HYPERSCALE DATA CENTERS: LOCATION DETERMINANTS AND ENVIRONMENTAL CONSIDERATIONS

by

SAM SCHROEDER

A THESIS

Presented to the Department of Economics
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

March 2025

An Abstract of the Thesis of

Sam Schroeder for the degree of Bachelor of Science
in the Department of Economics to be taken March 2025

Title: Hyperscale Data Centers: Location Determinants and Environmental Considerations

Approved: *Edward Rubin, Ph.D.*
Primary Thesis Advisor

Hyperscale data centers generate negative externalities in the form of environmental costs, which vary by location. I model counts of hyperscale data centers by county using a logistic regression to capture the extensive margin and a log-linear regression to describe the intensive margin. This analysis both formalizes assumptions about DC siting parameters and evaluates firms' environmental considerations. I find that the locations of hyperscale DCs are primarily influenced by sales tax rates, colocation benefits, population size, fiber optic connectivity, and renewable energy potential. These results suggest that firms are attracted to sites that minimize private costs. However, the log-linear model produces a significant estimate for renewable energy potential, showing that firms may consider reduction of their carbon footprint through siting. This analysis describes the siting process of hyperscale DCs to enhance siting efficiency by reduction of negative externalities.

Acknowledgements

I want to thank my primary thesis advisor, Professor Ed Rubin, for his guidance throughout the thesis process. I am grateful for Professor Rubin's enthusiasm and willingness to assist with a somewhat experimental and novel research topic. I would also like to thank my CHC representative, Dr. Angela Rovak, for offering four years of support and providing invaluable feedback on my thesis draft. It has been a privilege to share my thesis process with both Professor Rubin and Dr. Rovak.

Table of Contents

Introduction 6
Hyperscale data center locations 7
Background 9
Existing literature 9
Water footprint 9
Carbon footprint 10
Contributions 11
Siting parameters 11
Methods 13
Model data 13
Model input data 14
Constructing the model: Negative binomial 16
Constructing the model: Extensive and Intensive Margins 18
Extensive margin 18
Intensive margin 19
Results 21
Logistic regression 21
Log-linear regression 22
Discussion 24
Environmental implications 24
Limitations and future research 25
Conclusion 26
Appendix 28
References 33

List of Figures and Tables

Figure 1. Map of hyperscale DC counts by county, CONUS 14
Figure 2. Polynomial effect of neighbor counts on DCs 16
Figure 3. Hyperscale DC counts by county histogram, CONUS 17
Figure 4. Negative binomial residuals: Scatter plot and histogram 18
Figure 5. Logistic model residuals: Scatter plot and histogram 19
Figure 6. Log-linear model residuals: Scatter plot and histogram 20
Figure 7. Bivariate map of DC counts and energy prices, WA and OR 28
Figure 8. Bivariate map of DC counts and fiber optic connectivity, WA and OR 29
Figure 9. Bivariate map of DC counts and renewable energy potential, TX 30
Table 1. Model results 31
Table 2. Exponentiated estimates if logit model 32

Introduction

"Our inventions are wont to be pretty toys, which distract our attention from serious things. They are but improved means to an unimproved end."

— *Walden*, Henry David Thoreau (*Economy*)

Mount Hood, the tallest point in Oregon, towers over the Columbia River gorge at 11,245 feet. Around its peak, glaciers persist year-round. Although removed from the populated river valley below, this snowpack provides an essential service to the people and economies of the region (Bakken-French et al. 2024). Below the snowy peak, beginning at 6,000 feet, the mountain is flanked by dense forests. Dry slopes of ponderosa pine shade the east side of the mountain. Thin streams of snow melt snake downhill and converge to form larger rivers such as the Dog River and the South Fork Mill Creek (Shannon 2018). Rivers in this region provide critical habitat for a variety of species including salmon and steelhead while also offering valuable ecosystem services.

Eventually, at 2,600 feet, the South Fork Mill Creek reaches the Crow Creek Dam, filling the Crow Creek Reservoir. Sixteen miles Northeast of the reservoir, the city of The Dalles sits on the bank of the Columbia River. Crow Creek Reservoir serves as the city's primary water source, providing for its 15,000 residents ("The Dalles" 2023). Considering melting glaciers, decreased snowpack, and extreme drought, residents of The Dalles are concerned about the future of their water supply (Selsky and Valdes 2025). One resident, however, plays a larger role than most. In 2021, Google accounted for 29% of The Dalles water consumption, or 355 million gallons (Rogoway 2022b). This water was used to cool the thousands of servers stored in Google's hyperscale data center (DC) facilities. In addition to water use, DCs also consume significant amounts of energy to power servers, cooling systems, and backup storage. In The Dalles, Google capitalizes off cheap hydroelectric power from dams on the Columbia River. From the summit of Mount Hood to the Columbia River Gorge, water drains to cool and power the digital world.

DCs are the foundation to our data-driven decisions, economy, and livelihoods. DCs vary in size, purpose, and location but can be sorted into five categories: enterprise, colocation, hyperscale, edge, and container (Industry 2022). This analysis will focus on hyperscale DCs. With areas larger than ten thousand square feet, hyperscale DCs have the capacity to store large

quantities of data and perform cloud services. These facilities house servers, cooling systems, and IT equipment which provides data storage, trains artificial intelligence models, or simply allows us to look up directions to the nearest grocery store. The hyperscale market is dominated by three firms: Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (IBM 2024).

As reliance on digital solutions and artificial intelligence increases, so does the demand for DC capacity and computational power. The DC industry has expanded rapidly in the past decade and is projected to continue growing by 10% a year reaching $49 billion in new construction spending in 2030 (Schaap 2024). Recently, hyperscale providers have also invested heavily in development of A.I. models which require six times the computational capacity of standard DCs to train and operate (Font 2023). The expansion of A.I. services will continue as president Trump announced a $500 billion private sector investment in A.I. related infrastructure (Jacobs 2025). With modern industry reliance on digital solutions and growing demand for A.I., the growth of the hyperscale DC industry will accelerate.

While hyperscale DCs offer many benefits, they may also impose significant negative externalities. Negative externalities are costs generated from a market that are imposed on external agents.[1] The presence of supply-side negative externalities implies that the social cost of hyperscale DCs is higher than the private cost. With an elevated social cost, the private market equilibrium calls for a higher quantity of facilities than is socially optimal. Decreasing the number of facilities to reach the socially optimal equilibrium is not feasible due to the significant investment and growth in demand. Instead, by mitigating the external costs, the gap between the social and private market equilibriums can be diminished, improving social outcomes. The primary externalities associated with hyperscale DCs are carbon footprint and water footprint.

Hyperscale data center locations

Firms will locate DCs in areas that minimize private costs. However, the magnitude of the external costs associated with a hyperscale DC is largely determined by its location. Siting parameters that are attractive to firms may differ from those which minimize external costs. This

[1] An example of a negative externality is cigarette smoke in a room. Smoke is generated by the actions of the smoker, but the cost is imposed on others in the room.

results in socially inefficient DC siting. Optimizing siting to mitigate negative externalities can increase the social benefit of DCs.

The first step in improving DC siting is to identify the key factors that influence location decisions. Hyperscale firms will choose to site DCs in strategic locations that minimize capital and labor costs (Goiri et al. 2011). Multiple parameters may affect a firm's location decision including, but not limited to, energy prices, fiber optic network proximity, natural disaster risk, and tax rates (Greenstein and Fang 2022; Covas, Silva, and Dias 2013). To model the siting process, I conduct an econometric analysis to model hyperscale DC counts by county in the contiguous United States (CONUS). I divide the DC siting process into two margins (extensive and intensive) and utilize separate regression models to describe each. The independent variables included in the regressions capture the private costs associated with DC locations. This both formalizes and quantifies the effects of proposed siting parameters from the literature on DCs. I also include average temperature and renewable energy potential to test firms' sensitivity to environmental outcomes. Through inclusion of both private costs and environmental considerations, this analysis provides a comprehensive overview of hyperscale DC siting in the CONUS.

