Regional Determinants of Data center Location in Europe At NUTS3 level

A Socio-Economic, Infrastructure, and Environmental Analysis

Group 9 - The Oracle — Data Science, 2026

KEYWORDS: data centers, NUTS3, logistic regression, random forest, European digital infrastructure, GDP, internet connectivity

Abstract

The geographic distribution of data centers has become a foundational question for digital infrastructure policy across Europe. These facilities do not merely store data — they anchor cloud services, underpin financial systems, and support public digital services while concentrating significant capital investment within specific regions. As governments and planners increasingly seek to guide where this infrastructure lands, understanding which regional characteristics actually attract data center development becomes not just academically interesting, but practically urgent, particularly at a resolution fine enough to inform sub-national policy decisions. Despite its growing importance, this question remains poorly characterised at fine geographic resolution. Prior work has largely approached siting decisions through national-scale, multi-criteria frameworks which is useful for broad comparisons but blind to the considerable variation that exists within countries. More critically, existing studies have tended to examine economic, connectivity, and environmental determinants in isolation rather than within a unified empirical model, making it difficult to assess the relative contribution of each factor or draw consistent conclusions about what actually drives location choices across Europe's diverse regional landscape. To partially address these limitations, this study makes three contributions. First, we compile a dataset covering nearly all documented data centers across Europe, enriched with economic, connectivity, and environmental variables harmonised at NUTS3 level across 1,345 regions. Second, we frame the problem as binary classification and jointly test four theoretically grounded hypotheses: that economic scale drives demand (H1); that proximity to network infrastructure has a strong positive effect (H2); that water stress and high temperatures act as operational deterrents (H3); and that electricity prices have limited influence in Europe, where renewable procurement decouples siting from local energy costs (H4). We test these hypotheses by deploying both an interpretable logistic regression and a tuned random forest classifier, with both models achieving AUC-ROC above 0.76 on a held-out test set. Third, we provide consistent empirical evidence confirming H1 and H2 — regional GDP and internet connectivity emerge as the dominant predictors of data center presence — while H3 and H4 also hold, as environmental variables including water stress, ambient temperature, and electricity prices contribute negligible explanatory power across both modelling approaches.

1. Related Work

The literature on data center siting converges on three broad categories of determinants: economic, connectivity, and environmental. Economic factors consistently emerge as dominant, as investment and operational costs — particularly electricity, water, and maintenance — set hard boundaries on financial viability (Daim et al., 2013; Kheybari et al., 2020). Social considerations add a further layer of spatial selectivity, with information security and political stability acting as implicit filters that steer operators toward institutionally predictable environments (Kheybari et al., 2020). Connectivity represents an equally binding constraint, with network connection costs, power grid access, and infrastructure capacity collectively narrowing the range of geographically viable sites (Daim et al., 2013; Kheybari et al., 2020). Environmental conditions, particularly free cooling potential and energy saving, influence operational efficiency but have received less systematic empirical treatment than economic and infrastructure factors.

The closest empirical precedent is Bagnasco (2026), who applies a binary outcome model at U.S. county level and finds that GDP per capita and population are the strongest predictors of data center presence, while electricity prices lose significance once structural regional characteristics are controlled for. While methodologically close to our approach, this work is confined to the U.S. context, leaving open whether the same structural logic holds across NUTS3 regions with markedly different institutional and infrastructure conditions.

To sum up, previous work has not yet addressed the joint empirical modelling of economic, connectivity, and environmental determinants of data centre location at NUTS3 resolution across Europe.

2. Data and Exploratory Analysis

2.1 Dataset and Feature Engineering

The dataset covers 1,345 NUTS3 regions, drawing from Eurostat (GDP, population), Ookla Speedtest Open Data (download speed, latency, test counts), the European IXP Directory, Copernicus climate reanalysis (wind speed, GHI, SPEI), Eurostat energy statistics, and the E-PRTR pollution register (air and water releases). Sources were selected to capture the economic (H1), connectivity (H2), and environmental (H3) dimensions of the analysis. The binary target has_datacenter equals 1 if any data center is recorded in the region. Of the 1,345 regions, 593 (44.1%) are positive, yielding a near-balanced class distribution that makes accuracy a meaningful supplement to AUC. As shown in Figure 1, the class split is 752 regions without a data center (55.9%) and 593 with at least one (44.1%), confirming that no resampling or class-weighting corrections were required. Missing values were handled during data cleaning prior to feature engineering. Skewed variables — GDP, population, industrial releases, IXP counts — were log1p transformed prior to modelling.

