What Drives the Location of Data Centres in Europe?

An Empirical Analysis at the NUTS3 Level

LI ZECHANG
S336896
Georesources and Geoenergy Engineering
Politecnico di Torino
Turin, Italy

FAN XINRUI
S354487
Georesources and Geoenergy Engineering
Politecnico di Torino
Turin, Italy

WANG JUNYI
S357972
Georesources and Geoenergy Engineering
Politecnico di Torino
Turin, Italy

FAN ZIXUAN
S352845
Georesources and Geoenergy Engineering
Politecnico di Torino
Turin, Italy

ABSTRACT

Data centres are critical infrastructure for the digital economy, yet their spatial distribution across Europe remains highly uneven. This study investigates the socio-economic, infrastructure, and environmental factors that drive data centre location at the NUTS3 level. Using a dataset of 1,345 European NUTS3 regions and 45 variables, we frame the problem as a regression task and train two complementary models: an interpretable OLS regression (statsmodels) to identify the main drivers, and a Random Forest regressor (scikit-learn) to maximise predictive performance. After five rounds of data cleaning, including VIF-based removal of multicollinear variables, we find that GDP is the dominant driver of data centre count, with IXP connectivity, electricity price, and network latency as additional significant factors. The OLS model explains 71.5% of the variance, while feature importance from Random Forest confirms GDP as the primary predictor (58.6% of total importance). Cross-model consistency strengthens our findings: data centres cluster where economic demand, digital connectivity, and affordable energy converge.

CCS CONCEPTS

Computing methodologies → Machine learning approaches → Regression analysis

KEYWORDS

Data centres, NUTS3, regression analysis, multicollinearity, spatial analysis, Europe

1 INTRODUCTION

Data centres are the backbone of the digital economy, powering cloud services, AI workloads, and digital communications. Their numbers grow rapidly across Europe, yet their spatial distribution is highly uneven - some regions host dozens, others have none.

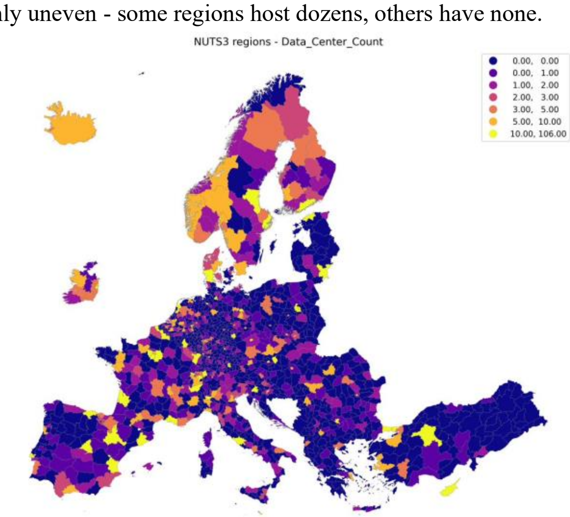

Figure 1: Distribution of data centres across European NUTS3 regions.

Complication. Despite their economic importance, the factors driving this uneven distribution remain poorly understood at the regional level. Most analyses operate at national or city scale, but site selection depends on regional characteristics - local energy costs, network infrastructure, economic conditions. Without systematic regional evidence, policy-makers and investors lack a basis for decisions.

Question. Research Question: What factors explain the number of data centres in each European NUTS3 region?

Answer. Using a dataset of 1,345 NUTS3 regions and 45 variables, we train an OLS regression (statsmodels) for interpretability and a Random Forest regressor (sklearn) for prediction. GDP emerges as the dominant driver, with IXP count, electricity price, and latency as additional significant factors. The interpretable model explains 71.5% of variance.

1.1 Hypotheses

Based on existing literature, we formulate the following hypotheses:

1. Higher GDP regions host more data centres (market demand).
2. Larger populations attract more data centres (user base).
3. Lower electricity prices increase data centre presence (operating cost).
4. Greater IXP presence and connectivity attract data centres.
5. Lower network latency is associated with more data centres.