Background

Existing literature

Despite the rapid increase in DC development, academic literature on DC siting is sparse. The existing literature consists of one econometric study that models DC location determinants (Greenstein and Fang 2022). Greenstein and Fang (2022) construct a logit model to measure the effects of demand, fixed costs, and variable costs on the probability of a DC existing in a given location. While no other econometric study of location determinants exists, Goiri et al. (2011) do build a cost optimization framework for DC siting. This framework includes a variety of variables that firms consider including average temperature, land availability, and network latency (Goiri et al. 2011). Similarly, Covas et al. (2013) and Gao et al. (2012) propose optimization frameworks for efficient DC siting that can minimize both private and social costs. Combining these existing cost optimization studies with the econometric model from Greenstein and Fang (2022) yields a large list of possible parameters that a firm may consider when siting a DC. Once selected, the location significantly influences a DCs environmental impact due to geographic variations in water scarcity, energy generation methods, and weather patterns (Siddik, Shehabi, and Marston 2021).

Water footprint

While research on the siting process is limited, there is a growing body of literature on the environmental effects of DCs. The water footprint of a hyperscale DC can be attributed to two main sources: cooling systems within the facility and water used in off-site electricity generation (Lei et al. 2023). The metric for understanding the water efficiency of DCs is “water usage effectiveness” (WUE). WUE describes the ratio between a DC's total water use and the energy that goes to IT equipment (Ristic, Madani, and Makuch 2015). A lower WUE indicates a more water efficient DC. WUE depends on factors such as the cooling technology and location of a DC. Advanced cooling technology and cooler climates have been shown to reduce WUE (Lei et al. 2023). This implies that strategic DC siting can minimize WUE, lowering the water footprint associated with the facility.

Lei et al. (2023) perform a geospatial analysis on hyperscale DC water usage in the United States. The authors focus on the effects of regional electricity generation methods and

cooling technologies on a DC's water footprint. Ristic, Madani, and Makuch (2015) conclude that water usage for electricity generation accounts for much of the water footprint of a DC when considering indirect usage. If only the direct water footprint of a DC is considered, the HVAC system and heat/chemical pollution are the primary sources of water usage (Ristic, Madani, and Makuch 2015).

Carbon footprint

In addition to water footprint, there is a growing body of literature on the carbon footprint of DCs. The carbon footprint of hyperscale DC operations can be attributed to the electricity generation required to power the facility. Data centers use significant amounts of electricity to power their operations. As of 2021, the United States DC industry accounted for 1.8% of total energy use in the US (Siddik, Shehabi, and Marston 2021). To measure energy efficiency, and subsequent carbon footprint of DCs, researchers use "power usage effectiveness" (PUE). PUE is the ratio of total power usage to the amount of power being used for IT equipment (Mytton 2021). A perfectly efficient facility with all input energy being used for IT equipment would have a PUE of 1. Efficiency gains in cooling and chip technology have improved PUE across the industry helping to reduce marginal increases in energy usage (Masanet et al. 2020). While a positive sign for the sustainability of this industry, improved efficiency does not fully offset demand. (Shehabi et al. 2018). Moreover, these studies do not capture the new demand for A.I. services.

Siddik et al. (2021) link DC energy usage metrics with research on electricity generation to find that DCs account for 0.5% of total greenhouse gas emissions in the United States. Hyperscale providers such as AWS have increased their investment in renewable energy generation both on and off site to reach net-zero carbon by 2040 (AWS 2024). Despite sustainability efforts by hyperscale providers, multiple studies have explored models to minimize these emissions. DCs have the potential to lower their emissions with improved siting if renewable energy proximity, underlying grid dynamics, and other siting parameters are weighed (Abdennadher et al. 2022). While DCs may face a tradeoff between minimizing electricity costs and emissions, there exist specific locations that would decrease emissions with no effect on energy price or network latency (Gao et al. 2012). The literature shows that the carbon footprint of a DC depends on various factors including its size, PUE, and location (Siddik, Shehabi, and Marston 2021).

Contributions

This paper bridges the gap between environmental assessments of DCs and literature on traditional siting decisions by including both renewable energy potential and average temperature data as model inputs. Additionally, this analysis is specific to hyperscale DCs, setting it apart from much of the existing literature. By joining environmental considerations with traditional siting parameters, this study expands the understanding of hyperscale DC locations.

Siting parameters

I selected DC siting parameters (independent variables in the model) based on existing research, news articles, and industry information. When siting a DC, firms will minimize both up-front and operational costs. Firms may also attempt to minimize environmental costs to align with their sustainability goals. In this analysis, I utilize energy prices, sales tax rates, colocation counts,[2] renewable energy potential, average temperature, natural disaster risk, fiber optic connectivity, and population to model the number of DCs by county. Based on existing literature and economic assumptions, I predict that attractive sites for DC are ones which minimize energy prices, sales tax rates, natural disaster risk, and average temperature while maximizing colocation benefits, renewable energy potential, fiber optic connectivity, and proximity to higher populations.

Low energy prices are a primary requirement as energy usage accounts for the largest cost of a hyperscale DC (Covas, Silva, and Dias 2013). Figure 7 in the appendix explores the relationship between energy prices and DC counts in the Pacific Northwest (PNW).[3] Firms may also seek low sales tax rates as this lowers the cost of purchasing the necessary capital such as servers and other IT equipment (Weise and Tamayo 2024). Areas with lower risk of natural disasters are attractive to firms as this helps mitigate future threats to the facility (Development 2009). Finally, firms may wish to site DCs in areas with lower ambient temperatures as this has

[2] Colocation is the process of clustering with similar firms to tap into favorable and shared labor, knowledge, and other conditions. This effect may decrease with higher DC counts as explored in the methods section.

[3] I chose to illustrate this region due to its relatively high count of counties with DCs.

been shown to lower cooling costs (Goiri et al. 2011). By minimizing these spatial variables, hyperscale providers can reduce the costs associated with their DCs.

Attractive sites for DCs will also have higher values of some variables. Firms may seek areas with high counts of exiting DCs as these areas have an established DC labor force, and reduced costs. However, as regions become overcrowded, this effect may diminish and even invert. Increased renewable energy potential may also be a desirable condition for hyperscale DC providers. Renewable energy generation allows firms to reduce their carbon footprint and meet their sustainability goals. Figure 9 in the appendix illustrates the relationship between renewable energy potential and DC counts in Texas to provide a snapshot of the country. Fiber optic networks are also crucial to hyperscale DC success as this infrastructure allows data to be transported across space seamlessly. Firms will locate DCs in areas with available or cheap connection to fiber optic networks. (Development 2009). Figure 8 in the appendix explores the interaction between fiber optic coverage and DC counts in the PNW. Finally, firms may site DCs closer to population centers to minimize the distance between DC and consumer (Goiri et al. 2011). The effect of population may be less significant for hyperscale firms as their services differ greatly from those of smaller, third-party providers (Greenstein and Fang 2022). Firms can derive the greatest benefit from hyperscale DCs by siting them in locations that minimize costs and maximize desirable variables.

Methods

Model data

This analysis relies entirely on existing, publicly available data. Data on DC locations is difficult to obtain. Often, companies do not disclose exact locations for their facilities and, as a result, most publicly available data rely on pieced together observations from a variety of sources. Datacenters.com is one such site that tracks global DC locations of all categories.

I scraped datacenters.com for all DC locations in the United States of all types. This yielded a dataset of facilities each with a name, owner, and address. I then converted these observations into spatial objects with latitude/longitude coordinates using the Google geocoding API. I filtered the facilities by provider to capture only the largest hyperscale providers.[4] Due to the constant evolution of this market and overall poor data availability, this was still an incomplete dataset. To address critical data gaps, I combed through each major hyperscale providers site list manually to identify and insert any missing facilities. Missing locations tended to be newer DCs or facilities with minimal press coverage.[5] After wrangling and cleaning the data, I was left with 414 hyperscale DCs in the CONUS. Using a spatial join, I aggregated DC observations at the county level (Figure 1).

[4] List of major hyperscale providers: Amazon AWS, Digital Realty, CyrusOne, EdgeConnex, CoreSite, H5 Data Centers, QTS Data Centers, Prime Data Centers, Google, Iron Mountain Data Centers, Aligned Data Centers, Facebook, Switch Data Centers, Oracle, Microsoft Azure, T5 Data Centers, Apple Inc., NTT Communications, Vantage Data Centers.

