Politecnico di Torino

POLITECNICO DI TORINO

Department of Management and Production Engineering

Master's Degree in Engineering and Management

Spatial Determinants of AI Data Center Location

A County-Level Econometric Analysis in the United States

Supervisor:
Prof. Francesco Nicoli

Candidate:
Matilde Bagnasco

Academic Year 2025–2026

Acknowledgements

I would like to sincerely thank Professor Francesco Nicoli for supervising this thesis and for his valuable guidance, feedback, and support throughout the development of this research.

I am also grateful to Professor Daniele Quercia and Giordano Paoletti for the opportunity to present and discuss this work within the course *Data Science and Machine Learning for Engineering Applications / Reservoir Geomechanics*, part of the Master's program in Georesources and Geoenergy Engineering. The possibility to share the dataset and empirical framework developed in this thesis, and to contribute to a related research project extending the analysis to the European context, represents a valuable opportunity to further develop and continue this line of research.

Vorrei ringraziare **mia mamma e mio papà** per l'immensa pazienza dimostrata durante questi anni. Mi hanno supportata in ogni momento, anche nei periodi di stress, cambiamenti d'umore improvvisi e momenti in cui probabilmente non ero la persona più facile da avere intorno. Li ringrazio anche per avermi mantenuta durante tutto questo percorso di studi... D'ora in poi, invece, dovrò continuare a provvedere alla mia sopravvivenza da sola — e so che non sono proprio così economica...

Speriamo anche che, dopo 5 anni, mio padre abbia finalmente imparato il nome del corso di studi che sto facendo... o almeno che adesso ci sia abbastanza vicino!

Ringrazio anche **mia sorella Nessie (Carlotta)**. Non siamo esattamente un esempio di armonia fraterna e probabilmente non lo saremo mai, ma i nostri litigi hanno sicuramente contribuito a formare il mio carattere. Dopotutto, alcune ricerche suggeriscono che il conflitto tra fratelli possa aiutare a sviluppare capacità di confronto e anche intelligenza sociale [1].

Partendo da chi è presente dagli anni della triennale, vorrei ringraziare **Ludo**, che c'è stata letteralmente dal giorno zero. Anche se nel tempo i nostri percorsi si

sono un po' divisi, sono davvero felice che sia ancora qui ora.

Un grazie enorme a **Lucrezia**, che è probabilmente una delle persone più piene di energia che conosca. Sempre sorridente, sempre spontanea, sempre con voglia di fare qualcosa di nuovo. La sua energia e il suo entusiasmo sono stati contagiosi in questi anni.

Ringrazio anche **Giulia S. e Giulia Ama.**, che ci sono state fin dall'inizio. Anche se con la magistrale ci siamo un po' disperse tra città diverse e percorsi diversi, siamo comunque rimaste presenti nella vita delle altre. Con loro porto dietro tantissimi ricordi: giornate infinite a studiare con la candela della sapienza, serate insieme, pet therapy e viaggi. In tutti questi anni sono state spesso la mia ancora, un punto fermo a cui fare riferimento.

Per gli anni della magistrale, invece, vorrei ringraziare i due **Leoni George e Matteo,** e **Marghe e Sara**. Tra lavori di gruppo, sessioni di studio e scambi continui di appunti (anche se spesso ero io la fornitrice ufficiale), ci siamo sostenuti a vicenda ininterrotamente e papà George ha fatto sì che non dimenticassimo nessuna scadenza... Non abbiamo solo studiato eh, ci siamo anche divertiti tremendamente. Un grazie anche a **Gigi e Lorenz(in)o**, per aver sopportato il mio essere semplicemente me stessa in questi anni: cambi di idea continui, momenti di caos e probabilmente anche un po' di sana follia.

Tra gli amici di lunga data, vorrei iniziare ringraziando **Chiara**, che è sempre stata super presente nella mia vita. Le voglio un bene dell'anima. Anche se negli anni siamo state spesso distanti, abbiamo sempre detto che un giorno saremmo tornate a vivere nella stessa città — chissà, magari succederà davvero. Abbiamo vissuto insieme mille avventure, ne abbiamo superate tante e, in qualche modo (non so bene come), ne siamo sempre uscite vive.

Un grazie enorme anche ad **Alessandra**. Ci conosciamo da, aiuto, circa 23 anni — praticamente da sempre. Senza di lei questo percorso avrebbe probabilmente avuto una sfumatura diversa. Ci sarebbero tantissime cose da dire sulla Q, ma la ringrazio soprattutto per il suo supporto costante e per il suo modo unico di ragionare, che spesso mi ha aiutata a vedere le cose da prospettive diverse e a riflettere meglio.

Ringrazio anche **Elena**, che in questi ultimi anni ha affrontato cambiamenti enormi, arrivando persino a reinventarsi completamente dal punto di vista lavorativo con la sua attività. È stata davvero forte e coraggiosa, e nonostante tutto è sempre

stata presente anche per me.

Un ringraziamento speciale va anche a **Nicolò**, probabilmente la persona più paziente tra tutti i miei amici. Ha ascoltato infinite volte gli stessi racconti, gli stessi dubbi, le stesse domande e i miei flussi di pensieri completamente caotici. Tra messaggi infiniti, momenti di overthinking e richieste di consigli a qualsiasi ora, ha sempre avuto la pazienza di esserci.

Grazie anche a **Simone e Carlo**, con cui ormai condivido un'amicizia che dura da una vita. Anche quando io ero in un'altra città, sapere che loro c'erano sempre per un'uscita, una chiacchierata o semplicemente per ritrovarsi ha fatto davvero la differenza. E un pensiero speciale a Simone per il gigantesco spavento che ci ha fatto prendere quest'anno: direi che può bastare così per un po'...

Infine, un ringraziamento speciale a **Catte**. Lo conosco da quando sono piccola ed è sempre stato una persona importantissima per me. Ha sempre saputo esserci in qualsiasi momento. Anche quando capita di sentirci meno, resta una di quelle persone su cui so di poter contare davvero: sia per farmi ridere con le sue battute completamente senza senso, sia per il doloroso dito della morte, sia per un discorso serio quando serve.

Un grazie anche a **Cristina e Noemi**, mie coinquiline dell'ultimo anno a Torino. Quest'ultimo anno è stato probabilmente uno dei più sereni vissuti in casa, e con loro sono stata davvero bene. Sono state delle scoperte formidabili: c'era sempre qualcosa che succedeva, ma in qualche modo riuscivamo sempre a risolvere tutto insieme e con tante risate. Hanno sempre saputo come tirare su il morale e inventarsi qualcosa di nuovo da fare, non ci si annoiava mai. Abbiamo condiviso davvero tanto e sono molto felice di averle incontrate e che siano ancora presenti nella mia vita.

A special thank you also to **Eliott, Mila and Andreea**, my flatmates during the six months I spent in Luxembourg. During that time we built a very close bond and shared an unforgettable experience together. They are extraordinary people, full of energy and always ready to make the most of every moment. Being far from home could have been much harder without them, but their presence made that adventure even more special. I truly hope to see them again very soon.

Un ringraziamento va anche ai miei vicini di casa e, più in generale, a tutte le persone che hanno fatto parte della mia **vita quotidiana a Torino**. In questi cinque anni mi avete fatta sentire davvero a casa, tra piccole abitudini, incontri

casuali e momenti condivisi. Anche se ero pronta a lasciare la città, salutare Torino non è stato così semplice, proprio perché lì mi sentivo ormai a casa — anche grazie a voi.

Un ultimo pensiero anche al **Politecnico**, che in questi anni mi ha insegnato molto, non solo accademicamente ma anche personalmente. Anche se nell'ultimo periodo ha portato con sé una buona dose di stress tra scadenze e procedure prima della laurea, chiudere questo capitolo fa comunque un certo effetto.

Infine, un ringraziamento lo devo anche a **me stessa**. Perché, nonostante tutto il supporto ricevuto lungo il percorso, una parte del lavoro l'ho dovuta fare anche io: tra momenti di dubbio, stanchezza, cambi di direzione e tentativi di capire cosa stessi davvero facendo.

Quindi mi faccio anche un personale in bocca al lupo per quello che verrà dopo. Perché se arrivare fin qui è stato impegnativo, capire il prossimo passo lo è forse ancora di più — soprattutto in un momento in cui trovare la propria strada non è così semplice.

What's next?

[1] Kramer, L. (2010). The essential role of sibling relationships in development. Annual Review of Psychology, 61, 611–636.

Abstract

The rapid expansion of artificial intelligence (AI) has dramatically increased global demand for data center infrastructure, turning these facilities into one of the most critical and energy-intensive pillars of the digital economy. Generative AI systems such as ChatGPT require a very large number of computational resources, including high-performance GPUs and advanced cooling systems, which significantly increase electricity consumption, water use, and overall environmental impact. Yet, despite the scale of this transformation, we still know little about how environmental and economic factors jointly influence where data centers are actually built. Understanding these dynamics is essential if we aim to develop AI in a way that is not only technologically advanced but also environmentally responsible. This thesis develops a quantitative, data-driven framework to analyze the determinants of data center placement across the United States. Using county-level data, the study integrates key variables such as electricity prices, water availability, GDP, and population to estimate their influence on the likelihood that the data center is present in a given area. Econometric modeling techniques are applied to identify the relative importance and interaction of these factors, providing empirical insight into how resource availability and regional characteristics shape the geography of AI infrastructure.

The study contributes to the emerging literature on data centers' infrastructure and sustainability by developing an empirical framework to analyze the economic and infrastructural determinants of data center location. By focusing on the U.S. county level, it provides a structured methodological approach that can be adapted to other geographic contexts, including Europe. In doing so, the research supports a deeper understanding of how economic conditions, resource availability, and infrastructure constraints interact in shaping the geography of digital infrastructure.

Contents

Chapter 1

Introduction

In recent years, the expansion of digital infrastructures has led to a significant increase in both size and number of data centers, which now constitute the backbone of modern digital services. These facilities support and enable a wide range of applications, including cloud computing, large-scale data storage, and artificial intelligence operations. However, their growing prevalence has raised increasing concerns about energy and water consumption, carbon emissions, and the overall sustainability of digital ecosystems. The proliferation of AI-driven applications, particularly large language models (LLMs) such as ChatGPT, has intensified these challenges by placing unprecedented computational and environmental pressure on global data center networks. Forecasts suggest that the energy demand for data centers will continue to increase sharply, underscoring the need for strategies that optimize their location and minimize their operational footprint. Unlike traditional cloud facilities, AI-oriented data centers must accommodate highly intensive training and inference workloads that depend on specialized hardware, including Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs). These accelerators enable efficient large-scale computation but substantially increase power use and operational costs. According to the International Energy Agency (IEA), data centers accounted for approximately 1-1.5% of global electricity demand in 2022, with AI workloads representing a substantial share of this consumption [1]. As AI adoption accelerates, this percentage is projected to rise further, making geographic optimization of data centers a key sustainability challenge. This study investigates the environmental and economic determinants that underlie the location of data centers. Although OpenAI does not operate its own dedicated data centers and instead relies on large cloud providers such as Microsoft Azure and, more recently, Amazon Web Services, the underlying infrastructure choices remain central to the environmental footprint of

AI and cloud systems. Understanding the factors that influence where data centers are built is therefore essential for assessing the sustainability of AI development. By combining county-level data on resource availability, infrastructure conditions, and economic characteristics, this study seeks to identify the determinants most strongly associated with data center presence and to draw implications for environmentally responsible infrastructure planning.

1.1 Research Problem and Objectives

The exponential growth of artificial intelligence (AI) and large-scale generative models such as ChatGPT has intensified the global demand for data center infrastructure. These facilities, which provide the physical backbone of AI operations, consume significant amounts of electricity and water since they require advanced cooling and high-performance networking systems. However, despite their environmental and economic relevance, the determinants driving the location of data centers remain underexplored in both academic and policy research.

This research aims to address the following central question:

> *How can the optimal placement of AI data center facilities be determined by jointly considering environmental sustainability and economic feasibility factors such as electricity pricing, water availability, renewable energy integration, cooling efficiency, network connectivity, and human capital?*

To answer this question, the study develops a data-driven empirical framework designed to identify key variables that influence the presence of data centers and to quantify their relative impact. The analysis focuses on the United States context, specifically at county-level. Using variables such as electricity prices, water availability, GDP, and population, the model employs econometric techniques to estimate how environmental and economic conditions interact to shape the spatial distribution of data centers. A binary variable indicates whether a county hosts at least one data center, enabling quantitative evaluation of the determinants behind localization patterns.

The research pursues three main objectives:

1. **Quantify** the relationship between environmental resource availability (electricity, water) and the likelihood of data center presence.

2. **Assess** how economic and demographic factors, such as GDP and population, mediate or reinforce environmental influences on site selection.

3. **Establish** a methodological foundation for extending the analysis to the European context in future research, comparing regional suitability for sustainable AI data center development and exploring potential optimization scenarios.

By combining environmental, infrastructural, and economic dimensions, this study contributes to a better understanding of the trade-offs underlying data center localization. The findings aim to support academic debate and offer empirical evidence that may be relevant for discussions on sustainable and strategically distributed data center networks.

1.2 Thesis Structure

The thesis is organized into four main parts.

The **first part** introduces the background and motivation of the study, highlighting the rapid expansion of artificial intelligence and the resulting growth in demand for data center infrastructure. It outlines the environmental implications associated with this development, particularly in terms of electricity consumption, water usage, and cooling requirements. This section also defines the research problem, presents the central research question, and situates the analysis within the broader debate on the sustainability of digital infrastructures.

The **second part** provides a review of the existing literature and establishes the theoretical framework of the study. It examines prior research on data center sustainability, infrastructure requirements, and site selection criteria, with particular attention to the emerging challenges associated with AI-driven computing systems. The section identifies key gaps in the empirical literature on data center location and uses these gaps to motivate the analytical approach adopted in the thesis.

The **third part** presents the research design and empirical methodology. It describes the construction of a county-level dataset for the United States and introduces the econometric framework used to analyze the determinants of data center presence. Through a quantitative analysis combining economic, demographic, and environmental variables, this section investigates the factors associated with the spatial distribution of data centers across U.S. counties. The methodological framework is also designed to be adaptable for future applications in other geographic contexts

Finally, the **fourth part** discusses the broader policy and regulatory context relevant to data center development, with particular reference to recent European initiatives concerning sustainability reporting and digital infrastructure governance.

The section reflects on how the empirical findings relate to ongoing policy discussions about the environmental impact and strategic planning of digital infrastructure.

The thesis concludes by summarizing the main findings of the empirical analysis and discussing their implications for the sustainable and geographically balanced development of data center infrastructure.

1.3 Background

Generative AI, particularly large language models such as GPT, has gained widespread popularity due to its ability to generate responses to a wide range of prompts quickly and with minimal user effort. While generative AI is often portrayed as broadly beneficial, its overall impact remains complex. Many organizations invest heavily in these technologies with high expectations, yet the outcomes do not always fully align with the initial optimism. At the same time, AI systems may generate unintended consequences, including the spread of misinformation and an increased reliance on highly energy-intensive digital infrastructures. A balanced perspective is therefore necessary to assess both the opportunities these technologies provide and the potential risks they introduce.

Although this issue has already been introduced in the previous section, it deserves further attention because it represents a central motivation of this research. The rapid expansion of AI applications has significantly increased the demand for data center infrastructure capable of supporting high-performance computing workloads. In practice, however, publicly available datasets rarely distinguish between facilities dedicated specifically to AI workloads and those supporting more general cloud computing services. For this reason, the empirical analysis conducted in this thesis focuses on data centers more broadly defined.

Nevertheless, the growing diffusion of AI workloads is progressively changing the technological profile of modern data centers. The deployment of high-performance computing hardware, such as GPUs and TPUs - essential for training and operating large language models (LLMs) - can substantially increase energy consumption and cooling requirements. These evolving technical characteristics represent an important shift in how data centers operate and scale, and they provide an important motivation for studying the environmental and economic factors associated with their geographic location.

In Europe, the situation is no less critical. Data centers consumed between 45 and 65 TWh in 2022, roughly 2% of total electricity usage across the region.

In countries like Ireland, the share reached up to 18% [2]. The rise of AI models further amplifies these trends. Research by Ren et al. estimates that each ChatGPT session could indirectly require up to 500 ml of water when including both direct and upstream cooling-related usage, yet water reporting remains sparse and largely unregulated [3].

Generative AI systems also present a distinct operational profile. Unlike many other digital services, large language models are continuously queried, making inference a significant and recurring source of energy consumption. In fact, inference emissions may exceed those associated with training over a model's lifetime [4].

Several mitigation strategies have emerged, including distributed training, advanced cooling systems, and power-capping techniques. However, despite the sustainability commitments announced by major cloud providers such as Microsoft, AWS, and Google, there is still no unified framework to enforce environmental best practices across data center infrastructures [5] [6] [7]. Many measures - such as zero-water cooling technologies, the use of recycled water, or on-site renewable energy generation - are implemented unevenly across regions and often remain voluntary.

The European Union has begun to respond to these challenges. Regulations such as the Corporate Sustainability Reporting Directive (CSRD) and the Energy Efficiency Directive (EED) call for more transparent reporting of key environmental indicators, including energy consumption and water withdrawals [8]. The forthcoming AI Act will also introduce environmental disclosure obligations for certain high-risk AI systems starting in 2025 [9]. However, enforcement remains uneven across Member States, with countries such as Germany and the Netherlands taking a more proactive role, while others - including Spain and Italy - are progressing more slowly.

A major obstacle to effective policy design is the limited availability of consistent and disaggregated data on AI-specific resource consumption. Differentiating between traditional cloud services and AI workloads remains challenging, as highlighted by the European Commission's Joint Research Centre. Moreover, widely used reporting frameworks, such as the Greenhouse Gas Protocol, do not yet fully capture the distinctive environmental footprint of rapidly evolving AI infrastructures [9].

Against this background, understanding the factors that influence where data centers are located has become increasingly important. Although the growing demand for generative AI workloads represents a key driver of infrastructure expansion, publicly available datasets rarely distinguish between AI-specific facilities and

general-purpose cloud data centers. For this reason, the empirical analysis conducted in this thesis focuses on data centers more broadly defined.

This thesis develops a data-driven framework to investigate the environmental and economic determinants associated with data center localization in the United States. By analyzing county-level variables, the study examines how factors such as electricity prices, water availability, and economic development interact in shaping the spatial distribution of data centers. The insights derived from the U.S. case provide an empirical foundation for future comparative research, which may extend the analysis to the European context to explore regional sustainability challenges and policy implications.

1.3.1 Evolution of OpenAI's Cloud Infrastructure

The rapid expansion of generative AI has significantly increased the demand for large-scale computing infrastructure. A prominent example of this trend is the evolution of OpenAI's cloud infrastructure, which illustrates the growing computational requirements associated with training and deploying advanced AI models.

Since its foundation, OpenAI has relied on large-scale cloud partnerships to support the intensive computational demands of model development and deployment. In 2020, Microsoft became its primary infrastructure partner through a multibillion-dollar agreement that positioned Azure as the main platform for training and serving models such as GPT-3, GPT-4, and DALL·E. As part of this collaboration, Microsoft developed specialized AI supercomputing clusters within several U.S. data centers, including facilities in Iowa, Texas, and Virginia. These infrastructures rely on large GPU clusters and advanced cooling systems designed to handle the high computational and energy demands associated with generative AI workloads [10] [11].

As the scale and usage of OpenAI's models continued to grow, the demand for computational capacity expanded accordingly. In late 2025, OpenAI announced a multi-year agreement worth approximately US$38 billion with Amazon Web Services (AWS) to secure additional computing resources, including large GPU clusters and AWS's proprietary Trainium and Inferentia processors [12] [13]. This development marked a transition from a predominantly single-provider infrastructure toward a more diversified multi-cloud strategy, aimed at expanding compute availability and increasing operational resilience.

More broadly, these developments reflect a structural transformation in the or-

ganization of AI infrastructure. The growing computational intensity of modern AI systems is increasing the scale, energy consumption, and technological complexity of data centers supporting these workloads. From a sustainability perspective, this evolution further highlights the importance of infrastructure planning, energy sourcing, and cooling technologies in shaping the environmental footprint of data center operations.

Chapter 2

Literature Review and Theoretical Framework

2.1 Introduction

Understanding the sustainability challenges associated with modern data centers requires an in-depth analysis of the existing literature on energy consumption, computational demands, and infrastructure development. Recent studies have documented the sharp rise in electricity usage associated with high-performance computing hardware such as GPUs and TPUs, as well as the growing water footprint linked to cooling systems that support large-scale AI workloads [3] [4]. Additional research highlights the rapid increase in overall data center energy demand across both global and European contexts [1] [2].

This literature review synthesizes academic research and industry reports to examine the environmental implications of the growing computational demand generated by artificial intelligence applications. Particular attention is given to data centers supporting large-scale AI workloads, including those associated with generative models such as ChatGPT. The objective is to provide a structured overview of the current state of knowledge and to identify the research gaps that motivate the empirical analysis developed in this thesis.

The review is organized into three main sections:

- **Environmental Impact and AI-Related Sustainability Challenges:** Empirical studies quantifying electricity consumption, water use, carbon emissions, and the resource requirements associated with AI training and inference workloads.

- **Optimization Strategies for Sustainable Data Centers:** Technological and operational solutions proposed to mitigate environmental impacts, including renewable energy integration, advanced cooling systems, and efficiency improvements reported by major cloud providers such as Amazon, Microsoft, and Google.
- **Site Selection Criteria and Analytical Approaches:** A review of geographic, economic, infrastructural, and environmental factors influencing data center location decisions, as well as analytical frameworks used in previous studies.

The chapter concludes by identifying the main gaps in the existing literature on data center location and sustainability. These gaps provide the foundation for the empirical investigation presented in the following chapters.