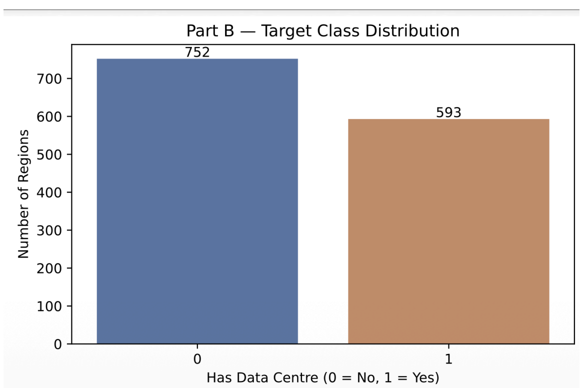


Figure 1. *Target Class Distribution across 1,345 European NUTS3 Regions*

2.2 Feature Selection

Feature selection followed absolute Pearson correlation with the target, computed on the full dataset as an exploratory measure consistent with the analysis scale. Table 1 shows the nine features retained. GDP is the strongest correlate (|r| = 0.513), followed by total internet tests and population. Environmental variables did not reach the top nine, providing early support for H3. log_ixp_count was excluded despite ranking tenth due to perfect collinearity with has_ixp, which produced numerically unstable coefficients in the logistic model.

Table 1. Features selected for modelling, ranked by absolute Pearson correlation with has_datacenter

Rank	Feature	|r|	Domain
1	log_gdp_eur_million	0.513	Economic
2	log_total_tests	0.430	Connectivity
3	log_population	0.362	Economic
4	log_climate_losses_million_eur_1980_2024	0.353	Environment
5	log_Water_Releases	0.339	Pollution proxy
6	log_Air_Releases	0.294	Pollution proxy
7	avg_download_mbps	0.284	Connectivity
8	has_ixp	0.257	Connectivity
9	log_avg_latency_ms	0.253	Connectivity

3. Methodology

All models share a stratified 80/20 split (random_state = 42), producing 1,076 training and 269 held-out test observations at identical class proportions. The shared split ensures valid comparisons between models on the test set.

Logistic Regression. We fit a binary logistic regression via maximum likelihood using statsmodels (sm.Logit), yielding coefficients, p-values, and confidence intervals. Features were standardised with StandardScaler fitted only on the training set. Country fixed effects were deliberately excluded to preserve coefficient interpretability. Stability was confirmed through five-fold stratified cross-validation (CV AUC = 0.827 ± 0.015). Predictions were thresholded at 0.5 for class assignment. **Random Forest.** A random forest classifier was tuned on the training set via five-fold cross-validation optimising AUC-ROC, varying tree count, maximum depth, and minimum samples per leaf. The final configuration: 500 trees, maximum depth of 10, improved CV AUC from 0.817 to 0.823. Unlike the logistic model, the random forest incorporates country dummies to capture macro-level fixed effects. Feature importances are reported as normalised mean decrease in impurity.

4. Evaluation

The random forest marginally outperforms the logistic regression on the held-out test set (AUC 0.775 vs. 0.764), though the difference is small (Table 2). That both models perform comparably, despite one using 54 features and country dummies and the other just 9, points to a relationship that is largely linear and well captured by the core economic and connectivity predictors alone. Both models achieve identical accuracy (0.710) and near-identical F1-scores. The logistic regression additionally yields a McFadden pseudo R^2 of 0.283, indicating reasonable explanatory power for a binary model of this type. Both models predict absence more reliably than presence, F1 of 0.74 vs. 0.67, reflecting the asymmetric structure of the outcome: absence requires only one necessary condition to fail, while presence demands all to be met simultaneously.

