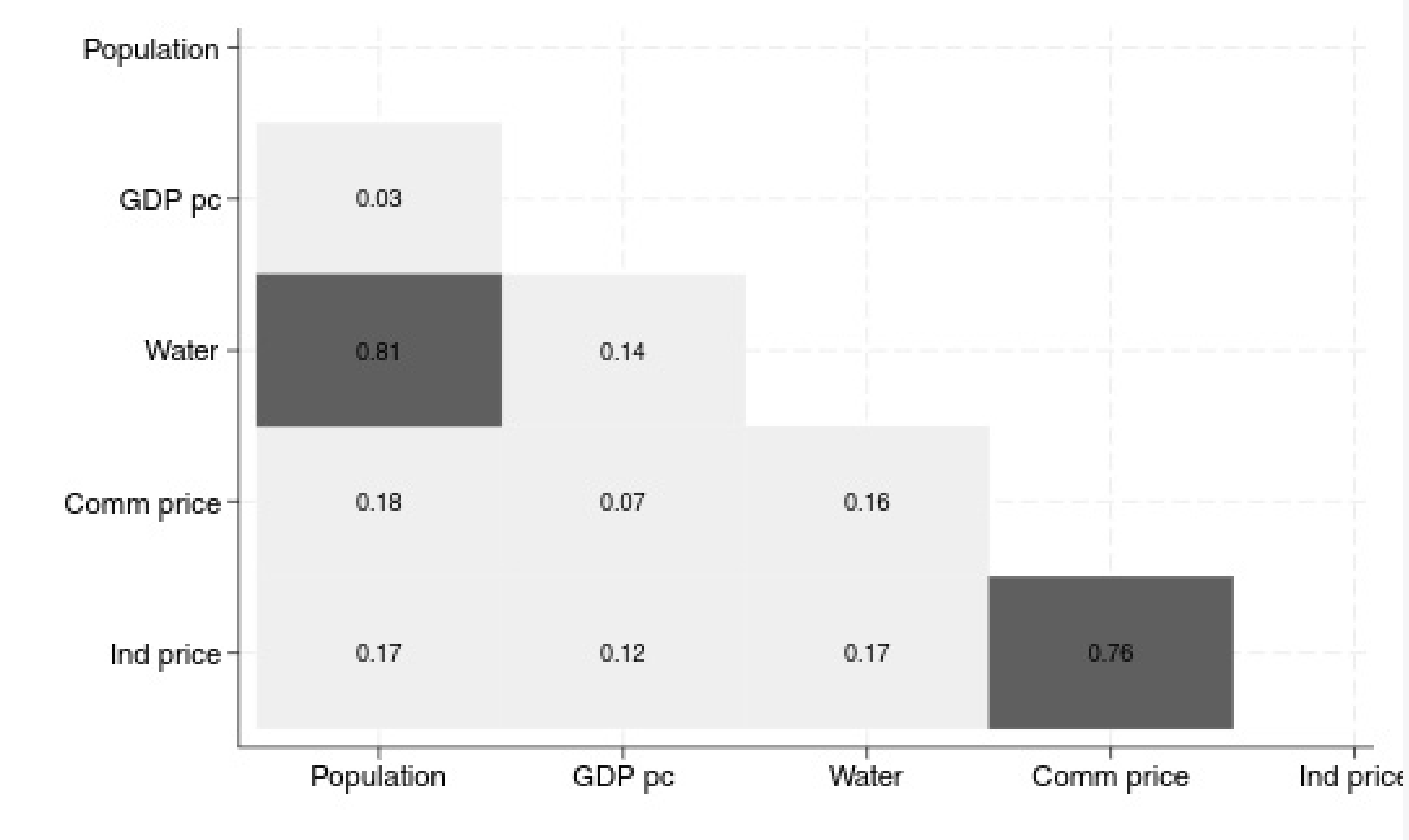
Excluding counties with:
- imputed commercial electricity prices
- fully approximated electricity tariffs
- neighbour-imputed industrial tariffs
- limited electricity price coverage (high share of imputed)
- with extreme electricity price values (> $0.40/kWh)

Robustness Checks - Key Takeaways

- **Results remain consistent across all robustness checks**
- Population, GDP per capita, and water availability remain significant determinants of data center location, while electricity prices remain insignificant
- Correlation analysis and VIF tests show **no multicollinearity concerns**

