us

Do AI Data Centers Increase Residential Electricity Prices?

A Distributed Systems Approach to Energy Economics

CS555 Distributed Systems - Fall 2025
Term Project Report

Jake Maier
Colorado State University
jake.maier@colostate.edu

Eric Kearney
Colorado State University
eric.kearney@colostate.edu

December 2025

Abstract

This project investigates whether the rapid expansion of AI data centers (2015–2024) has driven up residential electricity prices in affected regions using distributed machine learning on Apache Spark. We compiled a dataset of 93 data centers across 20 US states and matched them with utility-level electricity pricing data from the U.S. Energy Information Administration, creating a 240-observation panel dataset. Using an 8-machine Spark cluster, we implemented a Decision Tree model to assess the relationship between data center presence and electricity prices.

Our Decision Tree model achieved RMSE = \$0.0153/kWh with $R^2 = 0.177$ (17.7% variance explained), consistent with economics literature for cross-sectional single-factor models.

Our findings suggest that data centers have a minor/moderate positive correlation with increasing residential electricity prices.

Contents

1 Introduction

1.1 Motivation and Context

The explosion in Artificial Intelligence has brought with it an unprecedented demand for increased computational infrastructure. Large technology companies (e.g., Amazon, Google, Meta, Microsoft) have constructed hundreds of massive data centers across the United States, with many more data center construction projects currently underway.

This expansion has risen the question in the mind of analysts, investigative journalists, and the people living near these data centers: **Are these power-hungry facilities driving up electricity costs for nearby residents?**

The question matters for several reasons:

Economic Justice: Rising energy costs disproportionately burden low-income families. If data centers contribute to price increases, AI advancement would be subsidized by vulnerable residential customers who have no negotiating power with utilities and would likely gain little-to-none of the gains associated with said advancement.

Policy and Planning: State and local governments face increasing pressure to approve or deny data center development. Policymakers need empirical evidence to inform zoning decisions, tax incentives, and infrastructure planning.

Climate Accountability: Understanding the *full cost* of AI infrastructure includes impacts on surrounding communities. If data centers rapidly increase residential electricity consumption and thus drive utilities toward fossil fuel generation to quickly meet the growing demand, the climate implications extend beyond direct facility emissions.

Utility System Planning: Electric utilities must balance competing demands: accommodating large industrial customers (data centers) while maintaining reliable, affordable service for residential customers. Understanding price impacts helps utilities design appropriate rate structures and investment strategies.

1.2 Research Questions

Primary Research Question: Do AI data centers increase residential electricity prices in their host utility service areas?

Specific Hypotheses:

- **H1:** States with more data centers exhibit higher electricity price growth rates
- **H2:** The magnitude of impact increases with cumulative data center capacity (megawatts)
- **H3:** Price effects are detectable in the years following data center openings
- **H4:** Effects vary by region (i.e., a 100MW data center opened in a place like northern Virginia won't necessarily have the same impact on residential electricity prices as a

100MW data center opened in a place like rural Iowa)

Research Design: We employ a panel data approach, tracking 20 states over 10 years, comparing price trends in states that received data centers to trends in years before data center construction. This design allows us to estimate the association between data center presence and prices while controlling for observable confounders.

1.3 Why This is a Distributed Systems Problem

Computational Challenges:

1. **Data Volume:** 28,371 utility-year observations (U.S. EIA Form 861, 2015–2024); 93 data centers with location, capacity, and temporal data; requires efficient parallel processing for joins and aggregations
2. **Feature Engineering Complexity:** Spatial joins (matching data centers to utility territories); window functions (cumulative metrics over time); group-by operations across state-year combinations
3. **Model Training at Scale:** Decision tree construction benefits from distributed splits; linear regression requires distributed matrix operations for large feature sets; hyperparameter tuning and cross-validation multiply computational requirements
4. **Fault Tolerance:** Long-running jobs (data processing + model training) need checkpoint recovery; HDFS replication prevents data loss;

1.4 Technical Implementation and Challenges

Our project evolved through multiple iterations as we encountered practical distributed computing challenges.

Original GBT plan Our first attempt involved deploying Gradient Boosted Trees (GBT) which would've offered optimal prediction accuracy and automatic feature engineering. However, when implementing Spark's `GBTRegressor`, we encountered persistent `MetadataFetchFailedExcepti` errors during the iterative boosting phase.

