Spatial Determinants of Data Center Location in Europe: A NUTS3-Level Empirical Analysis

Data Reservoirs
Department of Georesources and Geoenergy Engineering.
Politecnico di Torino
Turin, Italy

Supervisor: Prof. Daniele Quercia
Dept. of Control and Computer Engineering
Politecnico di Torino, Turin, Italy

ABSTRACT

The rapid expansion of artificial intelligence has dramatically increased global demand for data center infrastructure. This paper investigates the spatial determinants of data center location across European NUTS3 regions, extending the methodological framework of Matilde work for U.S. counties to a novel geographic context. Using a rich dataset of socioeconomic, infrastructure, connectivity, environmental, and energy variables, we frame the problem as binary classification and regression. We apply interpretable models via statsmodels (logistic regression, pseudo $R^2 = 0.279$, LLR $p < 10^{-69}$) and scikit-learn (logistic regression, Random Forest, Gradient Boosting). Our key finding is that three features — GDP (gdp_eur_million), IXP count, and population — explain 56.6% of variance in data center counts via simple linear regression, outperforming all ensemble methods. Electricity prices are statistically significant and negative ($p = 0.041$) in the interpretable logistic model. We discuss transatlantic similarities and differences and draw policy implications for sustainable digital infrastructure.

CCS Concepts: • Information systems → Geographic information systems; • Applied computing → Economics; • Computing methodologies → Machine learning.

Keywords: data centers, NUTS3, machine learning, spatial econometrics, IXP, GDP, renewable energy, Europe, digital infrastructure

1. INTRODUCTION

The proliferation of cloud computing and AI-driven applications has turned data centers into one of the most strategically important and resource-intensive elements of modern digital economies. Globally, data centers consumed between 1–1.5% of electricity in 2022 [1], a share expected to grow significantly as large language models expand. In Europe, data centers accounted for an estimated 45–65 TWh in 2022 — roughly 2% of total electricity consumption, with Ireland alone reaching up to 18% [2].

Despite their economic and environmental significance, the spatial determinants of data center location remain underexplored in the academic literature, particularly in the European context. This paper addresses that gap by constructing a NUTS3-level dataset integrating socioeconomic, infrastructure, connectivity, environmental, and energy variables, and applying a supervised machine learning and econometric pipeline to identify the drivers of data center presence and count.

Our work is directly inspired by Matilde work [3], which applies logistic and count-data regression at the U.S. county level. We adopt a comparable methodology including statsmodels-based interpretable logistic regression — and discuss results in direct contrast, enabling a transatlantic comparison of the geography of digital infrastructure.

2. LITERATURE REVIEW AND HYPOTHESES

A growing body of research links data center siting to electricity cost, water availability, physical connectivity, climate, and broader economic geography. Omdia (2024) [4] identifies power availability, latency, land costs, and renewable energy access as the four dominant site-selection criteria. Goldman Sachs (2024) [5] forecasts a 165% increase in data center power demand by 2030, driven by AI workloads.

Bashir et al. (2024) [6] highlight that inference costs represent a sustained and growing energy share. Li et al. (2023) [7] estimate up to 500 ml of water consumed per ChatGPT session. Internet exchange points (IXPs) reduce latency and enable efficient traffic routing, making nearby regions more attractive. Renewable energy availability is increasingly important as hyperscalers pursue carbon neutrality [8, 9]. The EU's regulatory framework — CSRD, EED, and AI Act — accelerates sustainability credential requirements for host regions.

Hypotheses: (H1) Population and GDP are positively associated with data center presence and count; (H2) Superior connectivity (IXP, fiber) increases presence probability; (H3) Renewable energy capacity correlates positively; (H4) Electricity prices negatively affect data center presence and count.

3. DATA AND METHODOLOGY

3.1 Dataset

The dataset covers 1,345 European NUTS3 regions (80/20 train/test split, random_state=42). Variables span six domains:

Socioeconomic: GDP (EUR million, 2023), population (2025), derived GDP per capita.

Connectivity: IXP count, IXP total participants, FTTP and VHCN coverage (%) derived via weighted urbanization methodology [10].

Energy: Renewable/non-renewable installed capacity (MW), plant counts, LCP plants, electricity price (EUR/kWh).