[5] Data were originally pulled in May 2024, then updated in December 2024

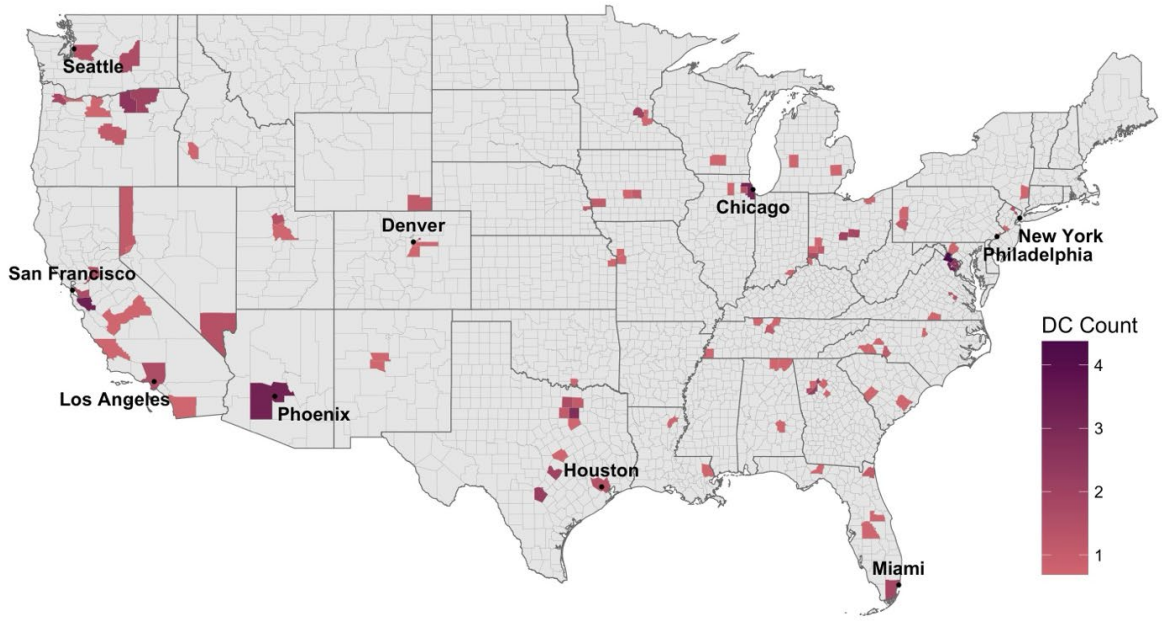

Figure 1. Map of hyperscale DC counts by county, CONUS

Source: Datacenters.com, 2024

Model input data

To model locations of hyperscale DCs, I used a variety of input parameters. These independent variables were selected based on the existing literature as discussed in the introduction and literature review.

Energy price data were obtained from the National Renewable Energy Laboratory (NREL 2024b). This dataset captures investor-owned utility (IOU) and non-IOU energy prices by zip code for residential, commercial, and industrial customers measured in dollars per kilowatt hour ($/kWh). I joined and averaged the observations by county to capture average commercial electricity price by county.

Temperature data by county were obtained from NOAA National Centers for Environmental Information (NOAA 2024). This dataset describes average temperature by county from 2020 to 2024 in Fahrenheit.

Sales tax rates were pulled from the Tax Foundation (Walczak 2024). These data are provided at the state level. The Tax Foundation uses both the state rate as well as local rates,

weighted by population to generate a combined rate. This combined rate is what I used in the analysis as it captures some effect of local rates.

Renewable energy data were pulled from NREL (NREL 2024a). I used NREL's technical potential measurements to capture renewable availability/future development opportunity. Technical potential defines an upper bound of a geographies maximum renewable capability based on its environment, topography, technology, and economic conditions. I pulled county level data for both solar and wind technical potential and state level data for hydroelectric potential (hydro data at the county level were not available). I normalized the three indices using Z-score normalization, meaning each field had a mean of zero and standard deviation of one. Lastly, I added the three indices resulting in one measure of renewable potential for each county.[6] This index is logged in the regression models to minimize its variance.

To capture **natural disaster risk**, I used the county level risk index from the Federal Emergency Management Agency (FEMA 2023). This index measures factors including likelihood, magnitude, community resilience, and existing infrastructure. The risk index, generated in 2023, accounts for 18 types of hazards.[7] While not all of the included hazards pose a significant risk to DCs, this index from FEMA is the best available option to capture natural disaster risk as the county level.

To describe **fiber optic connectivity** by county, I pulled data from the Federal Communications Commission (FCC 2024). These data describe the percent of units covered by a given technology and speed at the county level. I first filtered for fiber technology for business customers only. Then, I selected observations that measure connectivity to only the highest available speeds. Percentage of units covered is not an ideal measure of connectivity, but the inclusion of population in the model helps correct this limitation.

For **population**, I pulled estimates by county from the US Census ACS 5-year survey (U.S. Census Bureau 2023). This survey data includes margins of error for each observation which were not included for simplicity. Population estimates are logged in the models to account for high variance.

---

[6] I also added a small constant to the index to allow for log transformations

[7] Hazards: Avalanche, Coastal Flooding, Cold Wave, Drought, Earthquake, Hail, Heat Wave, Hurricane, Ice Storm, Landslide, Lightning, Riverine Flooding, Strong Wind, Tornado, Tsunami, Volcanic Activity, Wildfire, and Winter Weather.

Finally, to model the **colocation** patterns in the DC industry, I utilized a spatial lag to total hyperscale DC counts of all bordering counties for each county. The neighboring DC count is expressed as a polynomial term in the model. This term captures the diminishing (and eventually inverse) marginal effect of colocation. Initial DC growth may boost counts of DCs in neighboring counties, however, as the surrounding market and local grid becomes overcrowded, DC counts may decrease. Figure 2 formalizes this intuition as the data show a concave down, parabolic trend.

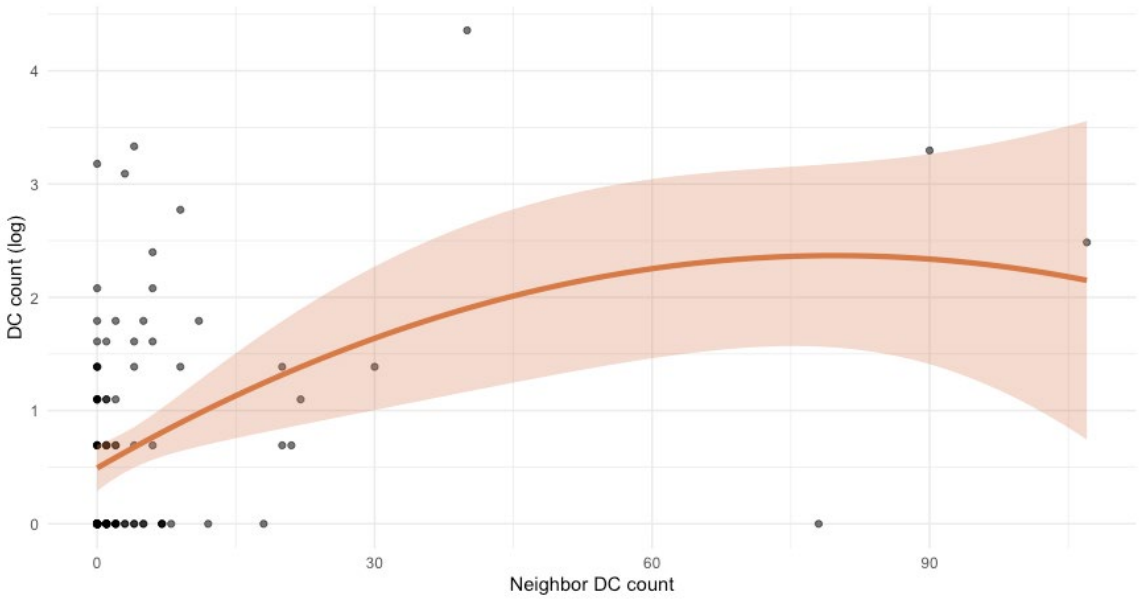

Figure 2. Polynomial effect of neighbor counts on DCs

Firms will leverage the benefits associated with colocation including established labor force, favorable regulations, and attractive conditions. Colocation benefits help to explain the clustering of DCs seen in certain regions such as Northern Virginia and the PNW. As areas become overcrowded and grid capacity is strained, providers may increase DC dispersion.