2.2 Environmental Impact and AI-Specific Sustainability Challenges

Major technology companies such as Microsoft, Meta, Amazon, and Google are rapidly expanding their computing capacity to support the growing demand for artificial intelligence applications. In 2024, Microsoft alone acquired approximately 485,000 NVIDIA Hopper chips, far exceeding Meta (224,000), Google (169,000), and Amazon (196,000). While these accelerators power advanced AI models and large-scale applications, companies are also accelerating the development of proprietary hardware to reduce reliance on external suppliers. Google has deployed more than 1.5 million Tensor Processing Units (TPUs), Meta has introduced its own AI accelerator, and Amazon continues to expand its Trainium and Inferentia processors. This surge in AI-related infrastructure investment contributes to record global server spending, reaching $229 billion in 2024, largely driven by the expansion of hyperscale data centers [14].

Generative AI is transforming many sectors of the digital economy, but its rapid diffusion also raises significant sustainability concerns. Training and operating large-scale models such as GPT-4 require substantial computational resources, increasing both energy consumption and cooling requirements within modern data centers [9]. As AI workloads become more widespread, the computational intensity of data center operations has grown considerably. AI computing clusters can consume seven to

eight times more electricity than traditional workloads, and their expansion is currently outpacing the growth of renewable energy supply. Consequently, many data centers continue to rely on electricity generated from fossil fuels, raising important environmental challenges [9].

According to the International Energy Agency, data centers accounted for approximately 1–1.5% of global electricity demand in 2022, with consumption potentially doubling and exceeding 1,000 TWh by 2026 - roughly equivalent to Japan's total electricity consumption [1]. Goldman Sachs Research similarly forecasts that global data center energy demand may increase by 50% by 2027 and by up to 165% by 2030 [15]. Global data center capacity currently stands at roughly 59 GW, about 60% of which is controlled by hyperscale operators and large wholesale providers. As demand for high-performance computing grows, capacity constraints are expected to tighten, with occupancy rates projected to rise from approximately 85% in 2023 to over 95% by 2026 before stabilizing as new facilities are deployed [15]. By 2030, global capacity may reach 122 GW, requiring substantial investments in electricity grid infrastructure estimated at around $720 billion [15].

In Europe, data centers consumed between 45 and 65 TWh of electricity in 2022, representing approximately 1.8–2.6% of total EU electricity demand [2]. This consumption is unevenly distributed across Member States. Germany, France, the Netherlands, and Ireland account for nearly two-thirds of total European data center electricity use despite representing less than 40% of the EU population. Ireland represents a particularly notable case, where data centers account for roughly 18% of national electricity consumption, significantly higher than in countries such as the Netherlands (5.2%), Luxembourg (4.8%), or Denmark (4.5%) [2]. EU-wide electricity consumption by data centers is expected to increase by approximately 30% by 2026, with Ireland and Denmark contributing a significant share of this growth [1].

A major challenge in evaluating the sustainability of AI systems lies in the limited availability of standardized data on energy consumption. Most companies do not clearly distinguish between electricity used for general cloud services and that attributable specifically to AI workloads. While early research focused primarily on the energy requirements of model training, more recent studies suggest that inference - the process of generating responses to user querie - may dominate long-term emissions. Each ChatGPT query, for example, is estimated to consume several times more electricity than a standard web search [9]. Generating a single AI-created image can consume as much electricity as fully charging a smartphone [16]. As a result,

the cumulative energy consumption associated with large-scale AI deployment may increasingly be driven by inference workloads rather than by the initial training process.

Recent research further highlights how energy consumption varies significantly across different types of AI tasks. Luccioni et al. (2024) analyzed 88 AI models and found that task type is a key determinant of energy use [4]. Text-to-image models are among the most energy-intensive applications, whereas simpler tasks such as text classification require substantially fewer computational resources. Differences in model architecture also affect efficiency. Masked Language Models (MLMs), such as BERT, typically exhibit lower emissions compared to decoder-only models such as GPT-style architectures, which may consume up to 30 times more energy for comparable tasks [4]. Output length also plays a role, as decoder-only models scale inefficiently when generating long responses.

Despite the growing relevance of AI workloads, the overall number of cloud data centers worldwide remains difficult to quantify precisely. Current estimates suggest that between 9,000 and 11,000 cloud data centers operate globally, yet only a limited number of companies disclose detailed information regarding the share of energy consumption attributable specifically to AI-related computing tasks. One of the few available disclosures comes from Google, which reported that machine learning workloads account for less than 15% of its total data center electricity consumption [17]. The absence of standardized reporting frameworks therefore makes it difficult to accurately assess the environmental footprint associated with the rapid expansion of AI systems.

The diffusion of large-scale generative AI applications further illustrates the scale of computational demand generated by modern AI systems. ChatGPT, for example, experienced extremely rapid adoption following its release in November 2022, reaching approximately 100 million users within two months and recording more than 1.7 billion visits by October 2023 [4]. At its peak, the system handled millions of daily users, generating a continuous stream of inference requests and associated computing workloads.

This rapid diffusion has important implications for energy consumption. Even under conservative assumptions, cumulative energy use associated with inference queries can surpass the energy required to train the underlying model within a relatively short period of time [4]. As a result, the environmental impact of AI systems increasingly depends not only on model training but also on the large-scale deployment and continuous use of AI applications.

Another important dimension of AI sustainability concerns water consumption. Cooling high-performance computing hardware requires substantial quantities of water, particularly in facilities that rely on evaporative cooling systems. Estimates suggest that approximately two liters of water may be required for every kilowatt-hour of electricity consumed for computing operations [9]. This can place additional pressure on local water resources, particularly in regions already affected by water scarcity [16]. Research by Shaolei Ren estimates that a single interaction with a model such as GPT-3 may indirectly consume around half a liter of fresh water when accounting for both direct and upstream cooling requirements [3].

Despite the increasing attention devoted to these issues, detailed reporting on water consumption related specifically to AI workloads remains limited. Many hyperscale data centers are located in regions where water resources are already under stress, and cooling operations may compete with local communities for access to freshwater. Even when alternative water sources such as recycled or desalinated water are used, these processes often require additional energy inputs, creating further environmental trade-offs [3]. For example, Microsoft's data center in Iowa, which supports large-scale AI training workloads, consumed approximately 11.5 million gallons of water in July 2022 alone, representing roughly 6% of the district's monthly consumption [18]. Over the same period, Microsoft's total water use increased by 34% between 2021 and 2022, while Google reported a 20% increase, partly associated with expanding AI-related operations particularly in Iowa and Las Vegas areas [18].

Beyond energy and water consumption, the rapid iteration cycle of AI models also raises concerns about hardware production and material use. The frequent release of increasingly large models requires continuous upgrades of specialized computing hardware, accelerating GPU manufacturing and encouraging further expansion of data center infrastructure. In line with the Jevons Paradox, improvements in computational efficiency may paradoxically increase total resource consumption as AI applications become more widely adopted [9].

More recently, the development of large-scale AI supercomputing clusters suggests a gradual shift toward increasingly specialized data center infrastructures. Projects such as Microsoft's planned "Stargate" infrastructure for OpenAI illustrate the emergence of hyperscale facilities designed specifically to support large GPU clusters, advanced cooling systems, and integrated renewable energy solutions [19] [20] [21].

Overall, these developments highlight how the rapid expansion of AI workloads

is reshaping the scale, energy intensity, and environmental footprint of modern data centers. As demand for high-performance computing continues to grow, understanding the factors that influence the geographic distribution of data center infrastructure becomes increasingly important.

2.3 Optimization Strategies for Sustainable Data Centers

In recent years, a wide range of strategies has been proposed to mitigate the environmental impact of data centers. These strategies focus primarily on improving energy efficiency, adopting renewable energy sources, optimizing cooling systems, and enhancing computational efficiency.

Reducing the environmental footprint of AI-driven data centers does not always require large capital investments. Even relatively small operational adjustments can lead to meaningful improvements in energy efficiency. Existing research suggests that straightforward optimization measures could reduce global data center electricity demand by approximately 10–20%, generating both environmental and economic benefits [16].

One of the most promising strategies is power capping, which limits the amount of power that processors and GPUs are allowed to consume. Instead of operating continuously at full capacity, systems can dynamically restrict power usage to more efficient levels. A study by Gadepally et al. from MIT Lincoln Laboratory shows that reducing GPU utilization to between 60% and 80% can decrease energy consumption by up to 23%, depending on the processor architecture. In addition to lowering energy demand, this approach reduces operational temperatures and improves overall system efficiency [15].

Another widely adopted technique is Dynamic Voltage and Frequency Scaling (DVFS), which dynamically adjusts processor voltage and clock frequency according to workload requirements. By scaling computational performance to actual demand, DVFS significantly improves energy efficiency. Empirical studies suggest that DVFS-based optimization can reduce energy consumption by up to 25% in certain computing environments [22].

Distributed training has also become essential for managing the growing computational requirements of modern artificial intelligence models. By distributing training workloads across multiple machines, techniques such as data parallelism,

model parallelism, and pipeline parallelism enable large-scale model training while improving computational efficiency and scalability [9].

However, large-scale AI workloads often generate uneven computational demand across servers, leading to inefficient resource utilization and energy spikes. Advanced workload scheduling algorithms can help address this problem by dynamically allocating tasks across distributed computing resources. Research indicates that AI-driven workload scheduling can reduce idle power consumption by up to 15%, improving both operational efficiency and environmental performance [22].

Data deduplication represents another important optimization strategy. By identifying and eliminating redundant data across storage systems, deduplication reduces both storage capacity requirements and associated energy consumption. Block-level deduplication techniques can lower storage-related energy use by up to 20% while maintaining data integrity and efficient retrieval performance [22].

Training deep learning models remains one of the most energy-intensive processes in AI development. To address this challenge, Gadepally and colleagues developed a training speed estimation tool capable of predicting a model's final performance after only 20% of computations. This allows researchers to terminate inefficient training runs early, reducing computational costs by up to 80% without compromising final model quality [16].

Another emerging approach for improving computational efficiency is the Mixture of Experts (MoE) architecture. In this framework, only a subset of specialized neural networks - referred to as experts - is activated for each task. By limiting the number of active parameters during computation, MoE architectures enable extremely large models to be deployed while keeping computational costs manageable [23]. Google's Switch Transformer represents a notable implementation of this approach.

Cooling systems remain one of the largest contributors to both energy consumption and water usage in data centers. To address these challenges, the European Union has proposed several policy and technological strategies aimed at improving cooling efficiency [8]:

- **Liquid Cooling:** Compared to traditional air cooling systems, liquid cooling technologies provide more efficient heat removal, reducing electricity consumption and extending hardware lifespan.

- **AI-Driven Cooling Systems:** Intelligent cooling management systems dynamically adjust cooling operations based on workload intensity and environ-

mental conditions. Companies such as Google report energy savings of up to 30–40% through these systems.

- **Strategic Location in Cold Climates:** Locating data centers in colder regions, such as Northern Europe, allows operators to benefit from natural cooling conditions and abundant renewable energy sources.
- **Water Use Reduction:** Recent EU initiatives promote improved transparency in water consumption and encourage the adoption of recycled water, dry cooling technologies, and immersion cooling systems, particularly in water-scarce regions.

Despite the availability of these technologies, many sustainability practices remain voluntary and are not yet widely implemented in the absence of binding regulatory frameworks [3].

Finally, the rapid hardware replacement cycles associated with AI systems contribute significantly to electronic waste (e-waste). Circular economy strategies - such as recycling and refurbishing obsolete GPUs and servers—can reduce hardware - related waste by up to 30% [22]. Emerging solutions, including blockchain-based hardware tracking systems, may further improve transparency and resource recovery within data center supply chains.

Overall, existing research highlights that optimizing AI deployment not only improves environmental sustainability but can also enhance economic performance. Energy-efficient computing strategies are therefore becoming increasingly important as the scale and computational intensity of AI systems continue to expand.

2.3.1 Real-World Sustainability Strategies in Data Centers: A Comparative Analysis of AWS, Microsoft, and Google

Major cloud service providers such as Amazon Web Services (AWS), Microsoft, and Google have developed comprehensive sustainability strategies to reduce the environmental footprint of their data center infrastructures. These strategies typically combine improvements in energy efficiency, large-scale renewable energy procurement, innovative cooling technologies, water conservation initiatives, and circular economy practices. Although each company adopts distinct technological and operational approaches, several common patterns emerge across their sustainability frameworks.

Energy efficiency represents one of the most critical dimensions of sustainable data center operations. All three companies have introduced significant hardware and operational optimizations to reduce energy consumption.

- **AWS** has developed proprietary processors such as Inferentia2 and Graviton4, which offer up to 50% higher energy efficiency compared to conventional processors. In addition, AWS has extended the operational lifespan of its servers from five to six years, reducing hardware replacement frequency and the associated environmental footprint [5].

- **Microsoft** has implemented low-power states for underutilized servers, reducing energy consumption by up to 25%. The company has also optimized CPU core allocation, decreasing overall hardware requirements by 1.5% and lowering power infrastructure needs by approximately 7%. As a result, Microsoft's data centers achieved an average Power Usage Effectiveness (PUE) of 1.12 in 2023, indicating significant efficiency improvements [6].

- **Google** remains a global leader in power efficiency, with data centers estimated to be approximately 1.8 times more efficient than the industry average. Google reports an average PUE of around 1.10. Furthermore, its latest Tensor Processing Unit (TPU), known as Trillium, is approximately 67% more energy-efficient than previous generations, substantially improving the sustainability of AI workloads [7].

In parallel, the transition toward carbon-free energy sources has become a central pillar of data center sustainability strategies.

- **AWS** achieved its target of sourcing 100% renewable energy in 2023, seven years ahead of its initial schedule. The company has also become the largest corporate purchaser of renewable energy worldwide [5].

- **Microsoft** has made large investments in renewable energy sources including solar, wind, and nuclear power, with the objective of achieving carbon-negative operations by 2030. Additionally, the company has introduced low-carbon construction materials in its data center infrastructure, such as alternative concrete mixtures capable of reducing embodied carbon emissions by up to 65% [6].

- **Google** has adopted an even more ambitious objective by targeting 24/7 carbon-free energy (CFE) by 2030. This approach aims to ensure that every unit of electricity consumed by Google's data centers is matched in real time by carbon-free energy generation rather than through offset mechanisms. In 2023 alone, Google signed contracts for approximately 4 GW of clean energy capacity across regions including Texas, Belgium, and Australia. The company also employs AI-driven demand response systems to align computing workloads with periods of high renewable energy availability [7].

Cooling systems represent another critical component of data center sustainability due to their substantial energy and water requirements. Each provider has implemented different strategies to improve cooling efficiency and reduce water consumption.

- **AWS** has improved its Water Usage Effectiveness (WUE) to approximately 0.18 L/kWh, representing a 28% reduction since 2021. The company increasingly relies on recycled water and rainwater harvesting for cooling operations. In addition, AWS reuses cooling water for agricultural irrigation and supported 15 water restoration projects across 10 countries, returning approximately 3.5 billion liters of water to the environment in 2023 [5].
- **Microsoft** has introduced zero-water cooling technologies for certain AI workloads and expanded rainwater harvesting initiatives across several European facilities [6].
- **Google** has committed to becoming water-positive by 2030, aiming to replenish more water than it consumes. In 2023 the company replenished approximately one billion gallons of water, equivalent to 18% of its freshwater consumption [7].

Beyond operational energy and water consumption, cloud providers are also addressing the environmental impact associated with hardware manufacturing, transportation, and electronic waste.

- **AWS** has reduced transportation emissions by shifting from air freight to ocean freight logistics, avoiding approximately 65,000 metric tons of CO_2 emissions in 2023. The company also promotes component reuse and hardware refurbishment to extend equipment lifecycle [5].

- **Microsoft** reports a recycling rate of approximately 89.4% for its cloud hardware. The company employs an AI-based Intelligent Disposition and Routing System (IDARS) to optimize server reuse and refurbishment processes [6].
- **Google** has implemented a zero-waste-to-landfill initiative, with approximately 29% of its data centers already meeting this target. Extensive server refurbishment and reuse programs further contribute to minimizing electronic waste [7].

Finally, Google has increasingly integrated artificial intelligence directly into its sustainability management systems.

- **AI-powered energy optimization**: Machine learning models are used to predict energy demand and dynamically optimize power allocation across data center infrastructure.
- **Carbon-Intelligent Computing**: Workloads are shifted temporally and geographically toward locations and time periods with higher availability of carbon-free energy.
- **AI-driven cooling management**: Real-time monitoring and predictive control systems optimize cooling operations, significantly reducing excess energy and water consumption [7].

These industry practices illustrate how energy costs, cooling efficiency, and resource availability have become central factors in data center operations. These same factors are reflected in the empirical analysis of this thesis, which examines how electricity prices, water availability, and regional economic characteristics influence the geographic distribution of data centers.

2.4 Site Selection Criteria and Optimization Models

Choosing the optimal location for data centers involves multiple interdependent factors. As AI-specific infrastructure intensifies resource consumption, selecting suitable sites becomes critical for achieving environmental, operational, and economic efficiency. This section summarizes key site selection criteria drawn from academic studies and industry reports [2].

Energy Availability and Cost Energy is the largest operational cost for data centers, and research consistently shows that data centers are often located in regions with low electricity prices and stable power grids [2]. Proximity to renewable energy sources, such as hydropower and wind farms, is also increasingly important for sustainability, especially to meet the high and constant power demands of AI workloads. Some regions promote on-site renewable generation, while others offer Power Purchase Agreements (PPAs) for long-term clean energy contracts [2].

Water Availability Cooling systems, particularly those used in AI data centers, require substantial water. Therefore, access to non-potable or recycled water is a growing consideration in site evaluation. It is recommended to avoid drought-prone areas and to choose locations with sufficient water rights or reclaimed sources [2].

Climate Conditions Natural climate significantly impacts cooling efficiency. Cooler regions, such as Northern Europe or the Nordic countries, offer free-air cooling options for most of the year, reducing the need for water-intensive and energy-heavy artificial cooling [24]. Studies show that building in colder climates can reduce data center energy use by 20-50% [2].

Connectivity and Network Infrastructure High-speed fiber-optic connections and proximity to major network exchange points are critical for minimizing latency, particularly for AI applications requiring real-time interaction. Locations with redundant network routes and strong peering infrastructure are preferred. Locations with dense fiber-optic infrastructure, such as Northern Virginia (known as "Data Center Alley"), offer exceptional connectivity, making them highly attractive. Similarly, the 2Africa subsea cable has revolutionized connectivity across Africa, positioning the region as an emerging data center hub [24].

Human Capital and Technical Expertise The availability of skilled labor, especially in computer science, electrical engineering, and data center operations, is another key factor. Regions hosting technical universities, AI research centers, or existing digital infrastructure clusters are often prioritized for investment.

Land Availability and Local Regulations Land cost and zoning policies can either support or hinder data center development. Regions with streamlined construction permits, tax incentives, or support for sustainable building practices are

increasingly attractive. On the other hand, areas with moratoriums or strict environmental regulations may discourage large-scale deployment.

Disaster Resilience Locations with low natural disaster risk are preferred to ensure uninterrupted operations. For example, Dallas, Texas, has become popular due to its minimal risk of earthquakes and hurricanes, along with its robust infrastructure [24].

2.4.1 Global Data Center Hubs and Emerging Markets

The highest concentrations of data center power capacity today are found in the following regions:

- **Asia-Pacific:** Beijing, Shanghai
- **North America:** Northern Virginia, San Francisco Bay Area

These regions have high compute demand, large-scale corporate investments, and strong infrastructure for AI development [15].

North America: The World's Largest Data Center Hub Northern Virginia remains the most significant hub due to its high connectivity and tax incentives. Dallas, Texas, is rapidly growing due to its low disaster risk, renewable energy access, and central location. Chicago, Illinois, houses over 70 data centers, benefiting from low latency, strong fiber-optic networks, and reliable power supply. Silicon Valley, California, remains a premium location, though its high earthquake risk and operational costs pose challenges [24].

Europe: The FLAP-D Region (Frankfurt, London, Amsterdam, Paris, Dublin) The FLAP-D region dominates the European market, with global giants like Equinix and AWS leading investments. Madrid is emerging as a key hub due to its connectivity to Latin America and North Africa. Nordic countries, known for sustainable energy sources and naturally cool climates, are also becoming preferred destinations [24].

Middle East & Africa: Emerging Growth Markets Nairobi, Kenya, is quickly becoming Africa's new tech hub, with companies like Digital Realty expanding operations. Dubai, UAE, is solidifying its status as a data center leader in the Middle East, attracting major investments from Microsoft and Oracle [24].

Asia-Pacific: Rapidly Expanding Market Singapore remains the dominant hub, but land scarcity and regulatory limits are shifting investments to Johor, Malaysia, and Batam, Indonesia. India's digital boom is driving massive investments in data centers, catering to its rapidly expanding online economy [24].

South America: Untapped Potential Brazil, Chile, and Argentina are witnessing increasing investments, with São Paulo and Rio de Janeiro leading South America's cloud market growth [24].

Chapter 3

Theoretical Framework and Research Gaps

3.1 Theoretical Framework

The location of large-scale data center infrastructure is influenced by a complex combination of economic, environmental, and infrastructural factors. In recent years, the rapid expansion of artificial intelligence (AI) applications has significantly increased the computational intensity of digital infrastructure, making the location of data centers an increasingly strategic decision.

Existing studies show that data centers tend to concentrate in regions offering favorable conditions for long-term operational efficiency, including access to reliable electricity, suitable climate conditions, and strong network connectivity. However, the growing importance of AI workloads introduces additional sustainability challenges, as these infrastructures require extremely high levels of energy consumption, cooling capacity, and computational resources.

Within this context, sustainability considerations increasingly overlap with economic objectives. Locations that provide access to renewable energy sources, sufficient water resources, and cooler climates can simultaneously reduce operational costs and environmental impacts. As a result, sustainability is not only an environmental objective but also a key determinant of operational efficiency and long-term infrastructure resilience.

Main Hypothesis

This study hypothesizes that data centers supporting AI-related workloads are more likely to be located in regions that offer a favorable combination of economic,

infrastructural, and environmental conditions. In particular, the literature suggests that the following factors play a central role in site selection decisions:

- **Electricity prices and energy availability**, given the extremely high and continuous energy demand associated with large-scale computing infrastructure.
- **Water availability and cooling capacity**, which are essential for maintaining hardware performance and preventing overheating.
- **Renewable energy availability and climate conditions**, which can reduce both operating costs and environmental impact through improved cooling efficiency and lower carbon emissions.
- **Network connectivity and digital infrastructure**, including high-capacity fiber networks and proximity to major internet exchange points.
- **Human capital and technical expertise**, particularly in regions hosting strong technology clusters or research institutions.
- **Regulatory and policy environment**, including permitting processes, taxation policies, and government incentives that influence infrastructure investment decisions.

