nature cities

Supplementary information https://doi.org/10.1038/s44284-026-00487-z

Urban infrastructure and fossil fuel industrial legacy drive US data center siting

In the format provided by the authors and unedited

Supplementary Discussion

Supplementary Figures

Supplementary Tables

Supplementary Table 1: **Negative binomial model for the number of data center at the city level (N = 910).** Entries are coefficients with standard errors (SE), z-statistics and P-values.

Predictor	Coeff.	SE	z	P
Nameplate capacity	1.324	0.041	32.668	< 0.001
IT employment	0.719	0.056	12.779	< 0.001
Broadband quality	0.444	0.052	8.560	< 0.001
Natural hazards	-0.476	0.062	-7.670	< 0.001
Water stress index	0.006	0.053	0.113	0.910
Energy expenditure	0.130	0.051	2.563	0.010
Retired coal plants	0.162	0.041	4.004	< 0.001
State tax incentive	0.126	0.053	2.382	0.017
Electricity-price variability	-0.041	0.049	-0.843	0.399
Energy grid stress	0.077	0.053	1.459	0.144
Constant	0.127	0.052	2.433	0.015

S1 Clustering analysis

To characterize the spatial distribution of data centers within cities, we cluster data center locations using the k-means algorithm [9]. Data center coordinates are first projected from WGS84 (EPSG:4326) to an equal-area coordinate system for the continental US (EPSG:5070), so that all distances are measured in meters. Clustering is performed independently within each city. To avoid degenerate or non-informative solutions, we restrict the analysis to cities with exact locations in the contiguous USA hosting at least three data centers, yielding a total of 148 cities for the analysis. For each of them, the number of clusters k is selected adaptively from 2 to 6 (a trade-off to identify multi-core configurations without over-partitioning), by maximizing the average silhouette score computed under Euclidean distance. For each candidate k, the k-means algorithm is run with multiple random initializations and the value $\hat{k}$ that maximizes the silhouette score is selected.

For data center j of coordinates $\boldsymbol{x}_j$ belonging to cluster c, we compute the within-cluster distance,

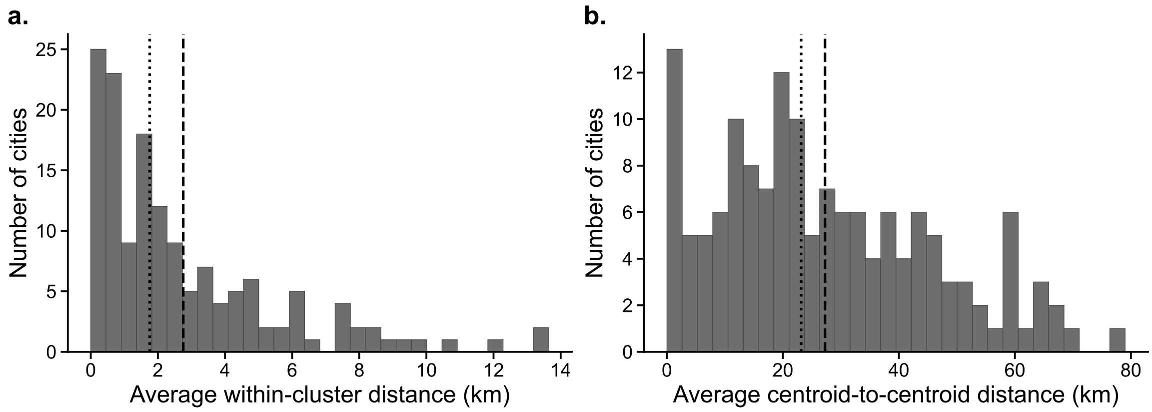
$$d_j = \|\boldsymbol{x}_j - \boldsymbol{\mu}_c\|, \tag{1}$$

where $\boldsymbol{\mu}_c$ denotes the centroid of the cluster and $\|\cdot\|$ is the Euclidean norm. Within each city, we summarize these distances by their mean and median. To characterize separation between clusters c and c', we compute all pairwise distances between cluster centroids,

$$D_{c,c'} = \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|. \tag{2}$$

Again, within each city, we summarize these distances by their mean and median. Supplementary Figure 1 reports the empirical distributions of (i) the average within-cluster distance and (ii) the average centroid-to-centroid distance across cities.

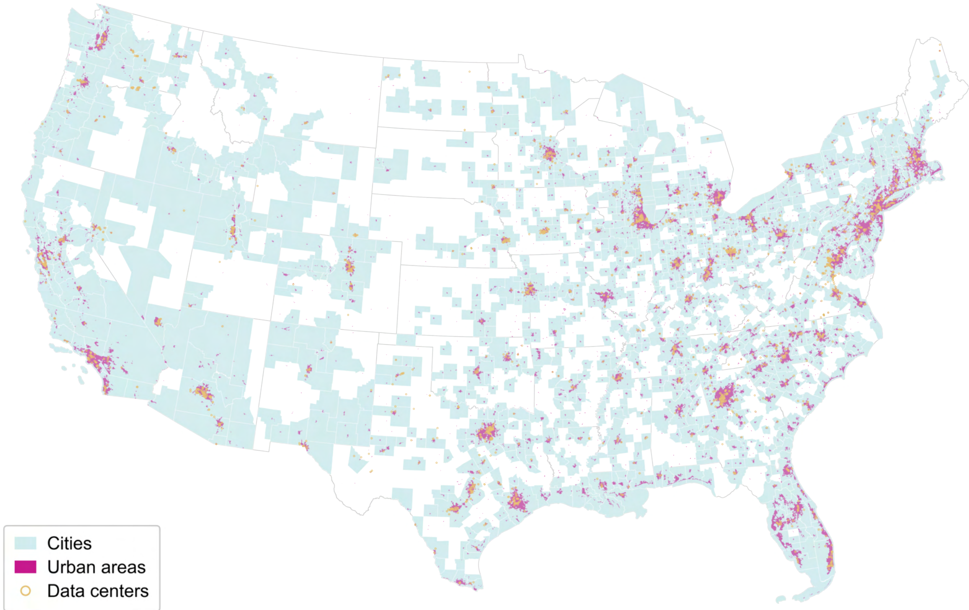
Across the 148 cities retained for analysis, the distribution of average within-cluster distances has a sample mean of 2.75 km and a median of 1.76 km (Supplementary Figure 1a). This indicates that most data center clusters are spatially compact, with a right-skewed tail reflecting a small number of more dispersed configurations. In contrast, the distribution of average centroid-to-centroid distances has a sample mean of 27.26 km and a median of 23.15 km (Supplementary Figure 1b), highlighting substantial heterogeneity in the spatial separation between clusters within cities.



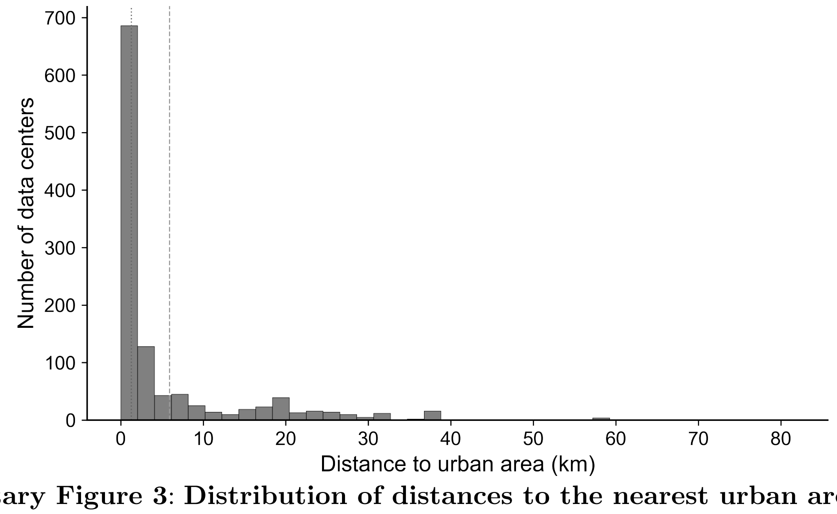
Supplementary Figure 1: **Distributions of clustering distances from k-means analysis**. **a.** Average within-cluster distance (km). **b.** Average centroid-to-centroid distance between clusters (km). Dashed and dotted vertical lines denote the mean and median of each distribution, respectively.

S2 Analysis of concentration of data centers in cities and urban areas

To quantify whether data centers are more likely to locate in cities (defined as MSAs and MicroSAs) rather than being equally distributed across counties, we partition counties in the contiguous USA into two categories: (i) counties within cities ("city counties") and (ii) counties outside cities ("rural counties"). A one-sided Welch two-sample t-test rejects the null hypothesis of equivalent distributions between the two groups ($t(3028.89) = 8.28$, $p < 0.001$).