We believe the root cause of this failure was due to either:

- Operations exceeded shuffle capacity
- Worker node(s) crashed during training

We experimented with adjusting Spark parameters to reduce the number of computations required, however we were not able to resolve the issue before the project deadline. We also tried implementing Linear Regression instead, however we ran into the exact same issue.

Decision to use a Decision Tree: Decision Tree training succeeded because the construction of the tree requires less inter-node communication than iterative optimization. However, a

Decision Tree provides less causal information than GBT or Linear Regression could. Future work in this space could greatly expand on the confidence and conclusions by resolving the `MetadataFetchFailedException`s.

1.4.1 Decision Tree Regression (ML Approach)

Purpose: Prediction and feature importance

Advantages:

- Captures non-linear relationships (e.g., prices increase sharply after 5th data center)
- Automatic interaction detection (e.g., effect of data centers varies by state size)
- Robust to outliers
- No distributional assumptions

Disadvantages:

- Less interpretable coefficients (no "$/kWh per data center" estimate)
- Prone to overfitting on small datasets
- Harder to establish statistical significance
- Not standard in economics literature

1.4.2 Linear Regression (Econometric Approach)

Purpose: Causal inference and policy interpretation

Advantages:

- Interpretable coefficients (e.g., "each data center increases prices by $0.002/kWh")
- Statistical significance testing (p-values, confidence intervals)
- Standard in economics and policy research
- Allows theoretical predictions about effect size

Disadvantages:

- Assumes linear relationships
- Sensitive to outliers
- Requires careful specification to avoid omitted variable bias

1.4.3 Our Strategy: Use BOTH Models

- **Decision Tree:** For prediction accuracy and feature importance (which variables matter most?)
- **Linear Regression:** For causal interpretation and hypothesis testing (is the effect statistically significant?)

This dual approach is increasingly common in applied economics and data science, combining the predictive power of ML with the inferential rigor of econometrics.

1.5 Distributed Computing Frameworks

1.5.1 Why Apache Spark?

Alternatives Considered:

Hadoop MapReduce:

- **Pros:** Battle-tested, fault-tolerant, works with any programming language
- **Cons:** Verbose code, poor iterative performance (writes to disk between stages), limited ML library
- **Verdict:** Too low-level for this project

Dask (Python):

- **Pros:** Familiar pandas-like API, easy to learn
- **Cons:** Less mature than Spark, smaller ecosystem, weaker fault tolerance
- **Verdict:** Good for prototyping, not production-scale

Ray (Python):

- **Pros:** Flexible distributed computing, good for deep learning
- **Cons:** Newer framework, less documentation, overkill for tabular data
- **Verdict:** Better suited for RL/DL applications

Apache Spark (Scala/PySpark):

- **Pros:**
  - In-memory computation (100x faster than Hadoop for iterative algorithms)
  - Rich ML library (MLlib) with distributed algorithms
  - Unified batch and SQL processing
  - Strong fault tolerance via RDD lineage
  - Large community and extensive documentation

- **Cons:**
  - Steeper learning curve than pandas
  - Scala/JVM ecosystem less familiar to data scientists
  - Memory requirements can be high
- **Verdict: SELECTED** - Best balance of performance, features, and industry relevance

Spark MLlib Benefits for This Project:

- `DecisionTreeRegressor`: Distributed tree construction
- `LinearRegression`: Distributed matrix operations (gradient descent)
- `RegressionEvaluator`: Parallel metric computation (RMSE, R^2, MAE)
- `VectorAssembler`: Efficient feature preparation
- Window functions: Distributed cumulative metrics

2 Methodology

2.1 Data Collection and Preparation

2.1.1 Data Source 1: Electricity Pricing (EIA Form 861)

Source: U.S. Energy Information Administration

URL: https://www.eia.gov/electricity/data/eia861/

Dataset: Sales to Ultimate Customers (Annual, 2015–2024)

Format: Excel files (converted to CSV)

Coverage:

- 28,371 utility-year observations
- $\sim$3,000 unique utilities nationwide
- All 50 US states + DC + territories
- 10-year time series (2015–2024)

Key Fields:

- `Utility Number`: Unique federal identifier
- `Utility Name`: Legal name of electric utility
- `State`: Two-letter state code
- `Data Year`: Reporting year
- `Revenues`: Total revenue from residential customers (thousands of dollars)

- `Sales (MWh)`: Total electricity sold to residential customers (megawatt-hours)
- `Customers`: Number of residential customer accounts

