Drivers of Data Center Locations Across European NUTS3 Regions

An Econometric and Machine-Learning Analysis of Regional Siting Factors

Group_03

S355122 GUL ZUBAIR
S359686 KUMAR MUKESH
S360593 KUNWAR SANJAYA
S351877 RANKIRIPATHIRANAGE MESHAN CHANUKA APPUHAMY
S314337 TOGHRUL NABIYEV

Abstract

Situation. Data centers are core infrastructure for cloud computing, digital services, financial systems and public-sector digitization. Their location choices matter because facilities require reliable electricity, high-capacity network access, land, cooling resources and institutional stability. At regional scale, data-center investment can reshape electricity demand, water planning and local economic development.

Complication. Previous studies identify plausible siting criteria, but they do not fully test which socio-economic, network, energy and environmental factors distinguish data-center regions from non-data-center regions across Europe at NUTS3 scale.

Proposal. To address this gap, this project uses a cross-sectional dataset of 1,345 European NUTS3 regions and makes three contributions: it builds a cleaned literature-guided modelling pipeline, estimates an interpretable country-clustered logistic regression, and validates the results using sklearn predictive models plus a Negative Binomial count robustness check.

1. Introduction

Data-center location matters because facilities support cloud computing, digital services, financial systems and public-sector digitization, while also requiring reliable electricity, land, cooling resources, high-capacity network access and institutional stability. At regional scale, new facilities can reshape electricity demand, water planning and local economic development.

NUTS3 analysis is useful because it compares sub-national regions where market size, network infrastructure, environmental pressure and energy conditions can vary within the same country. The dependent variable is binary: has data center=1 when Data Center Count is at least one, and 0 otherwise.

The research question is: "What socio-economic, digital-infrastructure, energy, and environmental factors explain whether European NUTS3 regions host at least one data center?" The hypothesis is: "Data-center presence is expected to be positively associated with economic prosperity and digital-network infrastructure, while environmental pressure, especially water-abstraction intensity, and power-cost constraints are expected to have negative or weaker independent associations after controlling for demand and connectivity."

The project contributes by building a cleaned literature-guided modelling pipeline, estimating an interpretable country-clustered logistic regression and validating the results with sklearn predictive models plus a Negative Binomial count robustness check.

2. Related Work and Gap

Firm-location theory links siting to market access, input costs and agglomeration economies (Weber, 1929; Krugman, 1991). For data centers, these mechanisms map onto demand concentration, power reliability, cooling conditions, fiber connectivity and proximity to Internet exchange ecosystems. More specialized work adds energy and risk dimensions: Covas et al. (2013) use GIS-based site selection for sustainable data centers; Depoorter et al. (2015) show that European location affects cooling demand and renewable-energy feasibility; and Gazzola et al. (2023) emphasize territorial risks and business continuity. Industry evidence also ranks power availability, network connectivity, land and regulatory stability among key siting criteria (JLL, 2023).

To sum up, previous work identifies plausible data-center siting criteria, but it has not fully tested these factors at European NUTS3 regional scale while combining interpretable econometric inference with predictive machine-learning validation.

3. Methodology

Data and unit of analysis. Each observation is a NUTS3 region. The dependent count, Data Center Count, is zero-inflated: 55.9% of regions have no data center and the count variance is about 18 times the mean. The main task is therefore binary classification, has data center=1 if the count is at least one, with a count-model robustness check.

Table 1: Thematic feature groups used in the analysis.

Group	Examples of variables
Demand and scale	Population, GDP, GDP per capita
Energy system	Electricity price, installed MW, renewable share
Network ecosystem	IXP count and participants, nearest IXP distance, speed, latency, test density
Water and drought	Water abstraction, WEI+, SPEI, SPI, soil moisture
Climate and pollution	Heat/cold mortality, losses, fatalities, air/water/LCP releases
Renewable potential	Solar GHI and wind-speed statistics

Cleaning and feature selection. Column names were standardized; year-metadata variables were removed from modelling; country code was derived from the NUTS prefix for cluster-robust inference; and skewed numeric variables were transformed with log(1 + x). Candidate variables were selected from literature-based groups covering demand, energy, network connectivity, water stress, climate risk, pollution and renewable potential. After a 75/25 stratified train-test split, correlation pruning and VIF pruning were learned only on the training fold, and median imputation and scaling were also fitted only on the training fold before model evaluation. High correlations were pruned using the training fold only ($|r| > 0.85$); iterative VIF pruning then removed remaining multicollinearity (threshold 10). The final model matrix contains 29 predictors.