2 RELATED WORK

The literature identifies several categories of factors driving data centre location:

1. **Energy and cost:** Electricity is the largest operating cost. CBRE (2026) reports energy pricing as a key concern; EUDCA (2026) finds 67% of operators cite power access as their greatest challenge. Jung et al. (2021) show low ambient temperatures significantly reduce cooling energy demand, favouring cooler climates.
2. **Network connectivity:** The FLAPD markets (Frankfurt, London, Amsterdam, Paris, Dublin) dominate Europe due to superior connectivity and IXP presence (Newmark, 2025).
3. **Economic geography:** Harvard Business School Working Paper 21-042 (Where the Cloud Rests) shows data centres cluster around economic centres but are largely decoupled from specialised human capital. Wang et al. (2025) provide an empirical analysis of data centres in China, identifying jointly supply-, demand-, and environment-side drivers.
4. **Urban and governance dynamics:** Monstadt et al. (2025) document data centre concentration in Dutch metropolitan areas (the "digital gateway to Europe"), highlighting growing tensions over land, water, and energy use.
5. **Climate:** Cooler regions reduce cooling costs (Jung et al., 2021). However, our dataset lacks direct temperature variables; drought indices (SPEI/SPI) are not reliable proxies.

3 METHODOLOGY

3.1 Dataset

The dataset contains 1,345 NUTS3 regions with 45 variables covering socio-economic indicators, digital infrastructure (IXP, network speed, latency), energy (electricity prices, capacity), environmental conditions, and climate risk. The target variable is Data_Center_Count - the number of data centres in each region. We frame the problem as regression.

3.2 Data Cleaning

Five rounds reduced the dataset from 45 to 23 features:

- **Round 1:** Removed 4 mortality variables - irrelevant to site selection.
- **Round 2:** Removed 4 year columns - constant across regions, no discriminating power.
- **Round 3:** Median imputation - robust to outliers (e.g., GDP is right-skewed by major cities).
- **Round 4:** Removed 11 redundant variables via VIF + correlation: GHI (VIF ~4900, corr ~0), SPEI/SPI (VIF 50-115, corr ~0), wind max/min/std (kept mean).
- **Round 5:** Removed avg_upload_mbps (redundant with download).

3.3 Feature Selection for OLS

From the 23 cleaned features, we selected 6 for the interpretable OLS model based on two criteria: (1) literature-driven hypotheses (energy, network, economic demand), and (2) high correlation with the target while maintaining VIF < 10. The selected features are GDP, IXP count, population, electricity price, average download speed, and latency. Random Forest uses all 23 features, as tree-based models are robust to multicollinearity.

We chose 6 features instead of all 23 because: (1) statsmodels-style interpretation requires few enough variables to discuss each one individually, (2) starting simple lets us iteratively explore feature combinations as recommended by the instructor, and (3) Random Forest below uses all 23 features and can handle the rest.

3.4 Models and Setup

1. **Interpretable:** OLS regression (statsmodels). We tested three feature combinations iteratively to explore multicollinearity, as recommended in class.
2. **Predictive:** Random Forest regressor (sklearn). We compared a baseline (default 100 trees) against a tuned model (300 trees, max_depth=15, min_samples_leaf=5).

Data split 80/20 (1,076 train / 269 test, random_state=42 for reproducibility).

4 RESULTS

4.1 Interpretable Models (OLS)

Multicollinearity check. Removing GDP (Model 2) reduces R^2 from 0.7094 to 0.6015, confirming GDP's critical role. The change in population's coefficient between models reveals that GDP, population, and IXP share variance - they all reflect regional economic scale.

Table 1: Comparison of OLS model specifications.

Model	Features	R^2	Key Finding
OLS 1	GDP, IXP, Population	0.7094	All three significant
OLS 2	IXP, Population (no GDP)	0.6015	R^2 drops; GDP critical
OLS 3 (final)	GDP, IXP, Pop, Elec, DL, Lat	0.7146	Best fit; 5 of 6 significant

Model 3 coefficients (R^2 = 0.7146):

1. **GDP (+0.0002, $p < 0.001$):** Strongest positive driver. Each million EUR of GDP adds 0.0002 data centres - small per unit, but compounded across high-GDP regions, this explains hub concentration. Confirms H1. For example, a region with 50,000 million EUR GDP (like Inner London) would have approximately 10 more data centres than a region with 0 GDP, holding other factors constant.
2. **IXP count (+2.70, $p < 0.001$):** Each additional IXP is associated with ~2.7 more data centres. Largest marginal effect; confirms H4.
3. **Electricity price (-10.44, $p = 0.001$):** A 1 EUR/kWh increase reduces count by ~10. Large coefficient reflects electricity's role as the dominant operating cost. Confirms H3.
4. **Latency (-0.04, $p = 0.008$):** Lower latency attracts data centres, confirming H5.
5. **Population (-0.0000013, $p < 0.001$):** Negative when controlling for GDP - a larger population at fixed GDP implies lower per-capita wealth, associated with fewer data centres.
6. **Download speed (-0.002, $p = 0.084$):** Not significant at 5%, likely correlated with latency and IXP.