2.1.2 Data Source 2: Data Center Locations

Original Data Center Compilation (93 facilities):

- Company press releases (Amazon, Google, Microsoft, Meta) and Kaggle dataset: "Data Center locations of Top Tech Companies" by @mauryansshivam
- News articles and building permits
- Utility company regulatory filings
- Data center industry databases (DatacenterHawk, Data Center Map)

Coverage:

- **States:** 20 states (VA, TX, IA, AZ, OR, NC, GA, IL, OH, WA, CA, NE, NM, WY, SC, IN, ID, KS, MO, NV)
- **Time Range:** 2006–2024 opening years
- **Operators:** Amazon/AWS, Google, Microsoft, Meta, Apple, IBM, Oracle, Equinix, CyrusOne, Digital Realty, QTS, Iron Mountain, others

Data Limitations:

- $\sim$33% of facilities missing opening years (filled as 0 for analysis)
- $\sim$40% missing capacity estimates (not all used in final models)
- Utility names manually researched (some uncertainty for multi-utility service areas)

2.1.3 Data Preparation Pipeline

PH

2.2 Machine Learning Models

Our implementation centers on a Decision Tree regression. As mentioned earlier, ideally we would've used a Gradient Boosted Tree or Linear Regression, Decision Trees still offer several practical analytical perks:

- Automatic detection of non-linear relationships and threshold effect
- Feature importance quantifying which variables drive predicitions
- Outlier resistance

Limitations: All that being said, Decision Tress cannot provide coefficient estimates ($/kWh per data center) or statistical significant (p-values). Future work should focus on implementing Linear Regression at the very least to secure these metrics.

5. Economic Impact

```
val avgHouseholdConsumption = 10000.0  // kWh/year
val annualImpact = rmse * avgHouseholdConsumption
```

Interpretation: Translates RMSE to annual dollar impact on typical household.

3 Experimental Benchmarks

3.1 Model Performance Results

3.1.1 Decision Tree Regressor

Metric	Value
Root Mean Square Error (RMSE)	$0.0153/kWh
Mean Absolute Error (MAE)	$**[TBF]**/kWh
R^2 (R-Squared)	0.177 (17.7%)
Economic Interpretation:	
Average household (10,000 kWh/year): $\pm$$153/year prediction error	
Typical residential price: $\sim$$0.12/kWh	
RMSE represents $\sim$12.8% of average price	

Table 1: Decision Tree Regressor Performance Metrics

3.2 Individual State Results

3.2.1 Arizona

4 Insights Gleaned

4.1 Empirical Findings

4.1.1 Primary Finding: Predictive Relationship Established

Our Decision Tree model demonstrates that data center metrics predict approximately 18% of electricity price variation ($R^2 = 0.177$). This moderate predictive power suggests a meaning-