Constructing the model: Negative binomial

As seen in Figure 3, most counties in the CONUS have a hyperscale DC count of 0. The maximum county count, however, is 78 facilities in Loudoun, VA. With a mean count of 0.1386 and a variance of 3.278 this data structure is significantly over dispersed. To model these data, I first used a negative binomial regression model. A negative binomial model was selected due to

its ability to handle over dispersion in count data. This model is an extension of a Poisson regression model as it includes an additional parameter to account for over dispersion.

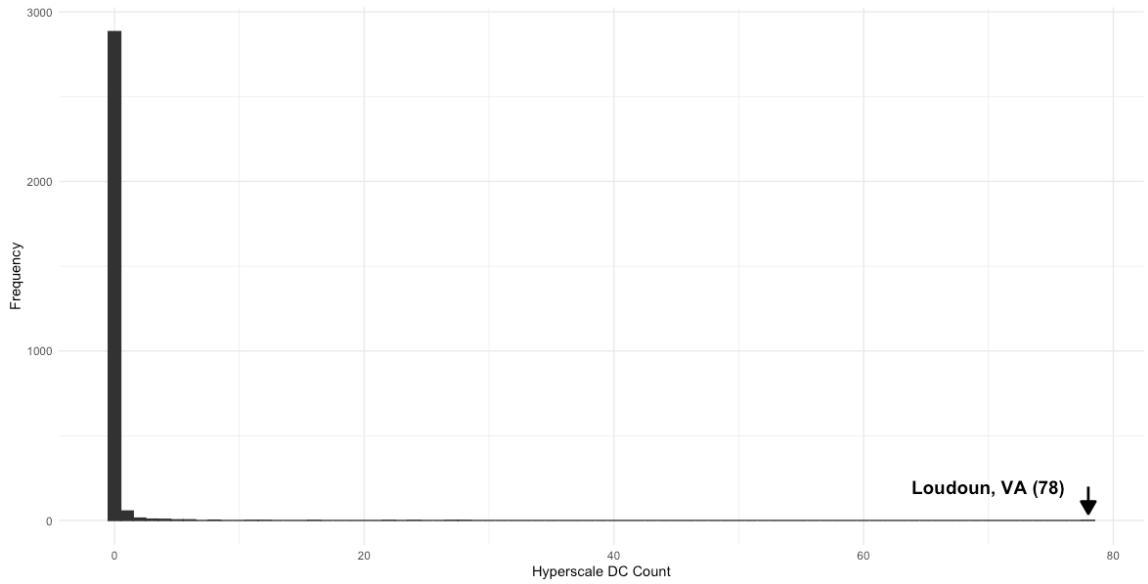

Figure 3. Hyperscale DC counts by county histogram, CONUS

This histogram illustrates the over dispersion of the data as well as the high zero count. The majority of counties have zero DCs, while a select few have high counts. For example, Loudoun County, VA has 78 hyperscale DCs.

Figure 4 shows an analysis of the model residuals. These residuals exhibit a downward, linear trend. The mean value is not constant around zero and decreases across higher fitted values. The trend of the residuals suggests that the negative binomial model is overestimating DC counts for many counties. This may be a result of the model struggling to describe both the high number of zeros as well as the few positive counts. Ideally, the residuals would follow the dotted line, maintaining a mean of zero. The histogram reveals that the residuals are non-normally distributed, further illuminating the issues with this model. The negative binomial regression does not adequately fit the data.

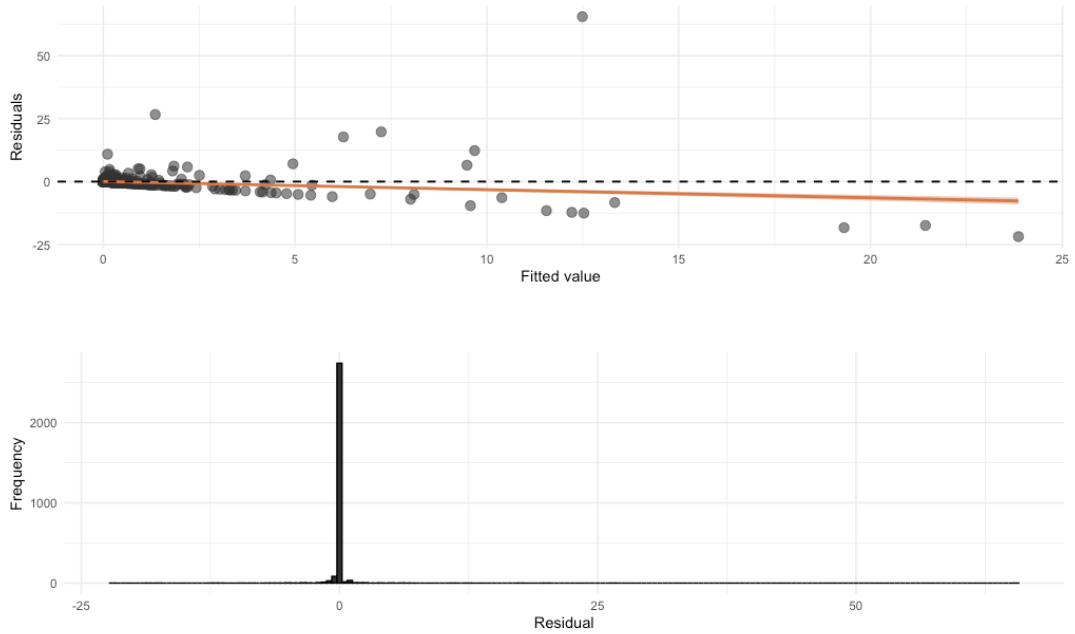

Figure 4. Negative binomial residuals: Scatter plot and histogram

Constructing the model: Extensive and Intensive Margins

The poorness of fit with a negative binomial model reflects the difficulty of modeling overdispersal and zero-inflated count data. To address this issue, I describe the data using two separate processes: extensive and intensive margins. The extensive margin describes a binary outcome of if there is at least one DC in a county or not. The intensive margin describes how many DCs are in counties that have at least one facility. I model the extensive margin using a logistic (logit) regression model and the intensive margin using a log-linear regression model. By dividing the analysis, I describe hyperscale DC locations as a product of two different processes.

Extensive margin

The extensive margin describes which counties have hyperscale DCs and which do not. I construct an indicator variable for counties with at least one facility and utilize a logit model to predict the probability that a hyperscale DC will exist in each county.

Figure 5 shows an analysis of the model residuals. The scatter plot shows two distinct, downward sloping lines. The upper line shows residuals from actual values of 1 as it remains positive, and the lower line shows residuals from actual values of 0. The high density of points

with fitted values of zero on the upper line suggest that this model is often predicting zero DCs when there is a DC in the county. These residuals have a constant, zero mean.

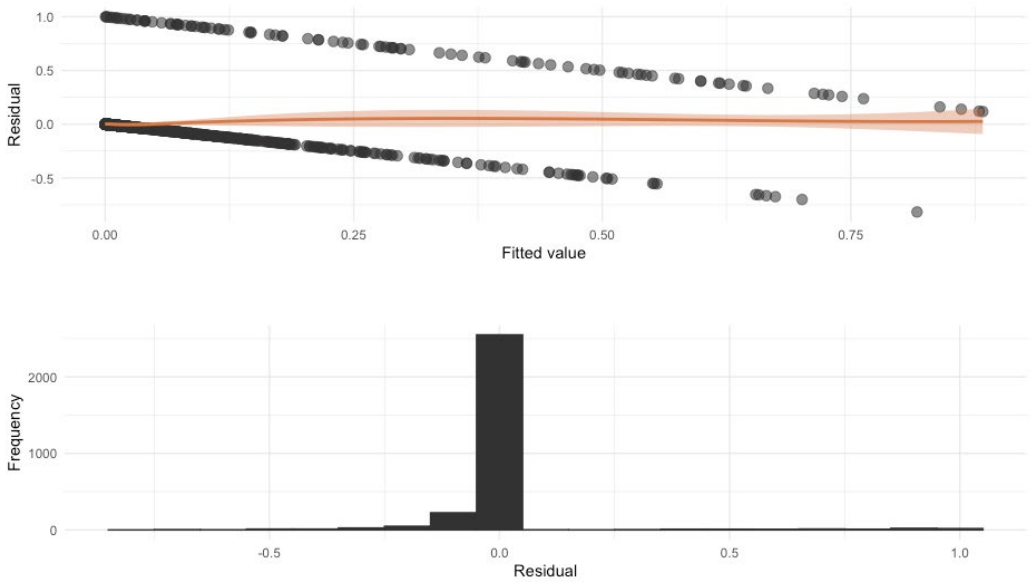

Figure 5. Logistic model residuals: Scatter plot and histogram

Intensive margin

The intensive margin describes the number of DCs within counties that have at least one DC. To model these data, I use a log-linear regression model, estimated with ordinary least squares (OLS)[8]. Logging the dependent variables shrinks the variance of DC counts making them simpler to model. The residuals of the log-linear model are shown in Figure 6. These residuals have a constant, zero mean, and no significant skew in the histogram. While improved, these residuals do not reflect an ideal model. There is a linear, downward pattern suggesting that the model is failing to capture a key trend in the data. Furthermore, these residuals are heteroskedastic, meaning their variance increases over fitted values of the dependent variable. A White's test confirms heteroskedasticity with a p value of .000009. To account for

[8] OLS minimizes the squared residuals (distance between actual and fitted values) of the model

heteroskedasticity, I set vcov = “het” in my model which ensures that standard errors are robust and can have differing variance across observations.