Building on these theoretical considerations, the empirical analysis of this thesis examines how key economic and infrastructural variables influence the geographic distribution of data centers across U.S. counties.

3.2 Research Gaps and Contribution

Although a growing body of literature examines the environmental impact of data centers, relatively few studies investigate the sustainability challenges associated with AI-driven digital infrastructure. Most existing research focuses on the energy consumption and water usage of traditional data centers, without fully considering the increasing computational intensity and infrastructure requirements associated with modern AI workloads.

A first research gap therefore concerns the determinants of data center location in the context of increasingly AI-intensive computing infrastructure. While several studies highlight the importance of electricity prices, climate conditions, and

connectivity for data center placement, relatively little empirical research examines how economic, environmental, and infrastructural factors jointly influence location decisions across different regions of the United States.

In particular, limited attention has been given to the interaction between economic development, infrastructure availability, and environmental constraints in shaping the spatial distribution of data centers. Existing studies often focus on individual determinants - such as electricity prices or water consumption - without providing a comprehensive empirical framework that considers multiple structural factors simultaneously.

Moreover, most empirical analyses are conducted at relatively aggregated geographic levels, such as countries or large regions. This limits the ability to capture local heterogeneity in economic conditions, infrastructure availability, and environmental constraints that may influence infrastructure investment decisions.

This thesis contributes to the literature by developing a quantitative empirical framework to analyze the determinants of data center location across U.S. counties. Using county-level data, the study examines how structural variables - including electricity prices, water availability, population size, and economic development - are associated with the probability that a county hosts data center infrastructure.

By focusing on a fine geographic scale and combining economic and infrastructural variables within a unified empirical framework, this study provides new evidence on the factors shaping the spatial distribution of large-scale digital infrastructure in the United States. While the analysis does not aim to develop an optimization model for data center placement, it establishes a methodological foundation that future research may extend to evaluate alternative infrastructure deployment strategies in other geographic contexts, including Europe.

Chapter 4

Methodology

4.1 Dataset Construction

This study constructs a multi-variable dataset at the U.S. county level to investigate the environmental and economic determinants driving the localization of data centers. The dataset integrates variables considered relevant for hosting high-energy-intensity infrastructures, including:

- **Electricity costs** (commercial and industrial rates), representing key operating expenses;
- **Water Public Supply**, used as a proxy for water availability, cooling capacity and long-term sustainability;
- **GDP**, capturing the economic development of each county;
- **Population**, included as a control for county size, infrastructure availability, and potential demand;
- **Data center presence**, a binary variable indicating whether a county currently hosts at least one facility.
- **Data center number**, a variable indicating how many data centers a county currently hosts.

Constructing this dataset required substantial effort due to the limited transparency and uneven availability of county-level information in the United States. Many relevant indicators, such as electricity tariffs, water withdrawals, and economic data, are reported at the state or ZIP-code level rather than directly at the county level, making geographic harmonization a necessary yet non-trivial task.

The resulting dataset provides a geographically consistent and analytically robust representation of key determinants across U.S. counties, forming a solid empirical foundation for the statistical analysis. The following sections outline the data sources, transformation procedures, and methodological choices applied to each variable.

4.2 Electricity Cost Dataset Construction

A central component of this study is the construction of county-level indicators for electricity costs. Because electricity prices are a key determinant of data centre operating expenses, and potentially of their location choices, it is essential to build a measure that is both geographically consistent and transparent in terms of data limitations and imputation choices.

The starting point is the set of ZIP-code-level electricity pricing files for 2023 published by the National Renewable Energy Laboratory (NREL) through the OpenEI platform [25]. These datasets report tariffs charged by investor-owned utilities (IOUs) and non-IOU providers, disaggregated by sector (residential, commercial, industrial) and linked to ZIP codes. Since the unit of analysis in this thesis is the county, a ZIP-to-county harmonisation step is required. This is performed using the ZIP Code Crosswalk file provided by the U.S. Department of Housing and Urban Development (HUD) [26], which specifies, for each ZIP code, the share of addresses belonging to each county. The `TOT_RATIO` field in this crosswalk is later used as a weight to aggregate ZIP-level tariffs into county-level averages.

Data Integration and Record Classification

Two separate files are used for electricity prices: one containing IOU tariffs and one containing non-IOU tariffs. Each record includes a ZIP code, a commercial tariff (`comm_rate`), an industrial tariff (`ind_rate`), and the ZIP-to-county weight `TOT_RATIO` (after merging with the HUD crosswalk).

These datasets are first combined into a single ZIP–county file. A variable `source` distinguishes IOU from non-IOU providers. To assess data quality and resolve overlaps, IOU records are then classified according to the availability of tariffs:

- **valid IOU** (`validIOU`): at least one of `comm_rate` or `ind_rate` is strictly greater than zero and not missing;
- **invalid IOU** (`otherIOU`): both `comm_rate` and `ind_rate` are missing;

- **non-IOU**: all remaining records, belonging to municipal utilities, cooperatives or other non-IOU providers.

Because more than 70% of U.S. electricity customers are served by IOUs and their tariffs are reported through standardised regulatory channels (such as FERC Form 1 and EIA Form 861) [27], IOU data are taken as the primary benchmark for price information. Non-IOU data, while valuable, tend to be less systematically reported and are therefore used as a secondary source.

On this basis, each ZIP–county record is assigned a priority level:

- **Priority 1**: valid IOU entries (at least one positive tariff);
- **Priority 2**: non-IOU entries;
- **Priority 3**: invalid IOU entries with no usable tariff information.

A unique key combining ZIP code and county FIPS is used to sort records by priority (from 1 to 3). Whenever multiple records refer to the same ZIP–county pair, only the highest-priority entry is retained. Priority-3 records are not used to construct prices, but - as discussed below - they are kept to diagnose where electricity data are structurally missing.

Spatial Imputation of Electricity Tariffs

Even after resolving overlaps and excluding priority-3 records from price construction, the ZIP–county dataset still contains missing tariffs, especially for the industrial sector. Commercial tariffs are relatively well reported, but industrial rates display substantial gaps due to limited industrial activity or incomplete reporting by utilities. To avoid dropping ZIP codes and to maintain broad coverage, a two-step spatial imputation strategy is implemented: first within counties, then, in a very small number of cases (19 counties), across neighbouring counties.

Within-county spatial imputation For each ZIP–county observation, missingness is identified using the indicators `comm_missing` and `ind_missing`. County-level averages are then computed separately for commercial and industrial tariffs using only non-missing values:

- **avg_comm_county**: mean of `comm_rate` among ZIP codes in the county with observed commercial tariffs;

- **avg_ind_county**: mean of `ind_rate` among ZIP codes in the county with observed industrial tariffs.

No restriction is imposed on the availability of the other tariff type: for example, industrial tariffs are not required when computing `avg_comm_county`. This ensures that each average uses the maximum amount of relevant information.

Using these averages, two spatially completed ZIP-level series are constructed:

- **comm_rate_spatial**: equal to the observed `comm_rate` if available, and to `avg_comm_county` otherwise;
- **ind_rate_spatial**: equal to the observed `ind_rate` if available, and to `avg_ind_county` otherwise.

These variables preserve genuine within-county variation whenever tariffs are reported and use the county average only as a fallback for missing ZIP codes. Two indicators, `comm_imputed_spatial` and `ind_imputed_spatial`, flag cases in which the county average was used in place of a missing ZIP-level tariff. In later robustness checks, these indicators help assess whether counties with a higher reliance on imputed values systematically influence regression estimates.

Across-county imputation for counties with no industrial data A more severe form of missingness affects a small group of counties (specifically 19 counties) for which no industrial tariff is reported at any ZIP code. In these cases, `avg_ind_county` cannot be computed and within-county imputation is not feasible. Rather than excluding these counties entirely, an across-county spatial procedure is applied.

For each affected county, a set of neighbouring counties is identified - typically within the same state and directly adjacent. The industrial tariff for the missing county is then defined as the average industrial price observed in its neighbours:

$$\texttt{avg\_ind\_neighbor} = \text{mean}\left\{\texttt{avg\_ind\_county} \text{ for neighbouring counties}\right\}.$$

A dummy variable `ind_neighbor_imputed` is set to 1 for these counties and 0 otherwise. This indicator is later used to re-estimate the model excluding counties whose industrial tariffs are inferred from neighbouring areas, providing an additional robustness check.

Priority-3 Observations as Data Quality Indicators

Although priority-3 ZIP–county records (those with neither commercial nor industrial tariffs reported) are excluded from the construction of electricity prices, they are informative about the completeness of the underlying dataset. They do not merely correspond to a different class of provider: they signal cases where no tariff information is available at all, regardless of whether the utility is IOU or non-IOU.

To capture this dimension, I compute for each county c the share of ZIP–county records that are classified as priority 3:

$$\texttt{share\_p3}_c = \frac{\text{number of priority3 ZIP-county observations in county } c}{\text{total number of ZIP-county observations in county } c}.$$

This measure summarises heterogeneity in data completeness across counties. A diagnostic dummy `high_p3` is defined as equal to 1 when $\texttt{share\_p3}_c > 0.60$, identifying areas where more than 60% of ZIP-level entries lack any electricity information. Although priority-3 observations never enter directly into the price calculations, `share_p3` and `high_p3` are merged into the final county-level dataset and used in robustness checks - for example, by excluding `high_p3` counties.

A special case is represented by the 22 counties with $\texttt{share\_p3}_c = 1$, for which all ZIP–county observations are priority 3 and no tariff information exists at any level. These counties do not appear in the intermediate electricity dataset based on priority-1 and priority-2 records and would be dropped when merging electricity prices with water, GDP, population and data centre presence. To preserve geographical coverage while avoiding arbitrary assumptions, electricity prices for these counties are estimated using an across-county procedure similar to that described above: for each county, a set of neighbouring counties within the same state is selected, and its commercial and industrial tariffs are set equal to the average tariffs of those neighbours.

The resulting observations are appended to the county-level electricity dataset. Because their electricity data are entirely inferred from neighbouring counties, all related imputation indicators are set to 1 (`county_comm_imputed`, `county_ind_imputed`). The main regressions are also estimated excluding the counties to verify that results are not driven by these approximations.

County-Level Aggregation

After imputing missing ZIP-level tariffs, commercial and industrial prices are aggregated to the county level using the ZIP-to-county weights from the HUD crosswalk.

For each ZIP–county observation, two weighted variables are defined:

$$\texttt{comm\_rate\_weighted} = \texttt{comm\_rate\_spatial} \times \texttt{TOT\_RATIO}$$

$$\texttt{ind\_rate\_weighted} = \texttt{ind\_rate\_spatial} \times \texttt{TOT\_RATIO}$$

For each county, I then compute:

- the sum of weighted commercial tariffs, $\sum \texttt{comm\_rate\_weighted}$;
- the sum of weighted industrial tariffs, $\sum \texttt{ind\_rate\_weighted}$;
- the sum of ZIP–county weights, $\sum \texttt{TOT\_RATIO}$.

The county-level average commercial and industrial electricity prices are obtained as weighted averages:

$$\texttt{avg\_comm\_rate\_county}_c = \frac{\sum_{z \in c} \texttt{comm\_rate\_spatial}_{zc} \cdot \texttt{TOT\_RATIO}_{zc}}{\sum_{z \in c} \texttt{TOT\_RATIO}_{zc}},$$

$$\texttt{avg\_ind\_rate\_county}_c = \frac{\sum_{z \in c} \texttt{ind\_rate\_spatial}_{zc} \cdot \texttt{TOT\_RATIO}_{zc}}{\sum_{z \in c} \texttt{TOT\_RATIO}_{zc}}.$$

This procedure ensures internal consistency with the underlying geography: ZIP codes that cover only a small portion of a county contribute proportionally less to its average tariff. After this aggregation, no anomalous or extreme values were detected, suggesting that the combination of prioritisation, spatial imputation and crosswalk-based weighting yields a robust set of county-level electricity cost indicators for the empirical analysis.

4.3 Water Supply Withdrawal Dataset Construction

To complement the electricity cost data with an environmental sustainability indicator, water availability was incorporated into the empirical dataset. Water resources play a critical role in data centre operations, particularly for cooling high-performance computing systems, and are therefore an essential dimension of the analysis.

County-level data on public-supply water withdrawals were obtained from the U.S. Geological Survey (USGS) dataset *Estimated Use of Water in the United States: County-Level Data for 2015* [28]. The dataset reports, for each county, the volume

of water withdrawn by public-supply systems, expressed in million gallons per day (MGD), and includes standard FIPS codes for geographic identification.

From this source, I construct the variable `water_PublicSupply_mgd`, which captures the absolute volume of water withdrawn for public supply in each county. This measure does not represent total natural water resources, but it provides a reasonable proxy for the operational water that is accessible through utility networks—precisely the type of resource that large data centres rely on in practice.

The electricity dataset used in this study is constructed from ZIP–level tariffs and requires a ZIP–to–county crosswalk provided by HUD, which applies only to the 50 U.S. states and the District of Columbia. Because Puerto Rico and the U.S. Virgin Islands do not participate in this ZIP-based geographic system in a way that allows county-level electricity prices to be computed consistently, they cannot be integrated using the same procedure and so they are excluded.

4.3.1 Handling Missing Water Withdrawal Data

When merging the USGS water withdrawal data with the electricity pricing dataset, several inconsistencies emerged. A small subset of counties appeared in the electricity dataset but had no matching observation in the USGS public-supply water withdrawal file. To avoid dropping these counties and to maintain geographical coverage, a set of harmonisation steps was carried out.

Connecticut: transition from counties to planning regions A major source of mismatch concerned the State of Connecticut. Beginning in 2022–2023, the U.S. Census Bureau officially replaced traditional counties with *Planning Regions* as the new county-equivalent geographic units. As a result, the USGS water dataset (which still reports data using the former counties) did not align with the electricity dataset (which already adopts the new Planning Regions) [29] [30].

To reconcile the two sources, I manually reconstructed the correspondence between the former counties (e.g. Fairfield County, Hartford County) and the new Planning Regions (e.g. Western Connecticut, Capitol, Northwest Hills) using the Federal Register notice and the associated geographic tables and maps. Each new Planning Region was assigned the water withdrawal value of its predecessor county. This procedure removed all cases of missing water data for Connecticut (Appendix .1).

Alaska: split of Valdez–Cordova into two new census areas A second source of inconsistency arose in Alaska. In 2019, the former Valdez–Cordova Census Area (FIPS 02261) was split into two new county-equivalent units [31]:

- Chugach Census Area (FIPS 02063),
- Copper River Census Area (FIPS 02066).

The USGS water dataset still reports withdrawals for the pre-2019 area (02261), whereas the electricity dataset uses the two new post-2019 census areas. To align the two, the water withdrawal value for 02261 was proportionally redistributed across 02063 and 02066 using their relative population:

$$\text{water}_{02063} = \text{water}_{02261} \times \frac{\text{population}_{02063}}{\text{population}_{02063} + \text{population}_{02066}},$$

$$\text{water}_{02066} = \text{water}_{02261} \times \frac{\text{population}_{02066}}{\text{population}_{02063} + \text{population}_{02066}}.$$

This population-weighted allocation preserves the total water withdrawal for the former Valdez–Cordova area while ensuring full consistency with the county definitions used in the electricity dataset.

Summary Through the combination of manual county-to-region remapping in Connecticut and population-based redistribution in Alaska, all cases of missing water withdrawal data for counties with available electricity information were resolved. This harmonisation step guarantees full geographical alignment between the electricity and water datasets and allows the empirical analysis to be conducted on a consistent set of county-level observations.

4.3.2 Limitation: Water Availability vs. Water Withdrawals

A methodological limitation concerns the interpretation of the water variable used in this study. The USGS dataset provides estimates of *public-supply water withdrawals*, which measure the volume of water actually withdrawn by utilities for domestic, commercial and industrial uses. In the empirical model, this variable is used as a proxy for water availability, but it does not capture the full hydrological capacity or the total natural water resources of each county.

In other words, water withdrawals reflect *current usage* rather than *potential availability*. This implies that densely populated counties with high demand may

appear to have "more water" simply because they withdraw more, whereas sparsely populated or rural counties may have abundant natural resources but low withdrawal levels.

This distinction matters for data centre siting, where long-term access to cooling water is more important than present consumption volumes. Public-supply water withdrawals are strongly correlated with population levels ($\rho = 0.81$), indicating that this variable partly reflects population-driven infrastructure demand. Accordingly, the water coefficient is interpreted with caution and population size is controlled for throughout the analysis. Interaction terms between population and public water supply are also explored and found to be statistically insignificant, suggesting that water infrastructure plays a broadly similar enabling role across counties of different sizes

This limitation does not invalidate the use of the USGS variable, but it requires cautious interpretation of the results. Future research could improve upon this approach by incorporating alternative hydrological indicators that more directly capture local water availability.

4.4 Population per County Dataset Construction

To incorporate demographic factors into the empirical analysis, county-level population estimates for the year 2023 were added to the dataset. These estimates are produced annually by the U.S. Census Bureau using the standard demographic balancing equation [32]:

$$\text{Population Estimate} = \text{Population Base} + \text{Births} - \text{Deaths} + \text{Net Migration}.$$

The dataset employed in this study corresponds to the resident population as of July 1, 2023 and was taken from the file `co-est2023-alldata.csv`. This release provides the most recent and methodologically consistent population series available, relying on intercensal demographic components compiled by the Census Bureau.

To ensure complete geographic compatibility with the electricity and water datasets, each observation was matched using a unified FIPS identifier constructed by concatenating the two-digit state FIPS code with the three-digit county FIPS code. This harmonisation step guarantees that all variables used in the empirical model refer to the same geographic units.

Consistency Checks Across Datasets A set of cross-checks was performed to verify alignment between the population dataset and the other county-level datasets. The checks confirmed the following:

- **Population vs. Water Dataset:** All counties appearing in the USGS water withdrawal dataset also appeared in the population dataset. Counties present in the water dataset but not in the population file were limited to Puerto Rico and the U.S. Virgin Islands - territories already excluded from the analysis due to missing electricity information. No additional inconsistencies were found.
- **Population vs. Electricity Dataset:** All counties with electricity information were also present in the population dataset, ensuring that no demographic data were missing for counties included in the electricity analysis.

Summary Following these consistency checks, the population dataset was fully aligned with both the electricity and water datasets, with discrepancies arising only in cases already excluded. The final integrated dataset therefore provides a complete and coherent set of demographic, environmental and economic variables for all U.S. counties included in the empirical analysis.

4.5 Datacenter Presence Dataset Construction

To complement the environmental and economic variables with a measure of digital infrastructure, a county-level indicator for data centre presence was constructed. Since no standardized national dataset provides county-level inventories of data centres, a multi-step procedure combining manual data collection and geographic matching was required.

Data Collection Data centres were identified through a manual search using the platform *datacentermap.com* [33], which offers updated listings and geolocations of operational facilities across the United States. Because the website does not report county identifiers, the names of all cities hosting at least one data center was extracted.

City-to-County Matching Each identified city was then matched to its corresponding county using the *SimpleMaps US Cities Database* [34], which provides standardized information on U.S. cities, including county names, state identifiers,

and county FIPS codes. This step enabled the conversion of city-level observations into county-level indicators suitable for merging with the rest of the dataset.

Variable Construction Based on the matched dataset, two county-level variables were created:

- `datacenter_presence`: a binary variable equal to 1 if at least one data center is located in the county, and 0 otherwise;
- `datacenter_count`: the number of identified data center within the county.

Both variables were merged into the main analytical dataset using county FIPS codes to ensure full geographic consistency.

Note on Data Stability Because the number and spatial distribution of data centres evolve rapidly - particularly with the recent expansion of cloud and AI-related infrastructure - the counts obtained represent a snapshot at the time of collection. Updated inventories may differ, but this study retains the original dataset to preserve internal consistency and avoid introducing time inconsistencies into the empirical model.

4.6 GDP Variable Dataset Construction

County-level economic activity was incorporated through measures of Gross Domestic Product (GDP), using the *Gross Domestic Product by County, 2023* release published by the U.S. Bureau of Economic Analysis (BEA) [35]. GDP values are expressed in thousands of chained 2017 dollars, ensuring real (inflation-adjusted) comparability across counties (Appendix A).

The BEA dataset required several harmonisation steps before integration into the main dataset, as it does not include FIPS codes and reports GDP for some combined or legacy geographic units that differ from current county boundaries.

Connecticut: Transition to Planning Regions The BEA dataset reports GDP for the former eight Connecticut counties, which ceased functioning as county-equivalent units when the Census Bureau replaced them with *Planning Regions* in 2022–2023. Following the procedure applied to the water dataset, the GDP values associated with former counties were reassigned to their corresponding Planning Regions. The mapping was derived from Federal Register documentation and official

geographic correspondence tables, ensuring alignment with the geography used in the electricity and population datasets (Appendix .1).

Hawaii: Maui and Kalawao Counties In the BEA dataset, Maui County and Kalawao County are reported as a single aggregated unit. To restore county-level consistency, the combined GDP value was split proportionally across the two counties based on their relative population shares:

$$\text{GDP}_i = \text{GDP}_{\text{combined}} \times \frac{\text{population}_i}{\text{population}_{\text{Maui}} + \text{population}_{\text{Kalawao}}}.$$

This procedure ensures that total GDP is preserved while maintaining consistency with the county-level geography used throughout the analysis.

Other Combined Geographic Areas In additional cases where the BEA reports GDP for multi-county or mixed city–county areas, the same population-based redistribution method was applied:

$$\text{GDP}_i = \text{GDP}_{\text{total}} \times \frac{\text{population}_i}{\sum_{j \in \text{group}} \text{population}_j}.$$

This allocation strategy preserves the relative demographic weight of each constituent county and ensures that the resulting GDP values remain comparable across all observations.

Consistency Checks After harmonisation and FIPS assignment, all counties present in the electricity dataset were also found in the GDP dataset.

