# Data Centers and Local Labor Markets

Dany Bahar[a,b] and Greg C. Wright[c]

[a]Center for Global Development, 2055 L Street NW, Fifth Floor, Washington, DC 20036, USA
[b]Brown University, Providence, RI 02912, USA
[c]University of California, Merced, 5200 North Lake Rd., Merced, CA 95343, USA

Dany Bahar: dbahar@cgdev.org
Greg C. Wright: gwright4@ucmerced.edu

May 5, 2026

**Abstract**

We study how the geography of digital infrastructure investment shapes local labor markets. Using a novel dataset of approximately 770 US data center facilities and a synthetic control design, we find that data center entry raises total private employment by 4–5 percent and information-sector employment by 22 percent. The information-sector effect is concentrated in counties hosting hyperscale (cloud and AI) operators rather than colocation operators. This indicates that local incidence of capital-intensive digital infrastructure depends on operator type, not on capital alone.

JEL: H25, J23, R11, R23
Keywords: data centers, local labor markets, agglomeration, place-based policy

# 1 Introduction

Digital infrastructure is becoming a central feature of local economic geography. Cloud computing and artificial intelligence run through large physical facilities that are geographically concentrated, costly to build, and tied to specific power and fiber networks. Yet these facilities employ few workers on site. The central question is therefore not simply whether a data center is a large investment, but whether this kind of capital-intensive digital infrastructure can anchor local agglomeration outside established technology hubs.

This paper argues that the identity of the investing firm, and in particular whether it anchors a local supply chain of specialized inputs, is a first-order determinant of the multiplier, and that capital intensity *per se* is not. We make that case using data centers as a laboratory. Data centers are an attractive setting for two reasons. First, they sit at an extreme point on the capital-to-labor ratio spectrum, so whatever multiplier they generate cannot be attributed to sheer headcount at the facility itself. Second, the industry splits cleanly into two operator types that hold capital intensity roughly fixed but differ sharply in the local ecosystem they sustain: *hyperscale* operators (Amazon, Google, Microsoft, Meta) build campuses to run their own cloud and AI workloads and, as we document below, appear to procure fiber construction, managed services, and network operations locally, while *colocation* operators (Equinix, Digital Realty) build landlord facilities leased to remote tenants whose inputs are not sourced in the host county. The contrast lets us hold the capital shock nearly fixed while varying the scope for local supplier linkages.

The United States now hosts more than 1,200 large data center facilities, annual investment in new capacity exceeds $100 billion, and a single hyperscale campus typically costs $500 million to $2 billion, draws 100–500 megawatts of continuous electrical load, and typically employs fifty to two hundred operations staff on site.[1] This scale makes data centers a useful setting for studying the spatial incidence of capital-intensive digital infrastructure.

[1] Estimates from the IM3 Data Center Atlas (Pacific Northwest National Laboratory), Synergy Research Group, and operator 10-K filings.

Moreover, at least 35 states offer tax incentives targeted specifically at data center construction [Cardinal News, 2026], individual deal packages routinely exceed $100 million per facility, and cumulative subsidies are nearly $20 billion [Good Jobs First, 2024].

Identification in this setting is not straightforward because data center locations are not randomly assigned. Operators screen locations on a small set of binding constraints and only after passing those screens compete on costs, incentives, and local negotiations. The set of technically feasible counties is narrow, the firms doing the choosing are sophisticated, and the relevant cost shifters are difficult to observe directly. Cross-county comparisons of treated and untreated counties therefore conflate the causal effect of data center entry with the structural advantages that put a county in the feasible set in the first place.

This paper uses the siting problem as the basis for identification. We construct a novel facility-level dataset of approximately 770 dated US data center openings, covering 93 counties that receive their first large data center between 2008 and 2024, and classify each facility as *hyperscale* (built by cloud and AI operators such as Amazon, Google, Microsoft, and Meta to run their own workloads) or *colocation* (built by data center landlords such as Equinix and Digital Realty who lease space to remote tenants). We link these facilities to county-year employment and wage data from the Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW) covering 2003 to 2024, and we estimate the effect of data center entry on total private employment as well as employment in construction, information, and professional services using a synthetic control (SCM) design. We supplement this approach with IV and staggered-DiD designs.

Our primary empirical approach is the SCM design that compares each treated county to a weighted combination of never-treated counties matched on its pre-treatment employment trajectory [Abadie et al., 2010]. SCM is well suited to the data center setting because treated counties differ from the average never-treated county in the level of employment, wages, and pre-trend growth, and because staggered treatment timing makes two-way fixed effects unreliable when treatment effects vary across cohorts. Matching on the outcome

trajectory at the unit level absorbs both problems directly. We also report a shift-share IV that instruments cumulative installed megawatts with the interaction of a county's pre-existing substation count and the leave-one-state-out national data center rollout, along with stacked DiD and Callaway–Sant'Anna estimators. We treat the heterogeneity-robust estimators as the primary cross-check on the SCM. Placebo-sector tests reveal that the shift-share IV does not cleanly satisfy the exclusion restriction, so we use it as a sign-of-effect check rather than as identification. We address the possibility that host-county gains reflect within-region reallocation by re-running the SCM on counties contiguous to treated hosts and by aggregating the panel to commuting zones.

We obtain three sets of results. First, the synthetic control estimates indicate that data center entry raises total private employment by 4 to 5 percent, construction employment by 11 percent, and information-sector employment by 22 percent at $t = 6$. For the median treated county, these correspond to roughly 4,400 additional private-sector jobs, 700 additional construction jobs, and 500 additional information-sector jobs six years after first entry. The effects build gradually from $t = 0$, with no evidence of anticipation, and are similar across heterogeneity-robust estimators. A complementary shift-share IV based on pre-existing high-voltage substation density also delivers positive coefficients.

Second, the information-sector effect is concentrated in counties hosting hyperscale facilities. Estimating SCMs separately by facility type, hyperscale-only counties show a 43 percent information-sector gain at $t = 6$ while colocation counties show no effect. Because hyperscale and colocation counties differ on pre-treatment size, we restrict to a size-matched SCM that holds pre-treatment information-sector size fixed and find a 36 percentage point hyperscale–colocation differential, statistically significant at the 1 percent level. A TWFE triple-difference yields a similar 31 percentage point conditional differential, with information-sector agglomeration requiring a *cluster* of facilities to materialize. A direct test of the firm-entry margin reinforces this: hyperscale entry raises the count of information-sector and construction establishments in the host county, while colocation entry does not.

Taken together, these results are consistent with hyperscale operators drawing in concrete local suppliers such as fiber contractors, managed service providers, and network operations centers that do not follow colocation entry. Construction employment, by contrast, rises by a similar amount in both facility types, reflecting the physical build-out of the campuses themselves.

Third, wages rise alongside employment. Average weekly wages in treated counties rise by approximately 3 percent. Using the Quarterly Workforce Indicators from the Census Bureau's LEHD program, we decompose this wage effect into within-job and compositional channels: incumbent workers see earnings rise by 3.3 percent, while new hires earn 3.8 percent more than new hires in control counties. That both groups benefit indicates actual local labor market tightening, not compositional reshuffling. FHFA house prices show no statistically significant response, in contrast to the fracking literature [Bartik et al., 2019], though the point estimate is a modest positive drift of 2.0 percent and the FHFA index is known to understate rent capitalization in rural counties. We therefore interpret the housing evidence as suggesting that rent growth is small relative to wage gains, not that it is zero.

Contiguous non-host counties show small and weakly *positive* spillover effects, the opposite sign from a reallocation story, and commuting-zone aggregations yield a positive total employment effect of 1.7 percent. We read these as evidence that host-county estimates are not inflated by within-CZ reallocation. At the local labor-market scale, the evidence is more consistent with net job creation than with pure displacement.

These findings contribute to three literatures. The first is the economic geography of digital infrastructure and local agglomeration. Prior work shows that communications infrastructure can reshape the spatial distribution of economic activity, but effects depend on local absorptive capacity and the organization of production [Forman et al., 2012, Hjort and Poulsen, 2019]. We show that the same is true for the physical infrastructure of cloud computing: the local effect depends less on installed capital alone than on whether the operator brings a localized supplier network with it. The second is the local labor market multi-

plier and place-based employment shocks literature [Moretti, 2010, Kline and Moretti, 2014, Greenstone et al., 2010, Bartik et al., 2019, Slattery and Zidar, 2020]. Our setting offers a case in which the siting margin is tightly linked to observable power and fiber infrastructure, allowing us to ground identification in a known feasibility constraint. The third is the small but growing literature on data centers themselves [Hicks, 2024, Fang and Greenstein, 2025], where the hyperscale-colocation distinction we document is, to our knowledge, one of the first systematic decompositions of how operator type shapes local labor demand. Methodologically, the paper demonstrates that OpenStreetMap edit histories can be used to date the construction of large physical structures, a technique that may be useful for other settings where official permitting data are sparse.

Section 2 reviews the related literature. Section 3 describes the data. Section 4 lays out the institutional setting and the synthetic control design. Section 5 presents the main SCM results. Section 6 examines heterogeneity by facility type. Section 7 tests for local displacement using contiguous-county and commuting-zone designs. Section 9 presents additional robustness checks. Section 10 concludes.

## 2 Related Literature

A large literature in urban and regional economics studies how localized shocks reshape employment, wages, productivity, and the spatial organization of activity. Kline and Moretti [2014] find that the Tennessee Valley Authority produced large and persistent effects on manufacturing employment. Greenstone et al. [2010] use a "Million Dollar Plants" design and find statistically significant TFP increases among incumbent plants. In a closely related setting, Pathania and Netessine [2026] examine the openings of Amazon fulfillment centers and document positive county-level employment effects in mid-sized counties, using a comparable winners–losers design. Our setting differs in that data centers are far more capital-intensive and labor-thin than fulfillment centers, and our hyperscale-versus-colocation decomposition

isolates the role of operator-supplied local supplier networks rather than direct on-site employment. Busso et al. [2013] evaluate federal empowerment zones and find substantial employment and wage gains. Related work on place-based incentives emphasizes that local effects depend on both the targeted activity and the incidence of the policy [see Neumark and Simpson, 2015, for a survey]: Bartik [2020] synthesizes the evidence on business incentives and finds wide variation in cost-effectiveness, Slattery and Zidar [2020] document that state tax incentives increasingly flow to large firms, and Suárez Serrato and Zidar [2016] show that the incidence of state corporate tax cuts falls largely on firm owners rather than workers.

The closest analogue to our setting is Bartik et al. [2019], who study the local economic consequences of hydraulic fracturing, another capital-intensive investment that arrives in rural counties with specific geological endowments. They find employment and income gains but also increased housing costs and health externalities. Our data center setting differs in that the investment is driven by power and fiber infrastructure rather than geology, and the externalities (electricity and water consumption) are less well-documented.

Moretti [2010] estimates a local multiplier of approximately 1.6 non-traded jobs per additional traded-sector job, with larger values for high-technology industries. Moretti and Thulin [2013] report a value around 2.5 for high-tech. Section 8 translates our SCM estimates into an implied multiplier of roughly 2.3 at a hyperscale median county under midpoint assumptions about direct on-site employment, placing data centers between Moretti's manufacturing and high-tech benchmarks despite their extreme capital-to-labor ratio. On the digital infrastructure side, Forman et al. [2012] find that the internet raised wages in already-large cities but not elsewhere, while Hjort and Poulsen [2019] document employment gains from fast internet in Africa. Our finding that information sector agglomeration requires a cluster of facilities is consistent with this literature's emphasis on threshold effects in digital infrastructure.

Empirical evidence on the local economic effects of data centers remains thin. The most rigorous existing study is Hicks [2024], who examines data center construction in Texas using

county-level employment data and finds no measurable effect on total employment, with a possible exception for professional and technical services near very large facilities. Policy reports underscore the stakes on both sides of the debate: Goetzel et al. [2026] emphasize strategies for converting data center investment into broader local prosperity, while Ohio River Valley Institute [2025] questions whether data centers justify their incentive costs given their low direct employment. We advance this literature in three ways: first, we assemble a national dataset of approximately 770 dated facilities, far larger than any prior study; second, we apply synthetic control methods that more directly address the selection concerns inherent in TWFE designs; and third, we decompose effects by facility type (hyperscale versus colocation), showing that information-sector agglomeration is concentrated in hyperscale investment.