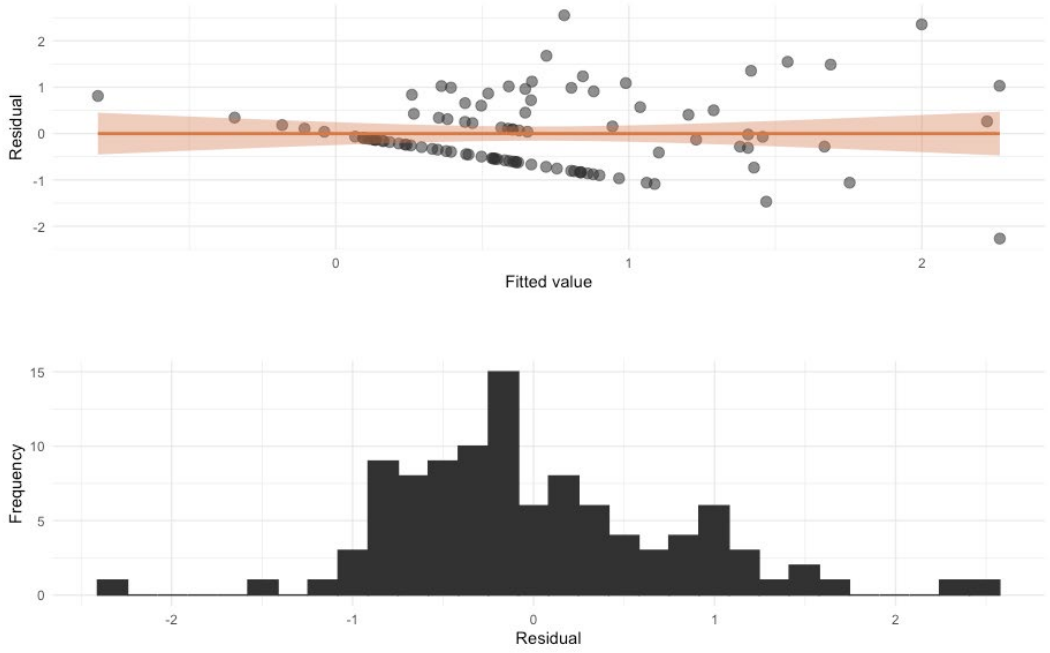

Figure 6. Log-linear model residuals: Scatter plot and histogram

Results

Logistic regression

Model results are presented in table 1 in the appendix. Table 2 shows exponentiated coefficients of the logit model to improve interpretation. The following results describe a percent change of the odds ratio of a DC locating in a given county due to a one unit change in the independent variable. Statistically significant coefficients from the logistic regression include neighboring counts, sales tax rates, fiber optic connectivity, and population. Specifically, an increase of one DC in a neighboring county is associated with a 2% increase in the odds of a DC locating in the given county, ceteris paribus.[9] This aligns with the literature and assumptions about colocation in the DC industry. Because DCs seek out similar geographical conditions and workforces, they are more likely to cluster together.

The model estimates a negative coefficient on sales tax. A 1% increase in the sales tax rate is associated with a 12% decrease in the odds of a DC being developed in a county, ceteris paribus. This finding is also consistent with the literature as we expect DCs to seek low sales tax rates. These lower rates allow firms to minimize capital costs such as servers and other IT equipment.

The logit model produces a large, positive estimate for the effect of fiber optic connectivity. A 1% increase in the percentage of units covered by high-speed fiber optic technology increases the odds of a DC by 271%, ceteris paribus. Managing high volumes of data requires high transport speeds to minimize costs and latency. This large, positive coefficient suggests that hyperscale DCs seek out locations in which connection to high-speed fiber optic technology is available.

A 1% increase in population is associated with a 286% increase in the odds of a DC, ceteris paribus. This result indicates a preference for counties with higher populations. This could be a result of hyperscale providers minimizing their distance to demand.

9 Holding all else equal.

Log-linear regression

The log-linear results describe the percent change of DCs in counties with at least one DC, based on a unit increase of the independent variable, ceteris paribus. These estimates describe the conditions that firms seek out when siting a DC in a county that contains existing hyperscale facilities. Significant estimates from the model include the linear coefficient of the polynomial neighboring count, the log of the renewable potential index, and the log of the population estimate.

The log-linear model estimates a positive coefficient for the linear term of colocation. A 1 unit increase in the number of DCs in neighboring counties is associated with a 3.28% increase in DCs in the given county, ceteris paribus. Additionally, while not significant, this effect decreases by 1.41% for each additional DC in a neighboring county. This finding aligns with the assumption that firms locate DCs in regions with established markets as they share favorable conditions, existing labor forces, and shared knowledge. The negative coefficient on the polynomial term suggests that this effect diminishes as areas become overcrowded.

The estimate of the coefficient for renewable energy potential is also statistically significant. A 1% increase in the renewable potential of a county is associated with a 0.45% increase in the number of DCs, ceteris paribus. This finding suggests that, out of counties with existing DCs, firms seek locations in which renewable development potential is high. While a small estimate, this is a positive indicator for the environmental impacts of DCs. With growing pressure for improving environmental outcomes, many leading firms such as AWS have made carbon neutrality and renewable development pledges to offset their emissions (AWS 2024).

A 1% increase in a county's population is associated with a 0.30% increase in the number of DCs, ceteris paribus. This estimate suggests that, out of counties with existing DCs, firms prefer counties with higher populations. This estimate is likely driven by high population counties with high DC counts such as Maricopa, AZ or Cook, IL which contain the cities of Phoenix and Chicago, respectively.

While not statistically significant, the energy price variable had the largest absolute value of its coefficient of any predictor, other than the intercept term. This large, negative estimate aligns with the literature as minimizing the cost of energy is at the top of this list for location determinants of hyperscale DCs.

The lack of significance could be a result of poor energy price data or the inability to capture price agreements between DCs and utility providers using publicly available data.

Discussion

Environmental implications

The results of both the logistic and log-linear regressions highlight important trends in the siting process of hyperscale DCs. The two variables that relate to environmental costs that were included in the models are average temperature and renewable energy potential. Lower ambient temperatures can lower the energy and water required to cool DCs which decreases both the private and external cost associated with water and power use (Lei et al. 2023). Higher renewable potential of a county indicates that there is a high ceiling for on-site or nearby renewable generation which can help reduce carbon emissions associated with powering DCs. This implies that colder climates with higher renewable potential would be optimal sites for hyperscale DCs from an environmental cost perspective. The logistic regression finds neither of these coefficients to be statistically significant. Moreover, the sign of the coefficients is the inverse of the environmentally optimal condition with a negative estimate for renewable potential and a positive estimate for average temperature. The results of this model suggest that firms primarily site their DCs based on taxes, colocation benefits, fiber optic connection, and population centers with no regard for renewable potential or temperature. This finding aligns with the assumption that firms will minimize their private costs. External costs such as carbon and water footprints are not imposed directly on a firm and therefore will not factor into their siting process in the absence of government intervention or social pressure.