Summary Through a combination of geographic reassignment, population-weighted redistribution, and merging through standardized FIPS identifiers, the GDP dataset was successfully aligned with the environmental and demographic variables used in the empirical model. The resulting GDP variable provides a consistent and fully harmonized measure of economic activity for all counties included in the analysis.

Chapter 5

Empirical Analysis

5.1 Empirical Strategy and Model Specification

This chapter presents the empirical analysis of the determinants of data center location across U.S. counties. The objective is to identify how county-level demographic, economic, and infrastructural characteristics are associated with the probability of hosting at least one data center.

The primary outcome of interest is a binary variable, *Datacenter Presence*, equal to one if a county hosts at least one data center and zero otherwise. Given the binary nature of the dependent variable, the empirical analysis relies on logistic regression models to estimate the likelihood that a county hosts a data center, conditional on a set of explanatory variables.

The baseline empirical specification is given by:

$$\begin{aligned}\Pr(\text{Datacenter}_c = 1) = \Lambda\Big(&\alpha + \beta_1 \ln(\text{Population}_c) + \beta_2 \ln(\text{GDPpc}_c) \\ &+ \beta_3 \ln(\text{Water}_c) + \beta_4 \text{Electricity}_c + X_c'\gamma\Big)\end{aligned} \tag{5.1}$$

where $\Lambda(\cdot)$ denotes the logistic cumulative distribution function, c indexes counties, and X_c represents additional controls and interaction terms introduced in extended specifications and robustness checks.

The analysis proceeds in a stepwise manner. First, parsimonious models examine the association between electricity prices and data center presence. Subsequently, population size and GDP per capita are introduced to control for county scale and economic development. Infrastructure constraints, proxied by public water supply withdrawals, are then incorporated to capture cooling-related requirements. Finally, interaction terms are included to explore whether the effects of electricity

prices, population, and economic development vary across counties with different characteristics.

Electricity prices are introduced sequentially, first using commercial electricity rates and subsequently industrial electricity rates, to assess whether results are sensitive to the type of tariff considered.

5.2 Average Marginal Effects and Interpretation

Coefficients from logistic regression models are expressed in log-odds units, which makes their direct interpretation in terms of probabilities difficult. To provide a more intuitive interpretation of the results, this study relies on average marginal effects (AMEs). Average marginal effects measure how the predicted probability of hosting a data center changes, on average, when one explanatory variable increases while all other variables are held constant. In practice, the marginal effect is computed for each observation in the sample and then averaged across all counties. Expressing results in probability terms allows a clearer interpretation of the magnitude of the effects and facilitates comparison across different model specifications. When interaction terms are included in the model, marginal effects are evaluated at different values of the interacting variables. This allows the analysis to capture how the effect of one variable varies depending on the level of another variable. These relationships are illustrated using marginal effects plots.

5.3 Descriptive Evidence and Outcome Distribution

This section presents descriptive statistics for the variables used in the empirical analysis and documents the distribution of data center activity across U.S. counties.

The final sample consists of 3,144 counties, of which 336 (10.69%) host at least one data center. Data center activity is highly unevenly distributed across space. While the mean number of data centers per county is 1.05, the median is zero, indicating that the majority of counties do not host any facility.

The distribution of data center counts is highly skewed, with a small number of counties hosting a large concentration of facilities. In the sample, the maximum number of data centers in a single county reaches 168, highlighting the strong geographic clustering of digital infrastructure.

This pronounced concentration motivates the use of a binary outcome model to study the extensive margin of data center location (presence versus absence).

Table 5.1: Datacenter presence across counties

Statistic	Value
Observations	3144
Datacenter presence = 1 (count)	336
Datacenter presence = 1 (%)	10.69
Datacenter presence = 0 (count)	2808
Datacenter presence = 0 (%)	89.31

The intensive margin, captured by the number of data centers within counties, is analyzed separately in additional robustness analyses.

Summary statistics for the log-transformed explanatory variables are reported in Appendix .2. The logarithmic transformation reduces the strong right-skewness of the original variables and facilitates the interpretation of regression coefficients in semi-elasticity terms.

County population exhibits substantial variation across the sample, reflecting large differences between rural and metropolitan counties. GDP per capita shows moderate dispersion but includes several extreme observations, indicating heterogeneity in local economic development. Public water supply withdrawals are also unevenly distributed across counties, with most observations concentrated at relatively low levels and a small number of counties exhibiting significantly higher infrastructure capacity.

5.4 Data Preparation and Validation

Prior to estimation, several data preparation and validation steps are performed to ensure consistency, interpretability, and transparency of the empirical analysis.

5.4.1 Variable Transformations

Gross Domestic Product (GDP) is originally measured in thousands of chained 2017 dollars at the county level. GDP per capita is constructed by dividing total county GDP by county population:

$$GDPpc_c = \frac{GDP_c}{Population_c}.$$

Since GDP is expressed in thousands of dollars, the resulting GDP per capita variable is measured in thousands of dollars per person.

To account for skewness and facilitate economic interpretation, key continuous variables are log-transformed. In particular, the logarithmic transformations are applied to population, GDP per capita, and public water supply withdrawals:

$$\ln(Population_c), \quad \ln(GDPpc_c), \quad \ln(Water_c).$$

These transformations allow estimated coefficients to be interpreted in percentage terms and reduce the influence of extreme values. Importantly, the choice of measurement units (thousands versus millions) does not affect statistical inference once variables are log-transformed.

To characterize the extent of electricity price imputation within each county, the shares of counties with spatially imputed electricity prices are expressed in percentage terms.

5.4.2 Validation of Electricity Price Data

A series of internal consistency checks are conducted to validate the construction of electricity price variables and the associated imputation indicators. Counties flagged as fully approximated are verified to exhibit imputation shares equal to one for both commercial and industrial tariffs. Similarly, counties for which industrial electricity prices are imputed using neighboring counties are confirmed to have a full industrial imputation share.

Based on the share of spatially imputed industrial electricity prices, a high-imputation indicator is defined for counties with imputation shares exceeding 60 percent. Cross-tabulations confirm perfect alignment between imputation flags and underlying imputation shares, with no misclassifications or missing values.

5.4.3 Descriptive Validation of Data Center Outcomes

As an additional validation step, the distribution of data center presence and data center counts is examined across counties. Notably, none of the counties characterized by fully approximated electricity prices host a data center, and only two

counties relying on across-county imputation for industrial electricity prices host any data center activity.

This descriptive evidence suggests that the most severe data quality issues are not concentrated in counties with substantial data center presence, mitigating concerns that imputation may mechanically drive the main empirical results.

Detailed validation tables and consistency checks are reported in Appendix .3.

5.5 Baseline Associations: Electricity Prices and Data Center Presence

This section examines the baseline relationship between electricity prices and the probability that a county hosts at least one data center. The objective is to document the unconditional association between electricity costs and data center presence, before introducing controls for county scale, economic development, and infrastructure.

5.5.1 Commercial Electricity Prices

The first baseline specification focuses on commercial electricity prices. In this model, the presence of the data center is regressed solely on the average commercial electricity rate at the county level.

The estimated coefficient on commercial electricity prices is positive and statistically significant at 1 percent level ($p<0.001$). This result indicates that in this simple specification, counties with higher commercial electricity prices are more likely to host at least one data center. Table 5.2 reports the results of the baseline specification where data center presence is regressed on commercial electricity prices. Although the estimated association is statistically strong, the general explanatory power of the model is limited. The Pseudo R^2 is equal to 0.0135, indicating that commercial electricity prices alone explain only a very small share of the variation in the presence of data centers in counties. This suggests that electricity prices, by themselves, are unlikely to capture the main forces shaping data center location patterns.

Table 5.2: Baseline Logistic Regression: Commercial Electricity Prices

	(1)	
	Coefficient	P-value
datacenter_presence		
avg_comm_rate_county	5.305	0.000
Constant	-2.819	0.000
Pseudo R^2	0.014	

5.5.2 Industrial Electricity Prices

This subsection replicates the baseline analysis using average industrial electricity prices as an alternative measure of electricity costs. The purpose of this exercise is to assess whether the baseline association observed for commercial rates is sensitive to the choice of electricity tariff.

When data center presence is regressed solely on the average industrial electricity rate, the estimated coefficient is again positive and highly statistically significant. The results indicate that counties with higher industrial electricity prices are, on average, more likely to host a data center in this unconditional specification.

As observed in Table 5.3, the explanatory power of the model remains very limited. The Pseudo R^2 of 0.0239 suggests that industrial electricity prices alone account for only a small share of the observed variation in data center presence across counties. Consequently, this baseline relationship should be interpreted as a descriptive correlation rather than evidence of a causal effect.

Taken together, these baseline results document a positive unconditional association between electricity prices and data center presence. In the next sections, additional controls are introduced to examine whether this relationship persists once differences in county size, economic development, and infrastructure availability are taken into account.

Table 5.3: Baseline Logistic Regression: Industrial Electricity Prices

	(1)	
	Coefficient	P-value
datacenter_presence		
avg_ind_rate_county	7.646	0.000
Constant	-2.852	0.000
Pseudo R^2	0.024	

5.6 County Scale and Economic Development: Population and GDP

This section examines the role of county scale and economic development in shaping data center location. County scale is proxied by population size, while economic development is captured by GDP per capita. These variables account for differences in market size, demand, and overall economic activity that may influence data center siting decisions.

5.6.1 Population and Commercial Electricity Prices

In the baseline specification, commercial electricity prices are positively correlated with data center presence, although the model explains only a small share of the overall variation. This result suggests that electricity prices alone are insufficient to characterize data center location patterns.

Once population size is introduced into the model, the estimated relationship between electricity prices and data center presence changes substantially. The coefficient on commercial electricity prices becomes negative and statistically insignificant (-1.5 and p=0.367), while population size emerges as a strong and highly significant predictor of data center presence (1.32 and p<0.001).

As showed in Table 5.4, the inclusion of population size leads to a marked improvement in model fit, with the Pseudo R^2 increasing from 0.0135 to 0.3729. This sharp increase indicates that county scale explains a large share of the variation in data center location across counties. Larger counties are significantly more likely to host at least one data center, reflecting the importance of scale, agglomeration, and demand-side factors.

These results suggest that the positive association between electricity prices and data center presence observed in the baseline specification is largely driven by omitted differences in county size.

Table 5.4: Logistic regression of data center presence on commercial electricity prices controlling for log county population.

	(1)	
	Coefficient	P-value
datacenter_presence		
avg_comm_rate_county	-1.501	0.367
ln_population	1.325	0.000
Constant	-16.787	0.000
Pseudo R^2	0.373	

5.6.2 Population and Industrial Electricity Prices

A similar pattern emerges when industrial electricity prices are considered. When population size is added to the model, population enters with a positive and highly statistically significant coefficient (1.31 and $p<0.001$), confirming that larger counties are substantially more likely to host data centers.

At the same time, the coefficient on industrial electricity prices becomes statistically insignificant once population is controlled for (-.083 and $p=0.984$). This indicates that industrial electricity prices do not have an independent effect on data center presence when differences in county scale are taken into account.

As in the case of commercial electricity prices, it is displayed in Table 5.5 the inclusion of population size substantially improves the explanatory power of the model. The Pseudo R^2 increases to 0.3723, further highlighting the central role of county scale in explaining data center location decisions.

Taken together, the results in this section underscore the importance of population size as a key determinant of data center presence and suggest that electricity prices play a limited role once differences in county scale are properly accounted for.

Table 5.5: Logistic regression of data center presence on industrial electricity prices controlling for log county population.

	(1)	
	Coefficient	P-value
datacenter_presence		
avg_ind_rate_county	-0.083	0.948
ln_population	1.311	0.000
Constant	-16.821	0.000
Pseudo R^2	0.372	

5.6.3 Economic Development and Data Center Presence: Commercial Electricity Prices

This subsection extends the analysis by introducing GDP per capita as an additional control for county-level economic development. The inclusion of GDP per capita allows to assess whether differences in economic prosperity explain variation in data center location beyond county size alone.

When GDP per capita is added to the specification together with population size and commercial electricity prices, it enters the model with a positive and highly statistically significant coefficient (1.187 and $p<0.001$). This result indicates that counties with higher levels of economic development are more likely to host at least one data center.

The inclusion of GDP per capita leads to a further improvement in model fit, with the Pseudo R^2 increasing to 0.3979. This substantial increase confirms that economic development, in addition to population size, plays an important role in explaining data center location decisions.

Importantly, as explained in Table 5.6, once population size and GDP per capita are controlled for, commercial electricity prices remain statistically insignificant (-2.25 and $p=0.183$). This finding reinforces the interpretation that the positive baseline association between electricity prices and data center presence reflects underlying differences in county scale and economic activity rather than a direct effect of electricity costs.

While one might expect wealthier counties to face higher land costs or regulatory constraints, the results suggest instead that higher-income counties are more likely to possess the infrastructure, connectivity, and institutional capacity required to support large-scale data center operations.

Table 5.6: Logistic regression of data center presence on commercial electricity prices controlling for county population and GDP per capita.

	Coefficient	P-value
datacenter_presence		
avg_comm_rate_county	-2.252	0.183
ln_population	1.226	0.000
ln_gdp_pc	1.187	0.000
_cons	-20.179	0.000
Observations	3144	
Pseudo R^2	0.398	

5.6.4 Economic Development and Data Center Presence: Industrial Electricity Prices

A similar pattern emerges when industrial electricity prices are considered. In this specification, population size, GDP per capita, and industrial electricity prices are included jointly to assess their relative contributions to data center location.

Population size remains positive and highly statistically significant, confirming that county scale continues to be a dominant determinant of data center presence. GDP per capita also enters the model with a positive and statistically significant coefficient (1.18 and $p<0.001$), indicating that economically more developed counties are more likely to host data centers (Table 5.7).

By contrast, industrial electricity prices remain statistically insignificant once both population size and GDP per capita are controlled for. This result suggests that industrial electricity costs do not exert an independent influence on data center location decisions beyond what is captured by county scale and economic development.

The explanatory power of the model increases further, with the Pseudo R^2 approaching 0.40. Taken together, these results show that the combination of population size and economic development explains a substantial share of the observed spatial distribution of data centers across U.S. counties, while electricity prices play a limited role once these structural factors are taken into account.

Table 5.7: Logistic regression of data center presence on industrial electricity prices controlling for county population and GDP per capita.

	Coefficient	P-value
datacenter_presence		
avg_ind_rate_county	-0.999	0.461
ln_population	1.215	0.000
ln_gdp_pc	1.179	0.000
_cons	-20.225	0.000
Observations	3144	
Pseudo R^2	0.397	

5.7 Infrastructure Constraints: The Role of Public Water Supply

This section examines the role of infrastructure constraints in shaping data center location, focusing in particular on the availability of public water supply. Water is a critical input for data center operations due to its role in cooling systems, and differences in water infrastructure may therefore influence where data centers can feasibly locate.

5.7.1 Public Water Supply and Commercial Electricity Prices

This Model extends the previous specification by introducing public water supply alongside commercial electricity prices, population size, and GDP per capita. The results in Table 5.8 indicate that water availability plays an independent and statistically significant role in explaining data center presence.

Once public water supply is included, the coefficient on commercial electricity prices becomes negative and statistically insignificant (-2.29 and p=0.179). This confirms earlier findings that electricity prices do not exert a meaningful influence on data center location once county scale and economic development are taken into account.

By contrast, population size and GDP per capita remain positive and highly statistically significant. Larger and more economically developed counties are consistently more likely to host data centers, reflecting the importance of market size, demand, and broader economic capacity.

Importantly, the coefficient on public water supply is positive and statistically significant (.22 and p=0.006). This result suggests that counties with greater water supply are more likely to host data centers, consistent with the role of water as a

Table 5.8: Logistic regression of data center presence on commercial electricity prices controlling for county population, GDP per capita and water supply.

	Coefficient	P-value
datacenter_presence		
avg_comm_rate_county	-2.288	0.179
ln_population	1.066	0.000
ln_gdp_pc	1.157	0.000
ln_water	0.224	0.006
_cons	-18.748	0.000
Observations	3144	
Pseudo R^2	0.401	

key infrastructural input for cooling and large-scale digital facilities.

The inclusion of public water supply leads to a further improvement in model fit. The Pseudo R^2 increases to 0.4014, indicating that population size, economic development, and water infrastructure together explain a substantial share of the observed variation in data center presence across counties.

5.7.2 Public Water Supply and Industrial Electricity Prices

A similar pattern emerges when industrial electricity prices are used instead of commercial rates. When public water supply is added to a specification that already includes population size and GDP per capita, industrial electricity prices remain statistically insignificant (Table 5.9).

Population size and GDP per capita continue to be strong and robust predictors of data center presence, confirming that county scale and economic development are the primary correlates of data center location.

Public water supply again enters the model with a positive and statistically significant coefficient (0.22 and p=0.006). This finding reinforces the interpretation that infrastructure capacity, as proxied by water availability, plays an important role in enabling data center siting beyond what is captured by population and income alone.

The inclusion of water leads to a modest additional increase in explanatory power, with the Pseudo R^2 rising from 0.3969 to 0.4004. While the main improvement in model fit is driven by population and GDP per capita, water availability provides additional explanatory power and captures an important infrastructural constraint.

Taken together, the results in this section highlight the role of public water supply

Table 5.9: Logistic regression of data center presence on industrial electricity prices controlling for county population, GDP per capita and water supply.

	Coefficient	P-value
datacenter_presence		
avg_ind_rate_county	-1.037	0.446
ln_population	1.055	0.000
ln_gdp_pc	1.149	0.000
ln_water	0.224	0.006
_cons	-18.793	0.000
Observations	3144	
Pseudo R^2	0.400	

as a relevant determinant of data center presence, while confirming that electricity prices—whether commercial or industrial—do not have an independent effect once structural economic and infrastructural factors are controlled for.

5.8 Average Marginal Effects

To facilitate interpretation of the estimated relationships in probability terms, this section reports average marginal effects (AMEs) for the main explanatory variables. AMEs measure the average change in the predicted probability that a county hosts at least one data center associated with a change in a given covariate, holding all other variables constant.

5.8.1 Average Marginal Effects with Commercial Electricity Prices

Table 5.10 reports the average marginal effects obtained from the specification including commercial electricity prices, population size, GDP per capita, and public water supply.

The AME associated with commercial electricity prices is negative but statistically insignificant. In the baseline specification (Model 1), a one-unit increase in the commercial electricity rate is associated with a 0.50 increase in the probability of hosting a data center ($p < 0.001$). However, once county characteristics are progressively introduced in the model, the estimated marginal effect becomes negative and statistically insignificant. In the full specification (Model 4), the AME is approximately -0.134 ($p = 0.179$), indicating that, once county scale, economic development, and water availability are controlled for, commercial electricity prices

do not have a statistically meaningful effect on the probability of hosting a data center.

By contrast, population size exhibits a large and statistically significant marginal effect across all specifications. In Model 2, the AME for ln(population) is 0.082 ($p < 0.001$), while in the full model it remains large at 0.063 ($p < 0.001$). This implies that increases in county population are associated with sizeable increases in the probability of hosting a data center. Since the variable is expressed in logarithmic form, these estimates indicate that a 1% increase in population is associated with approximately a 0.063 percentage point increase in the probability that a county hosts at least one data center.

GDP per capita also displays a positive and statistically significant marginal effect. In Model 3, the AME for ln(GDP per capita) is approximately 0.070 ($p < 0.001$), and remains similar in the full specification (0.068, $p < 0.001$). This indicates that economically more developed counties are systematically more likely to host data centers, reinforcing earlier findings that economic development captures structural conditions—such as infrastructure quality, digital demand, and agglomeration economies—that favor data center siting.

Public water supply enters with a positive and statistically significant marginal effect in the full model. The estimated AME for ln(water) is approximately 0.013 ($p = 0.006$), indicating that greater water availability is associated with a higher probability of data center presence. Although the magnitude of this effect is smaller than those associated with population size and economic development, it is consistent with the role of water infrastructure as an enabling factor for data center cooling and operations.

Table 5.10: Average marginal effects from logistic regression models estimated sequentially by adding control variables.

	(1)		(2)		(3)		(4)	
	Model 1		Model 2		Model 3		Model 4	
	AME	P-value	AME	P-value	AME	P-value	AME	P-value
avg_comm_rate_county	0.500	0.000	-0.093	0.366	-0.133	0.184	-0.134	0.179
ln_population			0.082	0.000	0.072	0.000	0.063	0.000
ln_gdp_pc					0.070	0.000	0.068	0.000
ln_water							0.013	0.006
Observations	3144		3144		3144		3144	

Figure 5.1 illustrates these results graphically. The marginal effects plot highlights that population size and GDP per capita are the dominant drivers in probability terms, while the confidence interval for commercial electricity prices crosses zero reflecting the lack of statistical significance.

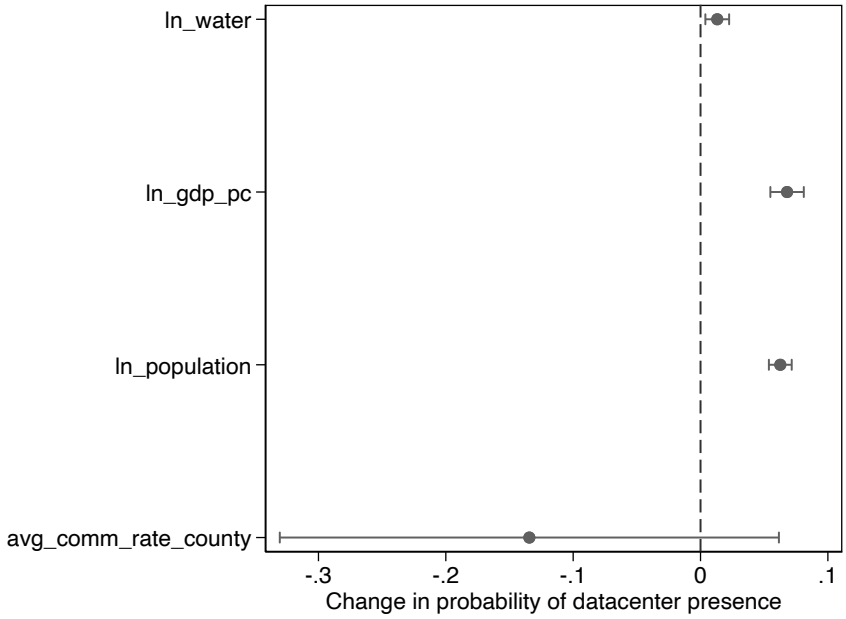

Figure 5.1: Average marginal effects on the probability of data center presence (95% confidence intervals).