# 3 Data

## 3.1 Data Center Locations: The IM3 Atlas

The foundation of the facility-level dataset is the IM3 Open Source Data Center Atlas, developed by the Pacific Northwest National Laboratory (PNNL) as part of the Department of Energy's Integrated Multisector, Multiscale Modeling (IM3) initiative. The atlas identifies 1,242 data center facilities across 46 US states, geocoded to county-level FIPS codes, with information on operator name, facility square footage, and geographic coordinates. The atlas is derived from OpenStreetMap building footprints tagged as data centers, supplemented with facility-level information from operator websites and industry databases.

Table 1 summarizes the atlas. The facilities range from small colocation sites under 10,000 square feet to hyperscale campuses exceeding 3 million square feet. The six largest operators, Amazon Web Services (176 facilities), Digital Realty (70), Meta (61), Google (60), Microsoft (53), and Equinix (38), account for 37 percent of all facilities and a substantially larger share of total floor space. The atlas is dominated by Northern Virginia, which hosts

292 facilities (24 percent of the total); as discussed below, these established hubs are excluded from our analysis sample.

Table 1: Summary Statistics: IM3 Data Center Atlas

| | All facilities | Large (>100K sq ft) |
|---|---|---|
| Number of facilities | 1,242 | 708 |
| Number of counties | 209 | 111 |
| *Square footage:* | | |
| Mean | 474,256 | 741,408 |
| Median | 132,388 | 211,010 |
| *Top operators:* | | |
| Amazon Web Services | 176 | 147 |
| Digital Realty | 70 | 45 |
| Meta | 61 | 61 |
| Google | 60 | 57 |
| Microsoft | 53 | 48 |
| Equinix | 38 | 17 |
| *Top states:* | | |
| Virginia | 292 | 219 |
| Oregon | 94 | 64 |
| Texas | 92 | 57 |

*Notes:* Source: IM3 Open Source Data Center Atlas (Pacific Northwest National Laboratory). The atlas geocodes 1,242 US data center facilities to county FIPS codes from OpenStreetMap building footprints, supplemented with operator websites and industry databases. "Large" facilities are those exceeding 100,000 square feet.

## 3.2 Opening Dates

The IM3 atlas does not include facility opening dates. We construct opening dates from multiple sources, covering 713 of the 1,242 atlas facilities and an additional 58 non-atlas hyperscale facilities. In the current analysis freeze, we date 617 of the 708 large facilities

($> 100,000$ sqft).

**OpenStreetMap Edit Histories.** The primary source of opening dates exploits the fact that the IM3 atlas facility identifiers correspond to OpenStreetMap (OSM) way identifiers. We query the OSM API for the edit history of each facility and record the timestamp when the building footprint was first added to the map. For facilities constructed after approximately 2014, when OSM coverage of commercial buildings became comprehensive, the first-mapped date typically falls within 0–2 years of actual construction completion. For earlier facilities, the lag may be larger. This approach yields approximate construction dates for 553 atlas facilities.

**State Incentive Records.** Several states that offer data center tax incentives maintain public registries of qualifying facilities. The most comprehensive is the Texas Comptroller's registry, which lists 132 qualifying data centers with effective dates, operator names, and occupant information. We also obtain facility-level records from the Nevada Governor's Office of Economic Development and the Ohio Tax Credit Authority.

**News and Trade Press.** We supplement the above sources with facility opening dates identified from Data Center Dynamics, Data Center Knowledge, company press releases, and state economic development announcements. This approach is most productive for the hyperscale operators (Meta, Google, Microsoft), whose facility openings are widely covered. We also conduct a targeted search for AWS facilities, whose parent company Amazon operates through subsidiary entities (principally Vadata, Inc.) that make individual facility identification difficult.

**Date Validation.** We validate the OSM-derived dates against the Texas Comptroller's registry of qualifying data centers, which records the effective date of tax exemption eligibility for 123 facilities. Matching 11 facilities that appear in both the Comptroller registry and the

IM3 atlas, we find a median lag of zero years (mean 0.8 years) between OSM and Comptroller dates, with 82 percent of matches falling within ±2 years. The two largest discrepancies (LinkedIn Richardson, 4 years; State Farm Richardson, 3 years) involve pre-2015 facilities, consistent with sparser OSM coverage of commercial buildings before 2014.

Because the OSM-derived dates may lag actual facility openings by up to two years, we further assess sensitivity to date measurement error by re-estimating the SCM with treatment timing shifted by $-1$ and $-2$ years. The total employment effect is stable across shifts (4.2, 4.2, and 3.5 percent), and the information sector effect varies from 17 to 29 percent (Table 19 in Section 9).

## 3.3 Employment Outcomes: QCEW

County-year employment and wage data come from the Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW). The QCEW is based on administrative records from the unemployment insurance system and covers approximately 95 percent of US employment. We obtain annual average employment levels and weekly wages for four industry categories at the county level: total private employment (NAICS 10), construction (NAICS 23), information (NAICS 51), and professional and technical services (NAICS 54).[2] The panel covers 2003–2024.

## 3.4 Analysis Sample

The analysis sample in the current paper consists of 93 treated counties that receive their first large data center (defined as a facility exceeding 100,000 square feet) between 2008 and 2024, and approximately 3,040 control counties that do not host a large data center during the sample period. We exclude counties with large data centers opening before 2008 to ensure at least five years of pre-treatment data, and exclude 17 control counties that

[2]The 2022 NAICS revision expanded NAICS 518 to explicitly include cloud computing and data center operators. Because our analysis uses 2-digit NAICS 51, this reclassification does not affect our series.

host large undated facilities (potential control contamination). The resulting panel contains approximately 69,000 county-year observations for total private employment, with somewhat fewer observations for sector-specific outcomes due to BLS disclosure suppression in small counties.

Of the 93 treated counties, 33 host only hyperscale facilities, 23 host only colocation facilities, 12 are "mixed" (hosting both types), and 25 host facilities by smaller or unclassified operators. This variation in facility type enables us to test whether the employment effects of data center construction depend on the type of operator, a distinction with direct policy relevance since state incentive programs do not typically differentiate between hyperscale and colocation investment.

Our design estimates the effect of *new data center entry* into previously untreated counties. We exclude established hubs that hosted large facilities before 2008, including Northern Virginia, the Chicago metro area, and parts of Dallas-Fort Worth, because their early adoption reflects deep structural advantages (proximity to internet exchange points, fiber backbone, federal agencies) that make them poor candidates for causal inference. The 93 treated counties in our sample represent the policy-relevant margin: locations where state and local incentives are deployed to attract new entry. Because roughly half of treated counties subsequently attract additional facilities, the estimand captures the full trajectory following first entry, including endogenous follow-on investment. Section 9 decomposes this by facility count using a dose-response design.

# 4 Institutional Setting and Identification

## 4.1 The Data Center Location Problem

Data center sites are selected from a narrow set of technically feasible locations. Operators screen first on four binding constraints: high-capacity electrical power, access to fiber-optic backbone, land with permitting pathways that admit heavy industrial use, and, for water-

cooled campuses, access to municipal or surface water at volume. Only after these engineering screens are passed do operators compete on costs, tax incentives, and local negotiations. This pattern is consistent with the broader location-choice literature [Alcácer and Chung, 2014, Alcácer and Delgado, 2016] and with recent evidence that colocation facilities value proximity to dense network ecosystems while cloud providers are more willing to trade proximity for cheaper land and scalable power access [Fang and Greenstein, 2025].

Of these screens, power is first-order. A modern hyperscale campus requires 100 to 500 megawatts of firm, continuous electrical load, an order of magnitude larger than a typical manufacturing facility. Delivering that load requires proximity to high-voltage transmission infrastructure, typically 345 kV or above, that can source from multiple generation points and absorb the demand without destabilizing local distribution. Counties with existing high-voltage substation infrastructure therefore enter the feasible set at low marginal cost, while counties without it can only enter after multi-year utility planning and substation construction.

This siting logic motivates two design requirements. The comparison group must approximate the narrow set of counties that could plausibly host a large facility, and even within the feasible set treated counties differ from controls on pre-trends, so the design must match on the pre-treatment outcome trajectory rather than rely on parallel-trends conditional on fixed effects. Both requirements point to a synthetic control design.

## 4.2 Synthetic Control

Our primary identification strategy is a synthetic control design. SCM constructs, for each treated county, a weighted combination of never-treated counties that matches its pre-treatment employment trajectory [Abadie et al., 2010, Abadie, 2021]. Weights are non-negative and sum to one, with no intercept, ensuring the counterfactual is a convex combination of real counties. The donor pool consists of the 200 nearest never-treated counties by pre-treatment employment size, and we exclude treated counties with pre-treatment RMSE

above 0.15 log points.

SCM is well suited to the data center setting for two reasons. First, treated counties differ from the average never-treated county on level employment, wages, and pre-trend growth (Table 2); a design that matches on the outcome trajectory itself absorbs these differences directly rather than asking fixed effects to do the work. Second, treatment timing is staggered across counties, and TWFE estimators with staggered timing can produce biased estimates when treatment effects are heterogeneous across cohorts [Goodman-Bacon, 2021, de Chaisemartin and D'Haultfœuille, 2020]. SCM sidesteps both concerns by matching at the unit level and aggregating only after each treated county has its own counterfactual. Section 9 reports stacked DiD and Callaway–Sant'Anna estimators that are robust to staggered-treatment heterogeneity, which we treat as the primary cross-check on the SCM. We also report a shift-share IV based on pre-existing high-voltage substation density, but placebo-sector tests indicate the exclusion restriction is not cleanly satisfied, so we use the IV as a sign-of-effect check rather than as a coequal identification design.

The SCM compares each treated host county to non-host donor counties. If data centers displace economic activity from donor counties into host counties, the host-county estimate overstates the true local effect. We test for this displacement directly in Section 7, using both the contiguous-county set (whether neighbors lose jobs) and the commuting-zone aggregation (whether the host gain is undone at the regional scale).

## 4.3 Threats to Identification

The main residual threat to the SCM design is selection on unobserved trends: conditional on the engineering and infrastructure screens described in Section 4.1, data centers may still locate in counties that are already growing for reasons we do not observe. Pre-treatment balance statistics (Table 2) show that treated counties are larger, have higher wages, and exhibit faster pre-treatment employment growth than the average control county. We address this threat in three complementary ways. First, the SCM design matches on the pre-treatment

employment trajectory itself, and the event study (Figure 1) provides a visual test for differential pre-trends. Second, we match treated counties to control counties within the same pre-treatment employment size quintile. Third, we present results separately for rural and urban counties and for hyperscale and colocation facilities, to assess whether the effects are driven by specific subgroups.

**Anticipation and operator secrecy.** A distinctive feature of data center site selection is the use of shell companies and non-disclosure agreements to conceal the operator's identity during the planning and construction phases. Amazon operates through Vadata, Inc.; Google uses subsidiaries such as Design LLC and Sharka LLC; and Virginia FOIA investigations have documented NDAs between localities and data center operators that suppress public information about incoming facilities. This secrecy limits the ability of local labor markets to anticipate the treatment: while a large construction project is visible to the community, the identity of the operator and the nature of the facility (hyperscale vs. colocation) are often unknown until after construction is complete. The gradual build-up of the information sector effect from $t = 0$ onward, rather than a pre-treatment jump, is consistent with this lack of anticipation: local IT firms and managed service providers cannot pre-position around a facility whose operator they do not yet know.