Supplementary Figure 2: **Spatial distribution of data centers relative to urban areas.** Data centers (yellow open circles; $N = 4,283$) are shown over cities (metropolitan and micropolitan statistical areas; $N = 917$) in light blue and Census-defined urban areas in magenta in the contiguous USA. Data are from the Data Center Map database [1] and the Census Bureau TIGER/Line Shapefiles 2020 and 2024 [2].

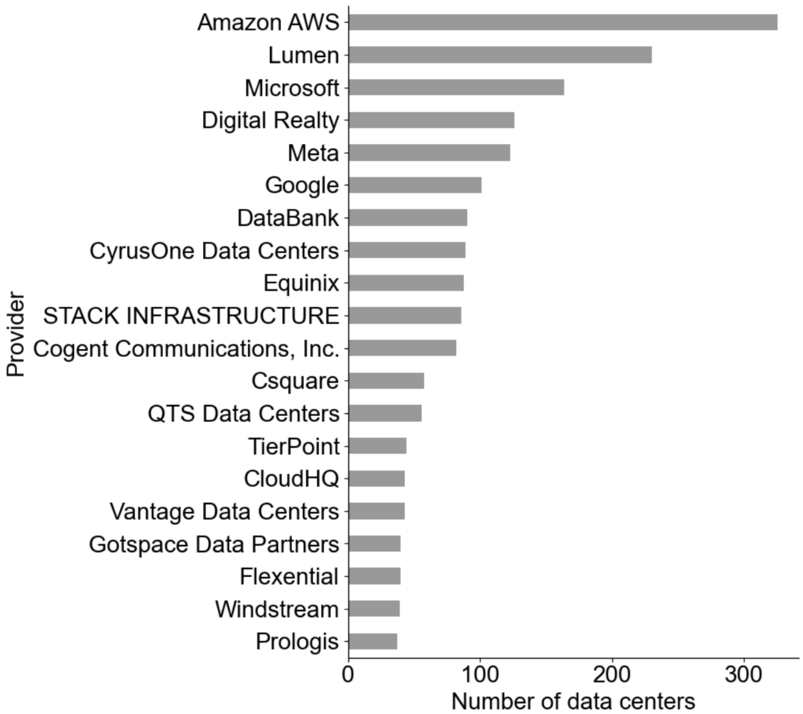


Supplementary Figure 3: **Distribution of distances to the nearest urban area.** Histogram of distances from data centers ($N = 4,283$) to the nearest Census-defined urban area. Vertical dashed and dotted lines indicate the mean (5.9 km) and median (1.26 km), respectively.

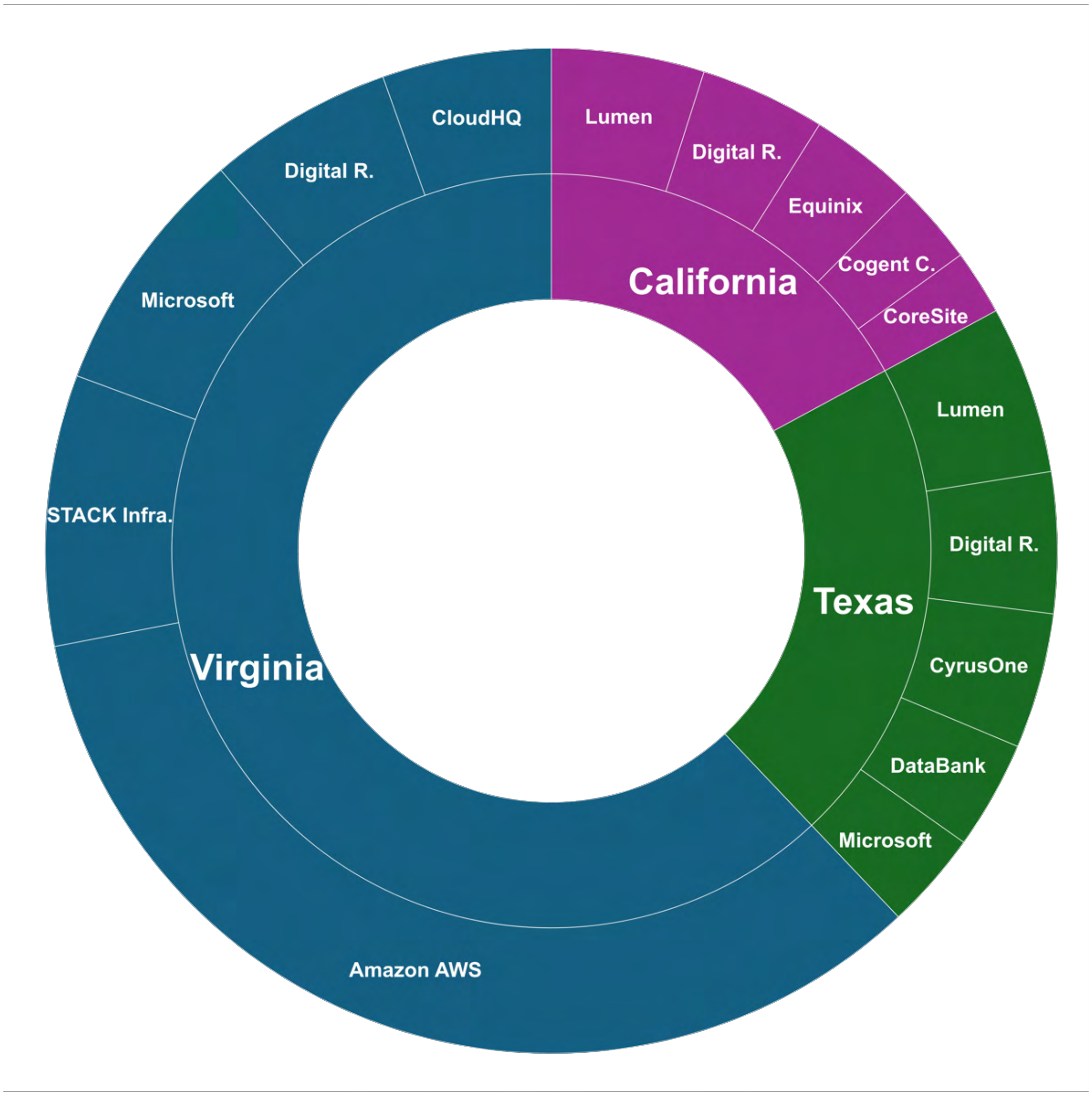
For the 2020 Census, an urban area comprises a densely settled core of census blocks that meet minimum housing unit density and/or population density requirements (at least 2,000 housing units or have a population of at least 5,000). The polygons of this geographical layer are taken from [2]. We show that 73.5% of data centers are located within these urban areas' boundaries (Supplementary Figure 2), increasing to 92.9% within 5 km, 95.1% within 10 km and 97.5% within 20 km (Supplementary Figure 3). The distribution of distances is strongly right-skewed, with a mean of 5.9 km and a median of 1.26 km (Supplementary Figure 3), indicating that data centers are embedded in or immediately adjacent to urban areas.

S3 Data centers providers across the US

Analyzing the data centers providers across the contiguous USA, we register a sharp contrast between national and local market structures. At the national level, the market is moderately concentrated: despite hundreds of distinct providers (Supplementary Figure 4 for the top 20 provider distribution), the top ten firms account for 34.2% of contiguous USA data center facilities and the largest provider accounts for 7.7% of the national stock (see Supplementary Figure 5 for the top 5 providers in the Virginia, Texas and California). At the city-level, instead, data centers are typically operated by a limited subset of providers (median: 2 and interquartile range: 1-5), resulting in high local concentration even in the absence of national-level dominance. A single provider operates more than 50% of data centers in 52.0% of cities with at least one data center, indicating that competitive diversity at the national level does not translate into evenly distributed local competition.



Supplementary Figure 4: **Distribution of the top 20 data center providers in the contiguous USA.** A bar chart illustrating the number of data centers operated by each provider; only facilities located within cities are included.



Supplementary Figure 5: **Sunburst chart of the top five data center providers in the three US states with the most data centers.** The inner ring represents the states of Virginia, Texas and California, while the outer ring depicts the top five providers within each state. Segment sizes indicate each provider's share among the top five. The totals correspond to the combined facilities of these leading providers: Virginia (356 of 667), Texas (120 of 429) and California (98 of 326). Virginia is primarily served by hyperscale providers, whereas Texas and California exhibit a more diverse mix of providers.