5.8.2 Average Marginal Effects with Industrial Electricity Prices

Table 5.11 reports average marginal effects for the specification using industrial electricity prices.

Table 5.11: Average marginal effects from logistic regression models estimated sequentially by adding control variables.

	(1) Model 1		(2) Model 2		(3) Model 3		(4) Model 4	
	AME	P-value	AME	P-value	AME	P-value	AME	P-value
avg_ind_rate_county	0.713	0.000	-0.005	0.948	-0.059	0.462	-0.061	0.446
ln_population			0.081	0.000	0.072	0.000	0.062	0.000
ln_gdp_pc					0.070	0.000	0.068	0.000
ln_water							0.013	0.006
Observations	3144		3144		3144		3144	

The marginal effect of industrial electricity prices is negative but statistically insignificant once additional controls are included. In the baseline specification (Model 1), the AME for industrial electricity prices is 0.713 ($p < 0.001$), suggesting a strong positive association when no controls are considered. However, after introducing county characteristics, this effect disappears. In Model 2 the AME drops to -0.005 ($p = 0.948$), and remains statistically insignificant in Model 3 (-0.059, $p = 0.462$) and in the full specification (Model 4: -0.061, $p = 0.446$). This indicates that industrial electricity costs do not exert an independent influence on the probability of hosting a data center once population size, GDP per capita, and water availability are taken into account.

Population size remains one of the strongest predictors of data center presence. The AME for ln(population) is 0.081 ($p < 0.001$) in Model 2, 0.072 ($p < 0.001$) in Model 3, and 0.062 ($p < 0.001$) in the full model. Since the variable is expressed in logarithmic form, these estimates imply that a 1% increase in county population is associated with approximately a 0.06 percentage point increase in the probability that a county hosts at least one data center.

GDP per capita also exhibits a positive and highly statistically significant marginal effect. In Model 3, the AME for ln(GDP per capita) is 0.070 ($p < 0.001$), remaining similar in Model 4 at 0.068 ($p < 0.001$). This suggests that counties with higher levels of economic development are systematically more likely to host data centers, reflecting the role of infrastructure quality, agglomeration economies, and demand for digital services.

Public water supply again displays a positive and statistically significant marginal effect in the full specification. The estimated AME for ln(water) is 0.013 ($p = 0.006$), confirming the relevance of infrastructure availability for data center siting. Although its magnitude is smaller relative to population and GDP per capita, water availability consistently contributes to explaining data center presence across specifications.

The marginal effects plot in Figure 5.1 visually reinforces these patterns. Population and GDP per capita display the largest positive effects with confidence intervals entirely above zero, indicating strong statistical significance. In contrast, the estimated effect of industrial electricity prices is statistically insignificant, highlighting the role of infratructure capacity in supporting data center operations.

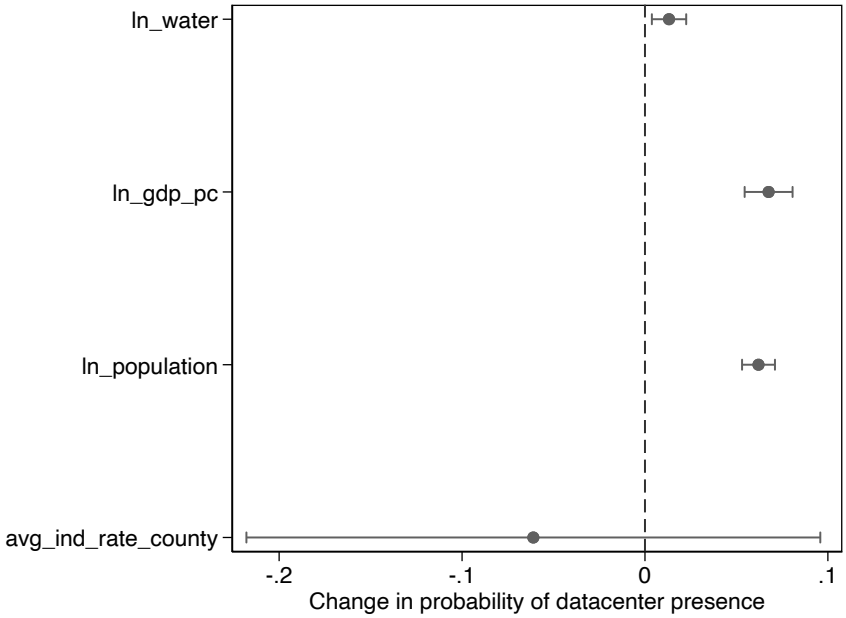

Figure 5.2: Average marginal effects on the probability of data center presence (95% confidence intervals).

5.8.3 Summary Interpretation of Main Effects

Taken together, the regression results and the corresponding average marginal effects point to a clear pattern. Once county scale, economic development, and infrastructure availability are controlled for, electricity prices do not exert a statistically meaningful influence on the probability that a county hosts a data center.

Population size emerges as the strongest predictor of data center presence, with larger counties exhibiting substantially higher predicted probabilities of hosting data centers. GDP per capita also displays a positive and statistically significant effect, indicating that economically more developed counties are more likely to attract data center investment.

Public water supply contributes positively to data center presence, although its magnitude is smaller relative to population and GDP per capita, consistent with its

role as an enabling infrastructure factor.

Overall, the results indicate that structural county characteristics—scale, economic development, and infrastructure capacity—are the primary drivers of data center location across U.S. counties. The next section examines whether these relationships vary across counties with different characteristics through the inclusion of interaction terms.

5.9 Interaction Effects

This section examines whether the effects of key explanatory variables on data center presence vary across counties with different characteristics. In particular, interaction terms are introduced to assess whether the impact of electricity prices depends on county scale, economic development, or infrastructure availability.

Given the nonlinear nature of the Logit model, the interaction effects are interpreted using average marginal effects evaluated at different values of the interacting variables and illustrated through marginal effects plots.

5.9.1 Population and Electricity Prices

To test whether the role of electricity costs differs between larger and smaller counties, I estimate specifications including an interaction term between county population size and electricity prices. The models control for GDP per capita and public water supply and are estimated with heteroskedasticity-robust standard errors.

Commercial electricity prices

In the specification interacting population and commercial electricity prices, as showed in Table 5.12 remains a strong predictor of data center presence ($\beta_{\ln(pop)} = 1.1769$, $p < 0.01$), while the main effect of commercial electricity prices is statistically insignificant ($\beta_{comm} = 7.9668$, $p = 0.594$). Importantly, the interaction term is also statistically insignificant ($\beta_{\ln(pop)\times comm} = -0.8396$, $p = 0.484$). These estimates provide no evidence that the association between electricity prices and data center presence varies systematically with county size.

Table 5.12: Logistic regression estimates with and without interaction between county population and commercial electricity prices.

	No interaction		Population × Comm. rate	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
avg_comm_rate_county	-2.288	0.179	7.967	0.594
ln_population	1.066	0.000	1.177	0.000
ln_gdp_pc	1.157	0.000	1.166	0.000
ln_water	0.224	0.006	0.225	0.006
ln_population × avg_comm_rate_county			-0.840	0.484
Constant	-18.748	0.000	-20.129	0.000
Pseudo R^2	0.401		0.402	
Observations	3144		3144	

Given the nonlinear nature of the Logit model, the interaction between population and commercial electricity prices is interpreted using average marginal effects (Table 5.13). Specifically, the marginal effect of ln(population) is evaluated at three representative values of the commercial electricity rate: the 10th percentile (low), the mean, and the 90th percentile (high). The estimated marginal effects are very similar across these values, equal to 0.065, 0.063, and 0.060 respectively, and all are highly statistically significant ($p < 0.001$). This indicates that the positive effect of population size on the probability of hosting a data center remains stable across the distribution of commercial electricity prices. Figure 5.3 provides a graphical representation of these results: the confidence intervals overlap substantially, confirming that there is no meaningful heterogeneity in the population effect across different electricity price levels.

Table 5.13: Average marginal effect of log population evaluated at different levels of commercial electricity prices.

	AME of ln(population)	P-value
ln_population		
Low (p10 = 0.090)	0.065	0.000
Mean (0.126)	0.063	0.000
High (p90 = 0.160)	0.060	0.000

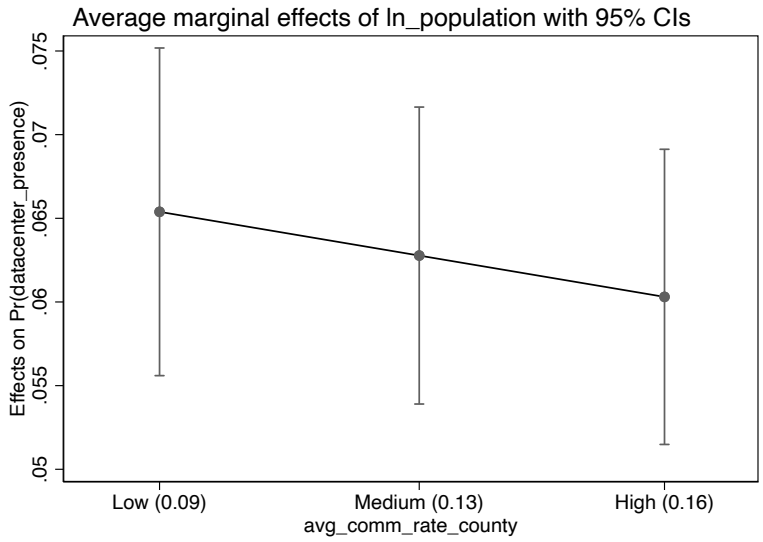

Figure 5.3: Marginal effect of population on the probability of data center presence evaluated at different levels of commercial electricity prices.

Industrial electricity prices

Results are consistent when industrial electricity prices are used (Table 5.14). Population remains positive and highly statistically significant ($\beta_{\ln(pop)} = 1.093$, $p < 0.001$), while the main effect of industrial electricity prices is statistically insignificant ($\beta_{ind} = 3.967$, $p = 0.779$). The interaction term between population and industrial electricity prices is also statistically insignificant ($\beta_{\ln(pop)\times ind} = -0.404$, $p = 0.723$), indicating no evidence that the effect of county population on data center presence varies across electricity price levels.

Table 5.14: Logistic regression estimates with and without interaction between county population and industrial electricity prices.

	No interaction		Population × Comm. rate	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
avg_ind_rate_county	-1.037	0.446	3.967	0.779
ln_population	1.055	0.000	1.093	0.000
ln_gdp_pc	1.149	0.000	1.154	0.000
ln_water	0.224	0.006	0.224	0.006
ln_population × avg_ind_rate_county			-0.404	0.723
Constant	-18.793	0.000	-19.280	0.000
Pseudo R^2	0.400		0.400	
Observations	3144		3144	

To interpret the interaction in the nonlinear Logit framework, average marginal effects are evaluated at three representative values of industrial electricity prices

(low, mean, and high). The estimated marginal effects of ln(population) are nearly identical across these values, equal to approximately 0.063, 0.062, and 0.061 respectively (Table 5.15).

The margins plot in Figure 5.4 visually confirms this pattern: the estimated effects are very similar and the confidence intervals overlap substantially, indicating that the positive effect of population on the probability of hosting a data center remains stable across the distribution of industrial electricity prices.

Table 5.15: Average marginal effect of log population evaluated at different levels of industrial electricity prices.

	AME of ln(population)	P-value
ln_population		
Low (p10 = 0.065)	0.063	0.000
Mean (0.089)	0.062	0.000
High (p90 = 0.113)	0.061	0.000

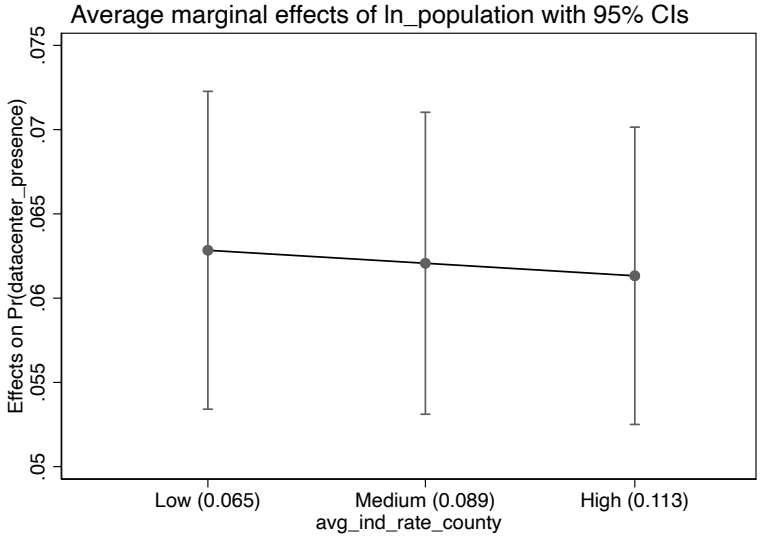

Figure 5.4: Marginal effect of population on the probability of data center presence evaluated at different levels of commercial electricity prices.

Implication. Across both commercial and industrial tariffs, the interaction results confirm that electricity prices are not significantly associated with data center presence once structural county characteristics are controlled for. By contrast, population size remains a strong and robust predictor, with no evidence that its effect varies across electricity price levels.

5.9.2 Population and Public Water Supply

This subsection tests whether the effect of county scale varies with infrastructure availability by introducing an interaction term between population size and public water supply. The specification controls for electricity prices and GDP per capita and is estimated with heteroskedasticity-robust standard errors. Given the nonlinear nature of the Logit model, the interaction is interpreted using marginal effects evaluated at different values of water availability.

Commercial electricity prices

As shown in Table 5.16, when interacting ln(population) and ln(water) in the specification including commercial electricity prices, population remains positive and highly statistically significant ($\beta_{\ln(pop)} = 1.075$, $p < 0.001$), while GDP per capita also remains positive and significant ($\beta_{\ln(gdp\_pc)} = 1.162$, $p < 0.001$). Commercial electricity prices remain statistically insignificant ($\beta_{comm} = -2.275$, $p = 0.181$).

Table 5.16: Logistic regression estimates with and without interaction between county population and water supply (commercial electricity prices).

	No interaction		Population × Comm. rate	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
ln_population	1.066	0.000	1.075	0.000
ln_water	0.224	0.006	0.293	0.582
avg_comm_rate_county	-2.288	0.179	-2.275	0.181
ln_gdp_pc	1.157	0.000	1.162	0.000
ln_population × ln_water			-0.006	0.895
Constant	-18.748	0.000	-18.882	0.000
Pseudo R^2	0.401		0.401	
Observations	3144		3144	

The interaction term between ln(population) and ln(water) is not statistically significant ($\beta_{\ln(pop)\times\ln(water)} = -0.006$, $p = 0.895$). Consistent with this result, the marginal effect of population evaluated at different levels of water supply remains broadly stable.

Specifically, Table 5.17 reports that the average marginal effect of ln(population) increases slightly from 0.053 at low water availability ($\ln(water) = 0.223$) to 0.061 at the mean ($\ln(water) = 1.477$) and 0.074 at high water availability ($\ln(water) = 3.217$). All marginal effects are statistically significant ($p < 0.001$), but the differences across water availability levels remain modest.

Table 5.17: Average marginal effect of log population evaluated at different levels of water supply.

	AME of ln(population)	P-value
ln_population		
Low (p10 = 0.223)	0.053	0.000
Mean (1.477)	0.061	0.000
High (p90 = 3.217)	0.074	0.000

Figure 5.5 illustrates this pattern. The marginal effect of population appears slightly larger in counties with greater water availability; however, the substantial overlap of confidence intervals indicates limited evidence of heterogeneous effects across water supply levels.

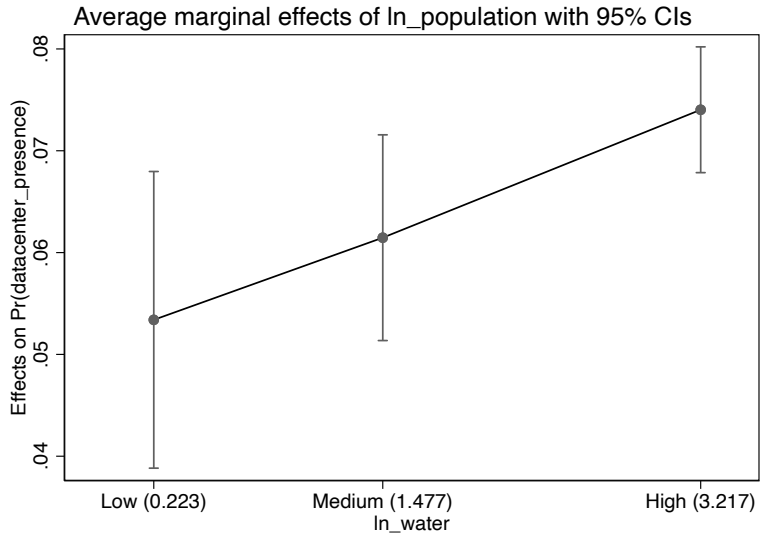

Figure 5.5: Marginal effect of population on the probability of data center presence evaluated at different levels of water supply.

Industrial electricity prices

Results reported in Table 5.18 are similar when industrial electricity prices are used. Population remains positive and highly statistically significant ($\beta_{\ln(pop)} = 1.067$, $p < 0.001$), while GDP per capita also remains positive and significant ($\beta_{\ln(gdp\_pc)} = 1.157$, $p < 0.001$). Industrial electricity prices remain statistically insignificant ($\beta_{ind} = -1.021$, $p = 0.452$).

Table 5.18: Logistic regression estimates with and without interaction between county population and water supply (industrial electricity prices).

	No interaction		Population × Water Supply	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
avg_ind_rate_county	-1.037	0.446	-1.021	0.452
ln_population	1.055	0.000	1.067	0.000
ln_gdp_pc	1.149	0.000	1.157	0.000
ln_water	0.224	0.006	0.319	0.551
ln_population × ln_water			-0.008	0.858
Constant	-18.793	0.000	-18.976	0.000
Pseudo R^2	0.400		0.400	
Observations	3144		3144	

The interaction term between ln(population) and ln(water) is again statistically insignificant ($\beta_{\ln(pop)\times\ln(water)} = -0.008$, $p = 0.858$), suggesting no evidence that the effect of county population on data center presence varies systematically with water availability.

Table 5.19: Average marginal effect of log population evaluated at different levels of water supply.

	AME of ln(population)	P-value
ln_population		
Low (p10 = 0.223)	0.053	0.000
Mean (1.477)	0.061	0.000
High (p90 = 3.217)	0.073	0.000

Average marginal effects evaluated at different levels of water availability are very similar. As shown in Table 5.19, the marginal effect of ln(population) increases slightly from 0.053 at low water availability ($\ln(water) = 0.223$) to 0.061 at the mean ($\ln(water) = 1.477$) and 0.073 at high water availability ($\ln(water) = 3.217$). Although marginal effects increase modestly with water availability, the differences remain small and all estimates are highly statistically significant ($p < 0.001$).

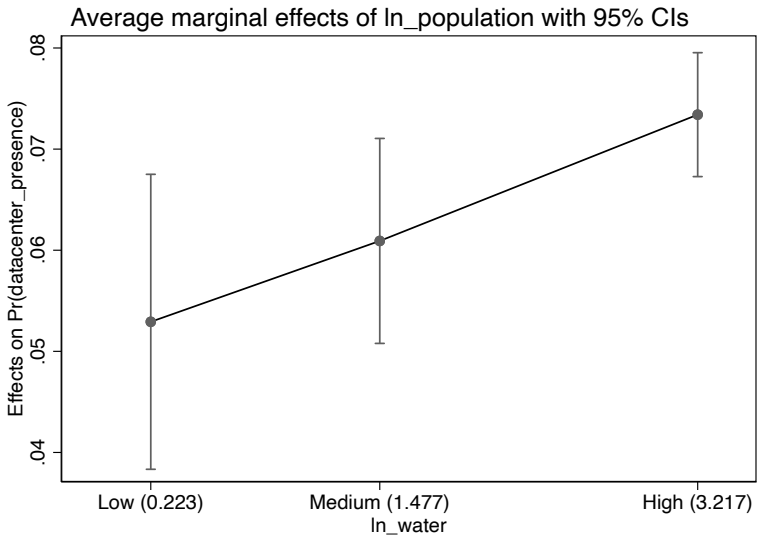

Figure 5.6: Marginal effect of population on the probability of data center presence evaluated at different levels of water supply.

Implication. Overall, the interaction analysis provides little evidence that the relationship between county scale and data center presence depends strongly on water availability. While marginal effects suggest a slightly larger population effect in counties with greater water supply, the interaction term is not statistically significant and the confidence intervals overlap substantially. This supports the interpretation of water availability as an enabling infrastructure factor with broadly similar relevance across counties, rather than a constraint that strongly amplifies or dampens the role of scale.

5.9.3 Economic Development and Electricity Prices

This subsection tests whether the association between electricity prices and data center presence varies with county-level economic development. To this end, I estimate specifications including an interaction term between electricity prices and GDP per capita, controlling for population size and public water supply. Given the nonlinear nature of the Logit model, the interaction is interpreted using marginal effects evaluated at different levels of GDP per capita.

Commercial electricity prices

In the model interacting commercial electricity prices and GDP per capita (Table 5.20), GDP per capita remains positive and statistically significant ($\beta_{\ln(gdp\_pc)} =$

0.878, $p = 0.003$), while population size and public water supply are also positive and significant ($\beta_{\ln(pop)} = 1.065$, $p < 0.001$; $\beta_{\ln(water)} = 0.226$, $p = 0.006$).

Table 5.20: Logistic regression estimates with and without interaction between county GDP per capita and commercial electricity prices.