The log-linear regression models the intensive margin and produced a significant and positive estimate for the effect of renewable energy on counties with at least one existing facility. This result shows that high renewable potential in counties with at least one data DC is associated with an increase in DC counts. While a small coefficient (.45%) this estimate indicates firms' intent to address and lower their environmental footprints. This result is reflected in reports directly from the DC industry. In 2023, Amazon matched its total operational power usage with renewable energy. "Matching" does not guarantee that every Amazon DC is directly powered with renewable energy, but the company has invested in over 500 solar and wind projects across the world making it the largest cooperate purchaser of renewable energy (Swinhoe 2024).

Neither the log-linear nor the logistic model produced a significant estimate for the effect of average temperature, suggesting that firms do not consider temperature on either the extensive or intensive margin. Lower average temperature decreases both internal and external costs making this result somewhat surprising. This finding reveals an opportunity for efficiency gains in this market. Siting facilities in cooler climates would lower operational costs due to decreased need for cooling while also lowering both the carbon and water footprint associated with cooling.

Limitations and future research

There are multiple limitations of this research, most of which stem from data availability issues. A primary concern with the data used in this analysis is the dependent variable, DC locations. I filtered observations by provider name to isolate hyperscale facilities. Filtering observations by square footage would increase sample accuracy, but these data were unavailable. Issues also persist with data availability for the independent variables. Rates such as taxes and energy prices are often negotiated privately between firms and local governments or utility providers. Private agreements play a large role in DC siting but will not be captured by any nationwide, publicly available dataset. Moreover, many of these data, such as tax rates and hydroelectric potential were only available at the state level despite the analysis being performed at the county level.

The regression model estimates may also suffer from omitted variable bias (OVB). OVB arises if an explanatory variable that influences both the dependent and independent variable is not included in the model. One potential omitted variable is water availability by county. This variable may affect DC counts as DCs will seek areas with cheap, accessible water for cooling. Water availability may also influence renewable energy potential as hydro-electric energy is incorporated into the index. Complete elimination of OVB would require extensive data gathering that was beyond the scope of this analysis.

Despite its limitations, this analysis illuminates important trends in hyperscale DC siting including key environmental considerations such as temperature and renewable energy potential. Future research of hyperscale DC locations should aim to include individually negotiated tax breaks and energy price discounts wherever applicable as these are important drivers of siting decisions. Overall, access to funding and additional time would yield improved data, minimizing the limitations of the model. With improved data, future analyses could explore how additional

environmental variables effect DC counts. These environmental variables could include current energy generation sources, water availability, and environmental regulations by county.

Conclusion

Hyperscale DCs are the backbone to an increasingly digital world. With capacity for data storage, processing, and A.I. training, these facilities are essential for the operation of our everyday lives. Furthermore, with increased demand for A.I. and data services, the hyperscale DC industry will continue growing at an accelerated rate. Along with their many benefits, hyperscale DCs also impose externalities in the form of environmental costs. These costs include the carbon and water footprint associated with both powering and cooling the thousands of servers housed within the facility. Despite significant gains in efficiency, the growing demand for DCs and the magnitude of their environmental costs requires immediate attention. The negative externalities of DCs are largely a product of their location. By optimizing DC siting, the external costs associated with their operations can be minimized. In this analysis, I use a variety of parameters to describe DC counts by county in the CONUS.

The results of this analysis show that there is opportunity to improve social outcomes in the hyperscale DC siting process. DCs can be sited in locations that minimize both private and external costs. However, without market adjustments or regulations, firms have no incentive to consider external costs in their siting decisions. Regulations around water usage and energy sources have the potential to significantly lower the magnitude of these externalities. Governments should consider a Pigouvian tax[10] on water, based on the local water scarcity index. This market adjustment would require DCs to pay a tax equal to the relative damage they impose on the water supply based on the area's water scarcity. Areas with increased water scarcity would see larger tax rates than those with abundant water. This policy would discourage DC development in water-scarce areas while incentivizing expansion in regions with abundant water resources. While this does not change the absolute size of water footprint, it does lower the social cost associated with that footprint.

To minimize the absolute size of the footprint, firms should consider the value of natural cooling. Leveraging cooler ambient temperatures would lower the cooling demand of the DC, subsequently reducing its carbon and water footprint. This would also decrease the private

[10] A Pigouvian tax is a tax levied on producers that is equal to the size of the externality associated with production.

operational costs to the firm. Government regulations prohibiting the development of hyperscale DCs in extremely hot climates would be an effective measure in facilitating this shift. However, this policy would require significant investment in fiber optic networks to ensure minimal latency to end users in hot climates. If the policy were not met with investment in fiber connectivity to locations with higher temperatures, it would have a disproportionately negative effect on those populations.

Since building its first DC in The Dalles in 2006, Google has saved over $240 million through local tax breaks. However, after the 2006 tax breaks, The Dalles has been taxing new Google DCs at an increasingly higher rate (Rogoway 2022a). Perhaps the initial enthusiasm for DCs has worn out as awareness of their external costs has grown. Further attention and research of hyperscale DCs is still needed to understand the optimal response to rapid industry growth. While DCs provide essential services, rushing to develop facilities without consideration for external costs would be unwise. Locally, towns must decide if the tax revenue and labor benefits of DCs outweigh the local negative externalities. Globally, we must consider if saving time in the short run through digitalization and A.I. is worth the potential costs in the long run. Without regulatory intervention, firms will continue to site hyperscale DCs in locations which minimize private costs, leaving communities to bear the external costs.

Appendix

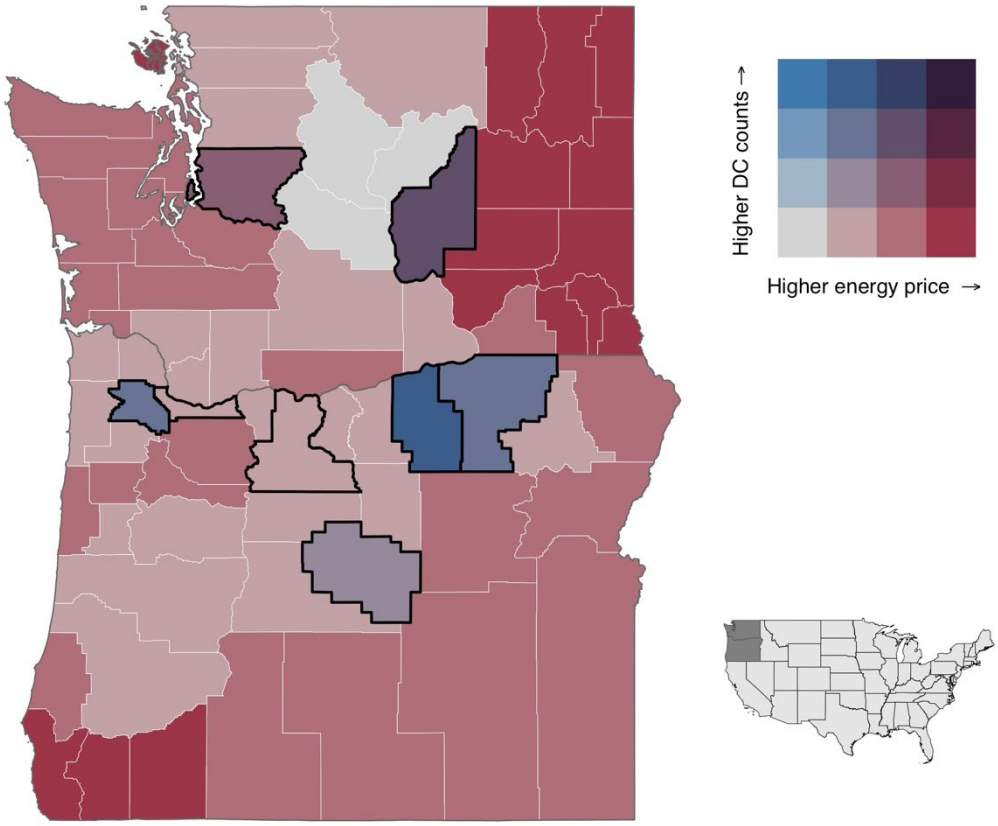

Figure 7. Bivariate map of DC counts and energy prices, WA and OR

This map shows the interaction between energy prices and DC counts by county for OR and WA. Counties with at least one hyperscale facility are outlined in black. Counties with high DC counts and low energy prices are shown in blue. These locations are primarily along the Columbia River in OR. In general, counties with high energy prices (dark red) do not have high DC counts.