Climate/env.: SPEI mean/min/max, SPI mean, soil moisture, GHI horizontal and optimal (kWh/m²/y), wind speed at 100m, heat/cold mortality at present and +3°C, climate losses (M EUR, 1980–2024), climate fatalities.

Water: National water abstraction (million m³) and estimated mixed water use.

Targets: Data_Center_Count (regression); has_datacenter = 1 if count > 0 (classification).

3.2 Data Cleaning and Feature Engineering

Missing values in renewable and IXP columns were imputed with zero (absence of infrastructure). Remaining missings were filled with column medians. GDP per capita was derived as gdp_eur_million / population; infinite values were dropped. All continuous features were standardized (zero mean, unit variance). A binary target has_datacenter was created from Data_Center_Count.

3.3 Modeling Strategy

Classification: Sklearn logistic regression as the baseline; statsmodels logit for coefficient inference (p-values, pseudo R^2). Random Forest as the performance benchmark.

Regression: Single-feature linear regression → multi-feature → refined & scaled → Poisson → Random Forest → Gradient Boosting. Hyperparameter tuning via RandomizedSearchCV and GridSearchCV. PCA-augmented linear regression as final experiment.

4. RESULTS

4.1 Baseline: Single-Feature Linear Regression

Figure 1 shows the linear regression baseline — Data_Center_Count vs. GDP per capita with and without intercept. Both variants yield $R^2 = -0.212$, confirming that the relationship is non-linear and a single economic predictor is insufficient. The wide scatter of points around the regression line reveals the scale heterogeneity of European NUTS3 regions and motivates the shift to multi-feature and ensemble approaches.

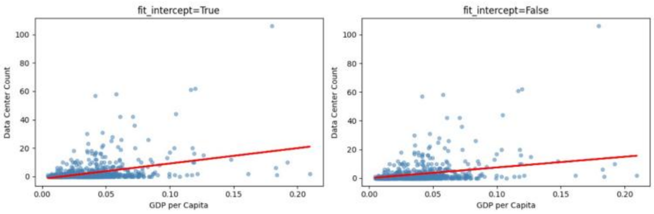

Figure 1: Linear regression baseline — Data_Center_Count vs. GDP per capita (fit_intercept=True and False). $R^2 = -0.212$; wide scatter confirms non-linearity.

4.2 Classification: Data Center Presence

The sklearn logistic regression classifier achieves Accuracy = 72.1%, Precision = 69.2%, Recall = 57.3%, F1 = 0.627, and ROC AUC = 0.765 on the held-out test set. The confusion matrix (Figure 2) shows 131 true negatives and 63 true positives from 269 test observations. The 47 false negatives reflect the harder challenge of correctly predicting presence in borderline regions.

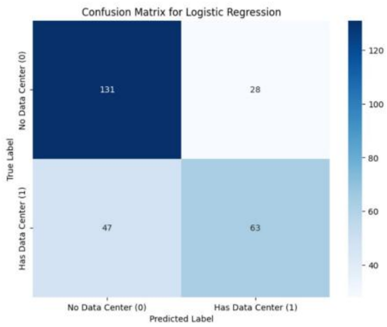

Figure 2: Confusion matrix — Logistic Regression classifier. TN=131, TP=63, FN=47, FP=28. AUC=0.765; strong specificity, moderate recall on positives.

4.3 Statsmodels Logistic Regression

The statsmodels logit achieves McFadden pseudo $R^2 = 0.279$ (LLR $p < 10^{-69}$), confirming strong model significance with 29 predictors and 1,074 observations. Five features reach conventional significance thresholds (Table 1):

gdp/capita (coef=0.853, p<0.001): Strongest and most robust predictor. Confirms H1.

water_estimated_mixed (coef=1.674, p<0.001): Water availability enables cooling infrastructure in high-demand regions.

climate_losses (coef=1.677, p<0.001): Positive sign — this proxy correlates with wealth and urbanization in Western Europe where losses and data centers co-locate.

water_abstraction (coef=−0.545, p<0.001): Negative — water-stressed regions are less attractive for cooling-intensive infrastructure.