Table 2. Predictive performance on the 20% holdout test set (n = 269 regions)

Metric	Logistic Regression	Random Forest
Test AUC-ROC	0.764	0.775
Test Accuracy	0.710	0.710
F1 — No DC (class 0)	0.74	0.740
F1 — Has DC (class 1)	0.67	0.670
5-Fold CV AUC	0.827 ± 0.015	0.823

Figures 2 and 3 illustrate the classification behaviour and discriminative power of the logistic regression on the held-out test set. As shown in Figure 2, the model correctly identifies 112 true negatives and 79 true positives at the default 0.5 threshold, with 38 false positives and 40 false negatives, confirming the pattern seen in Table 2, where absence is predicted more reliably than presence (F1 0.74 vs. 0.67). Figure 3 shows the ROC curve achieving an AUC of 0.764, with a notably steep rise at low false positive rates (reaching ~0.65 true positive rate at FPR = 0.1), indicating strong discrimination for clear-cut cases, before flattening across the mid-range where ambiguous borderline regions prove harder to separate.

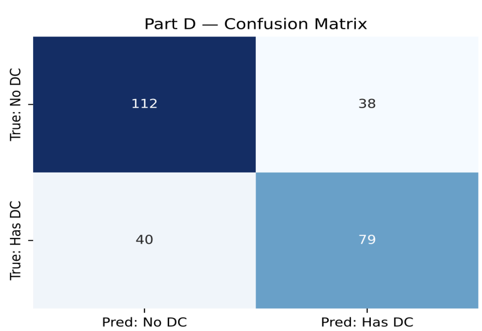


Figure 2. *Confusion Matrix — Logistic Regression (Test Set, n = 269)*

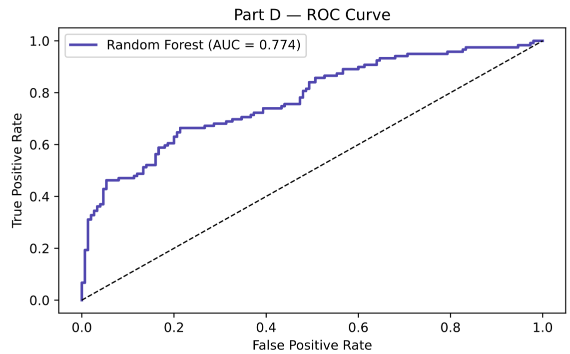


Figure 3. *ROC Curve — Logistic Regression (AUC = 0.764)*

Both models converge on the same top two predictors, GDP and total internet tests, consistent with H1 and H2 (Table 3). In the logistic regression, six of nine predictors are statistically significant ($p < 0.05$), with GDP (OR = 4.14), total tests (OR = 1.41), and IXP presence (OR = 1.28) showing the strongest positive effects, and latency (OR = 0.78) and population (OR = 0.68) the strongest negative ones. The key divergences concern population and pollution variables: the random forest assigns population positive importance (rank 4) while the logistic regression returns a negative coefficient once GDP is controlled for, and log_Water_Releases ranks third in the random forest but is not statistically significant in the logistic model.

Table 3. Top predictors by model

Rank	Logistic Regression	Random Forest
1	log_gdp_eur_million (+) OR = 4.14	log_gdp_eur_million (0.135)
2	log_total_tests (+) OR = 1.41	log_total_tests (0.094)
3	log_Air_Releases (+) OR = 1.34	log_Water_Releases (0.068)*
4	has_ixp (+) OR = 1.28	log_population (0.068)
5	log_avg_latency_ms (−) OR = 0.78	log_climate_losses_million_eur_1980_2024 (0.065)
6	log_population (−) OR = 0.68	log_Air_Releases (0.058)*
7	log_Water_Releases (n.s.)	log_dist_nearest_ixp_km (0.046)
8	avg_download_mbps (n.s.)	avg_download_mbps (0.043)
9	log_climate_losses (n.s.)	log_avg_latency_ms (0.038)

10	—	heat_mortality_per100k_presentC (0.037)