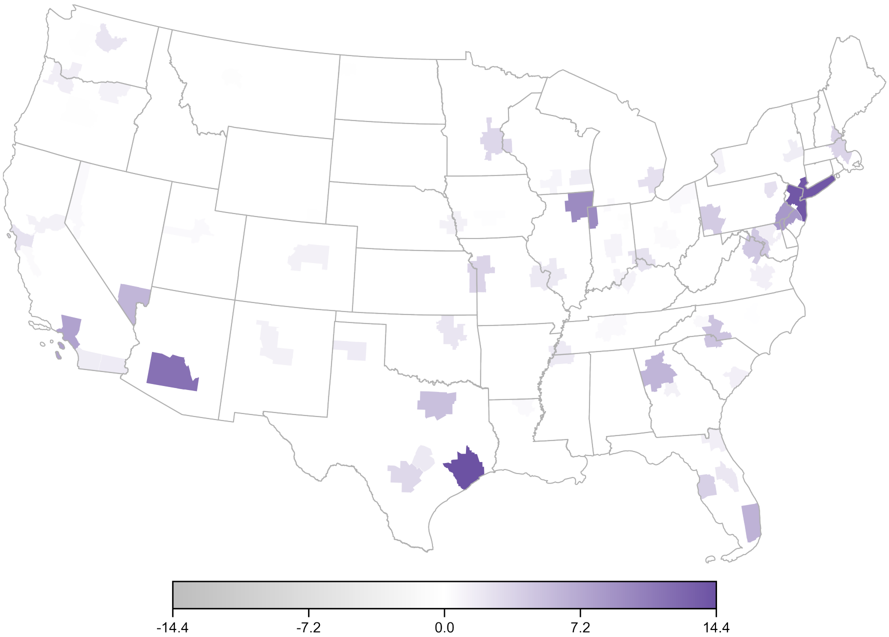
S4 Local contributions of the top three regressors

To further investigate the spatial heterogeneity underlying the top three predictors of data center siting, we compute city-level local contributions of nameplate generation capacity, IT employment and broadband quality. Consistent with the analysis in Section 2 of the main text, all regressors are standardized to zero mean and unit variance prior to estimation. For city i and predictor j, the local contribution is computed as $\beta_j z_{ij}$, where β_j is the estimated global model coefficient and z_{ij} is the standardized local value of the predictor. Rather than identifying the dominant predictor, we visualize the contribution associated with each regressor individually (Supplementary Figs. 6–8). For better visualization, different from Figure 2b of the main text, we restrict the analysis to cities hosting at least ten data centers. Alongside the visualizations, we also present the top ten cities for each of the top three regressor (Supplementary Table 2).

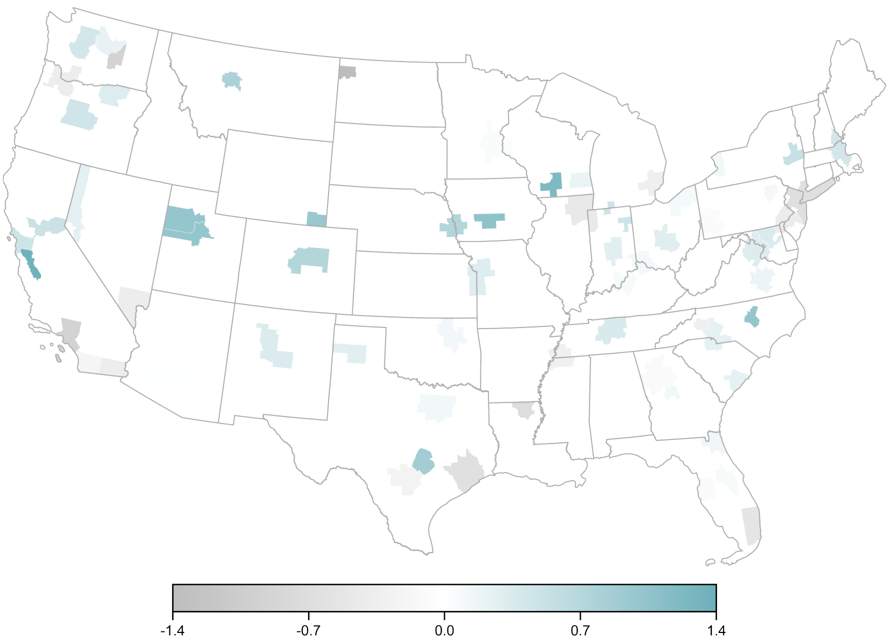
Nameplate generation capacity exhibits the strongest contributions in major energy-intensive metropolitan areas with strong knowledge-based service economy that cut across finance, insurance, healthcare and logistics, such as Houston-Pasadena-The Woodlands, New York-Newark-Jersey City, Phoenix-Mesa-Chandler, Chicago-Naperville-Elgin and Philadelphia-Camden-Wilmington. This evidence confirms the central role of large-scale electrical infrastructure in shaping data center geography discussed in the main text. IT employment metric highlights metropolitan areas characterized by strong technological specialization, including San Jose-Sunnyvale-Santa Clara, Madison, Des Moines-West Des Moines, Salt Lake City-Murray and Provo-Orem-Lehi. Several of these cities are associated with well-established technology ecosystems, including Silicon Valley (San Jose-Sunnyvale-Santa Clara) and the Silicon Slopes corridor (Salt Lake City-Murray and Provo-Orem-Lehi), while others are characterized by research-intensive environments (Madison) or major cloud-computing and digital-infrastructure clusters (Des Moines-West Des Moines). Importantly, IT employment metric used in the main model is a SAMI-normalized metric rather than an absolute count. In this vein, the strength of a contribution should be interpreted with respect to the nominal, baseline value set by population size. Broadband quality contributions are concentrated in major metropolitan areas with strong economies (finance, insurance, healthcare and logistics), including New York-Newark-Jersey City, Nashville, Philadelphia-Camden-Wilmington, Miami-Fort Lauderdale-West Palm Beach and Baltimore-Columbia-Towson.

Supplementary Table 2: Top cities ranked by local contribution of the three dominant regressors of data center siting. Local contributions were computed from the fitted negative binomial model; note that the IT employment variable corresponds to a SAMI-normalized metric. The analysis is restricted to cities hosting at least ten data centers ($N = 67$).

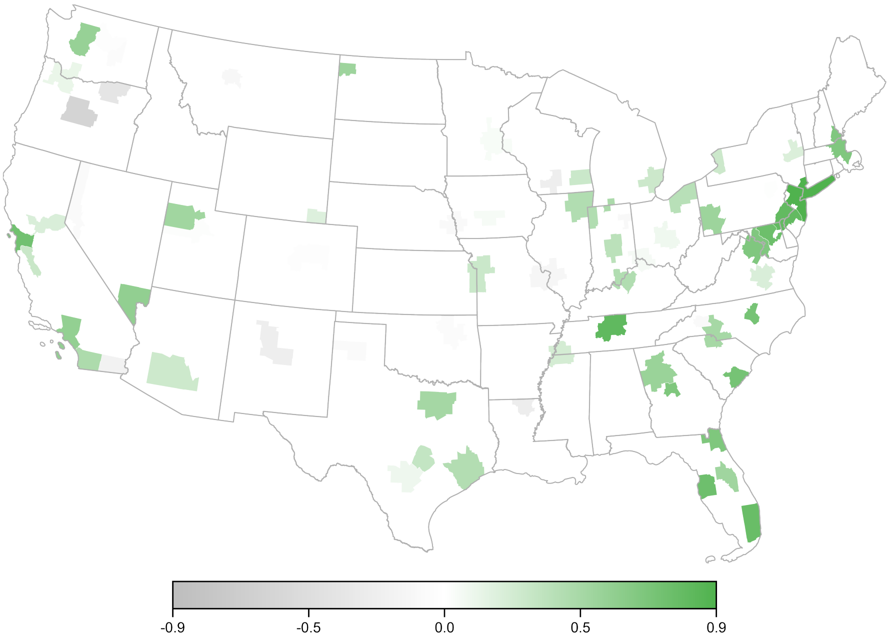
Regressor	Rank	City	Contribution
Nameplate capacity	1	Houston-Pasadena-The Woodlands, TX	14.43
	2	New York-Newark-Jersey City, NY-NJ	13.93
	3	Phoenix-Mesa-Chandler, AZ	11.80
	4	Chicago-Naperville-Elgin, IL-IN	9.59
	5	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	8.50
	6	Los Angeles-Long Beach-Anaheim, CA	7.73
	7	Miami-Fort Lauderdale-West Palm Beach, FL	6.39
	8	Atlanta-Sandy Springs-Roswell, GA	6.11
	9	Las Vegas-Henderson-North Las Vegas, NV	6.04
	10	Dallas-Fort Worth-Arlington, TX	5.19
IT employment	1	San Jose-Sunnyvale-Santa Clara, CA	1.33
	2	Madison, WI	1.20
	3	Des Moines-West Des Moines, IA	1.04
	4	Salt Lake City-Murray, UT	0.99
	5	Provo-Orem-Lehi, UT	0.98
	6	Raleigh-Cary, NC	0.98
	7	Cheyenne, WY	0.95
	8	Austin-Round Rock-San Marcos, TX	0.86
	9	Great Falls, MT	0.77
	10	Omaha, NE-IA	0.71
Broadband quality	1	New York-Newark-Jersey City, NY-NJ	0.94
	2	Nashville-Davidson-Murfreesboro-Franklin, TN	0.85
	3	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	0.83
	4	Miami-Fort Lauderdale-West Palm Beach, FL	0.80
	5	Baltimore-Columbia-Towson, MD	0.78
	6	Tampa-St. Petersburg-Clearwater, FL	0.77
	7	San Francisco-Oakland-Fremont, CA	0.75
	8	Charleston-North Charleston, SC	0.73
	9	Raleigh-Cary, NC	0.73
	10	Jacksonville, FL	0.69