Table 2: Pre-Treatment Balance: Treated vs. Control Counties

| | Treated | Control | Difference |
|---|---|---|---|
| Log pre-treatment employment | 11.58 | 9.49 | 2.09 |
| Pre-treatment employment growth (2003–07) | 13.6% | 7.8% | 5.8 pp |
| Average weekly wage ($) | 748 | 576 | 172 |
| Number of counties | 92 | 2,948 | |

*Notes:* Pre-treatment characteristics computed as county-level means over 2003–2007. Employment growth is the percentage change in total private employment from 2003 to 2007. One treated county with fewer than 3 pre-period observations is excluded from this table; the full analysis sample contains 93 treated counties.

# 5 Results

## 5.1 Main Results

A standard TWFE regression of log employment on a post-treatment indicator with county and year fixed effects yields large, statistically significant coefficients across all sectors: 15 percent for total private employment, 18 percent for construction, 41 percent for information, and 29 percent for professional services (Table 21 in Section 9). However, as documented in Section 4, treated counties were growing faster than controls before data center construction, inflating the TWFE estimates. We therefore turn to the synthetic control method as our preferred specification.

Applying the SCM procedure described in Section 4 yields 90 well-matched treated counties for total employment (the per-panel sample sizes in Figure 1 reflect the subset with observations at each event time). Figure 1 presents the results. All four panels show flat pre-treatment gaps, confirming that the synthetic controls match the treated counties' pre-treatment trajectories. Post-treatment, the gaps are positive and growing in all sectors.[3]

Total private employment shows a gradual increase reaching 4.2 percent by $t = 6$, comparable to the 1.5 percent TFP gain that Greenstone et al. [2010] document for "Million Dollar Plant" counties. Alternative estimators that are robust to heterogeneous treatment effects in staggered designs (Callaway–Sant'Anna, stacked DiD) yield similar estimates of 5 percent (Section 9). Construction employment jumps immediately at $t = 0$ and grows to 10.9 percent by $t = 6$, reflecting both the initial construction workforce and ongoing facility expansion.[4] Information sector employment builds gradually to 22.4 percent by $t = 6$,

[3]Throughout, percent effects from the SCM are reported as $100 \times (\exp(\bar{g}) - 1)$, where $\bar{g}$ is the average log-employment gap over the post-treatment period $t = 0$ through $t = +6$, except where "at $t = +6$" is stated explicitly and the figure is the gap at the long-run horizon (also exponentiated). The two conventions yield similar values for small effects but diverge for the per-type SCMs because the gap accumulates over the post period; we note this where it matters and use the exponentiated post-period mean as the default summary in line with applied SCM convention.

[4]Unlike a typical factory opening where construction employment spikes and then dissipates, data center campuses are built in phases over many years. Of the 93 treated counties, 47 host multiple large facilities, and many receive additional facilities within 2–4 years of the first. Ongoing construction activity, combined with ancillary commercial and residential development that accompanies large capital projects, likely explains

consistent with agglomeration of related IT activity around the data center. The 22 percent gain is not mechanical: NAICS 518 (data processing) accounts for only 14 percent of NAICS 51 employment at the median treated county, so even a large increase in 518 explains only a small share of the overall effect.[5] Professional services employment grows to 16.8 percent by $t = 6$.

To assess the plausibility of these magnitudes, we decompose the SCM estimates by facility count. Of the 93 treated counties, 46 host one large data center and 47 host two or more (median 2, mean 4.9). Single-facility counties show a 5.8 percent total employment effect and a 13.0 percent construction effect, both significant, but no statistically significant information sector effect (3.2 percent). Multi-facility counties drive the information sector result (25.8 percent) with a more modest total employment effect (2.5 percent). The pooled SCM estimates should therefore be interpreted as the effect of entering the data center market, which typically involves accumulation of multiple facilities over time, rather than a single facility opening in isolation.

## 5.2 Wage Effects

Figure 2 decomposes the wage effect by sector using the same SCM methodology. Average weekly wages across all private sectors rise by 2.8 percent by $t = 6$, with effects beginning immediately at $t = 0$. Information sector wages show the largest effect, consistent with data centers attracting high-paying technology jobs. Construction and professional services wages also rise, reflecting increased demand for building trades and skilled services. The wage effects are uniformly larger than the corresponding employment effects, suggesting that data centers raise local wages through a combination of direct high-wage job creation and tightening of local labor markets. To decompose the wage effect into compositional and within-job channels, we use the Quarterly Workforce Indicators (QWI) from the Census

why the construction effect persists and even grows through $t = 6$ rather than reverting to zero.

[5]NAICS 518 data are available from the QCEW for 59 of 93 treated counties; the remainder are suppressed by BLS for confidentiality. Among counties with available data, NAICS 518 ranges from 1 to 78 percent of NAICS 51 (mean 17 percent).

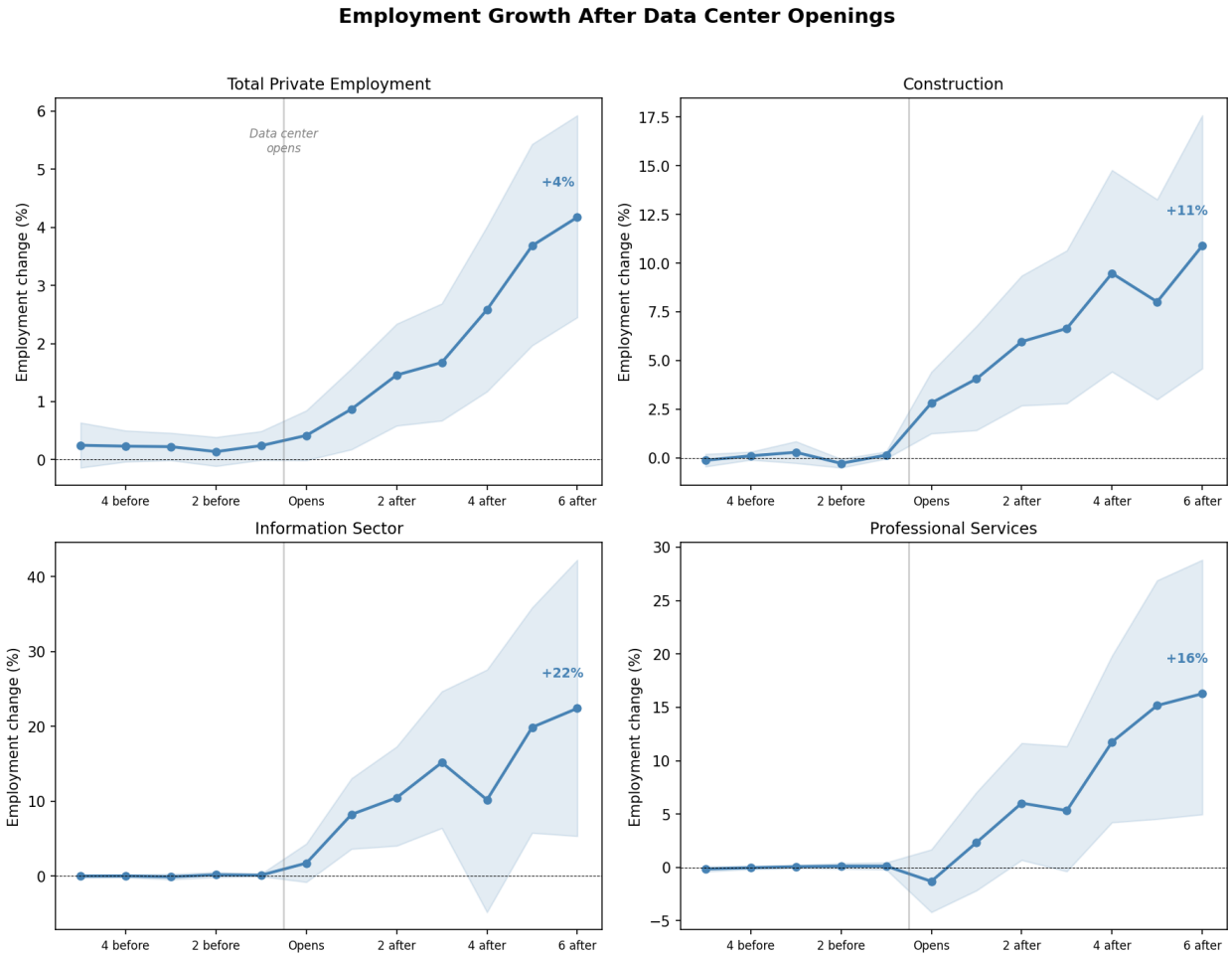


Figure 1: Synthetic Control Event Study: Employment Around Data Center Openings

*Notes:* Each panel plots the average gap between treated counties and their synthetic controls, with confidence intervals based on cross-county standard errors. Synthetic controls are constrained convex combinations of the 200 nearest donor counties by pre-treatment employment. Counties with pre-treatment RMSE $> 0.15$ are excluded. The three sector panels share a common $y$-axis. Event time 0 corresponds to the year of first large data center opening. Formal inference using the prediction interval approach of Cattaneo et al. [2021] is reported in Table 16 and Section 9.

Bureau's LEHD program, which reports average monthly earnings separately for *stable* (full-quarter, continuing) workers and *new hires*. The continuing-worker effect captures within-establishment wage growth, which could reflect promotions or role changes within firms as well as direct wage increases.

Applying the SCM methodology to QWI earnings, we find that both channels contribute to the wage gains. Continuing workers see earnings rise by 3.3 percent by $t = 6$, while new hires see a similar effect (3.8 percent). These magnitudes are consistent with the QCEW wage estimate of 2.8 percent and economically plausible. That both continuing workers and new entrants benefit indicates that data centers tighten local labor markets broadly, not merely through compositional shifts from high-wage firms entering the county.

# 6 Heterogeneity

We examine heterogeneity along four dimensions: facility type, facility size, county type, and treatment timing.

## 6.1 Heterogeneity by Facility Type

The most policy-relevant source of heterogeneity is *facility type*: do hyperscale data centers (operated by cloud and AI companies) generate different local employment effects than colocation facilities (operated by wholesale data center providers)? This distinction matters because hyperscale operators bring their own ecosystem of technology suppliers and managed service providers, while colocation providers primarily offer space and power to remote third-party tenants.

Hyperscale and colocation counties differ substantially in pre-treatment characteristics. Hyperscale counties are much smaller (median private employment of 36,000 versus 366,000 for colocation counties), have lower wages ($628 versus $896 weekly), and have smaller information sector employment shares (1.6 versus 3.5 percent). These differences reflect