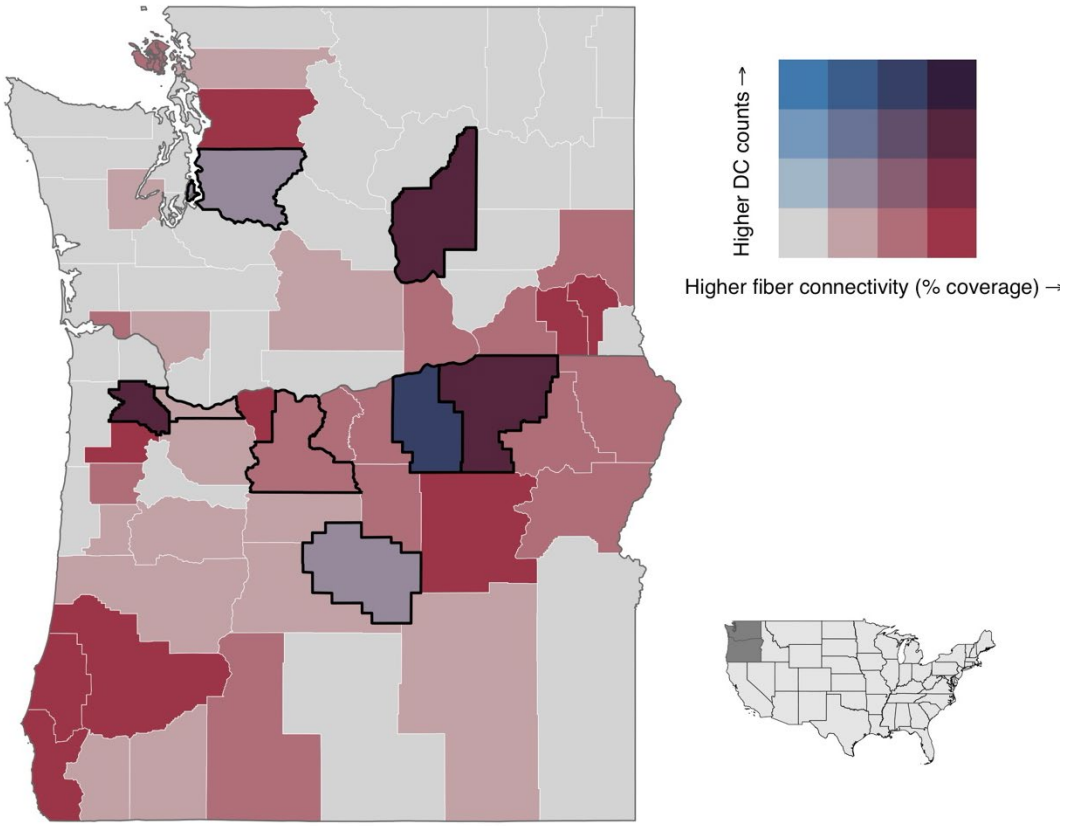

Figure 8. Bivariate map of DC counts and fiber optic connectivity, WA and OR

This map shows the interaction between fiber optic coverage (%) and DC counts by county for OR and WA. Counties with at least one hyperscale facility are outlined in black. Counties with high DC counts and high fiber optic coverage are shown in purple. In general, counties with low connectivity (grey and light red) do not have high DC counts.

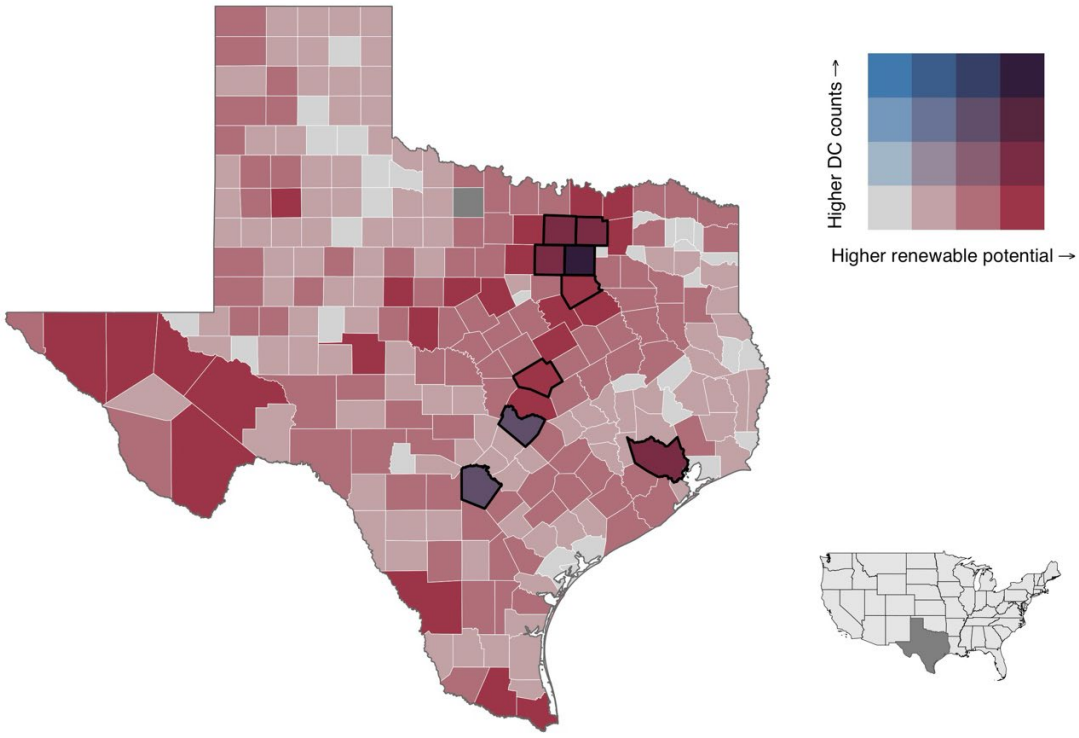

Figure 9. Bivariate map of DC counts and renewable energy potential, TX

This map shows the interaction between renewable energy potential and DC counts by county for TX. Counties with at least one hyperscale facility are outlined in black. All counties with hyperscale DCs have relatively high renewable potential (red). Moreover, DC counts tend to increase in counties with increased renewable potential. Counties with low renewable potential (grey) do not have any hyperscale facilities.

Logistic

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-17.716666	1.512208	-11.715757	< 2.2e-16	** *
Energy price	-4.480347	5.247461	-0.853812	0.393209	
Neighbor count	0.020638	0.009774	2.111472	0.034732	*
Sales tax	-0.122305	0.067409	-1.814376	0.06962	.
Avg. temp.	-0.006487	0.014245	-0.45542	0.648807	
log(renewables)	-0.129813	0.309543	-0.419371	0.674945	
Risk score	0.002394	0.008012	0.29875	0.765131	
Fiber speed	1.310353	0.516718	2.535913	0.011215	*
log(population)	1.349779	0.136741	9.871038	< 2.2e-16	** *

Log-Likelihood: -272.0 Adj. Psd. R: 0.375

Log-linear

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-2.951825	0.979217	-3.014475	0.0033177	**
Energy price	-4.650364	4.385153	-1.060479	0.2916729	
Neighbor count	0.045887	0.025969	1.767011	0.0805069	.
(Neighbor count) ^2	-0.000301	0.000247	-1.219040	0.2259121	
Sales tax	-0.060632	0.049462	-1.225828	0.2233585	
Avg. temp.	0.00823	0.012022	0.684541	0.4953359	
log(renewables)	0.446343	0.240038	1.859469	0.0661222	.
Risk score	-0.007629	0.005988	-1.273989	0.205842	
Fiber speed	-0.031681	0.342922	-0.092384	0.9265915	
log(population)	0.29548	0.089525	3.30052	0.0013694	**

RMSE: 0.779775 Adj. R2: 0.262482

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1. Model results

This table shows the regression output for both the logistic and log-linear regressions. Significance codes are provided in the last row.

	Estimate	e^estimate
(Intercept)	-17.716666	2.02185E-08
Energy price	-4.480347	0.011329481
Neighbor count	0.020638	1.020852436
Sales tax	-0.122305	0.884878439
Avg. temp.	-0.006487	0.993533995
log(renewables)	-0.129813	0.87825965
Risk score	0.002394	1.002396868
Fiber speed	1.310353	3.707482222
log(population)	1.349779	3.856573134

Table 2. Exponentiated estimates if logit model

This table provides exponentiated transformations of the estimates from the logit model. This improves the interpretation of the results by converting the estimates to be a percentage change in the odds ratio of the dependent variable.