Supplementary Figure 6: **City-level local contribution of nameplate generation capacity to data center siting estimated by the negative binomial model.** Local contributions were computed as $\beta_j z_{ij}$, where β_j is the standardized global coefficient associated with regressor j and z_{ij} is the standardized local regressor value for city i. Positive values indicate cities where the regressor contributes positively to the modeled prevalence of data center siting, while negative values indicate the opposite. The analysis is restricted to cities hosting at least ten data centers ($N = 67$). Data are from the Data Center Map database [1], the USA Energy Information Administration [3] and the Census Bureau TIGER/Line Shapefiles 2024 [2].



Supplementary Figure 7: **City-level local contribution of IT employment to data center siting estimated by the negative binomial model.** The IT employment variable corresponds to a SAMI-normalized metric. Local contributions were computed as $\beta_j z_{ij}$, where β_j is the standardized global coefficient associated with regressor j and z_{ij} is the standardized local regressor value for city i. Positive values indicate cities where the regressor contributes positively to the modeled prevalence of data center siting, while negative values indicate the opposite. The analysis is restricted to cities hosting at least ten data centers ($N = 67$). Data are from the Data Center Map database [1], the USA Bureau of Labor Statistics [4], the USA Census Bureau [5] and the Census Bureau TIGER/Line Shapefiles 2024 [2].



Supplementary Figure 8: **City-level local contribution of broadband quality to data center siting estimated by the negative binomial model.** Local contributions were computed as $\beta_j z_{ij}$, where β_j is the standardized global coefficient associated with regressor j and z_{ij} is the standardized local regressor value for city i. Positive values indicate cities where the regressor contributes positively to the modeled prevalence of data center siting, while negative values indicate the opposite. The analysis is restricted to cities hosting at least ten data centers ($N = 67$). Data are from the Data Center Map database [1], broadband quality score by [6] and the Census Bureau TIGER/Line Shapefiles 2024 [2].

S5 County-level robustness analysis

As a robustness check, we re-estimate the negative binomial model at the county level ($N = 2{,}137$; Supplementary Table 3). We confirm the dominant role of nameplate capacity in the spatial logic of data centers, identified in the main text (Fig. 2a; see also Supplementary Table 3). Also, we confirm that IT employment, broadband quality and retired coal plants are positively associated with the number of data centers (Fig. 2a; see also Supplementary Table 3). Different from the main text, we fail to identify an association between data centers and natural hazards, energy expenditure, or state level incentives (Fig. 2a; see also Supplementary Table 3). At the county-level, we also identify two new drivers that are not identified in the main text (Fig. 2a; see also Supplementary Table 3), namely, electricity-price variability and energy grid stress. In both cases, we document a negative association, suggesting that providers may tend to prefer counties with more stable energy prices and lower risk of power outages. Overall, these findings confirm the critical role of infrastructure and industrial legacy on county-level siting, while warning prudence against conclusions about the specific role of environmental and economic factors on data center logic.

To further assess the robustness of our results, we extend the county-level negative binomial analysis by including two additional regressors: land prices and community wealth. Land prices were estimated using the data by [10] which is based on property transactions, expressed as log-transformed USD per hectare (deflated to January 2017). Because land values scale with population, we control for this dependence using the scale-adjusted metropolitan indicator (SAMI), namely, the residual from the log–log scaling relationship between land values and county population size [11]. Community wealth was scored as the percentage of individuals below the 100% poverty threshold derived from the American Community Survey (ACS 5-year estimates, 2023; table S1701, variable S1701_C03_001E). Results are reported in Supplementary Table 4.

We find no statistically significant association between data center counts and land prices. Poverty exhibits a small, but statistically significant, negative association ($\beta = -0.097$, $p = 0.039$), indicating a slightly lower prevalence of data centers in counties with higher shares of low-income populations. This finding is consistent with the positive association we identify for IT employment, which can be interpreted as a proxy for local economic activity and demand. Taken together, these results suggest that, at the county level, land prices and community wealth have a less salient role than infrastructure and industrial legacy. We warn prudence, however, in drawing conclusions on this analysis due to some limitations: (i) both land prices and community wealth are likely to exhibit substantial within-county heterogeneity and their effects may therefore be more accurately captured at finer spatial resolution; (ii) the land price estimates are based on a model calibrated on 2010 data and may not fully capture temporal dynamics in land markets; and (iii) data centers are often located in leased or repurposed facilities, which may weaken the direct relationship between local land values and siting decisions.

Supplementary Table 3: Negative binomial model for the data center count at the county level ($N = 2{,}137$). Entries are coefficients with standard errors (SE), z-statistics and P-values.

Predictor	Coeff.	SE	z	P
Nameplate capacity	0.458	0.026	17.437	< 0.001
IT employment	0.302	0.035	8.578	< 0.001
Natural hazards	−0.059	0.042	−1.394	0.163
Broadband quality	1.026	0.039	26.300	< 0.001
Retired coal plants	0.104	0.028	3.750	< 0.001
Energy expenditure	−0.001	0.037	−0.026	0.979
State tax incentive	−0.043	0.039	−1.125	0.261
Energy grid stress	−0.489	0.037	−13.341	< 0.001
Electricity-price variability	−0.110	0.034	−3.249	0.001
Water stress index	−0.018	0.040	−0.459	0.646
Constant	−0.458	0.041	−11.11	< 0.001

Supplementary Table 4: **Extended negative binomial model at county level (N = 2,137) including land prices and poverty.** Entries are coefficients with standard errors (SE), z-statistics and P-values.

Predictor	Coeff.	SE	z	P
Nameplate capacity	0.463	0.027	17.32	< 0.001
IT employment	0.271	0.037	7.26	< 0.001
Natural hazards	−0.067	0.048	−1.40	0.161
Broadband quality	1.016	0.039	26.06	< 0.001
Retired coal plants	0.107	0.028	3.87	< 0.001
Energy expenditure	0.012	0.037	0.33	0.739
State tax incentive	−0.047	0.039	−1.20	0.230
Energy grid stress	−0.435	0.044	−9.84	< 0.001
Electricity-price variability	−0.096	0.034	−2.82	0.005
Water stress index	−0.009	0.040	−0.22	0.826
Land prices	0.040	0.040	0.99	0.322
Poverty (share below 100%)	−0.097	0.047	−2.07	0.039
Constant	−0.459	0.041	−11.09	< 0.001

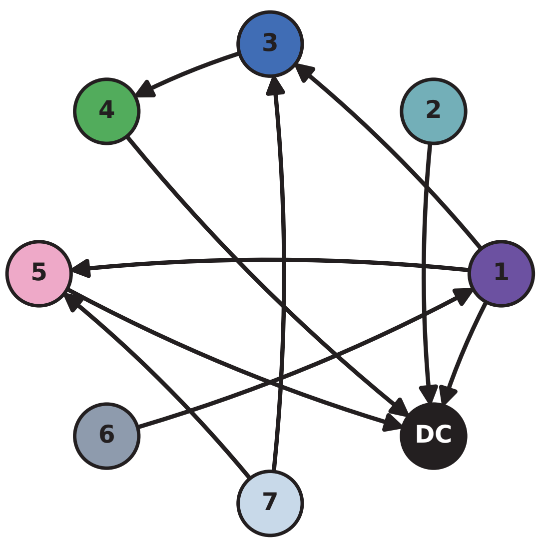
S6 Causal discovery

To complement the correlational finding from the negative binomial model, we apply a causal discovery algorithm that helps unveil the direction of the interactions. Specifically, we apply the Peter-Clark algorithm [12], which consists of two steps. First, it recovers an undirected skeleton by iteratively testing conditional independence between pairs of variables, starting from unconditional tests and gradually increasing the size of the conditioning set. Significance level is set to a conservative level of 1% to prune the graph from weak associations. Second, the remaining edges are oriented using standard propagation rules that identify v-structures and propagate directionality constraints, yielding a directed acyclic graph (DAG).