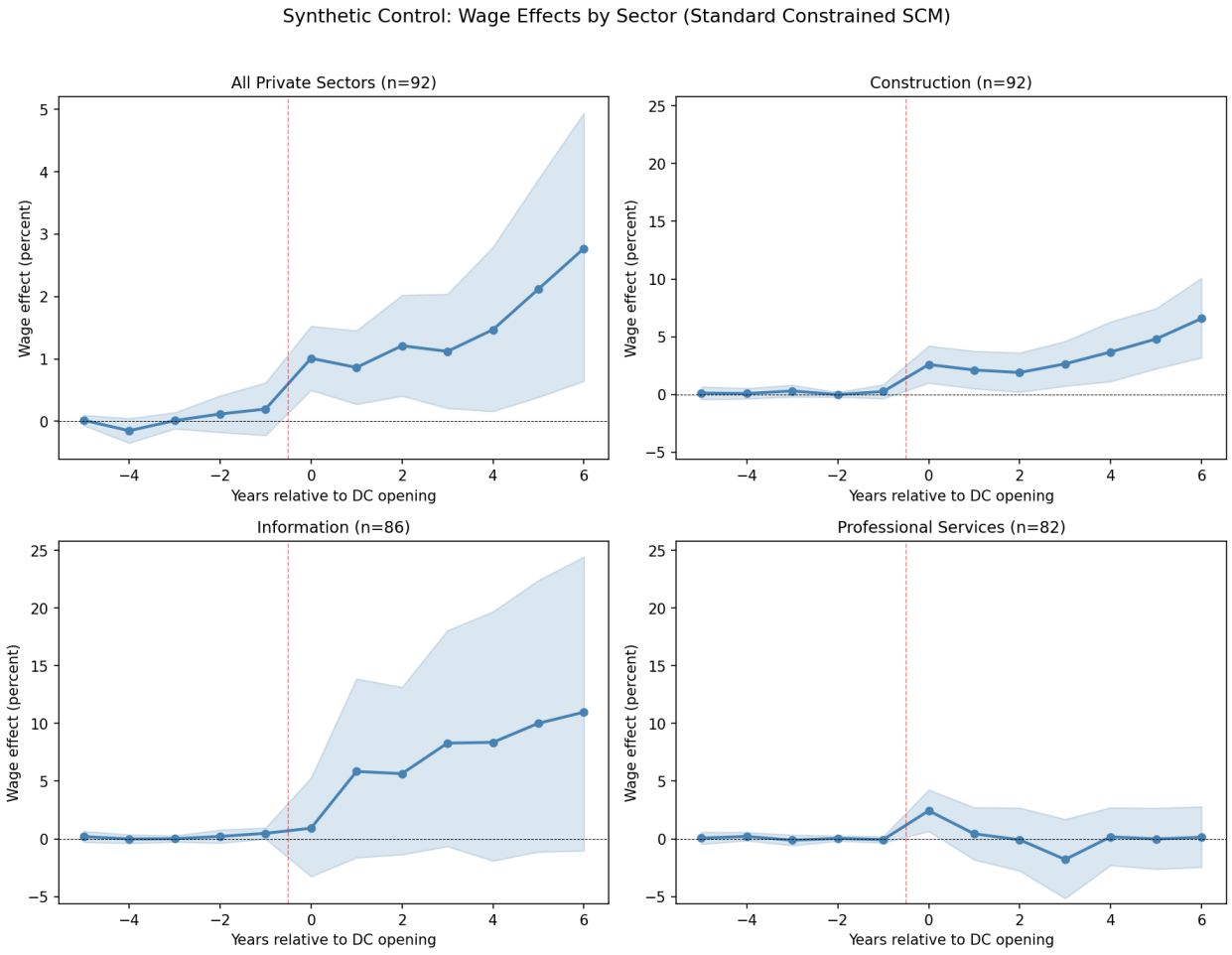


Figure 2: Synthetic Control: Wage Effects by Sector

*Notes:* Each panel plots the average gap in log average weekly wages between treated counties and their synthetic controls, with confidence intervals based on cross-county standard errors. Methodology as in Figure 1. The three sector panels share a common $y$-axis.

the location strategies of the two operator types: hyperscale facilities seek cheap land and power in less populated areas, while colocation providers locate near existing network infrastructure in major metros. Pre-treatment total-employment *growth rates* are balanced across types (12.2 versus 11.1 percent, with the difference far from statistical significance). Pre-treatment information-sector trajectories are noisier: both types show statistically significant pre-trends in info-sector employment relative to controls (Table 4).[6] This is part of the case for using the SCM-by-type design, which matches each treated county to its own pre-treatment information-sector trajectory, rather than a TWFE triple-difference as the primary type-comparison estimate.

Table 5 and Figure 3 present SCMs estimated separately by facility type. All types show positive total employment effects, but the information sector diverges sharply. Hyperscale-only counties show a 43 percent increase at $t = 6$ ($n = 25$), while colocation-only counties show a $-5$ percent decline ($n = 20$). Mixed counties fall in between (8 percent, $n = 11$).

The unconditional 48 percentage point gap is partly mechanical, because hyperscale counties have an order-of-magnitude smaller pre-existing information sector, so a given absolute number of new jobs translates into a larger percentage. Our preferred type-comparison estimates therefore come from a size-matched SCM that holds pre-treatment information-sector size fixed (Table 3).[7] Both restrictions preserve the qualitative pattern. On the common-support sample, the information-sector gap is 38 percent in hyperscale counties and $-4$ percent in colocation counties. On the matched-pair sample, the difference is 36 percentage points with $t = 5.0$, statistically significant at the 1 percent level. We treat the matched-pair SCM differential of 36 percentage points as the primary type-comparison estimate, because it is identified by the same SCM design as the headline aggregate effect

[6] The descriptive annualized pre-trend is 1.4 percent in hyperscale counties and $-0.2$ percent in colocation counties; the cross-type test that the lead $\times$ hyperscale interactions are jointly zero rejects only marginally ($F = 2.10$, $p = 0.08$).

[7] Panel A reports SCM effects on the common-support sample, retaining only counties with pre-treatment employment inside the overlap of the two type distributions (26 of 33 hyperscale and 18 of 23 colocation counties). Panel B reports SCM effects on size-caliper matched pairs, in which each colocation county is paired with the nearest hyperscale county in log pre-treatment employment, without replacement.

and explicitly addresses the base-rate concern. As a cross-check that does not match on size, a TWFE triple-difference specification interacting the post-treatment indicator with a hyperscale county indicator yields a similar 31 percentage point conditional differential. Size matching tightens rather than erases the hyperscale–colocation divergence, and the base-rate concern, that hyperscale gains look large because they start from a small information sector, does not survive restriction to counties whose information base is comparable to the colocation group.

Table 3: Hyperscale vs. Colocation on Size-Restricted Samples

| Total private | | Information | |
|---|---|---|---|
| Hyperscale | Colocation | Hyperscale | Colocation |
| *Panel A: Common-support sample* | | | |
| $+3.2^{***}$ | $+1.5^{***}$ | $+37.8^{***}$ | $-3.8$ |
| $n = 23$ | $n = 17$ | $n = 20$ | $n = 17$ |
| *Panel B: Size-caliper matched pairs* | | | |
| $+4.4^{***}$ | $+1.9^{***}$ | $+29.4^{***}$ | $-4.9$ |
| $n = 19$ | $n = 20$ | $n = 17$ | $n = 20$ |

*Notes:* Each cell reports the average post-treatment SCM gap (in percent, exponentiated from the log-employment gap) across counties in the indicated sample. Standard errors are computed from the cross-county distribution of average post gaps. Panel A restricts both groups to the common support of pre-treatment total private employment, [14K, 684K]. Panel B pairs each colocation county to its nearest hyperscale county in log pre-treatment employment, without replacement, yielding 22 matched pairs. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

One interpretation consistent with the SCM-by-type evidence is that hyperscale facilities draw in local suppliers, including network operations centers, managed service providers, and fiber and power contractors, that need physical proximity to the campus for installation, maintenance, and low-latency interconnection. This distinguishes data centers from technology investments where agglomeration operates through thick labor markets [Moretti,

Table 4: Information-Sector Pre-Trend Tests by Facility Type

| | Mean annual pre-trend (%) | Within-type treated $N$ | Joint $F$ on leads | $p$-value |
|---|---|---|---|---|
| *Panel A: Within-type pre-trend tests vs. controls* | | | | |
| Hyperscale only | +1.37 | 33 | 4.03 | 0.003 |
| Colocation only | -0.24 | 23 | 5.25 | 0.000 |
| Mixed | -0.06 | 12 | 2.51 | 0.040 |
| *Panel B: Cross-type difference test (hyperscale vs. colocation), four leads $t = -5$ to $-2$* | | | | |
| Lead × hyperscale interactions | — | 2,873 | 2.10 | 0.079 |

*Notes:* Panel A: For each facility type, the sample is treated counties of that type and all never-treated controls in the pre-period (event time $< 0$ for treated, all years for controls). The dependent variable is log information-sector employment. Regressors are event-time lead dummies for $t = -5$ through $t = -2$ (with $t = -1$ as the omitted base), with county and year fixed effects. The joint $F$-test is on the null that all four lead coefficients equal zero. "Mean annual pre-trend" is the descriptive annualized log change in information employment from event time $-5$ to $-1$ for treated counties of each type. Panel B: The sample combines hyperscale-only and colocation-only treated counties with controls in the pre-period. The specification adds a hyperscale-by-event-time interaction. The joint $F$-test is on the null that the four interaction terms are jointly zero, i.e., that pre-trends do not differ between hyperscale and colocation. Standard errors clustered by county throughout.

2010]. Colocation providers, by contrast, lease space to remote tenants whose operations staff typically remain at the tenant's headquarters, generating less demand for local IT employment.[8]

A direct check on the anchor-ecosystem interpretation is whether the information-sector workers entering hyperscale counties are higher paid, as predicted by a story in which the entrants are technical staff and on-site engineering rather than generic clerical or sales occupations. Table 6 reports SCM estimates of the effect of entry on log average wages by sector, split by facility type. The information-sector wage rises by 19.0 percent in hyperscale counties (SE = 6.0) and shows no effect in colocation counties ($-3.8$ percent, SE = 4.3); the hyperscale–colocation difference is 22.8 percentage points and is statistically significant

[8]The 43 percent figure reflects a small pre-existing information sector at the median hyperscale county; in absolute terms it corresponds to approximately 290 new information-sector jobs, large relative to the county's base but modest in absolute scale. The percentage effect would be smaller in counties with larger pre-existing information sectors.

Table 5: SCM Employment Effects by Data Center Facility Type

| | Total private | Information | $N$ |
|---|---|---|---|
| All DC counties | $4.2\%^{***}$ | $22.4\%^{**}$ | 90 / 80 |
| Hyperscale only | $3.3\%^{***}$ | $43.3\%^{***}$ | 30 / 25 |
| Colocation only | $1.9\%^{***}$ | $-4.9\%$ | 20 / 20 |
| Mixed | $1.1\%^{***}$ | $7.6\%^{**}$ | 12 / 11 |

*Notes:* Each cell reports the SCM gap between treated counties and their synthetic controls. The "All DC counties" row reports the gap at the long-run horizon $t = +6$. The per-type rows report the average gap over the post-treatment period ($t = 0$ through $t = +6$). Both are converted from the log-employment gap to percent via $100 \times (\exp(\bar{g}) - 1)$. Significance based on cross-county standard errors. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. $N$ reports the number of matched treated counties for total private / information sector. The raw classification yields 33 hyperscale-only, 23 colocation-only, 12 mixed, and 25 unclassified counties; sample sizes here are smaller because some counties drop out of the SCM due to convergence failures or pre-treatment RMSE above 0.15 log points, with attrition larger for the information-sector outcome. Hyperscale counties contain only hyperscale operators. Colocation counties contain only colocation/wholesale operators. Mixed counties contain both types.