5. Discussion

H1 — Economic Scale: Regional GDP emerges as the dominant predictor in both models - recording an odds ratio of 4.14 in the logistic regression and a feature importance of 0.135 in the random forest - confirming that larger economies generate stronger demand for digital services while offering the labour markets, institutional stability, and complementary infrastructure that large-scale investment requires. This is consistent with the broader literature: Kheybari et al. (2020) and Daim et al. (2013) consistently identified economic scale as the primary siting criterion, with operational costs and financial viability setting hard boundaries on where development is feasible. Most directly comparable to our work, Bagnasco (2026) found that GDP per capita and population size are the strongest predictors of data center presence at U.S. county level, with agglomeration forces and market size playing a central role in shaping the spatial distribution of digital infrastructure. Our results mirror this finding at NUTS3 resolution and extend it in two important ways: first, we confirm that economic scale remains the dominant driver not only across countries but within them - a finding the national-scale literature could not establish; second, unlike Bagnasco (2026), who operates in a single-country context with relatively homogeneous institutional conditions, our cross-European setting introduces substantial variation in regulatory environments, grid structures, and market maturity, making the consistency of the GDP effect across such diverse contexts a stronger and more generalizable result. The geopolitical dimension identified by Kheybari et al. (2020), whereby operators avoid politically unstable environments, further reinforces this pattern, as economically developed European regions tend to combine market size with institutional predictability. For policymakers, this implies that broadening the economic base of lagging regions is likely a more effective lever for attracting data center investment than targeted infrastructure incentives alone.

H2 — Digital Connectivity: Connectivity features emerge prominently in both models. IXP presence carries a positive coefficient (OR = 1.28) and latency a negative one (OR = 0.78), confirming that well-connected regions are significantly more attractive sitting locations - consistent with Kheybari et al. (2020), who identified network connection costs and grid accessibility as the primary investment barriers that narrow the range of geographically viable sites, and with Daim et al. (2013), who placed geographical and infrastructure considerations as the largest single criterion category in their hierarchical decision model. A notable additional finding is the negative population coefficient in the logistic model despite population's positive importance in the random forest, suggesting that economic density rather than raw headcount is the operative siting mechanism. This implies that mid-sized but economically active regions can compete with large population centres when connectivity infrastructure is sufficiently strong - a nuance that the prior literature, focused largely on aggregate infrastructure availability, does not explicitly capture. Moreover, for regional policy, this suggests that investing in internet exchange infrastructure and broadband capacity can attract data centre investment even in regions that lack the economic scale of major metropolitan centres.

H3 — Environmental Factors: No environmental variable achieves statistical significance in either model, and climate-related variables appear only in the lower ranks of the random forest, most plausibly reflecting co-location with economically active regions rather than genuine siting effects. This stands in partial tension with Kheybari et al. (2020), who identified free cooling potential and energy saving as the most important environmental siting criteria. However, our results suggest these factors matter at the operational and facility design level but do not function as binding constraints at the regional location decision level - where grid-level renewable procurement and power purchase agreements effectively decouple siting from local energy geography, a pattern that persists even as corporate sustainability commitments have tightened in recent years. **Pollution Variables as Urban Proxies.** Air and water releases rank in both models but carry no causal interpretation - they proxy for economic urbanisation and industrial activity rather than genuine environmental drivers. log_Water_Releases ranks third in the random forest despite 25.5% of non-EU regions reporting zero values, limiting its reliability as a cross-border predictor and warranting caution when applying the model outside the EU core.