	No interaction		GDP per capita × Comm. rate	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
ln_population	1.066	0.000	1.065	0.000
ln_water	0.224	0.006	0.226	0.006
avg_comm_rate_county	-2.288	0.179	-11.409	0.252
ln_gdp_pc	1.157	0.000	0.878	0.003
ln_gdp_pc × avg_comm_rate_county			2.168	0.355
Constant	-18.748	0.000	-17.569	0.000
Pseudo R^2	0.401		0.402	
Observations	3144		3144	

By contrast, commercial electricity prices remain statistically insignificant ($\beta_{comm} = -11.409$, $p = 0.252$), and the interaction term between electricity prices and GDP per capita is also statistically insignificant ($\beta_{comm \times \ln(gdp\_pc)} = 2.168$, $p = 0.355$). This indicates no evidence that the relationship between commercial electricity prices and data center presence varies systematically with the level of economic development.

Consistent with this result, the marginal effect of GDP per capita remains positive and very similar across the distribution of electricity prices. As reported in Table 5.21, the average marginal effect of $\ln(gdp\_pc)$ is approximately 0.066 at low electricity prices, 0.067 at the mean, and 0.069 at high electricity prices. All estimates are highly statistically significant ($p < 0.001$), indicating a robust positive association between economic development and the probability of hosting a data center.

Table 5.21: Average marginal effect of GDP per capita evaluated at different levels of commercial electricity prices.

	AME of GDP per capita	P-value
ln_gdp_pc		
Low (p10 = 0.09)	0.066	0.000
Mean (0.13)	0.067	0.000
High (p90 = 0.16)	0.069	0.000

Figure 5.7 provides a graphical representation of these results. The marginal effect of GDP per capita appears stable across different electricity price levels, and the substantial overlap of confidence intervals suggests little evidence of heterogeneous

effects.

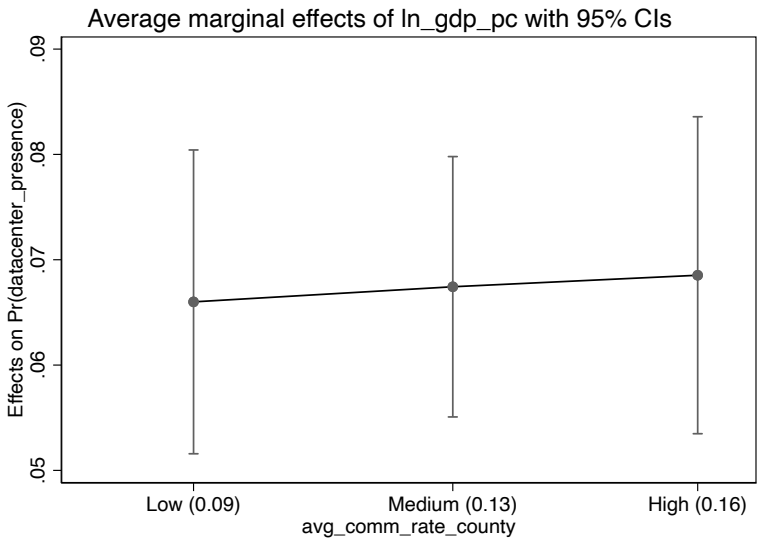

Figure 5.7: Marginal effect of GDP per capita on the probability of data center presence evaluated at different levels of commercial electricity prices.

Industrial electricity prices

Results reported in Table 5.22 are similar when industrial electricity prices are used. In the interaction model, GDP per capita remains positive and statistically significant ($\beta_{\ln(gdp\_pc)} = 1.183$, $p < 0.001$), while population size and public water supply are also positive and significant ($\beta_{\ln(pop)} = 1.056$, $p < 0.001$; $\beta_{\ln(water)} = 0.223$, $p = 0.006$). Industrial electricity prices remain statistically insignificant ($\beta_{ind} = 0.556$, $p = 0.973$), and the interaction between industrial electricity prices and GDP per capita is likewise statistically insignificant ($\beta_{ind \times \ln(gdp\_pc)} = -0.388$, $p = 0.923$).

Table 5.22: Logistic regression estimates with and without interaction between county GDP per capita and industrial electricity prices.

	No interaction		GDP per capita × Ind. rate	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
avg_ind_rate_county	-1.037	0.446	0.556	0.973
ln_population	1.055	0.000	1.056	0.000
ln_gdp_pc	1.149	0.000	1.183	0.000
ln_water	0.224	0.006	0.223	0.006
ln_gdp_pc × avg_ind_rate_county			-0.388	0.923
Constant	-18.793	0.000	-18.942	0.000
Pseudo R^2	0.400		0.400	
Observations	3144		3144	

Consistent with this result, the marginal effect of GDP per capita remains positive and very similar across the distribution of industrial electricity prices. As reported in Table 5.23, the average marginal effect of $\ln(gdp\_pc)$ is approximately 0.069 at low electricity prices, 0.067 at the mean, and 0.066 at high electricity prices. All estimates are highly statistically significant ($p < 0.001$), indicating that economically more developed counties are consistently more likely to host data centers.

Table 5.23: Average marginal effect of GDP per capita evaluated at different levels of industrial electricity prices.

	AME of GDP per capita	P-value
ln_gdp_pc		
Low (p10 = 0.065)	0.069	0.000
Mean (0.089)	0.067	0.000
High (p90 = 0.113)	0.066	0.000

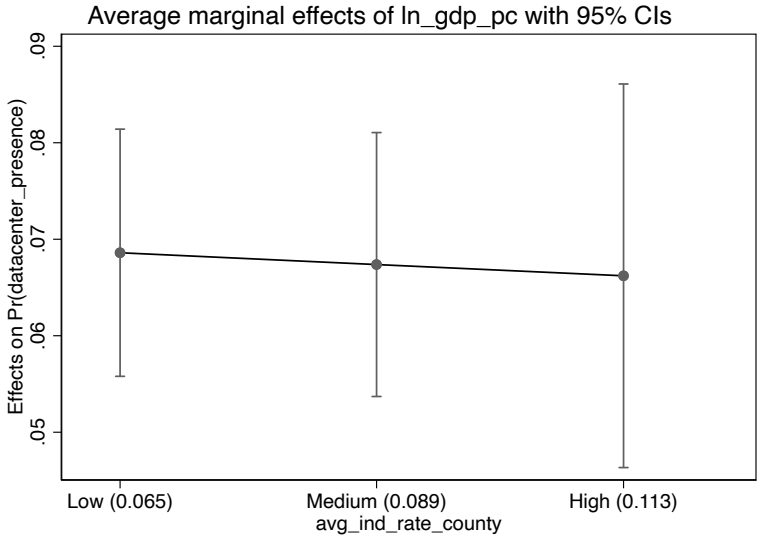

Figure 5.8: Marginal effect of GDP per capita on the probability of data center presence evaluated at different levels of industrial electricity prices.

Implication. Across both commercial and industrial tariffs, the interaction analysis provides no evidence that the effect of electricity prices on data center presence depends on economic development. Electricity prices remain statistically indistinguishable from zero across the GDP per capita distribution, while population size, GDP per capita itself, and water availability remain robust predictors. These results reinforce the main conclusion that structural county characteristics dominate

the spatial distribution of data centers, whereas electricity prices do not exhibit a first-order effect once these factors are controlled for.

5.9.4 Electricity Prices and Public Water Supply

This subsection investigates whether the association between electricity prices and data center presence varies with local infrastructure availability, proxied by public water supply. I estimate Logit specifications including an interaction term between electricity prices and ln(water), controlling for population size and GDP per capita. Given the nonlinear nature of the Logit model, the interaction is interpreted using marginal effects evaluated at different levels of water supply.

Commercial electricity prices

Table 5.24 reports the results of the model interacting commercial electricity prices and ln(water). The main effect of commercial electricity prices is statistically insignificant ($\beta_{comm} = 0.673$, $p = 0.865$). Public water supply remains positive and statistically significant ($\beta_{\ln(water)} = 0.366$, $p = 0.044$), while population and GDP per capita are also positive and highly significant ($\beta_{\ln(pop)} = 1.057$, $p < 0.001$; $\beta_{\ln(gdp\_pc)} = 1.159$, $p < 0.001$).

Table 5.24: Logistic regression estimates with and without interaction between county water supply and commercial electricity prices.

	No interaction		Water Supply × Comm. rate	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
ln_population	1.066	0.000	1.057	0.000
ln_water	0.224	0.006	0.366	0.044
avg_comm_rate_county	-2.288	0.179	0.673	0.865
ln_gdp_pc	1.157	0.000	1.159	0.000
ln_water × avg_comm_rate_county			-0.979	0.387
Constant	-18.748	0.000	-19.066	0.000
Pseudo R^2	0.401		0.402	
Observations	3144		3144	

The interaction term between commercial electricity prices and water supply is not statistically significant ($\beta_{comm \times \ln(water)} = -0.979$, $p = 0.387$), indicating no evidence that the effect of electricity prices differs systematically across counties with different levels of water availability.

Consistent with this result, the marginal effect of water supply remains positive and statistically significant across the distribution of electricity prices. As reported

in Table 5.25, the average marginal effect of ln(water) is approximately 0.016 at low electricity prices, 0.014 at the mean, and 0.012 at high electricity prices.

Table 5.25: Average marginal effect of water supply evaluated at different levels of commercial electricity prices.

	AME of Water Supply	P-value
ln_water		
Low (p10 = 0.09)	0.016	0.004
Mean (0.13)	0.014	0.004
High (p90 = 0.16)	0.012	0.012

Figure 5.9 illustrates this pattern. The marginal effect of water supply decreases slightly as electricity prices increase, but the confidence intervals overlap substantially, suggesting limited evidence of heterogeneous effects.

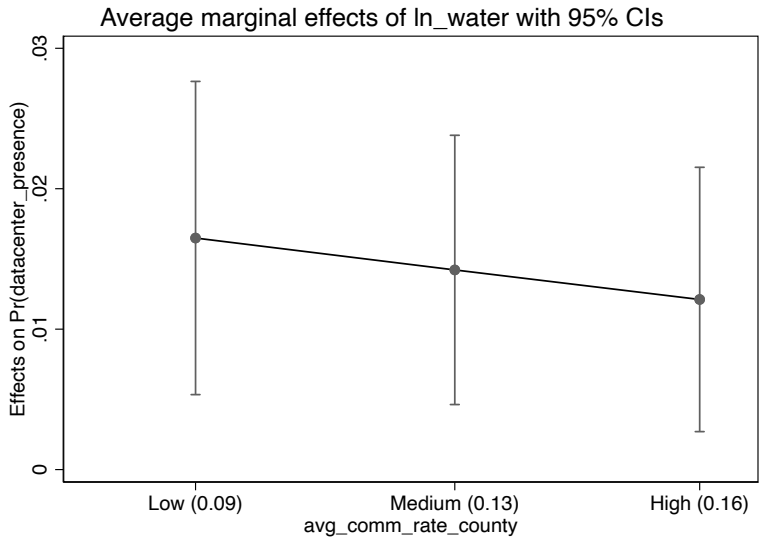

Figure 5.9: Marginal effect of water supply on the probability of data center presence evaluated at different levels of commercial electricity prices.

Industrial electricity prices

Results reported in Table 5.26 are similar when industrial electricity prices are used. Industrial electricity prices remain statistically insignificant ($\beta_{ind} = 1.326$, $p = 0.692$), while water supply remains positive and statistically significant ($\beta_{\ln(water)} = 0.309$, $p = 0.021$). Population and GDP per capita also remain positive and highly statistically significant ($\beta_{\ln(pop)} = 1.046$, $p < 0.001$; $\beta_{\ln(gdp\_pc)} = 1.151$, $p < 0.001$).

Table 5.26: Logistic regression estimates with and without interaction between county water supply and industrial electricity prices.

	No interaction		Water Supply × Ind. rate	
	Coefficient	P-value	Coefficient	P-value
datacenter_presence				
avg_ind_rate_county	-1.037	0.446	1.326	0.692
ln_population	1.055	0.000	1.046	0.000
ln_gdp_pc	1.149	0.000	1.151	0.000
ln_water	0.224	0.006	0.309	0.021
ln_water × avg_ind_rate_county			-0.764	0.439
Constant	-18.793	0.000	-18.946	0.000
Pseudo R^2	0.400		0.401	
Observations	3144		3144	

The interaction between industrial electricity prices and water supply is not statistically significant ($\beta_{ind \times \ln(water)} = -0.764$, $p = 0.439$), indicating no evidence that the relationship between water availability and data center presence varies systematically with electricity prices.

Consistent with this result, the marginal effect of water supply remains positive and very similar across the distribution of industrial electricity prices. As reported in Table 5.27, the estimated average marginal effects of ln(water) are 0.015 at low electricity prices, 0.014 at the mean, and 0.013 at high electricity prices, and all remain statistically significant.

Table 5.27: Average marginal effect of water supply evaluated at different levels of industrial electricity prices.

	AME of Water Supply	P-value
ln_water		
Low (p10 = 0.065)	0.015	0.004
Mean (0.089)	0.014	0.004
High (p90 = 0.113)	0.013	0.006

Figure 5.10 illustrates this pattern. The marginal effect of water supply decreases slightly as electricity prices increase, but the confidence intervals overlap substantially, suggesting limited evidence of heterogeneous effects across electricity price levels.

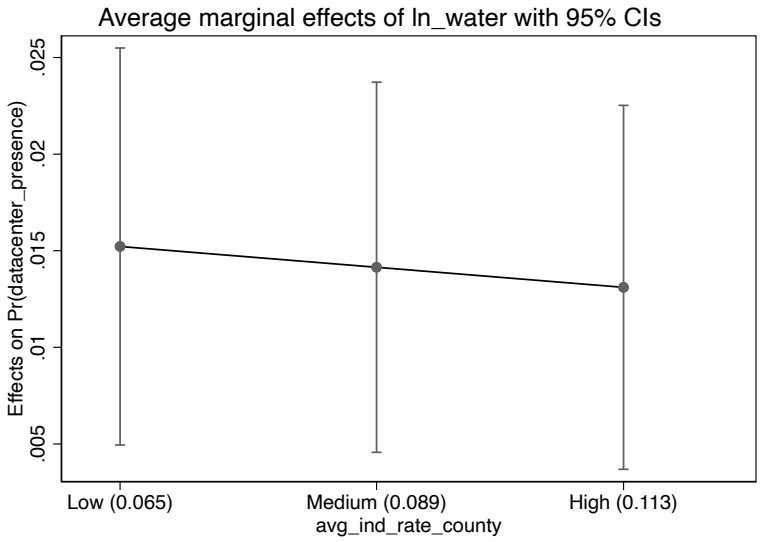

Figure 5.10: Marginal effect of water supply on the probability of data center presence evaluated at different levels of industrial electricity prices.

Implication. Across both commercial and industrial tariffs, the interaction analysis provides no evidence that the effect of electricity prices on data center presence depends on public water supply. Electricity prices remain statistically indistinguishable from zero, while population size, GDP per capita, and water availability remain the key predictors of data center presence.

5.10 Robustness Checks: Sample Integrity and Electricity-Price Imputation

A potential concern in this analysis is that county-level electricity prices are constructed from ZIP-code tariff information and, in a limited number of cases, missing values are addressed through spatial imputation procedures. To ensure that the main results are not driven by counties with imputed electricity tariffs, I perform a set of sample integrity checks that re-estimate the baseline specification on progressively more restrictive subsamples.

The benchmark specification throughout this section corresponds to the full model including electricity prices, population size, GDP per capita, and public water supply:

$$\begin{aligned} Pr(D_c = 1) = \Lambda(\alpha + \beta_1 Electricity_c + \beta_2 \ln(Population_c) \\ + \beta_3 \ln(GDPpc_c) + \beta_4 \ln(Water_c)) \end{aligned} \tag{5.2}$$

estimated with heteroskedasticity-robust standard errors.

5.10.1 Excluding counties with imputed commercial electricity tariffs (within-county imputation)

First, I exclude counties flagged as having imputed commercial electricity prices (`county_comm_imputed`=1). This restriction reduces the sample from 3,144 to 3,117 counties. Results remain highly consistent with the full-sample estimates.

As shown in Table 5.28, in the full sample, commercial electricity prices are statistically insignificant ($\beta = -2.288$, $p = 0.179$), while population size, GDP per capita, and public water supply are positive and statistically significant ($\beta_{\ln(pop)} = 1.066$, $p < 0.001$; $\beta_{\ln(gdp\_pc)} = 1.157$, $p < 0.001$; $\beta_{\ln(water)} = 0.224$, $p = 0.006$). When counties with imputed commercial prices are excluded, the coefficient on electricity prices remains statistically insignificant ($\beta = -2.271$, $p = 0.182$), and the coefficients on population, GDP per capita, and water supply remain stable in sign, magnitude, and statistical significance. The model fit is also essentially unchanged (Pseudo R^2 around 0.40).

Table 5.28: Robustness check excluding counties with imputed commercial electricity tariffs.

	(1)		(2)	
	Full sample		No imputation (comm_imputed=0)	
	Coeff.	P-value	Coeff.	P-value
datacenter_presence				
avg_comm_rate_county	-2.288	0.179	-2.271	0.182
ln_population	1.066	0.000	1.064	0.000
ln_gdp_pc	1.157	0.000	1.162	0.000
ln_water	0.224	0.006	0.225	0.006
Pseudo R^2	0.401		0.401	
Observations	3144		3117	

Overall, excluding counties with imputed commercial tariffs does not alter the core conclusions.

5.10.2 Excluding fully approximated electricity-price counties (across-county imputation for both tariffs)

Second, I exclude counties flagged as fully approximated electricity-price observations (`electricity_approx_tot`=1), i.e., cases where both industrial and commercial tariffs are reconstructed across counties. Under this restriction, the commercial-

price model is estimated on 3,122 counties. The results (Table 5.29) remain very similar to the full sample: commercial electricity prices remain statistically insignificant ($\beta = -2.279$, $p = 0.181$), while population, GDP per capita, and water supply remain positive and highly significant, with comparable magnitudes.

Table 5.29: Robustness check excluding counties with fully approximated electricity prices.

	(1)		(2)	
	Full sample		No imputation (elecricity_approx_tot=0)	
	Coeff.	P-value	Coeff.	P-value
datacenter_presence				
avg_comm_rate_county	-2.288	0.179	-2.279	0.181
ln_population	1.066	0.000	1.064	0.000
ln_gdp_pc	1.157	0.000	1.161	0.000
ln_water	0.224	0.006	0.226	0.006
Pseudo R^2	0.401		0.401	
Observations	3144		3122	

This check suggests that the main findings are not driven by the small subset of counties for which electricity prices are fully reconstructed.

5.10.3 Industrial electricity prices: excluding fully approximated counties and neighbour-imputed industrial tariffs

I replicate the same integrity checks for industrial electricity prices in Table 5.30. In the full sample, industrial electricity prices are statistically insignificant ($\beta = -1.037$, $p = 0.446$), while population size, GDP per capita, and water supply remain positive and significant ($\beta_{\ln(pop)} = 1.055$, $p < 0.001$; $\beta_{\ln(gdp\_pc)} = 1.149$, $p < 0.001$; $\beta_{\ln(water)} = 0.224$, $p = 0.006$).

When counties with imputed industrial prices are excluded, the coefficient on electricity prices remains statistically insignificant ($\beta = -1.05$, $p = 0.453$), and the coefficients on population, GDP per capita, and water supply remain stable in sign, magnitude, and statistical significance. The model fit is also essentially unchanged (Pseudo R^2 around 0.39).

Table 5.30: Robustness check excluding counties with imputed industrial electricity tariffs.

	(1)		(2)	
	Full sample		No imputation (ind_imputed=0)	
	Coeff.	P-value	Coeff.	P-value
datacenter_presence				
avg_ind_rate_county	-1.037	0.446	-1.054	0.453
ln_population	1.055	0.000	1.090	0.000
ln_gdp_pc	1.149	0.000	1.139	0.000
ln_water	0.224	0.006	0.199	0.027
Pseudo R^2	0.400		0.390	
Observations	3144		2907	

In Table 5.31 fully approximated counties are excluded (`electricity_approx_tot`=0), the industrial electricity coefficient remains statistically insignificant ($\beta = -1.067$, $p = 0.434$), and the structural covariates remain stable.

Table 5.31: Robustness check excluding counties with fully approximated electricity prices.

	Full sample datacenter_presence		No imputation (ellectricity_approx_tot = 0) datacenter_presence	
	Coeff.	P-value	Coeff.	P-value
avg_ind_rate_county	-1.037	0.446	-1.067	0.434
ln_population	1.055	0.000	1.053	0.000
ln_gdp_pc	1.149	0.000	1.154	0.000
ln_water	0.224	0.006	0.226	0.006
_cons	-18.793	0.000	-18.792	0.000
Observations	3144		3122	
Pseudo R^2	0.400		0.400	

Next, I further exclude counties where industrial tariffs are imputed using neighbouring counties (`ind_neighbour_imputed`=1). Under this stricter restriction (sample size 3,103), industrial electricity prices remain statistically insignificant ($\beta = -1.240$, $p = 0.365$), while population, GDP per capita, and water supply remain positive and statistically significant as we can see in Table 5.32 below.

Table 5.32: Robustness check excluding counties with fully approximated and across-county imputed electricity prices.

	(1)		(2)	
	Full sample		No approx & no neigh-impute	
	Coeff.	P-value	Coeff.	P-value
datacenter_presence				
avg_ind_rate_county	-1.037	0.446	-1.240	0.365
ln_population	1.055	0.000	1.056	0.000
ln_gdp_pc	1.149	0.000	1.118	0.000
ln_water	0.224	0.006	0.241	0.004
Pseudo R^2	0.400		0.401	
Observations	3144		3103	

Finally in Table 5.33, excluding only neighbour-imputed industrial counties (`ind_neighbour_imputed`=0; $N = 3,125$) yields again a statistically insignificant electricity coefficient ($\beta = -1.211$, $p = 0.376$) and stable estimates for the remaining covariates.

Table 5.33: Robustness check excluding counties with across-county imputed electricity prices.