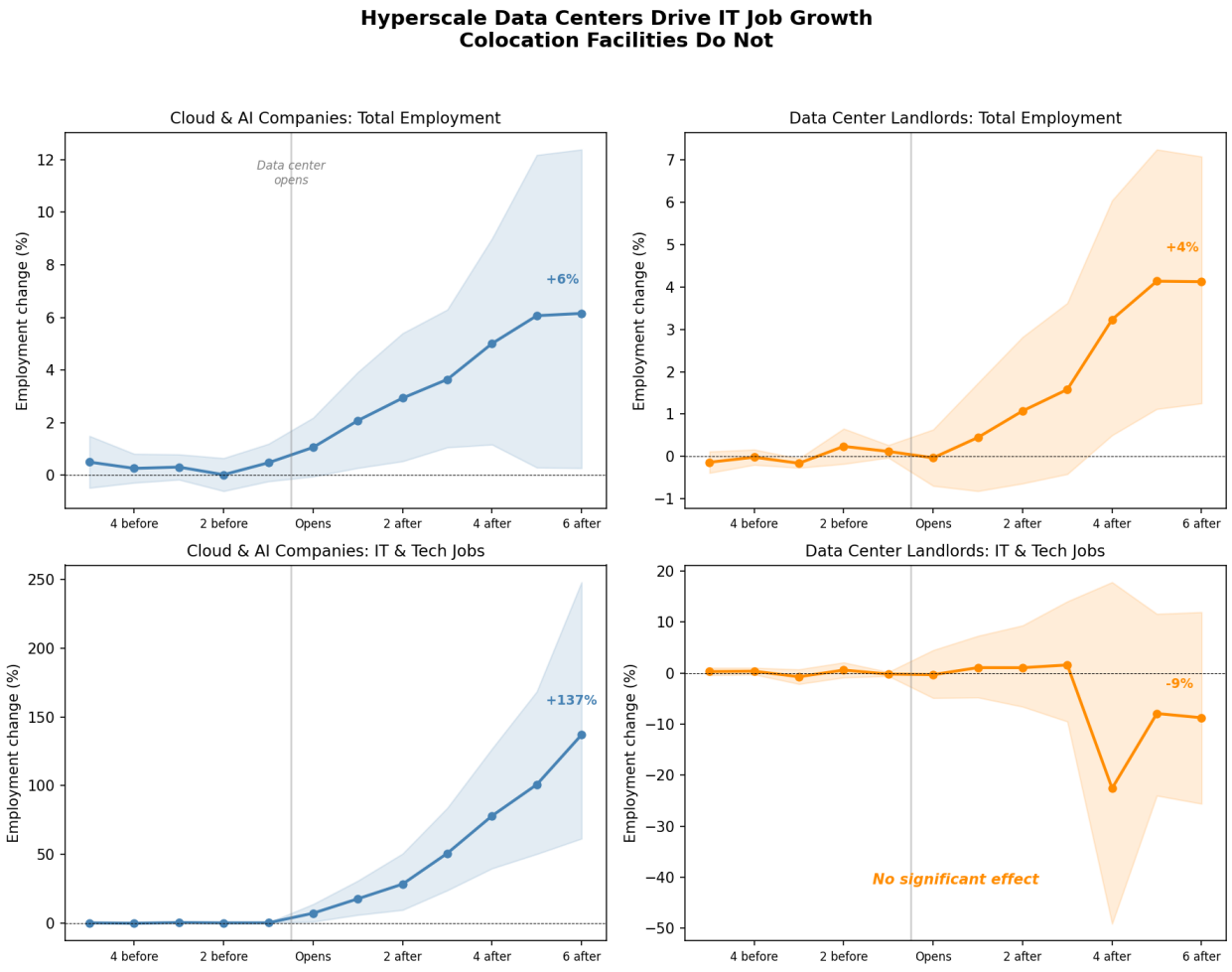


Figure 3: Synthetic Control by Facility Type: Hyperscale vs. Colocation

*Notes:* Each panel plots the average gap between treated counties and their synthetic controls, with 95 percent confidence intervals. Left column: hyperscale-only counties. Right column: colocation-only counties. Top row: total private employment. Bottom row: information sector employment. Per-panel sample sizes reflect counties with observations at each event time. Methodology as in Figure 1.

at the 1 percent level. Total private wages, by contrast, rise by a similar modest amount in both types (1.3 versus 2.0 percent) and do not differ significantly across types. The same facility-type split that produces a 36-percentage-point employment divergence in the size-matched information-sector SCM also produces a 23-percentage-point wage divergence in the information sector and essentially none in the aggregate. This is what one would expect if hyperscale entry pulls in higher-paid technical workers concentrated in information services. It is harder to reconcile with a story in which colocation simply fails to attract workers: colocation's zero employment effect is paired with a near-zero wage effect, so the type difference appears to operate through who enters rather than through wage compression at the bottom of the distribution.

Table 6: Wage Effects by Facility Type

| Outcome | Hyperscale | Colocation | Hyp − Col |
|---|---|---|---|
| Total private wage | +1.31%** | +1.99%* | −0.67% |
| | (0.52) | (1.07) | (1.19) |
| Information wage | +19.03%*** | −3.76% | +22.79%*** |
| | (6.03) | (4.32) | (7.42) |
| Information emp (ref.) | +26.88%*** | −4.67% | +31.54%*** |
| | (6.60) | (5.21) | (8.41) |

*Notes:* Each cell reports the average post-treatment SCM gap across treated counties in the indicated group. Standard errors are computed from the cross-county distribution of average post-treatment gaps. The third row re-reports the information-sector employment result from Table 5 for comparison. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

**Direct test of the anchor-firm mechanism: establishment entry.** The wage-premium result is consistent with an anchor-firm story but does not by itself distinguish firm entry from incumbent hiring. A sharper test is whether hyperscale entry raises the *count* of establishments in the information and construction sectors, not merely their headcount. If hyperscale operators are anchoring a local supply chain of fiber contractors, managed service providers,

and network operations centers, we should observe new physical establishments opening in the host county, not only existing establishments hiring more workers. If hyperscale entry instead reflects churn within incumbents or reclassification of existing workers, establishment counts should be unchanged. Table 7 applies the same SCM design to county-year log establishment counts from QCEW for NAICS 23 (construction), 51 (information), and 54 (professional services). Hyperscale-only entry raises the count of construction establishments by 18.1 percent at $t = 6$ (SE = 6.9) and the count of information-sector establishments by 26.2 percent (SE = 12.6), both statistically significant. Professional-services establishment entry is positive but imprecisely estimated. Colocation-only entry generates no establishment entry in any of the three sectors (Table 7). The information-sector establishment effect in hyperscale counties tracks the 22-percent information-sector employment effect closely, consistent with the employment gain being delivered by new firms entering the host county rather than solely by hiring at incumbents. QCEW establishment counts do not let us observe supplier relationships directly, so we read this as strong supportive evidence for the anchor-firm mechanism rather than a definitive supplier-linkage test.

Table 7: SCM Estimates for Establishment Counts, by Facility Type

| Sample | Construction (NAICS 23) | Information (NAICS 51) | Professional services (NAICS 54) |
|---|---|---|---|
| All treated | $+7.8^{***}$ (2.5) | $+9.4^{*}$ (4.9) | +1.8 (2.2) |
| Hyperscale-only | $+18.1^{***}$ (6.9) | $+26.2^{**}$ (12.6) | +7.5 (6.8) |
| Colocation-only | +1.5 (2.9) | −0.2 (7.5) | −1.6 (3.2) |
| $N$ at $t = 6$ (all / hyp / colo) | | 45 / 12 / 15 | |

*Notes:* SCM estimates at $t = 6$ for log county-year establishment counts from QCEW (annual singlefile, private ownership), for NAICS 23 (construction), 51 (information), and 54 (professional, scientific, and technical services). The same matching procedure used for employment outcomes is applied unchanged: nearest-200 donors by pre-treatment log employment, SLSQP-constrained weights, pre-period RMSE screen. Cross-treated standard errors in parentheses. Hyperscale-only counties host at least one hyperscale facility and no colocation facilities; colocation-only is the reverse. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

**Dose-response.** If hyperscale entry generates local information-sector agglomeration, larger local clusters should produce larger effects. Estimating the Callaway and Sant'Anna [2021] ATT separately by facility count, total private employment effects are flat across dose bins (4–5 percent for 1, 2–3, and 4+ facilities), suggesting that total employment gains are driven by first entry and do not scale with subsequent investment. Information-sector employment, by contrast, shows a clear dose-response gradient: 7 percent for single-facility counties (not statistically significant), 8 percent for 2–3 facilities, and 23 percent for 4+ facilities (significant at 1 percent). The information-sector agglomeration effect requires a *cluster* of facilities to materialize.

## 6.2 Other Sources of Heterogeneity

Table 8 examines heterogeneity by county characteristics and treatment timing using SCM.

Table 8: SCM Heterogeneity by County Characteristics and Cohort

| Subsample | Gap at $t$=+6 | SE | $N$ | Implied effect |
|---|---|---|---|---|
| *By county type:* | | | | |
| Rural (below-median pre-employment) | 0.055** | (0.021) | 45 | 5.6% |
| Urban (above-median pre-employment) | 0.032*** | (0.010) | 45 | 3.3% |
| *By treatment cohort:* | | | | |
| 2008–2014 | 0.029*** | (0.011) | 30 | 2.9% |
| 2015–2019 | 0.062*** | (0.020) | 21 | 6.4% |

*Notes:* Each row reports the average SCM gap at $t = +6$ for total private employment, estimated on the indicated subsample. SE is the cross-county standard error. The 2020–2024 cohort is omitted because insufficient post-treatment years are available for $t = +6$ estimates. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Two patterns emerge. First, *rural counties* show larger effects (5.6 percent) than *urban counties* (3.3 percent), consistent with the hypothesis that a single large facility represents a larger relative shock in less populated areas. Both estimates are statistically significant.

Second, the effects are positive across treatment cohorts. The 2015–2019 cohort shows a larger effect (6.4 percent) than the 2008–2014 cohort (2.9 percent), likely reflecting the rising scale of modern data center facilities.

# 7 Spillovers and Local Displacement

The host-county effect could reflect either genuine net job creation or spatial reallocation from nearby jurisdictions. If data centers pull construction crews, managed-service providers, and information-sector employment out of neighboring counties, the host-county estimate overstates the true local effect and the CZ-level or regional effect may be zero. We test for this in two ways: by re-running the SCM on counties contiguous to treated hosts, and by aggregating the panel to commuting zones.

**Contiguous-county spillovers.** We re-estimate the SCM design on the 155 never-treated counties that are geographically contiguous to hyperscale-only treated counties, using the same donor-pool and RMSE screening rules as the main estimates. A beggar-thy-neighbor pattern would appear as negative and significant contiguous-county effects. Instead, we find a small and weakly *positive* spillover. Total private employment in contiguous non-host counties rises by 1.1 to 1.4 percent at $t = 1$ through $t = 3$, fading toward zero by $t = 6$. This is the opposite sign from what reallocation would predict, and at the long horizon the point estimate is indistinguishable from zero.

**Commuting-zone aggregation.** We next aggregate the county-year panel to commuting zones using the Dorn [2009] crosswalk, reducing the sample to 59 treated and 682 control CZs (58 of these treated CZs reach the $t = 6$ horizon). The CZ-level total employment effect is 1.7 percent at $t = 6$. The CZ-level construction effect is near zero (0.3 percent, not significant), consistent with construction activity being geographically concentrated at the facility site. The CZ-level information effect is 5.3 percent and not statistically significant,

reflecting dilution of a highly concentrated effect across the larger CZ employment base. A pure reallocation story would predict $\tau_{CZ} \approx 0$ for total employment; our estimate is positive and larger than zero, corroborating the contiguous-county evidence. Table 9 reports CZ-level predictor balance. The CZ synthetic control matches treated CZs closely on pre-treatment growth, sectoral composition, and average wages; the only meaningful imbalance is in the level of pre-treatment employment, where the synthetic CZ is on average larger than the treated CZ. This level imbalance reflects that the SCM matches on the log outcome trajectory rather than on the level itself, and we read it as a feature of the design rather than a violation of the matching assumption.

Table 9: Predictor Balance: CZ-Level SCM (Treated vs. Synthetic Control CZs)

| | Treated | Synthetic | Controls | Diff. | Std. Diff. |
|---|---|---|---|---|---|
| Pre-treatment employment (CZ total) | 780,755 | 1,338,942 | 102,488 | −558,187 | −0.90 |
| Employment growth, 03–07 (%) | +7.5 | +6.9 | +4.5 | +0.6 | +0.09 |
| Average weekly wage ($) | 740 | 766 | 568 | −26 | −0.20 |
| Information share (%) | 2.3 | 2.1 | 1.3 | 0.2 | +0.24 |
| Construction share (%) | 5.7 | 5.8 | 4.6 | −0.0 | −0.02 |
| Prof. services share (%) | 4.8 | 4.5 | 2.1 | 0.3 | +0.12 |

*Notes:* Pre-treatment characteristics computed as CZ-level means over 2003–2007. Treated and synthetic CZs are restricted to the 58 treated CZs with pre-RMSE $\leq 0.15$. "Synthetic" is the weight-averaged value across each treated CZ's donor pool, then averaged across treated CZs. "Controls" is the simple mean across all never-treated CZs in the donor pool. Standardized difference is $(\bar{X}_{\text{treated}} - \bar{X}_{\text{synthetic}})$ divided by the pooled standard deviation across treated and never-treated CZs.