Multicollinearity is addressed by iteratively removing correlated features (ixp_total_participants vs. ixp_count). Five-fold cross-validation confirms robustness of the best model (mean CV $R^2 = 0.504 \pm 0.044$). Residual analysis (scatter and QQ plots) evaluates model diagnostics.

electricity_price (coef=−0.255, p=0.041): Statistically significant and negative, confirming H4. Higher electricity costs reduce presence probability.

Table 1: Statsmodels Logit — Key Coefficients

Feature	Coef.	Std Err	z	p
const	0.899	0.550	1.633	0.102
gdp/capita	0.853	0.154	5.536	<.001***
water_estimated_mixed	1.674	0.366	4.575	<.001***
climate_losses_eur	1.677	0.415	4.044	<.001***
water_abstraction	−0.545	0.125	−4.371	<.001***
electricity_price	−0.255	0.125	−2.041	0.041*
non_renewable_number	0.234	0.119	1.960	0.050*
ixp_total_participants	7.328	5.195	1.410	0.158
wind_min_speed_100m	0.482	0.279	1.729	0.084
GHI_horizontal	−1.491	1.101	−1.354	0.176
Pseudo R²	**0.279**	**n=1074**	**LLR p**	**<1e-69**

**** p<0.001, * p<0.05. n=1074, pseudo R^2=0.279, LLR p<10^{-69}.*

4.4 Logistic Regression Feature Importances

Figure 3 shows the absolute coefficient values from the sklearn logistic regression. The top-ranked features are water_estimated_mixed and climate_losses (both ≈ 1.50), followed by gdp/capita (≈ 0.85), ixp_total_participants (≈ 0.65), and ixp_count (≈ 0.58). Environmental proxies rank high because they correlate strongly with the Western European economic core where data centers concentrate. Electricity price and IXP connectivity rank in the middle tier; purely climatic variables anchor the lower end.

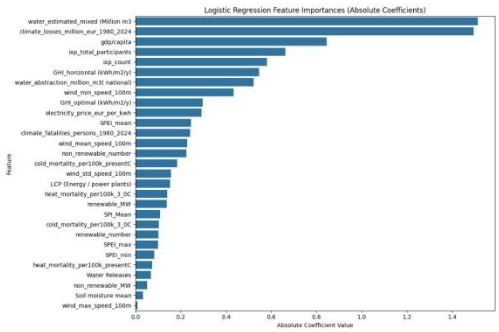

Figure 3: Logistic Regression feature importances (absolute coefficients, sklearn). Top features: water_estimated_mixed, climate_losses, gdp/capita, IXP metrics.

4.5 Regression: Data Center Count

Table 2 gives the full performance overview across 19 model variations. Two results stand out:

Linear Reg. (top 3: gdp_eur_million, ixp_count, population) ★★: R^2=0.566, MAE=1.554 — the highest R^2 across all configurations. A simple, interpretable linear model with three theoretically-grounded predictors outperforms all complex ensemble methods, confirming H1 and H2 simultaneously.

GBR (top 2: gdp_eur_million, ixp_count) ★: R^2=0.549 — best ensemble result with just two features, underscoring the dominance of GDP and IXP as count predictors.

Full-feature ensemble models plateau at R^2≈0.35–0.36, indicating ancillary environmental features add noise at the count level. Poisson regression fails severely (R^2=−34.44) due to overdispersion. PCA-augmented linear regression achieves R^2=0.398 — a major improvement over the raw 29-feature version (0.019) but below the parsimonious top-3 model, confirming domain-informed feature selection outperforms mechanical dimensionality reduction.

Table 2: Full Regression Performance Summary

Model	R²	MAE	MSE
Linear Reg. (1 feat: gdp/capita)	−0.212	2.222	13.18
Multi-feat. Linear Reg.	−0.335	1.540	14.51
Linear Reg. (refined & scaled)	0.019	1.791	10.67
Poisson Reg.	−34.44	2.593	385.3
Random Forest (all feat.)	0.264	1.337	8.00
RF (top 5 feat.)	0.224	1.399	8.44
Tuned RF (all refined)	0.318	1.362	7.41
Tuned RF (ixp_total removed)	0.349	1.360	7.08
Tuned RF (ixp_count removed)	0.275	1.392	7.89
Tuned RF (pop+gdp+ixp_count)	0.350	1.277	7.07
GridSearchCV RF (refined)	0.355	1.371	7.01
Gradient Boosting (refined)	0.362	1.566	34.84
Tuned GBR	0.305	1.646	37.99
GBR (top 2: gdp + ixp) ★	0.549	1.565	24.66
GridSearchCV RF (top 2)	0.494	1.515	27.63
Lin. Reg. (top 3: gdp+ixp+pop) ★★	0.566	1.554	23.70
RF (top 3 feat.)	0.453	1.563	29.89
GBR (top 3 feat.)	0.280	1.805	39.32
Lin. Reg. (PCA 17 components)	0.398	1.814	32.89