	(1)		(2)	
	Full sample		No imputation (ind_neighbour_imputed=0)	
	Coeff.	P-value	Coeff.	P-value
datacenter_presence				
avg_ind_rate_county	-1.037	0.446	-1.211	0.376
ln_population	1.055	0.000	1.058	0.000
ln_gdp_pc	1.149	0.000	1.112	0.000
ln_water	0.224	0.006	0.239	0.004
Pseudo R^2	0.400		0.401	
Observations	3144		3125	

Summary of robustness results. Across all restrictions, the coefficient on electricity prices - both commercial and industrial - remains statistically insignificant once population size, GDP per capita, and public water supply are controlled for. Conversely, population and GDP per capita remain consistently positive and highly significant across all samples, and public water supply remains positive and statistically significant. These integrity checks provide reassurance that the main results are not driven by counties with imputed electricity-price data, and that the conclusions are robust to alternative sample definitions.

5.10.4 Excluding counties with limited electricity-price coverage

As an additional robustness check shown in Table 5.34, I exclude counties where electricity-price information is largely incomplete. Specifically, I remove counties classified as `high_p3`, defined as those where more than 60% of ZIP–county records are priority-3 observations (i.e., cases where neither commercial nor industrial tariffs are reported).

These counties are characterized by a particularly low level of data availability in the underlying electricity dataset, and the corresponding county-level tariffs are therefore more likely to rely on imputation procedures based on neighbouring counties or within county imputation. Restricting the sample to counties with more complete tariff information provides a stricter test of the baseline results.

Table 5.34: Robustness check excluding counties with limited electricity-price coverage.

	Full sample		Excluding high_share == 1		Full sample		Excluding high_share == 1	
	datacenter_presence		datacenter_presence		datacenter_presence		datacenter_presence	
	Coeff.	P-value	Coeff.	P-value	Coeff.	P-value	Coeff.	P-value
ln_population	1.066	0.000	1.083	0.000	1.055	0.000	1.066	0.000
ln_water	0.224	0.006	0.242	0.004	0.224	0.006	0.242	0.004
avg_comm_rate_county	-2.288	0.179	-2.967	0.064				
ln_gdp_pc	1.157	0.000	1.120	0.000	1.149	0.000	1.114	0.000
avg_ind_rate_county					-1.037	0.446	-1.180	0.387
_cons	-18.748	0.000	-18.736	0.000	-18.793	0.000	-18.794	0.000
Observations	3144		3076		3144		3076	
Pseudo R^2	0.401		0.405		0.400		0.404	

The regression estimates remain broadly consistent with those obtained using the full sample. Population size, GDP per capita, and public water supply continue to display positive and statistically significant effects on the probability of hosting a data center across all specifications.

Commercial electricity prices remain statistically insignificant in the full sample ($\beta = -2.288$, $p = 0.179$), although the coefficient becomes marginally closer to statistical significance when counties with `high_share = 1` are excluded ($\beta = -2.967$, $p = 0.064$). Industrial electricity prices remain statistically insignificant in both specifications.

The magnitude and sign of the estimated coefficients are largely stable, and overall model fit changes only minimally, with the pseudo R^2 remaining around 0.40.

Overall, these results suggest that the baseline findings are robust to the exclusion of counties characterized by a high share of priority-3 observations, indicating that the main conclusions are not driven by areas with limited electricity-price coverage.

5.10.5 Robustness to Extreme Electricity Price Observations: Commercial Electricity Prices

Descriptive statistics reveal the presence of a small number of extreme values in the distribution of commercial electricity prices. In particular, a limited set of counties - primarily located in Alaska and Hawaii (state FIPS 02 and 15) - exhibit average commercial electricity prices above \$0.40 per kWh, with a maximum value of approximately \$0.55 per kWh. These observations are substantially higher than the national mean of \$0.13 per kWh and reflect the unique geographic and infrastructural characteristics of these states.

Table 5.35: Counties with commercial electricity prices above \$0.40 per kWh.

FIPS	Average Com Rate
02050	0.546
02070	0.546
02158	0.546
02188	0.546
02185	0.546
02180	0.546
02290	0.460
15001	0.446
15005	0.438
15009	0.438
02282	0.423

Although these counties represent less than 0.5% of the total sample, their extreme values may influence the estimated marginal effects in nonlinear models. To ensure that the main results are not driven by these outliers, the baseline specification is re-estimated after excluding counties with commercial electricity prices above \$0.40 per kWh.

The results remain in Table 5.36 qualitatively unchanged. Commercial electricity prices continue to exhibit a statistically insignificant association with the probability of hosting a data center, while population size, GDP per capita, and water availability remain positive and statistically significant determinants. This confirms that the main findings are not driven by extreme price observations in geographically isolated regions.

Table 5.36: Logistic regression estimates excluding counties with commercial electricity prices above $0.40 per kWh.

	(1)		(2)	
	Baseline		Excl. High-Price Counties	
	AME	P-value	AME	P-value
avg_comm_rate_county	-0.134	0.179	-0.211	0.025
ln_population	0.063	0.000	0.063	0.000
ln_gdp_pc	0.068	0.000	0.066	0.000
ln_water	0.013	0.006	0.013	0.005
Observations	3144		3133	

5.10.6 Robustness to Extreme Electricity Price Observations: Industrial Electricity Prices

To assess whether extreme values in industrial electricity prices influence the estimated marginal effects, the baseline specification is re-estimated after excluding counties with average industrial electricity prices above $0.40 per kWh. Only one county in the sample exceeds this threshold.

The results displayed in Table 5.37 remain virtually unchanged. The average marginal effect of industrial electricity prices is negative but statistically insignificant in both the full sample (AME = -0.061, p = 0.446) and the restricted sample excluding the high-price county (AME = -0.060, p = 0.453).

Population size and GDP per capita continue to exhibit positive and highly statistically significant marginal effects, while water availability remains positive and statistically significant. The stability of the estimates confirms that the main findings are not driven by extreme industrial electricity price observations.

Table 5.37: Logistic regression estimates excluding counties with commercial electricity prices above $0.40 per kWh.

	Baseline		Excl. High-Price Counties	
	AME	P-value	AME	P-value
avg_ind_rate_county	-0.061	0.446	-0.060	0.453
ln_population	0.062	0.000	0.062	0.000
ln_gdp_pc	0.068	0.000	0.068	0.000
ln_water	0.013	0.006	0.013	0.006
Observations	3144		3143	

5.10.7 Correlation structure

To assess potential multicollinearity among the main explanatory variables, I compute pairwise correlations between the core covariates used in the regressions in Figure 5.11. The results show that most correlations are relatively small.

Population size is strongly correlated with public water supply ($\rho = 0.81$), which is expected since larger counties naturally require greater water infrastructure. Commercial and industrial electricity prices are also highly correlated ($\rho = 0.76$), reflecting the common underlying electricity cost structure across sectors within each county.

In contrast, GDP per capita exhibits very low correlation with population size ($\rho = 0.03$), indicating that county economic productivity is largely independent of county scale. Electricity prices display only weak correlations with the other covariates, suggesting that multicollinearity is unlikely to be a concern in the regression analysis.

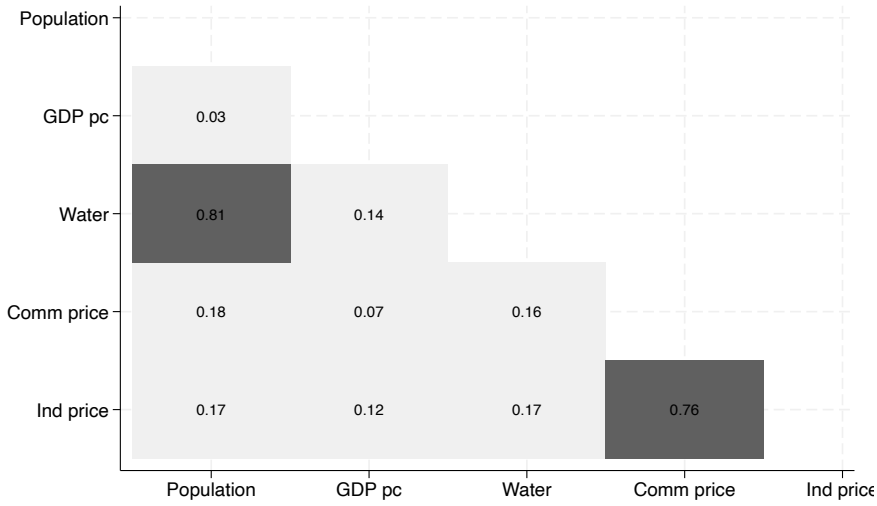

Figure 5.11: Correlation Heatmap between the core covariates used in the regressions.

5.10.8 Multicollinearity diagnostics

As discussed in the data section, public water supply is used as a proxy for local water availability. Since water infrastructure typically scales with the population served, a potential concern is that the variable may be correlated with county population.

Pairwise correlations confirm a relatively strong positive association between

ln_population and ln_water. To ensure that this relationship does not generate problematic multicollinearity in the regression model, variance inflation factors (VIF) are computed using auxiliary OLS regressions. The VIF measures how much the variance of a coefficient is inflated due to linear dependence among explanatory variables and is defined as $VIF_j = 1/(1 - R_j^2)$, where R_j^2 is obtained from a regression of variable j on all other regressors.

Table 5.38 reports the results. The highest VIF values are observed for ln_population and ln_water (around 3), reflecting their expected relationship. However, all VIF values remain well below commonly used thresholds (typically 5 or 10), indicating that multicollinearity is not a concern in the empirical specification. Consequently, the estimated coefficients can be interpreted without concerns that inflated standard errors or unstable estimates are driven by collinearity among regressors.

Table 5.38: Variance Inflation Factors

Variable	Commercial price model	Industrial price model
ln_population	3.00	2.99
ln_gdp_pc	1.05	1.06
ln_water	3.03	3.03
Electricity price	1.04	1.05
Mean VIF	2.03	2.03

5.11 Intensive Margins Analysis

5.11.1 Distribution of Data Center Counts

Before estimating the intensive-margin models, it is useful to examine the distribution of the number of data centers across U.S. counties. Table 5.39 reports descriptive statistics for the variable *datacenter_number*, which measures the total number of data centers located in each county.

Table 5.39: Distribution of data center counts

	N	Mean	SD	Min	Median	P75	P90	P95	P99	Max
datacenter_number	3144	1.05	7.71	0	0	0	1	2	28	168

The distribution is highly skewed and characterized by a large mass of observations at zero. The median county hosts no data centers, reflecting the fact that data center infrastructure is geographically concentrated in a relatively small subset of

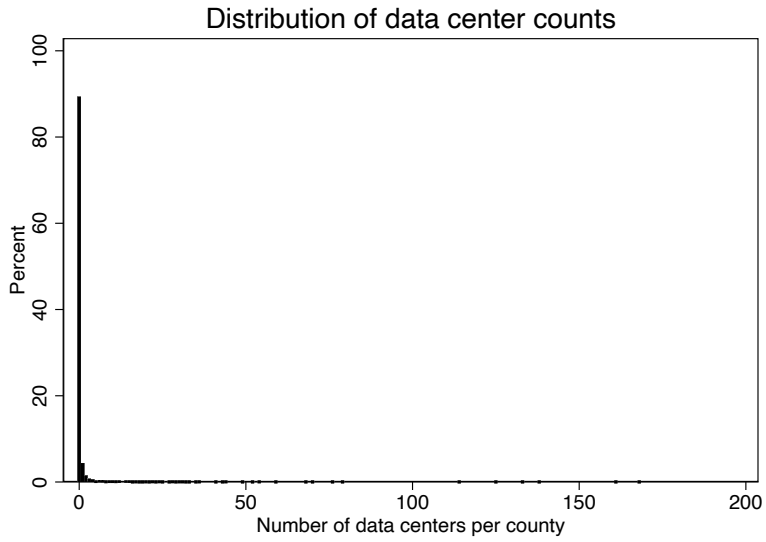

Figure 5.12: Distribution of data center counts across U.S. counties

counties. At the same time, the mean is higher than the median due to the presence of a small number of counties hosting multiple facilities.

This pattern is also visible in Figure 5.12, which shows the histogram of data center counts across counties. The distribution displays a strong right skew, with most observations clustered at very low values and a long tail extending toward counties hosting large data center clusters. A few counties host more than one hundred data centers, which significantly increases the dispersion of the variable.

Figure 5.13 further illustrates the presence of extreme outliers through a boxplot representation. While the majority of counties host very few or no data centers, a limited number of locations concentrate a large share of the infrastructure. This pattern is consistent with the well-known tendency of digital infrastructure to cluster in specific metropolitan or economically attractive areas.

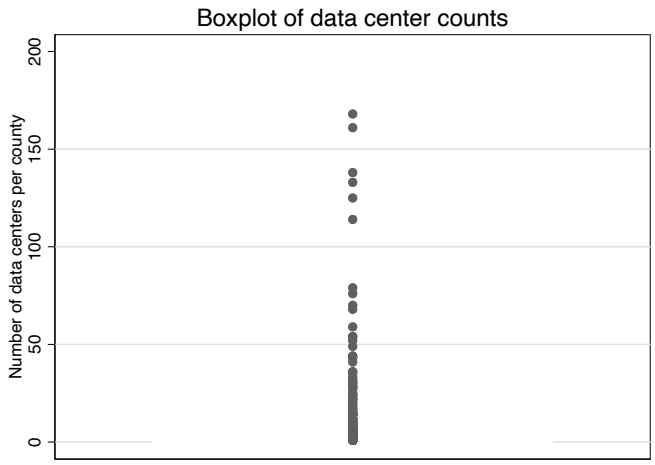

Figure 5.13: Boxplot of data center counts

Overall, the strong right-skewness and the large share of zero observations suggest that the number of data centers is best modeled using count-data techniques rather than standard linear regression. For this reason, the intensive-margin analysis that follows relies on Poisson and negative binomial models, which are well suited to handle non-negative integer outcomes and highly skewed distributions.

Poisson regressions for Data Center Counts

This section studies the intensive margin of data center location by modelling the number of data centers hosted in each county (`datacenter_number`). Since the outcome is a non-negative count variable with a highly right-skewed distribution and a large mass at zero, Poisson regression models are estimated. All specifications use heteroskedasticity-robust standard errors. Two alternative electricity-price measures are considered: the average commercial tariff and the average industrial tariff.

Table 5.40 reports Poisson coefficients and p-values. Across both electricity-price specifications, population size and GDP per capita are strongly and positively associated with data center counts, indicating that larger and more economically developed counties tend to host a higher number of facilities. In contrast, the water proxy (`ln_water`) is not statistically significant once scale and income are controlled for, suggesting that water availability may matter more for the extensive margin (whether a county hosts at least one data center) than for the intensity of clustering within hosting counties.

Table 5.40: Poisson regressions for data center counts (coefficients)

	Commercial rate		Industrial rate	
datacenter_number				
ln_population	1.131	0.000	1.100	0.000
ln_gdp_pc	1.250	0.000	1.204	0.000
ln_water	0.077	0.476	0.093	0.398
avg_comm_rate_county	-9.266	0.000		
avg_ind_rate_county			-6.916	0.000
Constant	-17.551	0.000	-17.569	0.000
Observations	3144		3144	
Log-likelihood	-4015.321		-4127.140	

Electricity prices enter with a negative and statistically significant coefficient in both models, consistent with higher operating costs discouraging data center concentration. Because tariffs are expressed in \$/kWh, coefficient magnitudes are best interpreted through incidence-rate ratios (IRR). Table 5.41 reports IRRs: for example, a one-cent increase in the electricity tariff (i.e., +0.01 \$/kWh) implies a multiplicative change in expected counts of $\exp(\beta \cdot 0.01)$, corresponding to an economically interpretable percentage effect.

Table 5.41: Poisson regressions for data center counts (IRR)

	Commercial rate	Industrial rate
datacenter_number		
ln_population	3.099	3.005
ln_gdp_pc	3.492	3.334
ln_water	1.080	1.097
avg_comm_rate_county	0.000	
avg_ind_rate_county		0.001
Observations	3144	3144

Exponentiated coefficients

Finally, goodness-of-fit diagnostics reject the equidispersion assumption of the Poisson model, consistent with substantial overdispersion in the data. For this reason, negative binomial specifications are estimated as robustness checks.

Goodness-of-fit diagnostics were performed using both deviance and Pearson tests. For both electricity price specifications, the tests strongly reject the Poisson equidispersion assumption (p-value < 0.001), indicating substantial overdispersion in the data. This result is consistent with the highly skewed distribution of data center counts across counties and suggests that the variance exceeds the conditional mean.

For this reason, negative binomial specifications are considered as additional robustness checks.

Negative Binomial Regressions

The goodness-of-fit diagnostics for the Poisson models indicate substantial overdispersion in the data, as both the deviance and Pearson tests strongly reject the equidispersion assumption.

To account for this issue, negative binomial regressions are estimated. The negative binomial model generalizes the Poisson specification by introducing an additional dispersion parameter:

$$Var(Y|X) = E(Y|X) + \alpha E(Y|X)^2$$

where α represents the dispersion parameter. When $\alpha > 0$, the data exhibit overdispersion and the negative binomial model provides a more appropriate specification than Poisson regression.

Table 5.42 reports the estimated coefficients and p-values for the negative binomial models.

Across both specifications, population size and GDP per capita remain positively and statistically significant determinants of the number of data centers. Counties with larger populations and higher levels of economic development tend to host a greater number of facilities, consistent with the role of market size and digital infrastructure demand.

Electricity prices display a negative and statistically significant association with data center counts. Higher electricity costs are associated with fewer data centers, reflecting the energy-intensive nature of data center operations and the importance of electricity prices in shaping the spatial distribution of digital infrastructure.

In contrast, the water supply proxy does not exhibit a statistically significant effect on the number of data centers once other county characteristics are controlled for. This result is consistent with previous findings suggesting that water availability may be more relevant for the initial location decision rather than for the intensity of data center clustering within counties.

Finally, the estimated dispersion parameter α is large and statistically significant in both specifications, confirming the presence of strong overdispersion in the data and supporting the use of the negative binomial model.

Table 5.42: Negative Binomial regressions for data center counts

	Commercial rate		Industrial rate	
datacenter_number				
ln_population	1.599	0.000	1.561	0.000
ln_gdp_pc	1.911	0.000	1.945	0.000
ln_water	-0.488	0.137	-0.519	0.148
avg_comm_rate_county	-9.935	0.000		
avg_ind_rate_county			-5.038	0.000
Constant	-24.196	0.000	-24.590	0.000
/				
lnalpha	1.895	0.000	1.970	0.000
Observations	3144.000		3144.000	
Log-likelihood	-1603.247		-1622.155	

To facilitate economic interpretation, Table 5.43 reports incidence rate ratios (IRR). These values indicate the multiplicative change in the expected number of data centers associated with a one-unit increase in the explanatory variables.

The IRR estimates confirm the strong role of county size and economic development. Population and GDP per capita substantially increase the expected number of data centers, while higher electricity prices reduce data center concentration. The magnitude of the electricity price effect should be interpreted relative to realistic tariff variations (e.g., changes of a few cents per kWh), given the units in which electricity prices are measured.

Table 5.43: Negative Binomial regressions (Incidence Rate Ratios)

	Commercial rate	Industrial rate
datacenter_number		
ln_population	4.948	4.764
ln_gdp_pc	6.760	6.994
ln_water	0.614	0.595
avg_comm_rate_county	0.000	
avg_ind_rate_county		0.006
/		
lnalpha	6.655	7.169
Observations	3144.000	3144.000

Exponentiated coefficients

5.12 Geographic Concentration of Data Centers Across U.S. States

To provide an additional descriptive perspective on the geographic distribution of digital infrastructure, Figure 5.14 reports the top 10 U.S. states hosting the largest number of data centers. The results reveal a strong spatial concentration of data center infrastructure across a limited number of states.

Virginia clearly dominates the ranking, reflecting the presence of the Northern Virginia data center cluster, one of the largest digital infrastructure hubs globally. Texas and California also host a large number of facilities, followed by states such as Illinois, Ohio, and Georgia. These states combine large economic markets, strong connectivity infrastructure, and favorable conditions for digital infrastructure deployment.

Overall, the distribution highlights the tendency of data centers to cluster in specific locations characterized by strong economic activity and advanced digital ecosystems. This descriptive evidence is consistent with the econometric results presented earlier, which identify market size and economic development as key determinants of data center location.

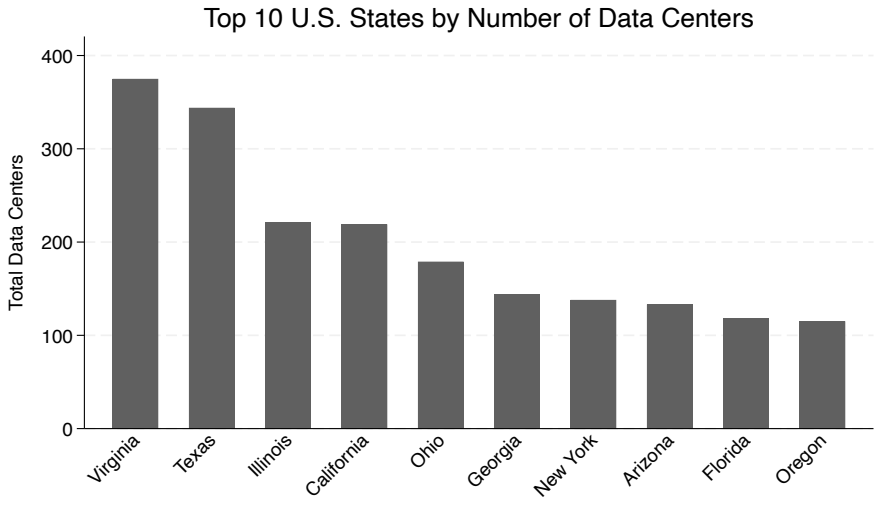

Figure 5.14: Top 10 U.S. states by number of data centers.

Chapter 6

Conclusions and Policy Implications

6.1 Summary of Findings

This thesis examined the determinants of data center location across U.S. counties, with the objective of understanding how economic, demographic, and infrastructural characteristics are associated with the spatial distribution of digital infrastructure.