**Reading the two tests together.** A pure within-CZ reallocation story predicts negative contiguous-county effects and a CZ-level effect close to zero. We find the opposite on both counts: the contiguous-county estimate is small and weakly positive, and the CZ-level estimate is positive and meaningfully larger than zero. We therefore read the host-county SCM estimates as net of within-CZ displacement to a first approximation. We cannot speak to cross-CZ leakage with this design; if data centers pull workers from counties outside the host

CZ, that channel is absorbed into the CZ-level estimate and would not be detected here.

# 8 Economic Magnitudes

The percentage effects reported so far are difficult to compare directly to the place-based-policy literature, where the natural units are jobs per direct employee (local multipliers) and public dollars per job (cost per job). This section translates the SCM estimates into both.

## 8.1 Implied Local Multiplier

Following Moretti [2010], we define the local multiplier as the number of indirect jobs created in the host county per direct job in the tradable sector that initiates the shock. At the median hyperscale-only treated county, pre-treatment total private employment is 36,400 and the post-treatment SCM gap implies approximately 1,640 additional jobs at $t = 6$. Direct on-site employment at a hyperscale facility is not directly observed; industry sources (operator 10-Ks, JLL 2024, CBRE 2024) place operational staffing at 50–200 per facility, with the upper end reflecting larger multi-building campuses. We report the implied multiplier under three calibrations of direct employment per facility, holding the total-emp effect fixed:

Under the midpoint assumption of 125 direct employees per facility, the implied multiplier is 2.27, which sits squarely in the range bracketed by Moretti [2010]'s manufacturing estimate (1.59) and the high-tech estimate in Moretti and Thulin [2013] (2.5). Under the high-end assumption (200 direct/facility, approximating the largest AWS and Meta campuses) the multiplier is 1.05, close to the manufacturing benchmark; under the low end (75/facility, closer to industry averages for older hyperscale sites) it is 4.46, above high-tech benchmarks. The range bracketing these calibrations is broadly consistent with the hypothesis that hyperscale data centers, despite their extreme capital intensity, generate local employment spillovers of a magnitude comparable to other technology investments, driven by the anchor-tenant mechanism documented in Section 6.

Table 10: Implied Local Multiplier at the Median Hyperscale County

| Sample | $N$ | Median pre-emp | Median facilities | Direct DC jobs | Total new jobs | Implied multiplier |
|---|---|---|---|---|---|---|
| Hyperscale-only (low, 75/fac) | 33 | 36,386 | 4 | 300 | 1,637 | 4.46 |
| Hyperscale-only (mid, 125/fac) | 33 | 36,386 | 4 | 500 | 1,637 | 2.27 |
| Hyperscale-only (high, 200/fac) | 33 | 36,386 | 4 | 800 | 1,637 | 1.05 |
| *Benchmarks from the literature* | | | | | | |
| Moretti [2010] manufacturing | | | | | | 1.59 |
| Moretti and Thulin [2013] high-tech | | | | | | 2.50 |

*Notes:* Total new jobs is the product of the median hyperscale-only county's 2003–2007 pre-treatment total private employment and the pooled SCM total-employment effect of +4.5 percent at $t = 6$. Direct DC jobs is the product of the number of large facilities at the median hyperscale county and the indicated per-facility operational staffing assumption. The implied multiplier is (total new jobs − direct DC jobs) / direct DC jobs. Benchmarks from Moretti [2010] (manufacturing) and Moretti and Thulin [2013] (high-tech).

## 8.2 Cost per Job

The fiscal cost of attracting a data center can be compared to the implied employment gain, though this calculation is an accounting benchmark rather than a welfare analysis or a but-for estimate of incentive effects. At the median treated county, the SCM total-employment effect implies approximately 4,000 new jobs at $t = 6$. For the ten states that publicly report the fiscal cost of their data center incentive programs, we allocate each state's annual subsidy across its treated counties and discount over a 10-year horizon at 3 percent.[9] The resulting distribution is bracketed by estimates in the literature: Greenstone et al. [2010] report Million Dollar Plant jobs at approximately $27,000 each, and Bartik [2020] documents typical state incentive costs in the range of $25,000–$50,000 per job. Relative to private construction outlay at $1,500 per square foot, state incentive PV represents about 2 percent of investment at the median hyperscale county and 62 percent at the median colocation county, consistent with power and fiber constraints playing the dominant role in hyperscale siting and fiscal

[9]Annual state subsidy estimates: Texas ($1,000M), Virginia ($732M), Illinois ($370M), Georgia ($296M), Iowa, Minnesota, Nevada, Ohio, and Washington ($150M each), and Tennessee ($100M). Sources: Good Jobs First (2025), Virginia JLARC, state comptroller reports. Twelve of the 32 states with data center incentive programs do not disclose their fiscal cost; treated counties in those states are excluded.

incentives potentially mattering more for colocation and smaller operators.

# 9 Robustness

We subject the main findings to a set of robustness checks, summarized in Table 14. The highest-leverage checks concern pre-existing differential trends, sensitivity to the TWFE specification, robustness to heterogeneity-robust estimators, a complementary shift-share IV (with placebo-sector diagnostics) based on pre-existing power infrastructure, and the role of feasible donor pools and housing capitalization in the interpretation of the estimates.

**Shift-share IV based on power infrastructure.** As a complementary design that does not rely on matching on the pre-treatment outcome, we instrument cumulative data center capacity with the interaction of a county's pre-existing 345+ kV substation count and the leave-one-state-out national time-path of installed megawatts. The shift is the county's standardized substation count as of the start of the sample; the shock is cumulative national capacity built outside the county's own state through year $t$, following the shift-share literature [Goldsmith-Pinkham et al., 2020, Borusyak et al., 2022]. Table 11 reports the main results. The just-identified power IV yields a total-employment coefficient of 0.165 (s.e. 0.051) and an information-sector coefficient of 0.126 (s.e. 0.070), both positive and directionally consistent with the SCM.

To assess the exclusion restriction directly, we run the same just-identified IV on a battery of placebo sectors with no plausible data center channel: agriculture (NAICS 11), manufacturing (31–33), retail (44–45), finance (52), health care (62), and accommodation and food services (72). The specification mirrors the main IV exactly, varying only the outcome. Table 12 reports the results. Construction (NAICS 23), included as a positive control, loads on the IV with a coefficient of 0.263 ($p < 0.01$), confirming that the instrument delivers signal in the expected direction. However, the placebo sectors are not uniformly null: manufacturing (0.31, $p = 0.03$), finance (0.41, $p < 0.01$), and accommodation and food (0.41, $p = 0.07$)

all load on the IV at conventional levels, while agriculture, retail, and health care are null. This pattern indicates that pre-existing 345 kV substation density is correlated with general employment growth in counties exposed to the national rollout, not solely with data center entry. We therefore do not interpret the IV as cleanly identifying the data center channel. We retain it as a directional cross-check on the SCM, while noting that the SCM, not the IV, is the load-bearing identification design in the paper.

Table 11: Complementary Shift-Share IV Evidence

| | Total private employment | Information employment |
|---|---|---|
| Power only (baseline) | $0.165^{***}$ | $0.126^{*}$ |
| | (0.051) | (0.070) |
| Lagged power shift $(t-1)$ | $0.164^{***}$ | $0.121^{*}$ |
| | (0.052) | (0.070) |
| Hyperscale power shift | $0.161^{***}$ | $0.115^{*}$ |
| | (0.050) | (0.069) |
| Matched power + connectivity | $0.161^{***}$ | $0.145^{**}$ |
| | (0.056) | (0.059) |

*Notes:* Each cell reports a 2SLS coefficient from a county-year specification with county and year fixed effects, instrumenting $\log(1 + \text{cumulative MW})$. "Power only" uses the standardized count of in-county 345+ kV substations interacted with leave-one-state-out national cumulative data center openings. "Lagged power shift" uses the one-year lag of that instrument. "Hyperscale power shift" interacts the same power exposure with national hyperscale openings only. "Matched power + connectivity" uses two instruments: power exposure × hyperscale openings and local connectivity exposure × colocation openings. Standard errors clustered by state in parentheses. LIML estimates are numerically identical to 2SLS in all specifications. Across the reported specifications, the first-stage partial $R^2$ ranges from 0.025 to 0.044 for total employment and from 0.031 to 0.049 for information employment. For the matched specification, Sargan $p$-values are 0.968 for total employment and 0.070 for information employment. Construction effects are positive in the just-identified power IV but are omitted here because overidentification tests reject in the overidentified specifications. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 12: Placebo-Sector Shift-Share IV (QCEW)

| Outcome | 2SLS coef. | First-stage $F$ | Partial $R^2$ | $N$ | Counties |
|---|---|---|---|---|---|
| Construction (positive control) | 0.263*** | 14.8 | 0.028 | 41,628 | 2,663 |
| | (0.097) | | | | |
| Agriculture | 0.043 | 23.5 | 0.029 | 53,623 | 3,202 |
| | (0.249) | | | | |
| Manufacturing | 0.309** | 23.3 | 0.029 | 53,766 | 3,204 |
| | (0.139) | | | | |
| Retail trade | 0.075 | 23.4 | 0.029 | 54,633 | 3,218 |
| | (0.053) | | | | |
| Finance and insurance | 0.406*** | 23.4 | 0.029 | 54,407 | 3,217 |
| | (0.153) | | | | |
| Health care | 0.389 | 23.3 | 0.029 | 54,406 | 3,219 |
| | (0.343) | | | | |
| Accomm. & food services | 0.414* | 23.4 | 0.029 | 54,504 | 3,218 |
| | (0.229) | | | | |

*Notes:* Each row reports a 2SLS coefficient from a county-year specification with county and year fixed effects, instrumenting log(1 + cumulative MW) with the standardized count of in-county 345+ kV substations interacted with leave-one-state-out national cumulative data center openings. The specification mirrors the just-identified "Power only (baseline)" column of Table 11 exactly, varying only the outcome. Construction is a positive control: substation-driven exposure to data center entry should generate construction employment if the IV identifies the data center channel. The remaining sectors have no plausible data center channel and are placebos. Standard errors clustered by state in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

**Pre-trends.** The regression-based event study reveals that treated counties exhibit differential pre-trends in total private employment. Relative to the omitted period $t = -1$, the lead coefficients at $t = -4$ through $t = -2$ range from $-0.025$ to $-0.011$ log points and are jointly significant ($F$-test $p < 0.001$), indicating that treated counties were already on a steeper upward trajectory than controls before data center construction. This motivates our preference for SCM over TWFE: the SCM matches pre-treatment trajectories directly, whereas TWFE conflates selection with causation when pre-trends exist.

**Specification robustness.** Within the class of TWFE specifications, the baseline estimates are stable across modifications that do not address the pre-trend concern (Table 14). Adding state $\times$ year fixed effects, matching on pre-treatment employment quintiles, and controlling for pre-treatment employment $\times$ year interactions all yield similar estimates. Employment weighting attenuates the effects, suggesting that the baseline effects are larger in smaller counties where a single data center represents a larger relative shock.

**Feasibility-restricted donor pool.** We re-estimate the baseline SCM with donors restricted to the 868 never-treated counties containing at least one 345+ kV substation, the same feasibility screen used by the shift-share IV. The feasibility-restricted estimates are within 0.2 percentage points of the baseline for total employment (3.94 vs. 4.08), move the information-sector effect slightly *upward* (23.4 vs. 20.2), and move construction modestly downward (7.7 vs. 10.3). Every treated county is matched under both donor pools, suggesting the baseline SCM was already selecting donor weights from the feasible subset (Table 13).