We infer causal associations between eight variables (the nodes of the DAG): the seven significant regressors identified from the negative binomial regression in the main text (Fig. 2a) – nameplate capacity, IT employment, natural hazard events, broadband quality, retired coal plant, state-level incentives and electricity expenditure – and the data center count. Independence tests are based on partial Spearman correlations and the algorithm is initialized from a fully connected undirected graph. Once the DAG is inferred, the sign of each directed edge is assigned based on the corresponding Spearman correlation computed on the full sample.

The inferred causal structure is highly asymmetric (Supplementary Figure 9). Data centers consistently appear as a sink node, receiving directed edges from a small subset of city characteristics. In particular, four strong predictors in the negative binomial regressions – nameplate capacity, IT employment, broadband quality and retired coal plant – directly point toward data center counts in the DAG. This convergence between regression results and causal discovery highlights the central role of infrastructure and industrial legacy as drivers for data center siting. Importantly, these very predictions are the ones that are stable to the selection of cities versus counties as the units of the analysis in the negative binomial model as shown in Supplementary Discussion S5). Energy expenditure, state tax incentives and natural hazards affect data center count indirectly.

The inferred graph does not identify causal effect magnitudes and does not rely on temporal ordering. Instead, it provides a structural hypothesis consistent with the conditional independence patterns observed in the city data, highlighting plausible pathways through which infrastructure and labor-market conditions jointly shape the spatial concentration of data centers.

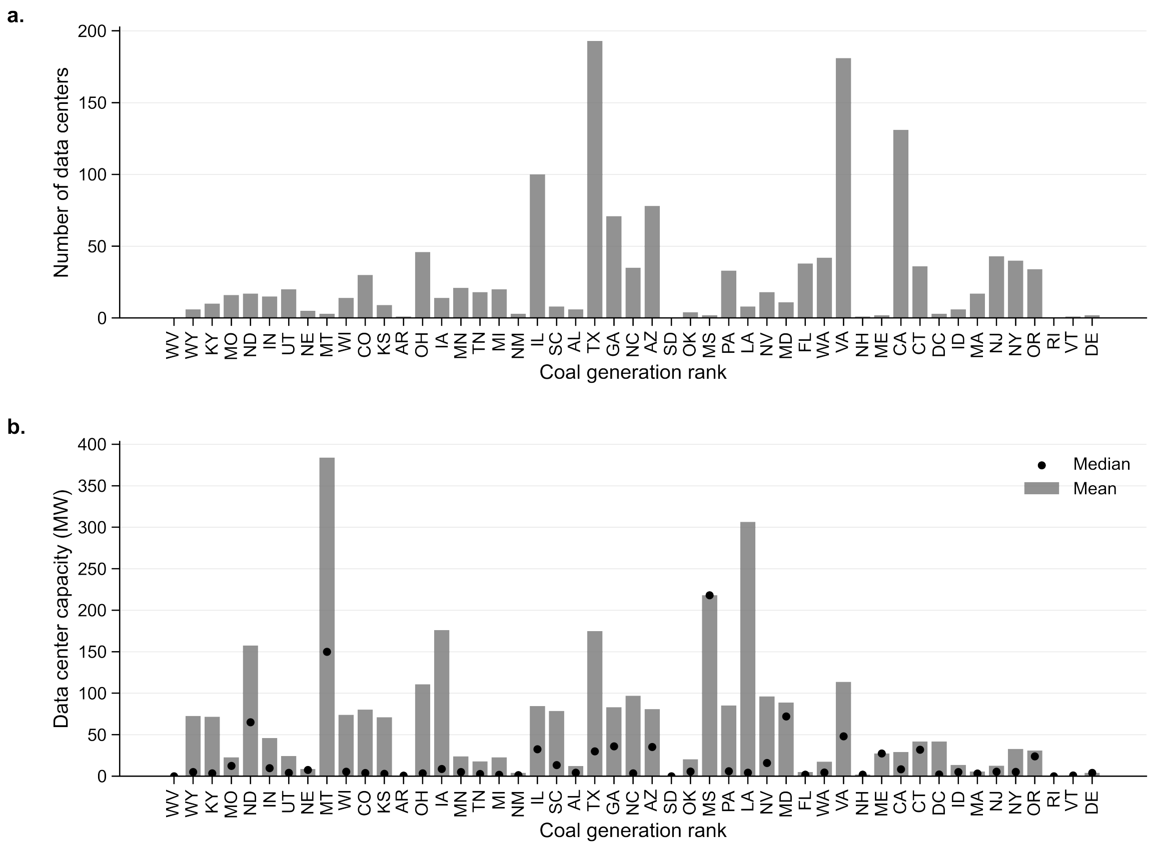


Supplementary Figure 9: **Directed acyclic graph inferred using the PC algorithm**. The nodes correspond to city-level ($N = 910$) measures of nameplate capacity (node 1), IT employment (node 2), natural hazards (node 3), broadband quality (node 4), retired coal plants (node 5), state-level incentives (node 6), energy expenditure (node 7) and data center count (DC). Arrows indicate directed edges inferred by the PC algorithm using pairwise Spearman correlation.

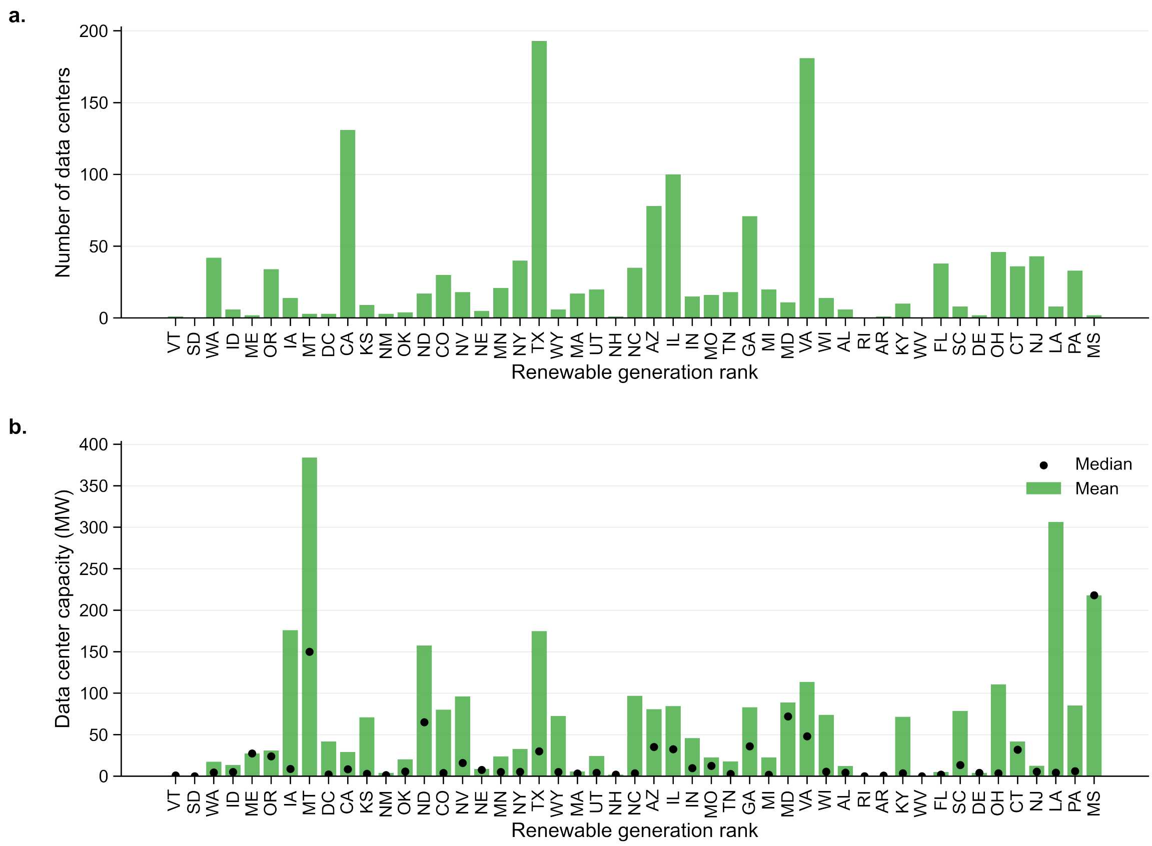
S7 State-level comparisons with electricity generation mix