To address this question, a novel county-level dataset was constructed by integrating multiple publicly available sources on electricity prices, population, economic activity, and water supply infrastructure. A significant part of the research involved the harmonization and validation of these datasets, which required aligning different geographic units, reconstructing missing identifiers such as FIPS codes, and implementing consistency checks across sources. This data construction process allowed the creation of a unified dataset covering 3,144 U.S. counties and combining economic and environmental indicators that are rarely analyzed jointly in the literature.

The empirical analysis distinguishes between two dimensions of data center location: the *extensive margin*, which captures whether a county hosts at least one data center, and the *intensive margin*, which examines the number of data centers located within counties that host such infrastructure.

The extensive margin is analyzed using logistic regression models, where the dependent variable indicates the presence of at least one data center. The results consistently show that population size and GDP per capita are the strongest predictors of data center presence. Counties with larger populations and higher levels of economic development are significantly more likely to host data center facilities. These findings suggest that agglomeration forces, market size, and the availability of complementary infrastructure play a central role in shaping the spatial distribution of digital infrastructure.

Water supply infrastructure, proxied through public water withdrawals, also exhibits a positive and statistically significant association with the probability of hosting a data center. Although this variable does not directly measure cooling water use, it captures broader infrastructure capacity and highlights the importance of resource availability for large-scale computing facilities.

In contrast, electricity prices do not exhibit a statistically significant relationship with the probability that a county hosts at least one data center once population size and economic development are taken into account. This suggests that the initial location of data center facilities is primarily associated with structural regional characteristics rather than local electricity cost differences.

The intensive margin analysis provides a complementary perspective. Because the distribution of data center counts across counties is highly skewed and characterized by a large share of zero observations, count-data models are used to examine how explanatory variables relate to the number of data centers located in each county. In these models, electricity prices enter with a negative and statistically significant coefficient, indicating that higher electricity costs are associated with lower levels of data center concentration. This result suggests that while electricity prices may not determine whether a county hosts a data center in the first place, they may influence the degree to which data centers cluster within already established locations.

Taken together, the empirical results indicate that structural characteristics such as population size, economic development, and infrastructure capacity play the dominant role in determining where data centers are initially located, while electricity costs may become more relevant for the intensity of infrastructure clustering once locations are established.

6.2 Policy Implications

The findings of this thesis carry several implications for energy and infrastructure policy.

First, the strong relationship between population, economic development, and data center presence highlights the importance of regional infrastructure and connectivity. Data centers tend to locate in economically developed areas where digital networks, skilled labor, and complementary services are already available. This suggests that the geography of digital infrastructure is closely linked to broader regional development patterns.

Second, the intensive-margin results indicate that electricity prices may influence

the concentration of data centers within existing clusters. As computing demand continues to expand—particularly with the growing adoption of artificial intelligence applications—regions with access to stable and competitively priced electricity may become increasingly attractive for large-scale infrastructure expansion.

These dynamics may place additional pressure on local electricity systems. Policymakers may therefore need to consider the interaction between digital infrastructure expansion and electricity system planning, particularly in regions experiencing rapid growth in computing demand.

Water infrastructure represents another relevant dimension. Although the water variable used in this study is only a proxy for infrastructure availability, the results suggest that access to sufficient water supply systems may represent an enabling condition for large computing facilities. This highlights the importance of integrating infrastructure planning with resource management considerations, particularly in regions where water scarcity may become more pronounced.

6.3 Relevance for the European Context

Although the empirical analysis focuses on the United States, the results offer insights that are relevant for the European policy debate on digital infrastructure sustainability.

In Europe, electricity prices, renewable energy availability, and grid carbon intensity vary significantly across countries and regions. As AI-driven computing demand continues to grow, these differences may increasingly influence where new data center capacity is developed.

At the same time, the European Union has begun introducing regulatory frameworks aimed at improving transparency regarding the environmental footprint of digital infrastructure. Instruments such as the Corporate Sustainability Reporting Directive (CSRD) and the Energy Efficiency Directive (EED) introduce reporting requirements related to energy use and sustainability indicators, while the AI Act is expected to introduce additional disclosure obligations for certain AI systems.

However, a major challenge remains the limited availability of standardized data on the environmental footprint of digital infrastructure. As highlighted throughout this thesis, even in the United States it is difficult to obtain detailed county-level information on electricity consumption, cooling technologies, or water usage associated with data centers. Improving transparency and reporting standards will therefore be essential for enabling more precise analysis of the environmental implications of

AI-driven infrastructure expansion.

6.4 Limitations

Several limitations should be acknowledged.

First, the dataset does not distinguish between different types of data centers. Facilities dedicated primarily to artificial intelligence workloads cannot be separated from more general cloud infrastructure due to the lack of publicly available data at the facility level.

Second, the analysis is constrained by the limited availability of environmental variables at the county level in the United States. While water supply withdrawals were used as a proxy for infrastructure capacity, other potentially relevant environmental indicators—such as water stress, local climate conditions, renewable energy availability, or grid carbon intensity—are difficult to obtain with consistent coverage across all U.S. counties.

6.5 Future Research

Future research could extend this work in several directions.

One promising avenue would be the construction of panel datasets tracking the evolution of data center infrastructure over time. This would allow researchers to analyze dynamic location patterns and evaluate the impact of changes in electricity markets, infrastructure investments, or regulatory frameworks.

Another important direction concerns the integration of more detailed environmental variables. Incorporating indicators such as renewable energy availability, local climate conditions, water stress indices, or grid carbon intensity could provide a more comprehensive understanding of the environmental determinants of data center location.

Finally, improvements in transparency and reporting standards may eventually allow researchers to distinguish between AI-specific data centers and more general-purpose cloud infrastructure. Such data would enable a more precise assessment of how artificial intelligence workloads are shaping the geography and environmental footprint of digital infrastructure.

6.6 Final Remarks

The rapid expansion of digital infrastructure and artificial intelligence systems is transforming the scale and resource requirements of data centers worldwide. Understanding the determinants of where such infrastructure is located is therefore increasingly important for both policymakers and researchers.

By constructing a harmonized county-level dataset and analyzing both the extensive and intensive margins of data center location, this thesis contributes to the growing literature on the economic geography of digital infrastructure. The results highlight the central role of regional economic development, population scale, and infrastructure capacity, while also indicating that electricity prices may influence the clustering of facilities within established locations.

As digital infrastructure continues to expand, the interaction between technological development, energy systems, and environmental sustainability will remain a critical area for future research and policy design.

References

[1] International Energy Agency (IEA). «Electricity 2024 – Analysis and Forecast to 2026». Paris: IEA, 2024. [Online]. Available: https://www.iea.org/reports/electricity-2024

[2] Kamiya, G., & Bertoldi, P. «Energy Consumption in Data Centres and Broadband Communication Networks in the EU». Publications Office of the European Union, 2024. DOI: 10.2760/706491. Available: https://data.europa.eu/doi/10.2760/706491

[3] Li, P., Yang, J., Islam, M. A., & Ren, S. «Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models». arXiv preprint arXiv:2304.03271, 2023. DOI: 10.48550/arXiv.2304.03271.

[4] Luccioni, A., Patterson, D., Thompson, M., & Henderson, P. «Power Hungry Processing: Watts Driving the Cost of AI Deployment?». 2024.

[5] Amazon Web Services. «2023 Amazon Sustainability Report». Seattle: Amazon, 2023. Available: https://sustainability.aboutamazon.com/2023-amazon-sustainability-report.pdf

[6] Microsoft. «2024 Environmental Sustainability Report». Redmond: Microsoft, 2024. Available: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Microsoft-2024-Environmental-Sustainability-Report.pdf

[7] Google. «Google Environmental Report 2024». Mountain View: Google, 2024. Available: https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf

[8] European Parliament & Council of the European Union. «Directive (EU) 2023/1791 on energy efficiency and amending Regulation (EU) 2023/955». Official Journal of the European Union, L 231/1, 2023. Available: https://eur-lex.europa.eu

[9] Bashir, N., Donti, P., Cui, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C., & Olivetti, E. «The climate and sustainability implications of generative AI». MIT Climate and Sustainability Consortium, 2024.

[10] Microsoft. «Microsoft and OpenAI Extend Partnership to Build AI Supercomputing Infrastructure on Azure». 2023. Available: https://blogs.microsoft.com/blog/2023/01/23/microsoft-and-openai-extend-partnership/

[11] Wired. «OpenAI's Future Runs on Microsoft's Cloud». 2024. Available: https://www.wired.com/story/openai-microsoft-cloud-infrastructure/

[12] Reuters. «Amazon and OpenAI Announce $38 Billion Cloud-Computing Partnership». 2025. Available: https://www.reuters.com/business/retail-consumer/amazons-38-bln-openai-deal-shows-it-is-no-longer-an-ai-laggard-2025-11-04/

[13] The Guardian. «OpenAI Signs Multi-Billion-Dollar AWS Deal to Expand Compute Infrastructure». 2025. Available: https://www.theguardian.com/technology/2025/nov/03/openai-cloud-computing-deal-amazon-aws-datacentres-nvidia-chips **[14]** Financial Times. (2024). *AI and the surge in data centre demand*. Available at: https://www.ft.com/content/e85e43d1-5ce4-4531-94f1-9e9c1c5b4ff1

[15] Goldman Sachs. (2024). *AI to drive 165% increase in data center power demand by 2030*. Available at: https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030

[16] MIT Sloan School of Management. (2024). *AI has high data center energy costs. There are solutions*. March 7. Available at: https://mitsloan.mit.edu/ideas-made-to-matter/ai-has-high-data-center-energy-costs-there-are-solutions

[17] Yale Environment 360. (2024). *Artificial intelligence and its climate and energy implications*. Available at: https://e360.yale.edu/features/artificial-intelligence-climate-energy-emissions

[18] Des Moines Register. (2023). *ChatGPT was built in Iowa using artificial intelligence*. Available at: https://eu.desmoinesregister.com/story/money/business/2023/09/10/chatgpt-was-built-in-iowa-using-artificial-intelligence-microsoft-west-des-moines/70819093007/

[19] The Information. «Microsoft and OpenAI Plan $100 Billion AI Supercomputer, Codenamed Stargate». April 2024. Available: https://www.theinformation.com

[20] Reuters. «Microsoft and OpenAI Planning $100 Billion Data Center, Supercomputer Project 'Stargate'». April 2024. Available: https://www.reuters.com

[21] Bloomberg. «Microsoft Building Massive AI Supercomputer Called Stargate». April 2024. Available: https://www.bloomberg.com

[22] Bharambe, P., Satkar, A., Mahandule, V., and Mahajan, K. (2024). *Sustainable Optimization of Data Processing Waste in Distributed Systems: Addressing Energy Consumption and Electronic Waste in Data-Intensive Applications*. MIT Arts, Commerce & Science College, Alandi, Pune, India.

[23] Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. (2017). *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-*

of-Experts Layer. Available at: https://arxiv.org/abs/1701.06538

[24] Omdia. (2024). *Best Data Center Locations 2024*. Available at: https://omdia.tech.informa.com/blogs/2024/may/best-data-center-locations-2024

[25] Huggins, J. (2024). *U.S. Electric Utility Companies and Rates: Look-up by Zipcode (2023)* [Data set]. Open Energy Data Initiative (OEDI), National Renewable Energy Laboratory (NREL). Available at: https://data.openei.org/submissions/6225

[26] U.S. Department of Housing and Urban Development (HUD). (2024). *USPS ZIP Code Crosswalk Files – Q4 2024* [Data file]. Available at: https://www.huduser.gov/portal/datasets/usps_crosswalk.html

[27] U.S. Energy Information Administration (EIA). (2019). *Investor-owned utilities served 72% of U.S. electricity customers in 2017*. Available at: https://www.eia.gov/todayinenergy/detail.php?id=40913

[28] Ivahnenko, M. A., Lovelace, T. I., Maupin, J. K., & Barber, N. L. (2018). *Estimated Use of Water in the United States: County-Level Data for 2015* (Version 2.0, June 2018) [Data set]. U.S. Geological Survey. https://doi.org/10.5066/F7TB15V5

[29] U.S. Census Bureau. (2023). Available at: https://www.census.gov/programs-surveys/acs/technical-documentation/user-notes/2023-01.html

[30] Federal Register. (2022). *Change to County Equivalents in the State of Connecticut*. Available at: https://www.federalregister.gov/documents/2022/06/06/2022-12063/change-to-county-equivalents-in-the-state-of-connecticut

[31] U.S. Census Bureau. (2019). *Geography Changes: 2019 ACS — County and Statistical Area Updates*. Available at: https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2019/geography-changes.html

[32] U.S. Census Bureau. (2023). *Methodology for the United States Population Estimates: Vintage 2023 — Nation, States, Counties, and Puerto Rico (April 1, 2020 to July 1, 2023)*. PDF saved for internal reference.

[33] DataCenterMap. (2024). *USA Data Centers*. Available at: https://www.datacentermap.com/usa/

[34] SimpleMaps. (2024). *US Cities Database*. Available at: https://simplemaps.com/data/us-cities

[35] U.S. Bureau of Economic Analysis (BEA). (2024). *Gross Domestic Product by County and Metropolitan Area, 2023*. Available at: https://www.bea.gov/news/2024/gross-domestic-product-county-and-metropolitan-area-2023

Appendix

.1 Geographic Harmonization of Connecticut Counties

Figure 1 illustrates the relationship between the former county boundaries and the new planning regions in Connecticut. This adjustment ensures that the geographic units used in the empirical analysis remain consistent with the broader county-level dataset.

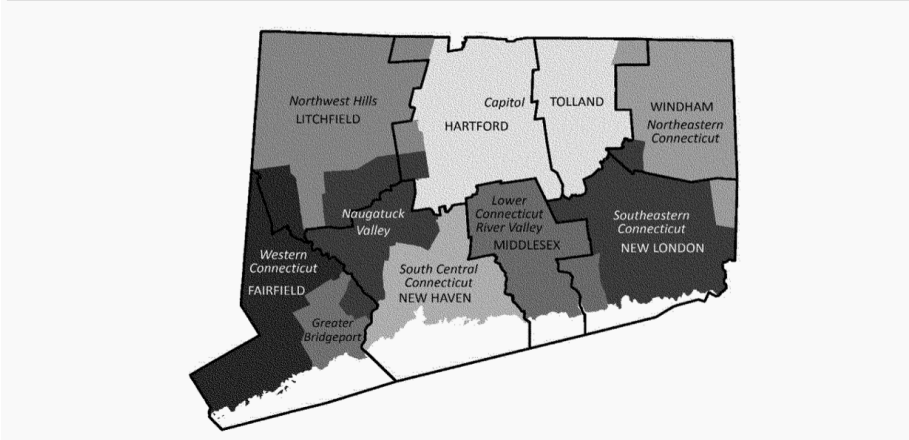

Figure 1: Relationship between historical counties and planning regions in Connecticut.

.2 Summary Statistics of Log-Transformed Variables

This appendix reports summary statistics for the log-transformed explanatory variables used in the empirical analysis.

Table 1: Summary Statistics of Log-Transformed Variables

	(1)						
	count	mean	sd	min	max	skewness	kurtosis
ln_population	3144	10.27583	1.522404	3.7612	16.08385	.2590354	3.350952
ln_gdp_pc	3144	3.781701	.587827	-5.573827	12.17022	.6986913	41.1694
ln_water	3144	1.476747	1.196729	0	7.136833	1.203987	4.351168
Observations	3144						

.3 Data Quality, Imputation, and Validation Checks

.3.1 Consistency between Imputation Flags and Imputation Shares

Table 2: Consistency between electricity approx. and imputation share.

	High share = 0	High share = 1
Approximation flag = 0	3076 (97.84%)	46 (1.46%)
Approximation flag = 1	0 (0.00%)	22 (0.70%)

Table 3: Consistency between across county approx. and imputation share.

	High share = 0	High share = 1
Approximation flag = 0	3076 (97.84%)	49 (1.56%)
Approximation flag = 1	0 (0.00%)	19 (0.60%)

Table 4: Consistency check for high-share imputation flag.

Check	Count
High_share=1 but share_ind ≤ 0.6 (violations)	0
High_share=0 but share_ind > 0.6 (violations)	0
Missing in either variable	0

.3.2 Counties Affected by Electricity Price Imputation

Table 5: Counties with across-county imputation for industrial tariffs (N=19)

FIPS	Datacenter presence	Datacenter count
02050	0	0
02070	0	0
02100	0	0
02158	0	0
02180	0	0
02185	1	1
02188	0	0
02198	0	0
02220	0	0
02230	0	0
02275	0	0
11001	1	7
24037	0	0
36103	0	0
46039	0	0
46051	0	0
46057	0	0
46109	0	0
48281	0	0

Table 6: Counties with fully approximated electricity data (N=22)

FIPS	Datacenter presence	Datacenter count
02013	0	0
02016	0	0
02060	0	0
02164	0	0
02195	0	0
02282	0	0
12033	0	0
27031	0	0
27077	0	0
31013	0	0
31015	0	0
31045	0	0
31065	0	0
31103	0	0
31165	0	0
46007	0	0
46075	0	0
46117	0	0
48009	0	0
48023	0	0
48103	0	0
48475	0	0

.3.3 Counties with the highest number of data centers

Table 7: Counties with the highest number of datacenters

FIPS	Datacenter count	Presence
48113	168	1
17031	161	1
51107	138	1
13121	133	1
04013	125	1
39049	114	1
51683	79	1
19153	76	1
06037	70	1
36081	68	1
51760	59	1
53033	54	1
17089	54	1
27053	52	1
48201	49	1

.3.4 Data Center Presence in Imputed Counties

Table 8: Data Center Presence in Fully Approximated Counties

	(1)
	datacenter_presence
	b
0	22
Total	22
N	22

Table 9: Data Center Presence in Neighbour-Imputed Counties

	(1)
	datacenter_presence
	b
0	17
1	2
Total	19
N	19

.4 Distribution of electricity prices

Electricity prices are strongly right-skewed, with a small number of counties exhibiting substantially higher tariffs than the national distribution.

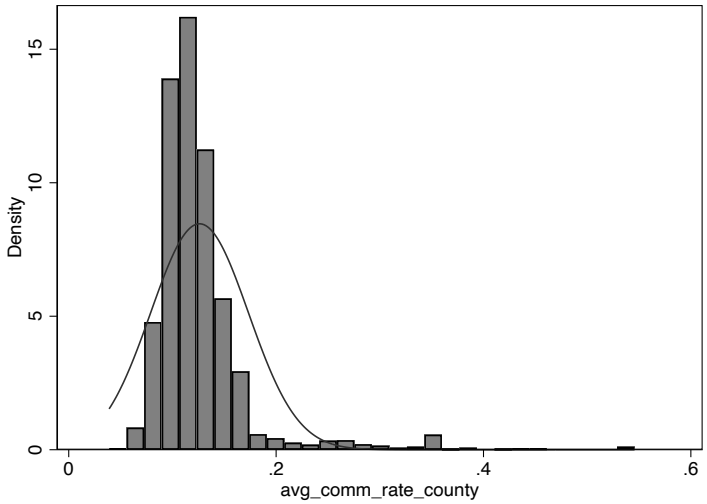

Figure 2: Distribution of average commercial electricity prices across counties.

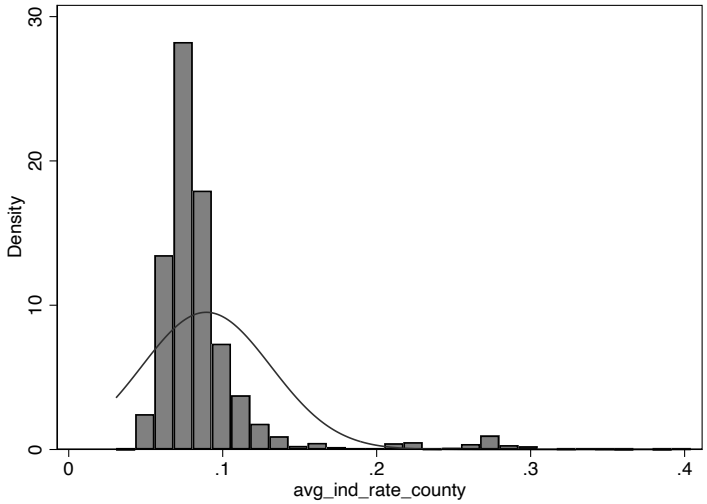

Figure 3: Distribution of average industrial electricity prices across counties.

.5 Infrastructure scaling

Figure 4 shows the relationship between public water supply and county population (both expressed in logarithmic form). The relationship is strongly positive and close

to linear, indicating that water infrastructure scales with population size.

This pattern reflects the fact that larger counties require greater infrastructure capacity to serve residential, commercial, and industrial users. In the context of data centre location, water supply may also capture the availability of cooling resources, which are relevant for data centre operations.

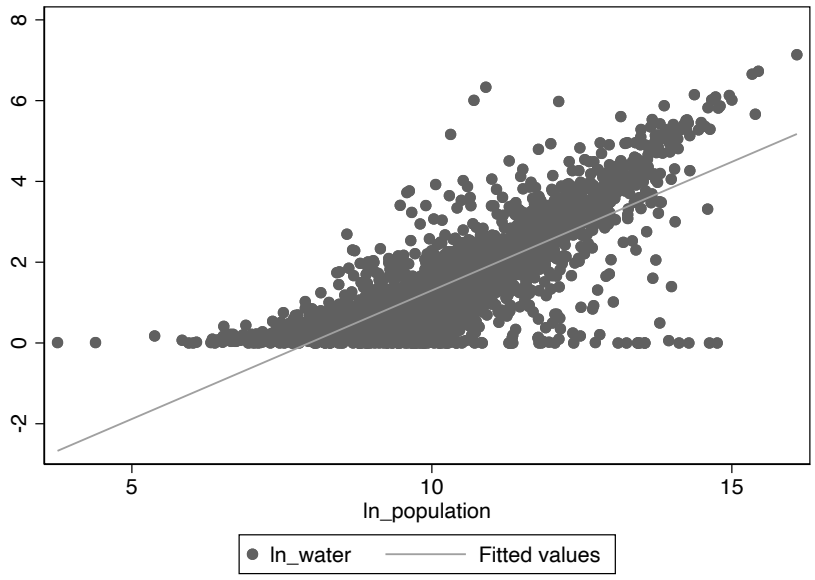

Figure 4: Public water supply and county population.