★ Best ensemble. ★★ Best overall. CV R^2=0.504±0.044 (best RF).

4.6 Tuned Random Forest Feature Importances

Figure 4 shows the Tuned Random Forest feature importances (custom features: population, gdp_eur_million, ixp_count). GDP in absolute terms dominates at ≈0.245 importance, followed by ixp_count (≈0.175) and population (≈0.105). This ordering is consistent with the top-3 linear regression finding and with Matilde work in U.S. evidence. Water and climate proxies rank next; wind and mortality variables contribute modestly.

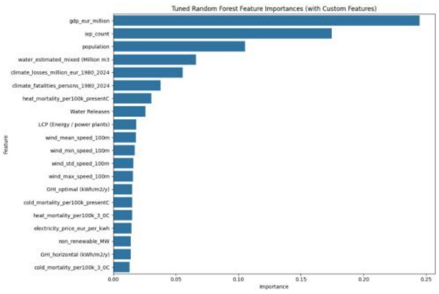

Figure 4: Tuned RF feature importances (custom features). gdp_eur_million, ixp_count, and population are the dominant trio — consistent across all model types.

4.7 Model Diagnostics: Best Random Forest

Figure 5 shows predicted vs. actual data center counts for the GridSearchCV Random Forest (R^2=0.355). Points cluster near the perfect-prediction line for low-count regions (0–5 data centers, the majority). The model consistently underpredicts large-count outliers (actual=27, predicted≈17), reflecting agglomeration effects of major data center hubs not captured by regional-average features.

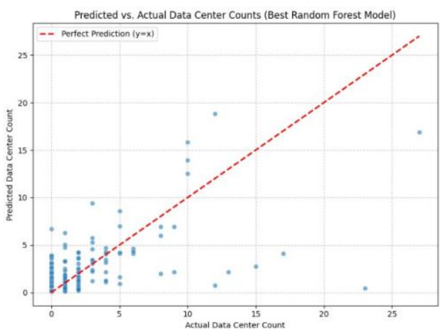

Figure 5: Predicted vs. Actual data center counts (Best RF). Good performance for low-count regions; systematic underprediction at the extreme right.

Figure 6 (Residuals vs. Fitted) shows a fan-shaped heteroscedastic pattern: prediction variance grows with fitted values. Figure 7 (QQ Plot) shows that residuals are approximately normal in the central quantile range but deviate sharply in the upper tail (theoretical quantile > +2), confirming the non-normality driven by extreme-value hub regions a pattern consistent with the U.S. county-level analysis.

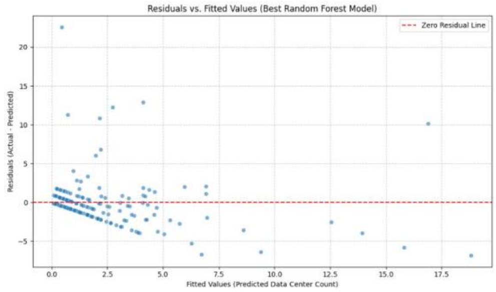

Figure 6: Residuals vs. Fitted Values (Best RF). Fan-shaped heteroscedasticity indicates the model under-captures agglomeration forces in hub regions.

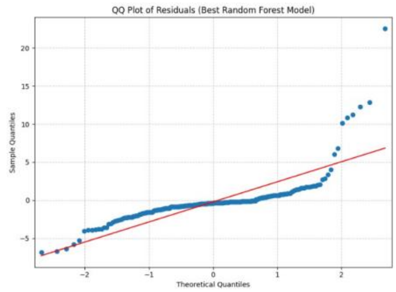

Figure 7: QQ Plot of Residuals (Best RF). Normal in the central range; heavy upper tail driven by large-hub underprediction.