4.2 Predictive Model (Random Forest)

Table 2: Random Forest model performance.

Model	R^2 (test)	MAE	Notes
RF Baseline (default)	0.3844	1.55	100 trees
RF Tuned (final)	0.3860	1.53	300 trees, max depth=15

OLS vs RF performance gap. The OLS training R^2 of 0.7146 is much higher than the RF test R^2 of 0.3860. This is not a contradiction: OLS R^2 (0.7146) is measured on the same data the model was trained on - like a student grading their own homework. RF R^2 (0.3860) is measured on new, unseen data - like the same student taking a real exam. The exam is always harder than the homework. The gap reveals the difficulty of generalising to unseen regions - particularly extreme outliers like Dublin (106 data centres). Feature importance from RF confirms OLS findings: GDP dominates with 58.6% of total importance, followed by IXP-related variables (~22.55%). This cross-model consistency strengthens our confidence in GDP as the primary driver.

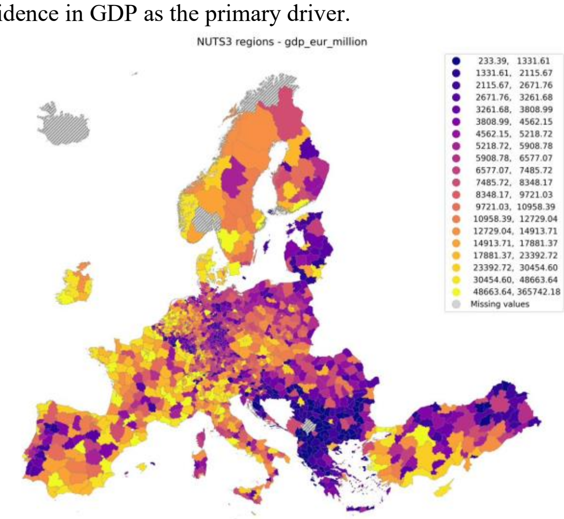

Figure 2: GDP distribution across NUTS3 regions. Comparing with Figure 1, high-GDP regions correspond closely to data centre clusters.

4.3 Hypothesis Review

1. Higher GDP → more data centres → Confirmed (strongest driver)
2. Larger population → more data centres → Complex (negative after controlling GDP)
3. Lower electricity price → more data centres → Confirmed
4. Greater IXP presence → more data centres → Confirmed
5. Lower latency → more data centres → Confirmed

5 DISCUSSION AND LIMITATIONS

Missing temperature data: Jung et al. (2021) show cooling costs are critical for data centres. The dataset documentation references temperature variables (Avg summer temp, No. of days tmax >35) from ERA5, but these were not present in the provided dataset. Drought indices do not proxy for temperature. Future work should incorporate ERA5 temperature directly.

Unobserved factors: Tax incentives, land costs, regulations, and corporate decisions are not captured. This explains the gap between

OLS training R^2 (0.7146) and RF test R^2 (0.3860) - regional characteristics alone cannot fully predict location.

Extreme values: Regions like Dublin (106 data centres) are high leverage points. Both models struggle with these outliers, suggesting hyper-concentration in established hubs operates under different dynamics.

6 CONCLUSION

GDP is the strongest driver of data centre location, confirmed by both OLS coefficients (highest significance) and Random Forest feature importance (58.6%). IXP count (+2.7 per IXP), electricity price (-10.4 per EUR/kWh), and latency are also significant. The interpretable model explains 71.5% of variance, while the predictive model achieves test MAE of 1.53. Data centres cluster where economic demand, digital connectivity, and affordable energy converge.

REFERENCES

[1] CBRE. 2026. Key Factors to Consider for Effective Data Centre Site Selection. CBRE Research.

[2] European Data Centre Association. 2026. Europe's Data Center Market Enters a Pivotal Phase. EUDCA Report.

[3] Harvard Business School. 2021. Where the Cloud Rests: The Economic Geography of Data Centers. Working Paper 21-042.

[4] Jung, M., et al. 2021. Optimized data center site selection - mesoclimatic effects on data center energy consumption and costs. Energy Efficiency, Springer.

[5] Monstadt, J., et al. 2025. How Data Centers Have Come to Matter: Governing the Spatial and Environmental Footprint of the 'Digital Gateway to Europe'. International Journal of Urban and Regional Research.

[6] Newmark. 2025. Data Center Site Selection Dynamics in Europe.

[7] Wang, L., Chen, D., Yao, M., & She, G. 2025. Spatial distribution and influencing factors of data centers in China: An empirical analysis based on the geodetector model. Energy and Buildings, 330.