To further contextualize the relationship between data center geography and energy systems, we compared state-level data center statistics against state electricity generation characteristics, including coal generation (Supplementary Figure 10) and renewable generation share (Supplementary Figure 11). Specifically, we examine the number of data centers, as well as the (mean and median) facility IT power capacity across states as a function of state rankings in coal-based and renewable electricity generation. For each electricity source, states (and the District of Columbia) in the contiguous US are ranked according to the fraction of total state electricity generation attributable to that source, computed using eGRID 2023 data. Coal generation share is defined as the ratio between annual coal-based net generation and total annual net generation. Renewable generation share is defined as the ratio between annual total renewable net generation and total annual net generation. Thus, rank 1 corresponds to the state with the highest coal generation (or renewable generation) share and rank 49 to the state with the lowest share.

The resulting patterns do not reveal meaningful associations between data center activity and state-level electricity composition. Large data center hubs are found in states characterized by substantially different energy mixes, including fossil-fuel-intensive, renewable-intensive and heterogeneous electricity resource mixes (Supplementary Figures 10a and 11a). Likewise, neither the mean nor median facility capacities exhibit systematic trends with coal generation share or renewable generation share (Supplementary Figures 10a,b and 11a,b). These findings are consistent with the interpretation offered in the main text that the relationship between data center expansion and industrial legacy is grounded in broader infrastructural conditions rather than in a direct preference for specific electricity generation mixes.

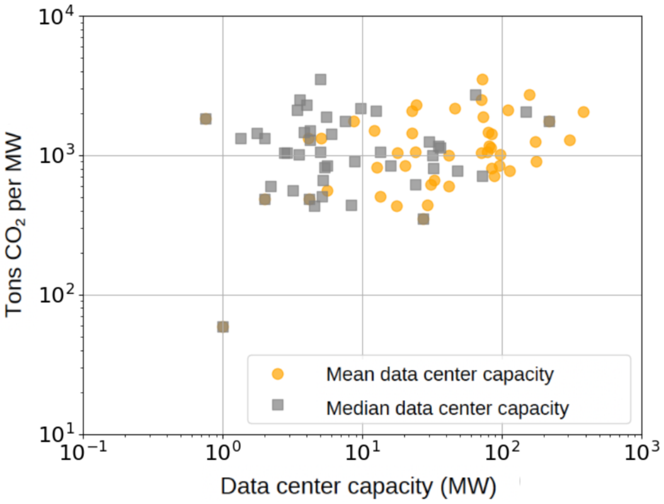


Supplementary Figure 10: **Number and capacity of data centers by state rank in coal generation. a.** Number of data centers. **b.** Facility IT power capacity (mean and median). States ($N = 49$ jurisdictions, corresponding to the 48 contiguous US states and the District of Columbia) are ordered from highest (rank 1) to lowest coal generation share and labeled using postal abbreviations.



Supplementary Figure 11: **Number and capacity of data centers by state rank in renewable energy generation. a.** Number of data centers. **b.** Facility IT power capacity (mean and median). States ($N = 49$ jurisdictions, corresponding to the 48 contiguous US states and the District of Columbia) are ordered from highest (rank 1) to lowest renewable generation share and labeled using postal abbreviations.

S8 Data centers' allocated emissions across the US



Supplementary Figure 12: **Mean and median data centers' capacity against state CO_2 emissions.** The mean and the median values of the data centers' IT power capacity (MW) are calculated by aggregating the IT capacity of all the facilities residing in a state ($N = 49$ jurisdictions, corresponding to the 48 contiguous US states and the District of Columbia). The state CO_2 emissions per MW of energy generated are computed as the ratio between the state annual CO_2 emissions per state and the state nameplate capacity in MW.

In Supplementary Figure 12, we compare the mean and median IT capacity of data centers in each state with the state's annual CO_2 emissions, in tons, per nameplate capacity, in MW. Across states, data center capacity spans a wider range than emissions intensity, indicating that variation in our estimated allocated CO_2 emissions is driven more by differences in facility size than by differences in state energy systems. We also observe substantial differences between the mean and median IT capacity within states, reflecting skewed distributions and limited availability of capacity data in some states.

Supplementary Table 5: **Data center emissions estimates by state in the contiguous USA.** Power capacity of data centers in each state (MW) and corresponding CO_2 emission estimates (tons). Values in parentheses indicate the measure used: mean or median. District of Columbia is included ($N = 49$).

State	Power capacity (mean)	Power capacity (median)	Emission estimate (mean)	Emission estimate (median)
Alabama	12.17	4.25	18207.17	6360.04
Arkansas	0.75	0.75	1379.96	1379.96
Arizona	80.72	35.30	93602.03	40935.44
California	29.27	8.31	12864.19	3652.16
Colorado	80.27	3.85	117106.84	5617.16
Connecticut	41.86	32.00	41771.32	31928.81
District of Columbia	41.90	2.20	25070.98	1316.38
Delaware	4.15	4.15	2023.55	2023.55
Florida	5.09	2.00	6796.23	2667.94
Georgia	83.05	36.00	94323.94	40885.67
Iowa	176.01	8.80	159610.97	7979.90
Idaho	13.56	5.12	6878.72	2600.13
Illinois	84.57	32.50	68354.00	26268.46
Indiana	46.03	9.77	99908.11	21208.84
Kansas	71.15	2.90	73614.89	3000.56
Kentucky	71.58	3.54	179051.92	8868.17
Louisiana	306.49	4.25	396707.73	5501.07
Massachusetts	5.62	3.20	3136.30	1785.61
Maryland	88.70	72.00	63102.69	51222.02
Maine	27.50	27.50	9633.13	9633.13
Michigan	22.69	1.75	32654.23	2518.51
Minnesota	23.95	5.00	25197.19	5259.55
Missouri	22.84	12.50	47561.59	26025.50
Mississippi	218.00	218.00	385703.55	385703.55
Montana	384.00	150.00	791551.96	309199.98
North Carolina	96.96	3.50	97733.03	3528.05
North Dakota	157.46	65.00	432097.54	178375.94
Nebraska	8.71	7.50	15401.79	13262.16
New Hampshire	2.00	2.00	970.07	970.07
New Jersey	12.72	5.40	10456.61	4438.14
New Mexico	4.12	1.35	5485.57	1798.91
Nevada	96.11	16.00	81222.44	13522.05
New York	32.86	5.25	21660.18	3461.12
Ohio	110.81	3.40	234196.42	7185.93
Oklahoma	20.41	5.62	17277.29	4761.04
Oregon	30.85	24.00	19013.21	14793.60
Pennsylvania	85.23	6.00	120752.06	8501.13
Rhode Island	NaN	NaN	NaN	NaN
South Carolina	78.71	13.45	83254.46	14226.11
South Dakota	NaN	NaN	NaN	NaN
Tennessee	17.89	2.75	18758.27	2882.84
Texas	174.78	30.00	219070.32	37602.66
Utah	24.26	4.00	55464.19	9145.34
Virginia	113.69	48.00	87518.35	36949.79
Vermont	1.00	1.00	58.82	58.82
Washington	17.60	4.50	7617.78	1948.25
Wisconsin	73.94	5.50	138940.76	10334.66
West Virginia	NaN	NaN	NaN	NaN
Wyoming	72.67	5.00	257140.14	17693.53

S9 The unfolding landscape of data centers within the Energy Community framework

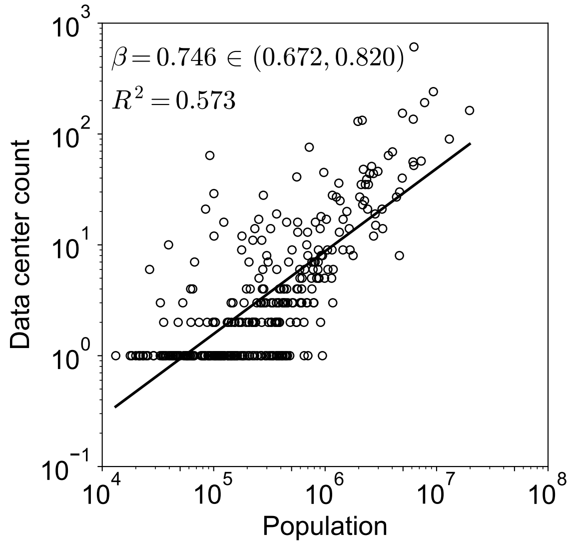
Of the 1,221 projects under development, 845 are in non-EC areas (69.21%), compared to 2,203 of 2,956 operational facilities (74.53%). Economic Revitalization ECs account for 281 facilities under development (23.01%) and 637 operational facilities (21.55%). Coal Closure EC areas account for 79 facilities under development (6.47%) and 90 operational facilities (3.04%). Areas overlapping both designations account for 16 facilities under development (1.31%) and 26 operational facilities (0.88%). Information is reported in a tabular form in Supplementary Table 6.