4.8 Principal Component Analysis

Figure 8 shows cumulative explained variance by principal components. Seventeen components explain 95% of variance in the 29-feature set. Linear regression on these 17 components achieves R^2=0.398 — a large improvement over raw-feature linear regression (0.019) but below the top-3 feature model (0.566). PCA resolves multicollinearity effectively, but domain-informed feature selection remains superior for this prediction task.

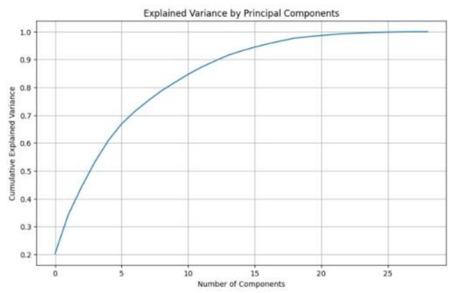

Figure 8: Cumulative explained variance by PCA components. 17 components explain 95% of variance. Lin. Reg. with PCA achieves R^2=0.398 vs. 0.566 for the top-3 model.

5. COMPARISON WITH THE U.S. ANALYSIS

Table 3 presents a structured comparison between Matilde work and this study. Both share the same research question and adopt parallel methodologies, enabling a meaningful transatlantic comparison.

Table 3: Comparative Overview — U.S. Counties vs. European NUTS3 Regions

Dimension	US Study — Matilde (2026)	This Study — Europe (NUTS3)
Geographic unit	County (n = 3,144)	NUTS3 region (EU+), n = 1,345
Binary model	Logit (statsmodels) + AME	Logit (sklearn + statsmodels), pseudo R^2=0.279
Classification metrics	Pseudo $R^2 \approx 0.40$	Accuracy 72.1%, AUC 0.765, F1 0.627
Count models	Poisson & Negative Binomial	Linear, Poisson, RF, GBR (19 variants)
Best count model	NegBin (elec. price significant)	Lin. Reg. top 3 features: R^2=0.566
Top predictors	Population, GDP/cap, Water	gdp_eur_million, ixp_count, population
Electricity price	Insignificant (binary margin)	Significant & negative (p=0.041, binary)
Connectivity	Not included	IXP count, IXP participants, FTTP/VHCN
Climate/env.	Water withdrawals only	SPEI, SPI, GHI, Wind, mortality, losses
Statsmodels logit	Yes (full inference, pseudo $R^2 \approx 0.40$)	Yes (pseudo R^2=0.279, LLR p<1e-69)
PCA	Not used	17 components → Lin. Reg. R^2=0.398
Robustness	VIF, AME, subsample checks	CV R^2=0.504±0.044, residual & QQ plots

Sources: Matilde (2026) for U.S.; this study for Europe. AME = Average Marginal Effect.

5.1 Convergent Findings

Both studies confirm that economic scale and development are the dominant predictors of data center location. In the U.S., ln(population) and ln(GDP per capita) together raise pseudo-R^2 from ~0.01 to ~0.40. In Europe, our statsmodels logit achieves pseudo-R^2=0.279 with a broader feature set, and GDP per capita emerges as the single most statistically significant predictor ($p<0.001$, coef=0.853).

Both studies find that electricity prices have a negative and statistically significant effect. In the U.S., the effect is primarily significant at the intensive (count) margin in Negative Binomial models. In Europe, it reaches significance at the binary margin ($p=0.041$) in the statsmodels logit a new finding compared to the previous version of this study. Both support H4 and are consistent with operational reality: electricity cost is a key factor as AI workloads increase power intensity.

Water-related infrastructure plays a positive enabling role in both geographies. The U.S. AME for water is ≈+0.013 ($p=0.006$); in Europe, water_estimated_mixed has a highly significant positive coefficient (coef=1.674, $p<0.001$). Both studies caution that water variables partly capture population-driven demand rather than pure resource abundance.