Correlation Structure and Multicollinearity diagnostics

Correlation Heatmap



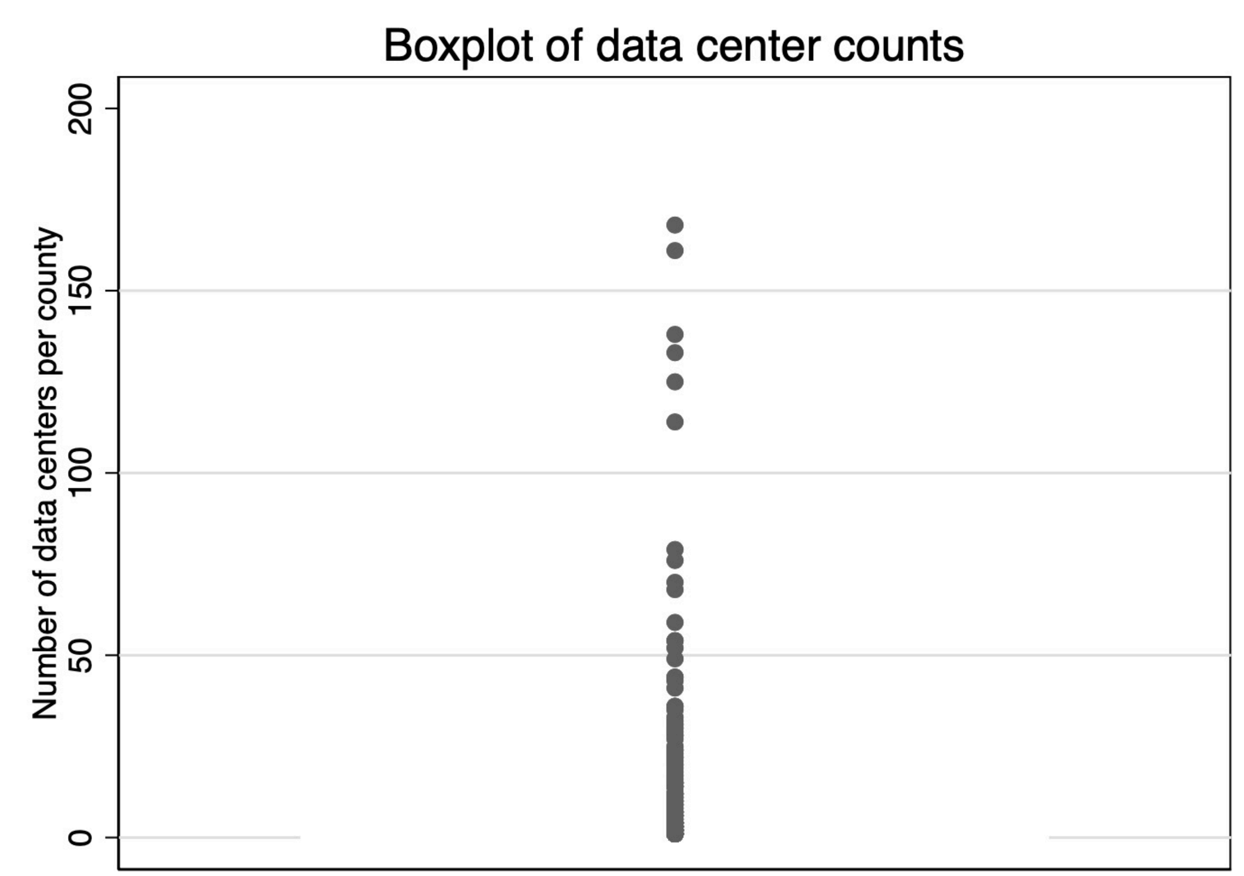
VIF Values

Variable	Commercial price model	Industrial price model
Population	3.00	2.99
GDP per capita	1.05	1.06
Electricity price	1.04	1.05
Water supply	3.03	3.03

Intensive Margin Results: Determinants of Data Center Counts

Distribution of Data Center

- highly skewed
- large share of zero observations



Negative Binomial Regression

- Larger and more economically developed counties host more data centers
- **Higher electricity prices reduce data center concentration**
- Water availability does not significantly affect clustering

Variable	Coefficient(Commercial)	Coefficient (Industrial)	IRR(Commercial)	IRR(Industrial)
Population	1.599***	1.561***	4.95	4.76
GDP per capita	1.911***	1.945***	6.76	6.99
Electricity price	-9.935***	-5.038***	0.00	0.01
Water supply	-0.488	-0.519	0.614	0.595

* p < 0.10, ** p < 0.05, *** p < 0.01, Observations: 3,144, variables expressed in logs