Supplementary Table 6: **Contingency table for the chi-square test. Number and share of operational and under-development facilities (N = 4,177 data centers) mapped against Energy Community (EC) categories**: Coal Closure EC areas (CC EC), Economic Revitalization EC areas (ER EC), areas classified as both (CC + ER) and non-EC (Non-EC) areas

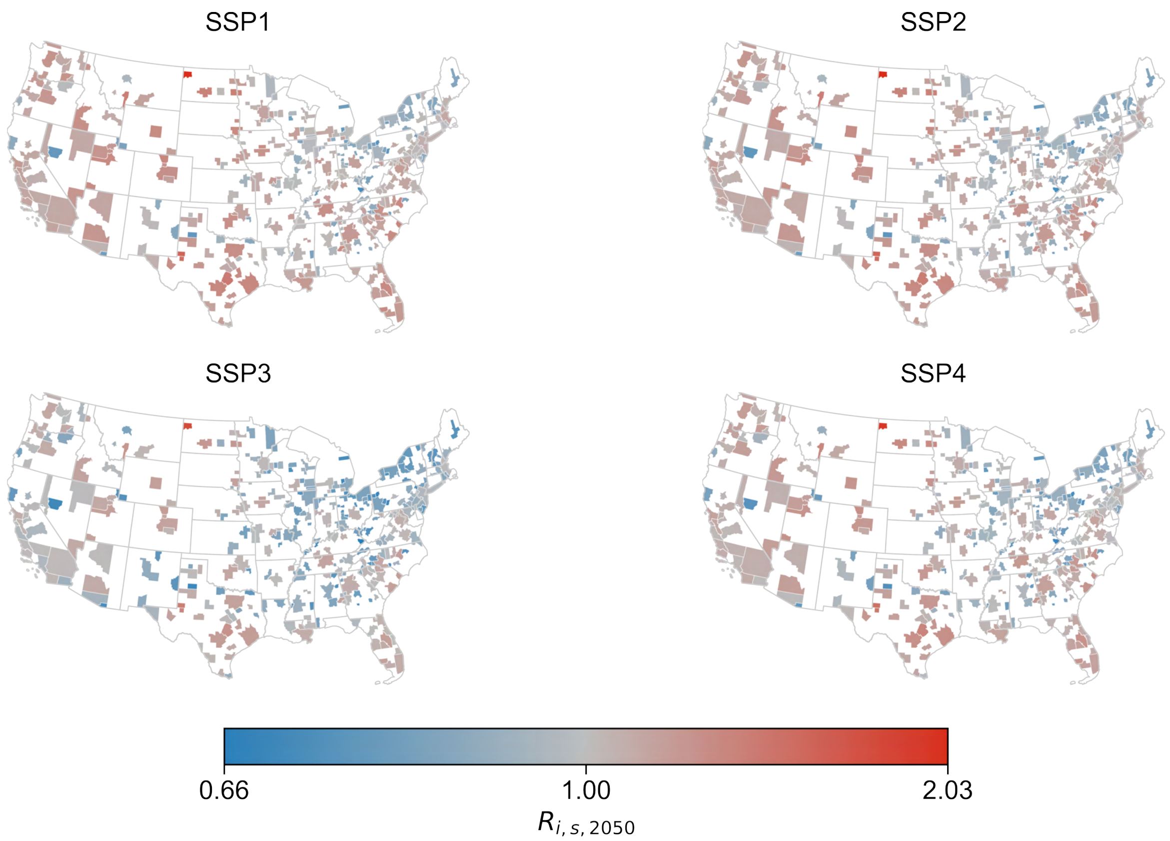
	CC EC	ER EC	CC + ER	Non-EC
Under development	79 (6.47%)	281 (23.01%)	16 (1.31%)	845 (69.21%)
Operational	90 (3.04%)	637 (21.55%)	26 (0.88%)	2203 (74.53%)

S10 Scaling of the number of data centers

We calculate the scaling relationship between the number of data centers and city population, across the 319 US cities hosting at least one facility in 2025 (Supplementary Figure 13), as explained in the main text. The slope of the scaling relationship is smaller than 1, indicating that larger cities host more data centers in absolute terms but fewer per capita than smaller ones.



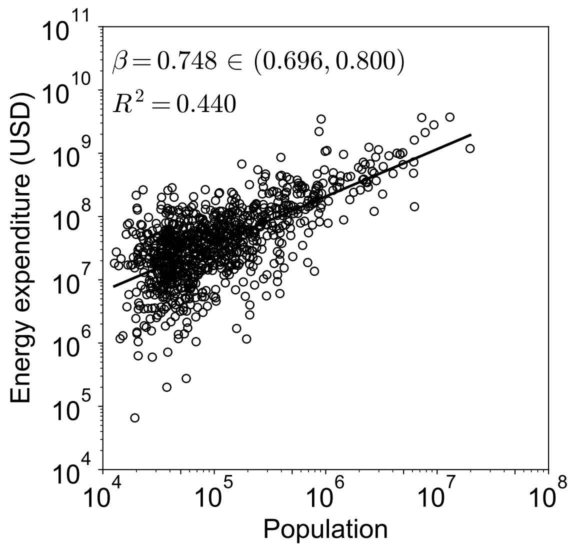
Supplementary Figure 13: **Scaling of the number of data centers in cities in the USA.** The dots identify the number of data centers as a function of population across US cities with at least one data center ($N = 319$); scaling exponent along with 95% confidence interval and R^2 are reported for completeness.



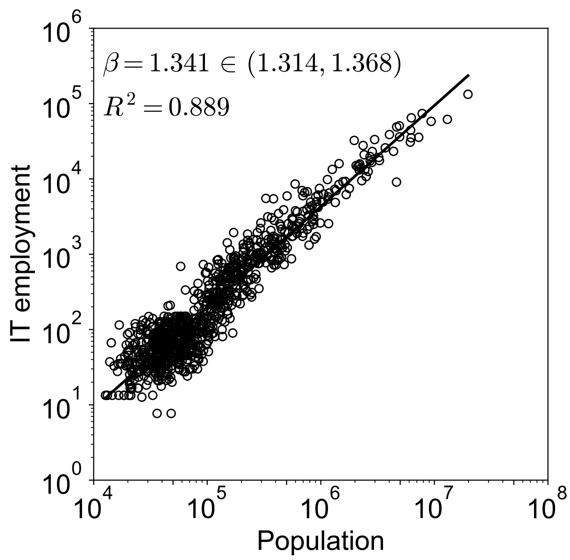
Supplementary Figure 14: **Spatial distribution of projected changes in data center presence under alternative socioeconomic scenarios (SSPs).** Maps report the ratio $R_{i,s,2050} = \mu_{i,s,2050}/\mu_{i,s,2025}$, where $\mu_{i,s,t}$ is the expected number of data centers in year t under scenario $s = \text{SSP1}, \text{SSP2}, \text{SSP3}, \text{SSP4}$, respectively, estimated for cities that host at least one data center in 2025 ($N = 319$). Values above 1 (red) indicate projected growth relative to 2025, while values below 1 (blue) indicate projected decline. Data are from the Data Center Map database [1], county population projections under the Shared Socioeconomic Pathways [7] and the Census Bureau TIGER/Line Shapefiles 2024 [2].