H4 — Electricity Prices: Contrary to expectations given the energy intensity of data centre operations, electricity prices show no significant effect in either model, suggesting European operators effectively decouple location decisions from local energy pricing through grid-level procurement. This finding is consistent with Bagnasco (2026), who found the same pattern at US county level

Authors:
Ahmad A. S. Metwaly (s358694)
Amir Khakimov (s358799)
Bengt Wesselink (s358632)
Mirdana Nurtayeva (s359925)
Nayib Huseynzade (s339277)

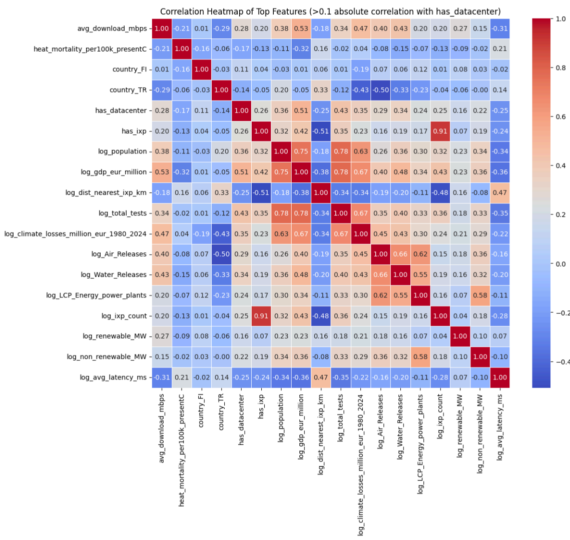
References

[1] Kheybari, S., Davoodi Monfared, M., Farazmand, H., & Rezaei, J. (2020). Sustainable location selection of data centers. International Journal of Information Technology and Decision Making, 19(3), 741–773.

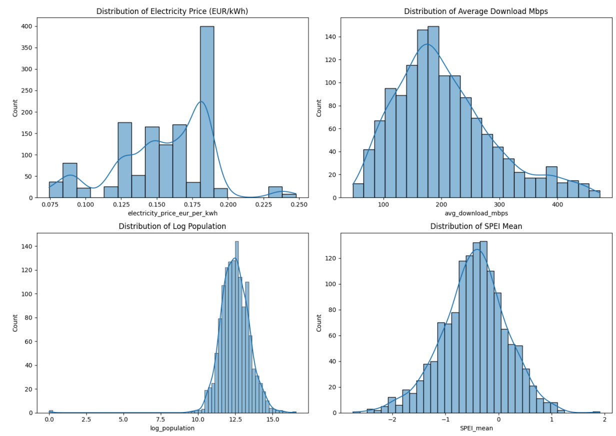
[2] Daim, T. U., Bhatla, A., & Mansour, M. (2013). Site selection for a data center — a multi-criteria decision-making model. International Journal of Sustainable Engineering, 6(1), 10–22.

[3]Bagnasco, M. (2026). Spatial determinants of AI data center location: A county-level econometric analysis in the United States [Master's thesis, Politecnico di Torino]. Politecnico di Torino Repository.

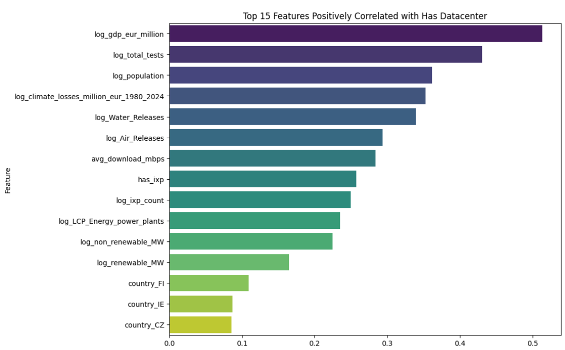
Appendix



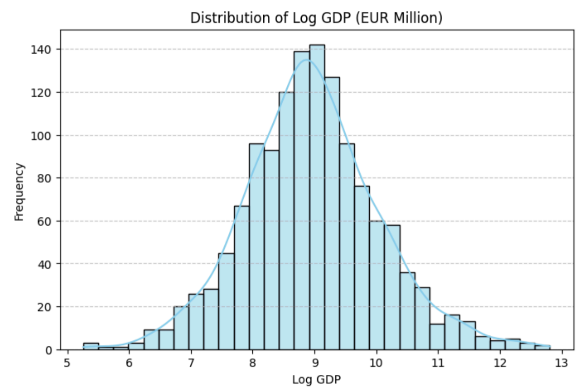
Correlation HeatMap of Top Features



Distribution of features



Top 15 features positively correlated



LogGDP distribution