References

Abdennadher, Yasmine, Julia Lindberg, Bernard C Lesieutre, and Line Roald. 2022. "Carbon Efficient Placement of Data Center Locations." In *2022 North American Power Symposium (NAPS)*, 1–6. https://doi.org/10.1109/NAPS56150.2022.10012198.

AWS. 2024. "AWS Cloud - Amazon Sustainability." 2024. https://sustainability.aboutamazon.com/products-services/aws-cloud.html.

Bakken-French, Nicolas, Stephen J. Boyer, B. Clay Southworth, Megan Thayne, Dylan H. Rood, and Anders E. Carlson. 2024. "Unprecedented 21st Century Glacier Loss on Mt. Hood, Oregon, USA." *The Cryosphere* 18 (9): 4517–30. https://doi.org/10.5194/tc-18-4517-2024.

Covas, Miguel T., Carlos A. Silva, and Luis C. Dias. 2013. "Multicriteria Decision Analysis for Sustainable Data Centers Location: International Transactions in Operational Research." *International Transactions in Operational Research* 20 (3): 269–99. https://doi.org/10.1111/j.1475-3995.2012.00874.x.

Development, Steve Kaelble, Staff Editor, Area. 2009. "Location Factors for Data Centers." Area Development. September 8, 2009. https://www.areadevelopment.com/siteselection/august09/data-centers-electricity-climate-space008.shtml.

FCC. 2024. "FCC National Broadband Map." FCC National Broadband Map. June 30, 2024. https://broadbandmap.fcc.gov.

FEMA. 2023. "Data Resources | National Risk Index." March 2023. https://hazards.fema.gov/nri/data-resources.

Font, Juan. 2023. "Working At Full Power: Data Centers In The Era Of AI." *Forbes*, 2023. https://www.forbes.com/sites/forbestechcouncil/2023/10/30/working-at-full-power-data-centers-in-the-era-of-ai/.

Gao, Peter, Andrew Curtis, Bernard Wong, and Srinivasan Keshav. 2012. "It's Not Easy Being Green." *ACM SIGCOMM Computer Communication Review* 42 (September):211–22. https://doi.org/10.1145/2377677.2377719.

Goiri, Inigo, Kien Le, Jordi Guitart, Jordi Torres, and Ricardo Bianchini. 2011. "Intelligent Placement of Datacenters for Internet Services." In *2011 31st International Conference on Distributed Computing Systems*, 131–42. https://doi.org/10.1109/ICDCS.2011.19.

Greenstein, Shane, and Tommy Pan Fang. 2022. "Where the Cloud Rests: The Location Strategies of Data Centers." *Harvard Business School*. https://www.hbs.edu/faculty/Pages/item.aspx?num=58964.

IBM. 2024. "What Is a Hyperscale Data Center? | IBM." May 8, 2024. https://www.ibm.com/topics/hyperscale-data-center.

Industry, Voices of the. 2022. “Understanding the Differences Between 5 Common Types of Data Centers.” Data Center Frontier. May 18, 2022. https://www.datacenterfrontier.com/sponsored/article/11427373/belden-understanding-the-differences-between-5-common-types-of-data-centers.

Jacobs, Jennifer. 2025. “Trump Announces up to $500 Billion in Private Sector AI Infrastructure Investment - CBS News.” January 21, 2025. https://www.cbsnews.com/news/trump-announces-private-sector-ai-infrastructure-investment/.

Lei, Nuoa, Jun Lu, Zhu Cheng, Zhi Cao, Arman Shehabi, and Eric Masanet. 2023. “Geospatial Assessment of Water Footprints for Hyperscale Data Centers in the United States.” *Journal of Physics: Conference Series* 2600 (17): 172003. https://doi.org/10.1088/1742-6596/2600/17/172003.

Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. 2020. “Recalibrating Global Data Center Energy-Use Estimates.” *Science* 367 (6481): 984–86. https://doi.org/10.1126/science.aba3758.

Mytton, David. 2021. “Data Centre Water Consumption.” *Npj Clean Water* 4 (1): 1–6. https://doi.org/10.1038/s41545-021-00101-w.

NOAA. 2024. “Climate at a Glance | County Mapping | National Centers for Environmental Information (NCEI).” https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/county/mapping/110/tavg/202409/48/value.

NREL. 2024a. “Renewable Energy Technical Potential | Geospatial Data Science | NREL | Geospatial Data Science | NREL.” 2024. https://www2.nrel.gov/gis/re-potential.

———. 2024b. “U.S. Electric Utility Companies and Rates: Look-up by Zipcode (2022).” National Renewable Energy Laboratory (NREL). https://catalog.data.gov/dataset/u-s-electric-utility-companies-and-rates-look-up-by-zipcode-2022.

Ristic, Bora, Kaveh Madani, and Zen Makuch. 2015. “The Water Footprint of Data Centers.” *Sustainability* 7 (8): 11260–84. https://doi.org/10.3390/su70811260.

Rogoway, Mike. 2022a. “After 15 Years, Google Will Pay Taxes on Its First Oregon Data Center.” Oregonlive. November 8, 2022. https://www.oregonlive.com/silicon-forest/2022/11/after-15-years-google-will-pay-taxes-on-its-first-oregon-data-center.html.

———. 2022b. “Google’s Water Use Is Soaring in The Dalles, Records Show, with Two More Data Centers to Come - Oregonlive.Com.” *Oregon Live*, December 17, 2022. https://www.oregonlive.com/silicon-forest/2022/12/googles-water-use-is-soaring-in-the-dalles-records-show-with-two-more-data-centers-to-come.html.

Schaap, Andrew. 2024. “Five Trends Driving The Booming Data Center Economy In 2024 (And Why Investors Are Taking Notice).” Forbes. 2024. https://www.forbes.com/sites/forbestechcouncil/2024/01/22/five-trends-driving-the-booming-data-center-economy-in-2024-and-why-investors-are-taking-notice/.

Selsky, Andrew, and Manuel Valdes. 2025. “Data Centers Intensify Water Fears in the Columbia River Gorge.” The Columbian. January 20, 2025. https://www.columbian.com/news/2021/oct/25/data-centers-intensify-water-fears-in-the-columbia-river-gorge/.

Shannon, Patrick. 2018. “Water, Wildlife and Wonder: The Mt. Hood National Forest.” National Forest Foundation. 2018. https://www.nationalforests.org/our-forests/light-and-seed-magazine/water-wildlife-and-wonder-the-mt-hood-national-forest.

Shehabi, Arman, Sarah J Smith, Eric Masanet, and Jonathan Koomey. 2018. “Data Center Growth in the United States: Decoupling the Demand for Services from Electricity Use.” *Environmental Research Letters* 13 (12): 124030. https://doi.org/10.1088/1748-9326/aaec9c.

Siddik, Md Abu Bakar, Arman Shehabi, and Landon Marston. 2021. “The Environmental Footprint of Data Centers in the United States.” *Environmental Research Letters* 16 (6): 064017. https://doi.org/10.1088/1748-9326/abfba1.

Swinhoe, Dan. 2024. “Amazon: All Our Operations Now Run on Renewable Energy.” Data Center Dynamics. July 11, 2024. https://www.datacenterdynamics.com/en/news/amazon-all-our-operations-now-run-on-renewable-energy/.

“The Dalles.” 2023. Wasco County Watersheds. 2023. https://www.wascowatersheds.org/the-dalles.

U.S. Census Bureau. 2023. “B01003: Total Population - Census Bureau Table.” https://data.census.gov/table/ACSDT1Y2023.B01003?q=B01003.

Walczak, Jared. 2024. “State and Local Sales Tax Rates, 2024.” Tax Foundation. February 6, 2024. https://taxfoundation.org/data/all/state/2024-sales-taxes/.

Weise, Karen, and Jovelle Tamayo. 2024. “A.I., the Electricians and the Boom Towns of Central Washington.” *The New York Times*, December 25, 2024, sec. Technology. https://www.nytimes.com/2024/12/25/technology/ai-data-centers-electricians.html.