Models. The interpretable model is a statsmodels logistic regression with country-clustered standard errors. Predictors are standardized, so odds ratios show the multiplicative change in odds for a one-standard-deviation increase. Predictive models are sklearn Logistic Regression, Random Forest and Gradient Boosting, benchmarked against a majority-class DummyClassifier. A Negative Binomial model on Data Center Count checks whether the binary results are consistent with count intensity. Cross-validation is used as a diagnostic after training-fold feature selection.

4. Evaluation and Results

Predictive performance. All fitted models outperform the majority baseline on the held-out test set (n = 337). ROC-AUC is around 0.77-0.78 for the econometric logit and the sklearn models, compared with 0.50 for the baseline. Similar performance across linear and tree models suggests that most useful signal is captured by transformed, approximately monotonic relationships rather than highly complex interactions.

Table 2: Held-out model performance. Accuracy and F1 use a 0.5 threshold; CV3 AUC for the three sklearn models is 0.819-0.825.

Model	Acc.	F1	ROC-AUC	PR-AUC	Brier
Dummy baseline	0.558	0.000	0.500	0.442	0.442
Logit, statsmodels	0.724	0.687	0.777	0.759	0.191
sklearn Logistic Regression	0.727	0.689	0.777	0.757	0.191
Gradient Boosting	0.691	0.658	0.776	0.763	0.195
Random Forest	0.703	0.671	0.770	0.756	0.194

Interpretable coefficients. The country-clustered logistic model has McFadden pseudo-R^2 = 0.311 and a highly significant likelihood-ratio test ($p = 7.69 \times 10^{-73}$). Four non-constant predictors are significant at the 5% level.

Table 3: Significant logit predictors; standardized inputs.

Predictor	Coeff.	OR	95% CI	p
`log_gdp_per_capita_eur`	1.200	3.32	[2.23, 4.95]	< 0.001
`log_water_abstraction_national`	-0.348	0.71	[0.56, 0.90]	0.004
`dist_nearest_ixp_km`	0.281	1.32	[1.04, 1.69]	0.022
`ixp_count`	0.670	1.96	[1.09, 3.51]	0.025

Driver interpretation. The results suggest that data centers are disproportionately located in prosperous and digitally connected regions. GDP per capita is the strongest positive factor, IXP count supports the role of interconnection ecosystems, and water-abstraction intensity has a negative association. These results should be interpreted as statistical associations, not causal effects. GDP per capita is the dominant demand-side factor: a one-standard-deviation increase is associated with 3.32 times higher odds of data-center presence. Water-abstraction intensity is negative, consistent with the hypothesis that environmental pressure can deter location, though the variable is national-level and should not be read as direct local water stress. The positive sign on distance to the nearest IXP is counter-intuitive; after controlling for a region's own IXP count, it may reflect secondary deployment zones where connectivity has expanded outside the primary exchange node.

The count robustness check supports the same broad story. The Negative Binomial model reports McFadden pseudo-R^2 = 0.234, dispersion α = 0.531, and test count MAE=1.551. Its strongest substantive effects are GDP per capita (IRR=2.04), population (IRR=1.97), water abstraction, electricity price, latency, IXP count and test density. Population is therefore clearer for data-center intensity than for simple binary presence once GDP per capita is controlled.

5. Discussion and Conclusion

The findings support the hypothesis with qualifications. Data-center location across European NUTS3 regions is most strongly associated with economic gravity and digital-network ecosystems. The presence of IXP infrastructure and network-performance proxies matters, but the results do not prove that building an IXP alone causes investment; data centers and network infrastructure may co-evolve. Environmental variables are not the main positive attractors, but water-abstraction intensity has a statistically significant negative association with presence. Electricity price has the expected negative sign and becomes significant in the count specification, suggesting that power cost may matter more for intensity than for the initial regional presence decision. Solar and wind potential do not emerge as independent positive drivers in this historical cross-section.