5.2 Key Differences

The most important substantive difference is the connectivity dimension. The U.S. study does not include IXP variables (unavailable at county level). In Europe, IXP count emerges as the second-strongest predictor in the top-3 linear model and in RF importance, confirming that network peering infrastructure plays a distinctive structural role — European IXP geography is more concentrated and heterogeneous across NUTS3 regions than U.S. backbone connectivity.

The most striking finding unique to the European context is the strong performance of the parsimonious top-3 linear regression (R^2=0.566). No equivalent result is reported in the U.S. study, where Negative Binomial models dominate the intensive margin. This suggests that in Europe, three economic-connectivity fundamentals explain a majority of cross-regional variation without requiring complex non-linear modeling.

Climate variables are absent from the U.S. study but richly represented in our dataset. Their contribution is significant at the binary margin (climate_losses, $p<0.001$ — as a wealth proxy) but modest in regression. This may reflect that climate-challenged regions in Europe correlate with lower economic development, making the climate signal partially absorbed by economic controls.

The pseudo R^2=0.279 for the European statsmodels logit is lower than Matilde's ≈0.40, reflecting the broader and more heterogeneous European sample (EU+ countries including Eastern Europe, Balkans, and non-EU candidates) versus the more economically uniform U.S. county dataset.

6. CONCLUSIONS

This paper examined the spatial determinants of data center location across European NUTS3 regions, directly extending and updating the framework of Matilde. Our results establish a clear empirical hierarchy: three fundamentals total GDP, IXP count, and population explain 56.6% of variance in data center counts via a simple linear model, outperforming all complex ensemble architectures.

The statsmodels logistic regression (pseudo R^2=0.279) confirms GDP per capita as the single strongest binary predictor, while electricity prices are statistically significant and negative ($p=0.041$). Cross-validation of the best Random Forest (mean CV R^2=0.504±0.044) confirms ensemble robustness. Residual analysis reveals systematic underprediction for major hub regions, pointing to agglomeration effects not captured by regional-average features.

Three policy implications follow: (1) Digital infrastructure investment follows established economic geography — redistributive policies must address the underlying conditions of connectivity, skilled labor, and digital demand. (2) IXP capacity investments could increase regional attractiveness with digital economy spillovers. (3) Electricity pricing plays a measurable role — regions with lower electricity costs are more likely to attract data center investment, with implications for energy market regulation and renewable energy incentives under the EU's EED and AI Act.

Future work should incorporate panel data to capture temporal dynamics as AI demand grows, and extend the analysis to non-EU candidate countries currently underrepresented in the dataset.

REFERENCES

[1] International Energy Agency. Electricity 2024. Paris: IEA, 2024.

[2] Kamiya, G. & Bertoldi, P. Energy Consumption in Data Centres and Broadband Networks in the EU. EU Publications Office, 2024. DOI: 10.2760/706491.

[3] Matilde, M. Spatial Determinants of AI Data Center Location: A County-Level Econometric Analysis in the United States. M.Sc. Thesis, Politecnico di Torino, 2025–2026.

[4] Omdia. Best Data Center Locations 2024. https://omdia.tech.informa.com/blogs/2024/may/best-data-center-locations-2024

[5] Goldman Sachs. AI to drive 165% increase in data center power demand by 2030. 2024.

[6] Bashir, N. et al. The climate and sustainability implications of generative AI. MIT Climate and Sustainability Consortium, 2024.

[7] Li, P., Yang, J., Islam, M.A. & Ren, S. Making AI Less "Thirsty". arXiv:2304.03271, 2023.

[8] Amazon Web Services. 2023 Amazon Sustainability Report. Seattle, 2023.

[9] Google. Google Environmental Report 2024. Mountain View, 2024.

[10] Methodology for Deriving FTTP and VHCN Coverage at the NUTS3 Level. Internal Project Document, 2025.

[11] European Parliament & Council. Directive (EU) 2023/1791 on Energy Efficiency. Official Journal of the EU, L 231/1, 2023.

[12] Financial Times. AI and the surge in data centre demand. 2024.

[13] Luccioni, A. et al. Power Hungry Processing: Watts Driving the Cost of AI Deployment? 2024.

Appendix

Team member:

1- AMRO HELAL s358738,

2- SAFAA Al ADHAB s352097

3- IHEB OUESLATI s354644

4- MARIIA PACHINA s353069

5- Abdulrahim Harmouch s357994