Table 13: SCM Estimates with Feasibility-Restricted Donor Pool

| Outcome (at $t = 6$) | Baseline donors | Feasible-only donors |
|---|---|---|
| Total private | $+4.1^{***}$ (1.0) | $+3.9^{***}$ (1.1) |
| Information | $+20.2^{**}$ (9.1) | $+23.4^{***}$ (8.8) |
| Construction | $+10.3^{***}$ (3.6) | $+7.7^{**}$ (3.7) |
| $N$ treated matched | 47 | 47 |

*Notes:* Compares baseline SCM estimates (unrestricted never-treated donor pool) to estimates obtained by restricting donors to the 868 never-treated counties with at least one 345+ kV substation at the start of the sample, the same feasibility set used to define the shift-share IV. All outcomes in logs; reported effects are at $t = 6$. Cross-treated standard errors in parentheses. The two specifications use identical matching procedures (nearest-200 donors by pre-treatment log employment, SLSQP-constrained weights, pre-period RMSE screen). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

**Heterogeneity-robust estimators.** We also estimate heterogeneity-robust alternatives to SCM for staggered adoption. Using a stacked difference-in-differences design [Cengiz et al., 2019], we obtain a total private employment effect of 5.4 percent and an information sector effect of 18.7 percent, both statistically significant at the 1 percent level. A Callaway and Sant'Anna [2021]-style estimator with bootstrap standard errors (200 replications, resampling counties with replacement, IQR-based SE robust to outlier draws) gives a total employment ATT of 5.2 percent (bootstrap SE = 0.008) and an information employment ATT of 15.5 percent (bootstrap SE = 0.041) using a ±5-year window. These estimates are consistent with the SCM estimates, reinforcing the conclusion that the true total employment effect is in the range of 4–5 percent and the information sector effect is 15–22 percent. Appendix Table 21 reports the full cross-estimator comparison.

**Housing prices.** If data centers raise local wages and attract workers, housing costs may rise, offsetting the welfare gains. Using the FHFA All-Transactions House Price Index (available for 92 of 93 treated counties), we apply the same SCM methodology to log house prices. The pre-treatment gaps are flat, and the post-treatment point estimate reaches a modest 2.0

percent by $t = 6$, but the effect is not statistically significant at any event time. We read this as suggesting that capitalization is small relative to wage gains, while noting that the FHFA index may understate rent changes in thin rural markets.

**Additional diagnostics.** Appendix A reports additional inference and sensitivity checks, including SCM prediction intervals and permutation inference, predictor balance, leave-one-out estimates, treatment-timing sensitivity, in-time placebos, reverse-causality checks, and population and migration results.

# 10 Conclusion

We match each treated county to a weighted combination of never-treated counties on its pre-treatment employment trajectory and track the resulting gaps through six years after first entry. The SCM yields total private employment gains of 4 to 5 percent, construction gains of 11 percent, and information-sector gains of 22 percent at $t = 6$, reinforced by stacked DiD and Callaway–Sant'Anna. Wages rise by approximately 3 percent in both incumbent workers and new hires. Contiguous non-host counties show small and weakly positive effects, the opposite sign from a reallocation story, and commuting-zone aggregation yields a positive total employment effect of 1.7 percent. We read these as evidence that the host-county estimates are not inflated by within-CZ reallocation, and the local labor-market reading remains positive rather than zero-sum.

The headline information-sector effect masks sharp differences by facility type. Hyperscale counties drive the result, with a 43 percent SCM gain in information employment, while colocation counties show no statistically significant gains. A size-matched SCM yields a 36 percentage point hyperscale–colocation differential, and a TWFE triple-difference recovers a similar 31 percentage point conditional differential. Hyperscale entry also raises the count of information-sector and construction establishments in the host county, while colocation entry does not. The central implication is that the local incidence of digital infrastructure

Table 14: Robustness of Main Results Across Specifications

| | Total private employment | Information employment |
|---|---|---|
| Baseline (county + year FE) | 0.136*** | 0.344*** |
| | (0.018) | (0.053) |
| State x year FE | 0.118*** | 0.328*** |
| | (0.016) | (0.050) |
| Callaway-Sant'Anna ATT | 0.049*** | 0.153*** |
| | (0.007) | (0.033) |
| Matched (size quintile) | 0.136*** | 0.344*** |
| | (0.018) | (0.053) |
| Pre-emp x year trend | 0.104*** | 0.333*** |
| | (0.019) | (0.055) |
| Dropping Virginia | 0.141*** | 0.366*** |
| | (0.019) | (0.055) |
| Employment-weighted | 0.091*** | 0.274*** |
| | (0.014) | (0.049) |
| *Alternative facility size thresholds:* | | |
| 50K sqft threshold | 0.126*** | — |
| | (0.019) | |
| 100K sqft (baseline) | 0.127*** | — |
| | (0.019) | |
| 200K sqft threshold | 0.123*** | — |
| | (0.019) | |
| Neighbor county spillover | 0.000 | −0.008 |
| | (0.003) | (0.011) |
| Total emp. excl. construction | 0.136*** | — |
| | (0.018) | |

*Notes:* Dependent variable: log employment. Each cell reports the coefficient on the post-treatment indicator from a separate regression. All specifications include county and year fixed effects unless otherwise noted. Standard errors clustered at the county level in parentheses. Baseline sample: 93 treated counties and approximately 3,040 never-treated counties, 2003–2024. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. Callaway-Sant'Anna estimates are cohort-size-weighted averages of cohort-specific ATTs estimated against the never-treated group, with bootstrap standard errors (200 replications).

investment depends on who operates the facility, not only on how much capital is installed. Colocation generates comparable construction employment but not the information-sector agglomeration that distinguishes data centers from other large capital projects. Section 8 translates these effects into an implied local multiplier of roughly 2.3 at the median hyperscale county and a cost per job consistent with standard benchmarks in the incentive literature. State incentives represent a small fraction of private investment in hyperscale counties but a much larger share in colocation counties, consistent with power and fiber constraints playing the dominant role in hyperscale siting rather than fiscal subsidies.

Three scope conditions frame the interpretation. First, our spillover tests address within-CZ reallocation only and do not speak to cross-CZ leakage or general-equilibrium effects outside the local labor market. Second, we cannot answer the "but-for" question of whether incentivized facilities would have been built elsewhere absent the subsidy. Third, our setting estimates the effect of *first entry* into a county that is in the technically feasible set and not already hosting one of the established clusters we exclude (Northern Virginia, Silicon Valley, Dallas, Phoenix), so the estimates speak to new-entry counties rather than mature ecosystems. Within these limits, first entry by a hyperscale operator generates real local labor demand and information-sector agglomeration, while otherwise similar colocation entry does not. The policy implication is narrow but important: if states aim to buy local agglomeration rather than construction alone, operator type is hard to ignore.

# A Additional Robustness and Inference

## A.1 IV diagnostics

Table 15: Shift-Share IV: First-Stage Diagnostics and Alternative-Industry Placebos

| *Panel A: Baseline IV first-stage diagnostics* | | | | |
|---|---|---|---|---|
| Outcome | 2SLS coef. | (SE) | First-stage stat. | $N$ |
| Total private employment | +0.165 | (0.051) | 14.7 | 42,180 |
| Information-sector employment | +0.126 | (0.070) | 22.4 | 37,759 |
| *Panel B: Alternative-industry 2SLS placebos* | | | | |
| Sector | 2SLS coef. | (SE) | $p$ | $N$ |
| Agriculture | +0.189 | (0.140) | 0.177 | 29,359 |
| Construction (positive control) | +0.377*** | (0.105) | 0.000 | 49,038 |
| Manufacturing | +0.128** | (0.064) | 0.046 | 48,445 |
| Retail trade | +0.050 | (0.043) | 0.239 | 53,556 |
| Finance & insurance | +0.190*** | (0.057) | 0.001 | 46,870 |
| Health care | +0.243*** | (0.067) | 0.000 | 34,633 |
| Accommodation & food | +0.113 | (0.070) | 0.106 | 42,401 |

*Notes:* Panel A reports the baseline just-identified shift-share IV first-stage diagnostics on the two headline outcomes. The instrument is the interaction of a county's standardized count of 345+ kV substations with the leave-one-state-out national time-path of cumulative data-center megawatts. The reported first-stage statistic comes from the excluded-instrument test in the just-identified first stage; LIML estimates (not shown) are numerically identical to 2SLS throughout. Panel B reports 2SLS coefficients when the same endogenous variable and instrument are used to predict log county-year employment in alternative 2-digit NAICS sectors. Construction is reported as a positive control because the data-center build-out passes directly through that sector. Sectors with no plausible data-center channel (agriculture, retail, accommodation) are null or marginal. Sectors that respond mechanically to local income and population (finance, health care) are positive and of the same order as the total-employment headline; we therefore treat the IV as supportive rather than definitive. Standard errors clustered at the state level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

## A.2 SCM inference

**SCM inference.** We address inference for the synthetic control estimates using two complementary approaches: the prediction interval framework of Cattaneo et al. [2021], which provides finite-sample valid event-time-specific inference, and the canonical permutation test of Abadie et al. [2010].

The prediction interval approach decomposes the uncertainty of a synthetic control prediction into two distinct sources: estimation error in the SCM weights (from using finite pre-treatment data to estimate donor weights) and unobservable stochastic shocks in the post-treatment period (idiosyncratic county-level variation that is unpredictable from the pre-treatment fit). We implement this using the `scpi` package [Cattaneo et al., 2024], running the procedure separately for each treated county, using simplex-constrained weights and 200 simulation draws.

Table 16 reports the detailed results. Two inferential objects are relevant. First, the *cross-county standard errors* (the standard deviation of county-level gaps divided by $\sqrt{N}$) measure uncertainty about the average treatment effect across all treated counties. These indicate statistically significant effects: total private employment reaches 4.0 percent at $t = 6$ and information sector employment reaches 23.5 percent. The information sector effect is significant from $t = 1$ onward, with a trajectory that builds gradually from 7.5 percent at $t = 1$ to 23.5 percent at $t = 6$. Second, the *prediction intervals*, which account for both weight estimation error and stochastic uncertainty at the individual county level, are wide: $[-9, 19]$ percent for total employment and $[-14, 62]$ percent for information at $t = 6$. This width reflects genuine heterogeneity across counties rather than the absence of an average effect: some counties gain substantially from data center entry while others do not, a pattern consistent with the hyperscale/colocation decomposition in Section 6.

As a complementary check, we implement the canonical permutation test [Abadie et al., 2010], randomly assigning placebo treatment dates to 300 never-treated counties and comparing their SCM gaps to the actual treated counties' gaps. We report the test in two forms: an

Table 16: SCM with Prediction Intervals (Cattaneo et al. 2021)

| | Total Private Employment | | | Information Sector | | |
|---|---|---|---|---|---|---|
| Event time | Gap | SE | 95% PI | Gap | SE | 95% PI |
| $t = -3$ | +0.2% | (0.0015) | — | -0.1% | (0.0023) | — |
| $t = -2$ | +0.1% | (0.0016) | — | +0.2% | (0.0022) | — |
| $t = -1$ | +0.3% | (0.0016) | — | +0.1% | (0.0011) | — |
| $t = +0$ | +0.4% | (0.0027) | [-4.4, +4.3] | +2.0% | (0.0148) | [-29.4, +13.7] |
| $t = +1$ | +0.9%* | (0.0045) | [-6.3, +6.7] | +7.5%*** | (0.0255) | [-22.1, +22.7] |
| $t = +2$ | +1.4%** | (0.0057) | [-7.0, +9.0] | +10.4%*** | (0.0364) | [-18.5, +30.1] |
| $t = +3$ | +1.6%** | (0.0063) | [-9.3, +11.2] | +17.2%*** | (0.0484) | [-15.7, +44.5] |
| $t = +4$ | +2.5%*** | (0.0089) | [-8.0, +13.1] | +11.6% | (0.0893) | [-37.4, +41.6] |
| $t = +5$ | +3.7%*** | (0.0107) | [-7.2, +16.7] | +21.1%** | (0.0770) | [-22.8, +59.0] |
| $t = +6$ | +4.0%*** | (0.0108) | [-8.7, +19.4] | +23.5%** | (0.0910) | [-14.0, +61.5] |

*Notes:* Gap is the average treatment effect across treated counties at each event time (in percent). SE is the cross-county standard error. 95% PI is the average 95 percent prediction interval across counties, computed following Cattaneo et al. [2021] with 200 simulation draws per county using the `scpi` package. Point estimates differ slightly from Figure 1 because `scpi` re-estimates weights using its own simplex-constrained procedure. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$ (based on cross-county SE).

average-post-gap statistic (averaged over $t = 0$ through $t = 6$) and horizon-specific statistics at $t = 1$, $t = 3$, and $t = 6$, where the effect has fully accumulated. Table 17 reports both, in unrestricted and feasibility-restricted donor pools (the feasibility restriction limits placebos and their donors to counties with at least one 345+ kV substation). At every horizon and in both pools the actual treated mean gap falls inside the placebo distribution at conventional levels: the smallest $p$-value across all panels is 0.29 (information sector, max-absolute-gap, feasibility pool); the $t = +6$ horizon-specific $p$-value is 0.63 for total employment and 0.57 for information. We read the permutation evidence as inconclusive at the unit level. The wide `scpi` prediction intervals reported in Table 16 carry the same message: pooled across treated counties the cross-county standard errors deliver statistically significant mean effects, but at the individual-county level the data do not let us reject zero. The gap between the pooled-mean cross-county standard errors and the unit-level permutation rank tests reflects the small number of treated units (90) and the high cross-county heterogeneity, not a contradiction between the two procedures: each addresses a different inferential question. We treat the pooled-mean cross-county standard errors as the primary inferential object for the average treatment effect, while disclosing the unit-level permutation evidence transparently.