The answer to the research question is therefore: data centers are located disproportionately in prosperous, digitally connected regions, with environmental water-pressure variables and power-cost variables acting as secondary constraints. For regional policy, the implication is that attracting data centers is less about a single cheap-input variable and more about a combined ecosystem: demand, interconnection, reliable infrastructure and credible sustainability planning.

Limitations and future work. The analysis is cross-sectional, so it cannot prove causality. Some important variables, including electricity price and water abstraction, are measured at country level rather than NUTS3 level. Missing environmental variables require imputation, and spatial dependence among neighbouring regions is not modelled. Future work should use spatial models and hurdle or zero-inflated count models.

.

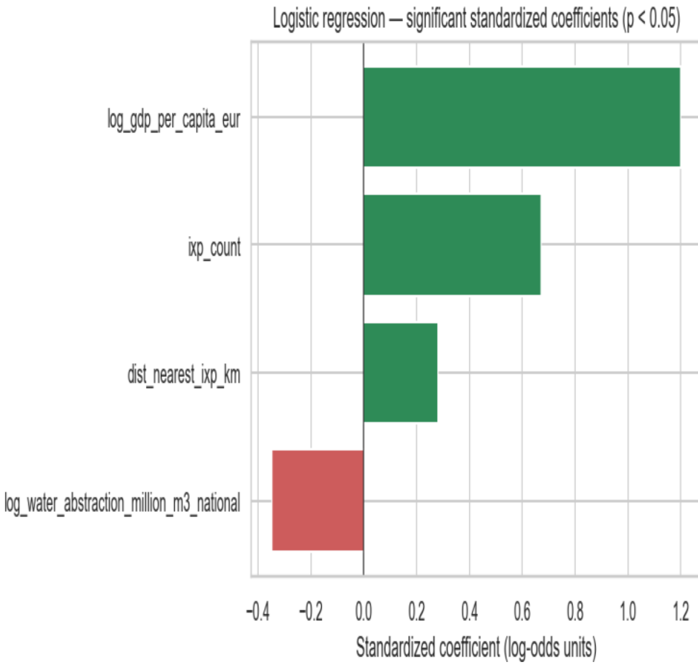

Figure 1: Significant standardized logit coefficients ($p < 0.05$). Positive values increase the odds of data-center presence; negative values reduce them

References

Covas, M. T., Silva, C. A., & Dias, L. C. (2013). On locating sustainable Data Centers in Portugal: Problem structuring and GIS-based analysis. Sustainable Computing: Informatics and Systems, 3(1), 27-35.

Depoorter, V., Oro, E., & Salom, J. (2015). The location as an energy efficiency and renewable energy supply measure for data centres in Europe. Applied Energy, 140, 338-349.

Gazzola, V., Menoni, S., Ghignatti, P., Marini, A., Mauri, R., & Oldani, G. (2023). Analysis of Territorial Risks and Protection Factors for the Business Continuity of Data Centers. Sustainability, 15(7), 6005.

JLL. (2023). Global Data Center Outlook 2023. Jones Lang LaSalle Research.

Krugman, P. (1991). Increasing returns and economic geography. Journal of Political Economy, 99(3), 483-499.

Weber, A. (1929). Theory of the Location of Industries. University of Chicago Press.

Eurostat. NUTS3 demographic and GDP statistics. https://ec.europa.eu/eurostat

European Environment Agency. Water Exploitation Index Plus (WEI+) and climate indicators. https://www.eea.europa.eu

JRC. PESETA climate impacts in Europe and Copernicus European Drought Observatory. https://edo.jrc.ec.europa.eu

PCH IXP Registry. https://www.pch.net/ixp

Ookla Open Data. Speedtest Global Fixed and Mobile Performance Maps. https://www.ookla.com/ookla-for-good/open-data

WRI Global Power Plant Database. https://datasets.wri.org/dataset/globalpowerplantdatabase

PVGIS. Photovoltaic Geographical Information System. https://re.jrc.ec.europa.eu/pvg_tools

Global Wind Atlas v3.0. https://globalwindatlas.info