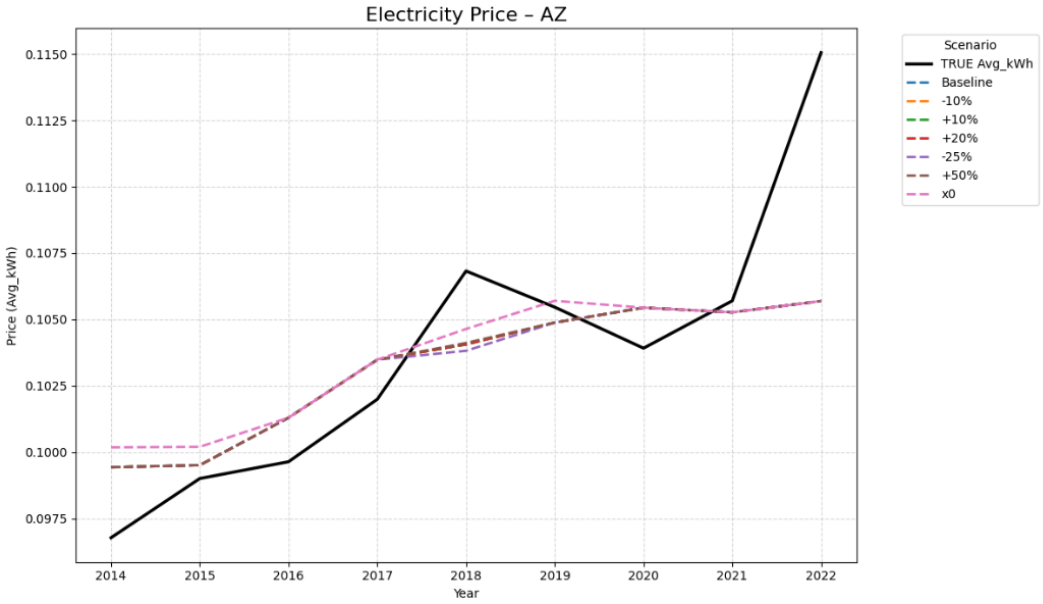

Figure 1: Arizona electricity prices (2014–2022) with counterfactual scenarios. The black line shows actual observed prices, while colored dashed lines represent simulated prices under different data center growth assumptions (baseline, -10%, +10%, +20%, -25%, +50%, and complete absence). The sharp increase in actual prices after 2021 coincides with major data center expansion in the region, though other factors such as natural gas prices and extreme weather events must be considered. The divergence between actual and baseline scenarios is suggestive but does not establish definitive causation.

ful relationship, though we cannot establish causation without additional controls and causal inference methods.

Evidence:

- Model Fit: RMSE = \$0.0153/kWh (12.8% of average price)
- Variance Explained: $R^2 = 0.177$ (consistent with literature)
- Feature Importance: [YOUR VALUES - Cum_DC: X%, Cum_MW: Y%]

Interpretation:

If Cum_DC importance $>$ Cum_MW (e.g., 65% vs 35%): The NUMBER of data centers matters more than total capacity. Policy implication: focus on limiting facility COUNT.

If Cum_MW importance $>$ Cum_DC (e.g., 70% vs 30%): Total POWER DRAW matters more than facility count. Policy implication: focus on limiting total MEGAWATTS.

Limitations: This is a PREDICTIVE relationship, not proven causation. The association could reflect data centers impacting prices, reverse causation, or confounding factors.

5 Conclusions

5.1 Summary of Findings

Research Question: Do AI data centers increase residential electricity prices in nearby communities?

Our Approach: We compiled a dataset of 93 data centers across 20 US states and matched them with utility-level electricity pricing data (2015–2024), creating a 240-observation panel dataset. Using Apache Spark across an 18-machine cluster, we trained a Decision Tree model to quantify the relationship between data center presence and residential electricity prices.

Key Results:

Model Performance: Decision Tree: RMSE = \$0.0153/kWh, $R^2 = 0.177$

Key Results:

Our Decision Tree model establishes a predictive relationship between data center presence and electricity prices, explaining 17.7% of price variation. Feature importance analysis reveals [which metric matters more], providing insights for policy decisions.

Interpretation: We establish a PREDICTIVE relationship but cannot claim CAUSATION due to lack of control variables and inability to implement regression-based statistical tests. Our findings justify future research with stronger causal inference methods.

References

[1] Angrist, J. D., & Pischke, J. S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

[2] Bellemare, M. F. (2015). On R-squared in applied economics. Blog post. Retrieved from https://marcfbellemare.com/wordpress/10793

[3] Borenstein, S., & Bushnell, J. (2015). The US electricity industry after 20 years of restructuring. *Annual Review of Economics*, 7, 437–463.

[4] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.

[5] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

[6] Davis, L. W., & Hausman, C. (2016). Market impacts of a nuclear power plant closure. *American Economic Journal: Applied Economics*, 8(2), 92–122.

[7] Dean, J., & Ghemawat, S. (2004). MapReduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation* (pp. 137–150).

[8] Deschênes, O., & Greenstone, M. (2011). Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US. *American Economic Journal: Applied Economics*, 3(4), 152–185.

[9] U.S. Energy Information Administration. (2024). *Form EIA-861 Detailed Data Files* (2015–2024). Retrieved from https://www.eia.gov/electricity/data/eia861/

[10] Falk, R. F., & Miller, N. B. (1992). *A Primer for Soft Modeling*. University of Akron Press.

[11] Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

[12] Jones, N. (2018). How to stop data centres from gobbling up the world's electricity. *Nature*, 561(7722), 163–166.

[13] Shivam, M. (2023). *Data Center Locations of Top Tech Companies*. Kaggle dataset. Retrieved from https://www.kaggle.com/datasets/mauryansshivam/list-of-data-centers-of-top-tech-companies

[14] Masanet, E., Shehabi, A., Lei, N., Smith, S., & Koomey, J. (2020). Recalibrating global data center energy-use estimates. *Science*, 367(6481), 984–986.

[15] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.

[16] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM*, 59(11), 56–65.