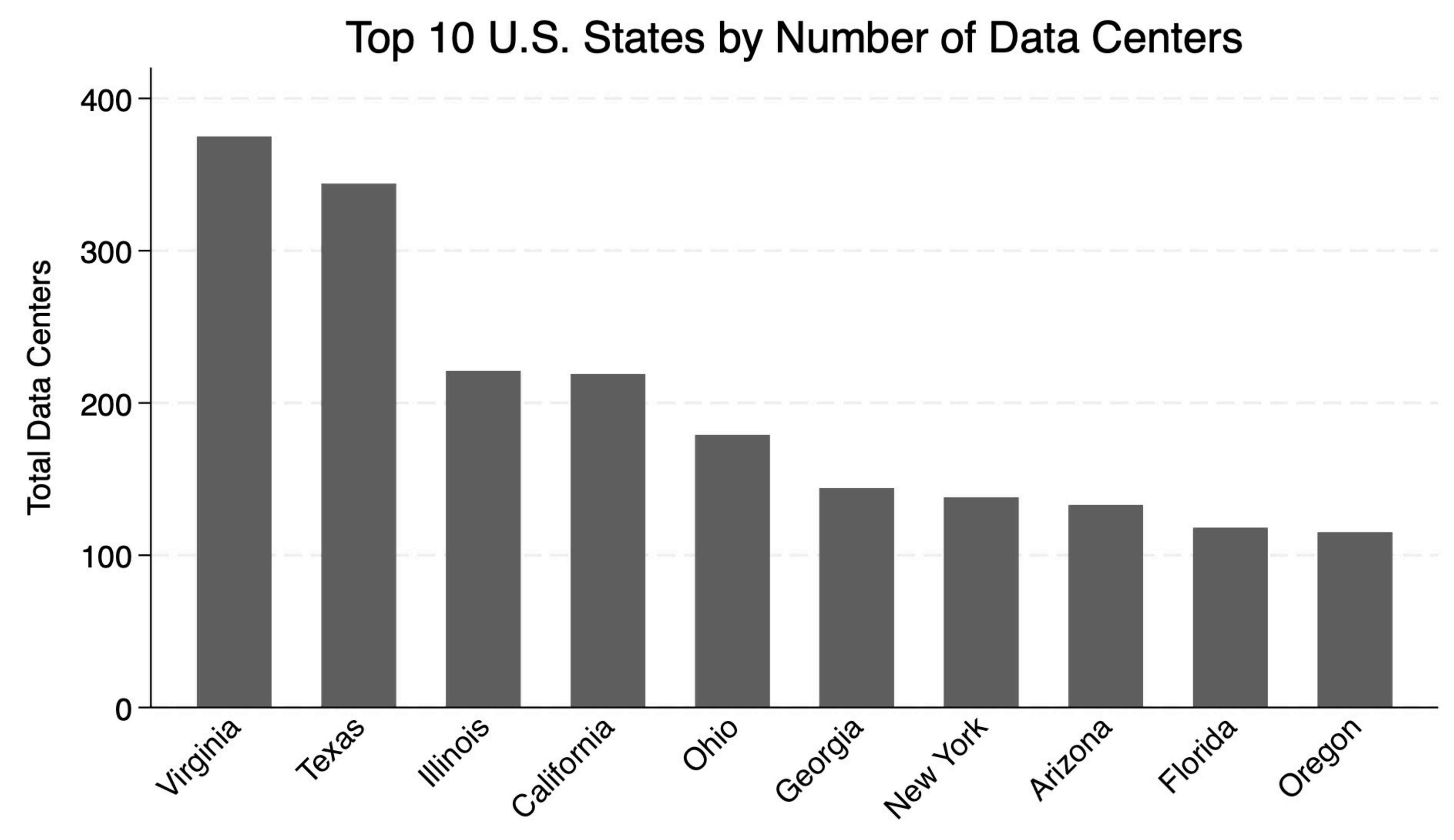
Geographic Concentration of Data Centers in the U.S.

Highly Concentrated Distribution

- Only **10.7%** of U.S. counties host at least one data center
- Majority of counties have zero data centers
- Data centers **cluster in a small number of states** and metropolitan areas

Key Insights

- Data center infrastructure is **highly geographically concentrated,** with a small number of states hosting the majority of facilities.
- Large digital infrastructure clusters emerge in economically dynamic regions, such as **Virginia, Texas, and California**.



Conclusions - What drives data center location across U.S. Counties?

Main Findings

- **Structural factors** dominate location decisions
- **Electricity prices affect clustering**, but not the initial location choice

Policy Implications

- Data center policy should consider **regional infrastructure and economic capacity**, not only energy costs

Limitations

- Limited harmonized county-level data
- **Complex dataset integration** across multiple sources
- **No distinction** between general and AI data centers

Future Research

- Include **sustainability indicators** (water stress, renewable energy, climate risk)
- **Comparative analysis** across regions (e.g., Europe vs U.S.)
- **Models for optimal placement** of digital infrastructure

Thank you!