S11 Scaling of predictors in the negative binomial regression

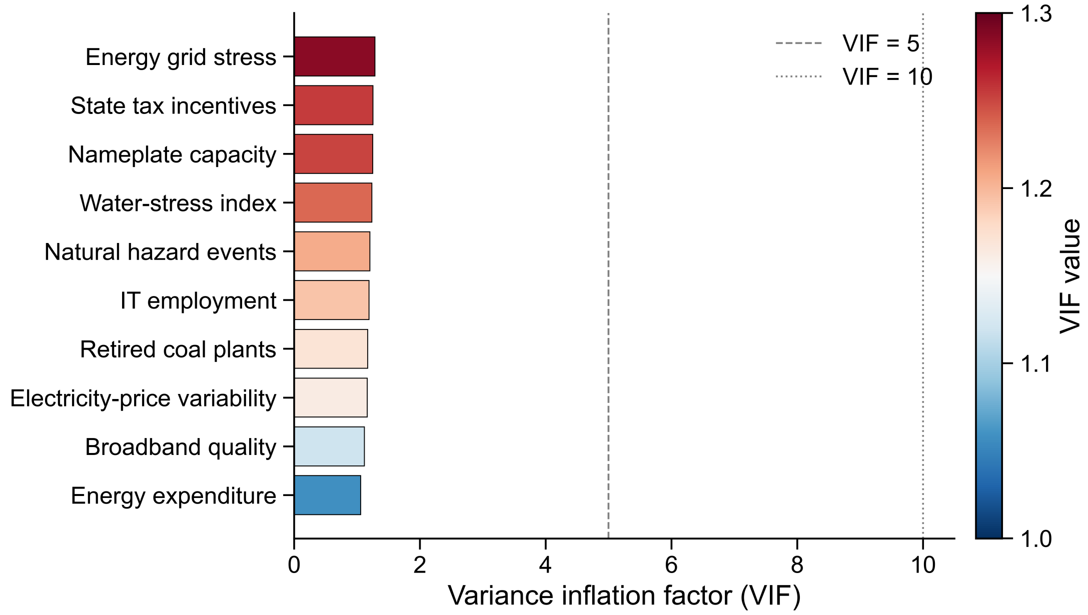
In the negative binomial regression presented in the main text, two of the regressors control for population size, namely, energy expenditure and IT employment. We use the same procedure as the one employed for the scaling of the data center count in the main text; results are reported in Figs. 15 and 16. As a further verification of the scaling laws, we repeated the analysis after removing the top and bottom 5% of cities by population size; energy expenditure retained sublinear behavior ($\beta = 0.748$, 95% confidence interval = $(0.696, 0.800)$) and IT employment remained superlinear ($\beta = 1.341$, 95% confidence interval = $(1.314, 1.368)$). Regressors are then calculated through the concept of scale-adjusted metropolitan indicators [11].



Supplementary Figure 15: **Scaling of the energy expenditure in cities in the USA.** The dots identify the number of data centers as a function of population across US cities ($N = 917$); scaling exponent along with 95% confidence interval and R^2 are reported for completeness.



Supplementary Figure 16: **Scaling of the IT employment in cities in the USA.** The dots identify the number of data centers as a function of population across US cities ($N = 910$); scaling exponent along with 95% confidence interval and R^2 are reported for completeness.



Supplementary Figure 17: **Variance inflation factors (VIFs) for the city-level predictors used in the regression model ($N = 910$).** Bars show VIF values for each regressor. All predictors exhibit low multicollinearity (VIF ≤ 2), indicating that the regressors capture largely distinct dimensions of environmental, infrastructural and economic conditions [8].

S12 Sensitivity of count-regression results to alternative model specifications.

Preliminary analyses considered a Poisson model ($\kappa = 0$) as a reference. Subsequently, we assessed the sensitivity of the results to the estimation of the overdispersion parameter using maximum likelihood ($\kappa \approx 3.8$). Results in Supplementary Table 7 indicate that the main findings are qualitatively similar for the models using the geometric distribution ($\kappa = 1$) and the estimated overdispersion. Sign and statistical strength of the associations are generally stable, although the significance of some individual coefficients depends on the assumed overdispersion. Given the large number of zero observations, we additionally repeated the analyses on the non-zero subset of the data for $\kappa = 1$ and estimated $\kappa \approx 1.1$, obtaining qualitatively similar conclusions. Overall, the geometric specification ($\kappa = 1$) adopted in the main text provides a reasonable baseline model.

Supplementary Table 7: **Sensitivity of negative binomial regression to model specification.** We compare five model specifications, varying the overdispersion parameter in the negative binomial model and the cities used in the analysis. For each specification, we report the value of all the regressors estimates with their corresponding standard error in parentheses. In addition to Cox and Snell pseudo R-squared, we report AIC (Akaike information criterion) and BIC (Bayesian information criterion) values. Asterisks are used to indicate the level of significance $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. All models were fitted using the Python generalized linear model implementation in `statsmodels`. The estimation of κ was carried out in a maximum likelihood sense using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) optimization.

	All cities			Cities with at least one data center	
	$\kappa = 0$	$\kappa = 1$	$\kappa = 3.816$	$\kappa = 1$	$\kappa = 1.133$
Constant	0.625*** (0.026)	0.128* (0.052)	0.037 (0.079)	1.408*** (0.076)	1.405*** (0.085)
Water-stress index	0.116*** (0.016)	0.006 (0.053)	0.023 (0.086)	-0.101 (0.067)	-0.100 (0.069)
Broadband quality	0.478*** (0.020)	0.444*** (0.052)	0.494*** (0.101)	0.080 (0.077)	0.080 (0.093)
Natural-hazard events	0.082*** (0.013)	-0.476*** (0.062)	-0.545*** (0.097)	-0.201* (0.081)	-0.205* (0.088)
Energy-grid stress	-0.197*** (0.016)	0.077 (0.053)	0.120 (0.099)	-0.044 (0.071)	-0.042 (0.089)
State tax incentives	-0.130*** (0.017)	0.126* (0.053)	0.159 (0.092)	0.001 (0.071)	0.004 (0.081)
Retired coal plants	0.239*** (0.006)	0.162*** (0.041)	0.122 (0.081)	0.125** (0.044)	0.125* (0.057)
Electricity-price var.	0.019 (0.017)	-0.041 (0.049)	-0.080 (0.076)	-0.043 (0.062)	-0.043 (0.062)
Energy expenditure	-0.186*** (0.019)	0.130* (0.051)	0.168 (0.101)	0.029 (0.081)	0.031 (0.093)
IT employment	0.574*** (0.021)	0.719*** (0.056)	0.740*** (0.101)	0.427*** (0.082)	0.428*** (0.095)
Nameplate capacity	0.412*** (0.006)	1.324*** (0.041)	1.550*** (0.150)	0.666*** (0.045)	0.670*** (0.074)
Observations	910	910	910	314	314
AIC	10270.564	2956.883	2635.668	1908.303	1907.860
BIC	10323.512	3009.831	2693.429	1949.546	1952.853
R^2	1.000	0.865	0.349	0.703	0.543

References

[1] DataCenterMap. U.S. data centers dataset (2025). URL www.datacentermap.com. Accessed: 2025-11-1.

[2] U.S. Census Bureau. TIGER/Line Shapefiles. https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html (2025). Accessed: 2025-12-3.

[3] U.S. Energy Information Administration. Power plants: Geospatial data. https://atlas.eia.gov/datasets/eia::power-plants/explore (2023). Accessed: 2025-12-2.

[4] U.S. Bureau of Labor Statistics. Occupational employment and wage statistics. https://www.bls.gov/oes/tables.htm (2023). Accessed: 2025-12-2.

[5] U.S. Census Bureau. American community survey 5-year population estimates. https://data.census.gov/table/ACSDT5Y2023.B01003 (2023). Accessed: 2025-10-07.

[6] Vicente, M. R. & Gil-de Bernabé, F. Assessing the broadband gap: From the penetration divide to the quality divide. *Technological Forecasting and Social Change* **77**, 816–822 (2010).

[7] Hauer, M. E. Population projections for U.S. counties by age, sex, and race controlled to shared socioeconomic pathway. *Scientific Data* **6**, 1–15 (2019).

[8] Thompson, C. G., Kim, R. S., Aloe, A. M. & Becker, B. J. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology* **39**, 81–90 (2017).

[9] Likas, A., Vlassis, N. & Verbeek, J. J. The global k-means clustering algorithm. *Pattern Recognition* **36**, 451–461 (2003).

[10] Nolte, C. High-resolution land value maps reveal underestimation of conservation costs in the United States. *Proceedings of the National Academy of Sciences* **117**, 29577–29583 (2020).

[11] Bettencourt, L. M., Lobo, J., Strumsky, D. & West, G. B. Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities. *PloS One* **5**, e13541 (2010).

[12] Pearl, J. *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, U.K., 2000).