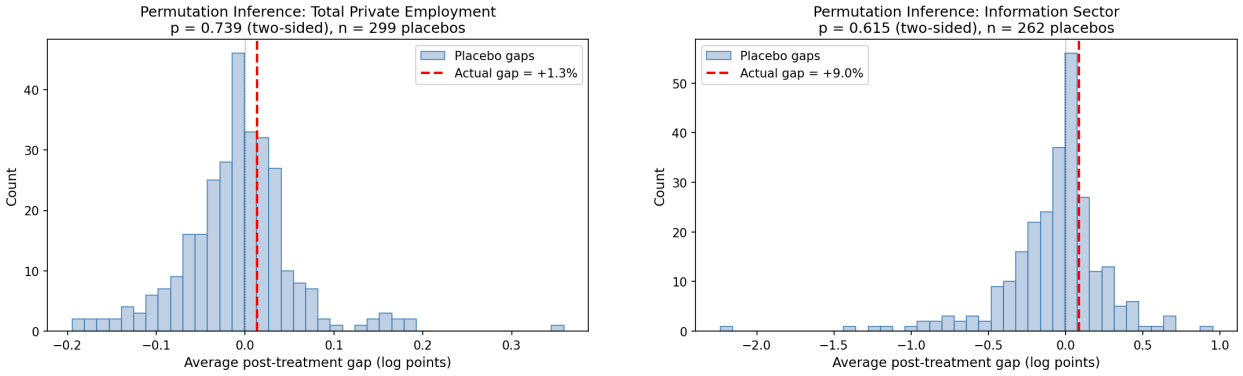


Figure 4: Permutation Inference: Distribution of Placebo Average Post-Treatment Gaps

*Notes:* Each histogram shows the distribution of average post-treatment gaps ($t = 0$ through $t = 6$) for 300 placebo counties drawn from the never-treated pool. The red dashed line marks the actual treated counties' average gap. $p$-values are two-sided: the fraction of placebo gaps with absolute value at least as large as the actual gap. As discussed in the text, the average-gap test statistic lacks power for gradually accumulating effects.

Table 17: Permutation Inference at Specific Horizons

| | | Total private | | Information | |
|---|---|---|---|---|---|
| Statistic | Donor pool | Actual | $p$-value | Actual | $p$-value |
| *Panel: Unrestricted donor pool* | | | | | |
| Avg. post gap | Unrestricted | +1.3% | [0.731] | +8.6% | [0.680] |
| Max \|gap\| (post) | Unrestricted | 5.5% | [0.521] | 39.4% | [0.364] |
| Gap at $t = 1$ | Unrestricted | +0.9% | [0.831] | +7.9% | [0.693] |
| Gap at $t = 3$ | Unrestricted | +1.7% | [0.751] | +14.1% | [0.585] |
| Gap at $t = 6$ | Unrestricted | +4.1% | [0.629] | +20.2% | [0.571] |
| *Panel: Feasible only donor pool* | | | | | |
| Avg. post gap | Feasible only | +1.3% | [0.715] | +8.6% | [0.620] |
| Max \|gap\| (post) | Feasible only | 5.5% | [0.472] | 39.4% | [0.285] |
| Gap at $t = 1$ | Feasible only | +0.9% | [0.784] | +7.9% | [0.642] |
| Gap at $t = 3$ | Feasible only | +1.7% | [0.714] | +14.1% | [0.555] |
| Gap at $t = 6$ | Feasible only | +4.1% | [0.581] | +20.2% | [0.573] |

*Notes:* Each "Actual" column reports the cross-treated-county mean of the test statistic at $t = +6$ (or as labeled). Each "$p$-value" column reports the rank-based two-sided permutation $p$-value comparing the actual statistic to the placebo distribution from $N = 300$ never-treated counties assigned random treatment years drawn from the actual treatment-year distribution. Each placebo county runs the same SCM procedure on the same outcome. "Unrestricted" uses the full never-treated pool; "Feasible only" restricts placebos and their donors to counties hosting at least one 345+ kV substation. "Avg.
post gap" averages $t = 0$ through $t = 6$. "Max |gap| (post)" uses the supremum statistic from
citetabadie$_d$*iamond*$_h$*ainmueller*2010.*Gapat*t =
k*isthecross* − *countymeangapathorizon*k*specifically.Stars* :$^*p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

## A.3 Predictor balance

Table 18 compares pre-treatment characteristics of treated counties, their synthetic controls, and the full control pool. The SCM substantially reduces imbalance on observables relative to the raw treated-versus-control comparison. Some residual differences remain, especially in pre-treatment growth rates, but the identifying object is the pre-treatment *trajectory* rather than exact equality in levels, and the pre-treatment RMSE below 0.15 log points for 90 percent of counties indicates that the path match is tight.

Table 18: Predictor Balance: Treated vs. Synthetic Control Counties

| | Treated | Synthetic | Controls | Diff. | Std. Diff. |
|---|---|---|---|---|---|
| Pre-treatment employment | 341,515 | 397,314 | 30,066 | −55,799 | −0.10 |
| Employment growth, 03–07 (%) | 13.6 | 10.3 | 8.5 | +3.3 | +0.18 |
| Average weekly wage ($) | 748 | 787 | 560 | −39 | −0.18 |
| Construction share (%) | 6.0 | 6.0 | 5.3 | +0.1 | +0.04 |
| Information share (%) | 2.4 | 2.0 | 1.3 | +0.4 | +0.27 |
| Prof. services share (%) | 5.1 | 4.7 | 2.3 | +0.4 | +0.13 |

*Notes:* Pre-treatment characteristics computed as county-level means over 2003–2007. "Synthetic" column reports the weighted average across synthetic control counties, averaged over all treated counties with pre-RMSE $\leq 0.15$. Industry shares are percent of total private employment.

## A.4 Sensitivity to fit and timing

**RMSE attrition.** The pre-treatment RMSE threshold of 0.15 log points excludes counties with poor pre-treatment fit. The baseline threshold retains 90 of 92 converged fits (98 percent), and tightening to 0.05 retains 88 with a nearly identical point estimate (4.2 percent), indicating that the results are not driven by poorly-matched counties.

**Leave-one-out.** To assess whether the hyperscale information sector result is driven by a single influential county, we re-estimate the SCM for information employment dropping each

county in turn from the broader set containing any hyperscale facility ($n = 39$, including hyperscale-only and mixed counties). No single county's removal changes the qualitative conclusion: the cross-county mean gap at $t = +6$ stays in the 50–80 percent range under every leave-one-out replication.

**Treatment timing sensitivity.** As a further check on the robustness of the SCM results to potential measurement error in facility opening dates (discussed in Section 3), we re-estimate the main SCM specification with treatment dates shifted $-1$ and $-2$ years (Table 19). If the results were an artifact of systematic dating error, shifting the treatment date should substantially alter the estimates. The $t = +6$ information sector gap is stable across timing shifts, confirming that the results are not sensitive to the precise dating of facility openings.

Table 19: Treatment Timing Sensitivity

| Outcome | Treatment timing | Gap at $t$=+6 (%) | SE | $p$ | $N$ |
|---|---|---|---|---|---|
| Total Employment | Actual † | +4.2%*** | (0.0101) | 0.000 | 90 |
| | -1 year | +4.2%*** | (0.0132) | 0.002 | 83 |
| | -2 year | +3.5%** | (0.0155) | 0.024 | 77 |
| Information Sector | Actual † | +22.4%** | (0.0913) | 0.027 | 80 |
| | -1 year | +29.1%*** | (0.0896) | 0.004 | 77 |
| | -2 year | +17.2% | (0.1267) | 0.210 | 72 |

*Notes:* Each row reports the average SCM gap at $t = +6$ under the indicated treatment timing. "Actual" uses the observed first data center opening year. "$-1$ year" and "$-2$ years" shift the treatment date earlier. † marks the baseline specification. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

**SCM sensitivity and in-time placebos.** We assess the stability of the $t = 6$ information sector estimate across three dimensions of the SCM procedure (Table 20). The estimate is robust to varying the pre-treatment RMSE threshold (0.05 to 0.20: 22–24 percent) and the pre-treatment matching window (3 to 7 years: 19–22 percent). Smaller donor pools (50 or 100 nearest counties) yield larger estimates (25–33 percent), likely because a narrow donor

pool is less representative; the estimate stabilizes at 22 percent with 200 or more donors. Across all specifications, the $t = 6$ information sector gap is significant at the 5 percent level or better.

As a final validation, we implement an in-time placebo test by shifting the treatment date three years earlier, capping the post-period before the true treatment date. For total private employment, the placebo shows no significant effect at any event time, supporting the identifying assumption. For the information sector, the placebo shows a mild positive drift reaching 5.7 percent by placebo $t = +2$ (corresponding to real $t = -1$), which could reflect anticipation effects from construction-phase announcements, consistent with the OSM date measurement lag discussed in Section 3. The magnitude is small relative to the 22 percent effect at real $t = +6$.

Table 20: SCM Sensitivity: Information Sector Employment at $t = +6$

| Parameter | Value | Gap at $t$=+6 | SE | $N$ |
|---|---|---|---|---|
| Donor pool size | 50 | +32.7%*** | (0.0850) | 74 |
| Donor pool size | 100 | +24.0%** | (0.0859) | 77 |
| Donor pool size | 200 † | +22.4%** | (0.0902) | 80 |
| Donor pool size | 500 | +16.3%* | (0.0869) | 85 |
| Pre-RMSE threshold | 0.05 | +23.7%** | (0.0923) | 78 |
| Pre-RMSE threshold | 0.1 | +22.4%** | (0.0902) | 80 |
| Pre-RMSE threshold | 0.15 † | +22.4%** | (0.0902) | 80 |
| Pre-RMSE threshold | 0.2 | +22.4%** | (0.0902) | 82 |
| Pre-treatment window | 3 years | +18.6%** | (0.0824) | 87 |
| Pre-treatment window | 5 years † | +22.4%** | (0.0902) | 80 |
| Pre-treatment window | 7 years | +21.8%** | (0.0833) | 79 |

*Notes:* Each row reports the average gap at event time $t = +6$ for information sector employment under a different SCM specification. SE is the cross-county standard deviation of gaps divided by $\sqrt{N}$. † marks the baseline specification. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

## A.5 Cross-estimator comparison and additional checks

**Cross-estimator comparison.** Table 21 summarizes the employment effects across all estimation approaches. SCM and Callaway–Sant'Anna converge at 4–5 percent for total employment and 15–22 percent for the information sector. TWFE without trend controls yields larger estimates (15 percent total), an upper bound inflated by pre-existing differential growth.

Table 21: Cross-Estimator Summary: Employment Effects at $t = +6$ (Percent)

| Method | Total | Construction | Information | Prof. Services |
|---|---|---|---|---|
| TWFE (baseline) | $14.6^{***}$ | $18.4^{***}$ | $41.0^{***}$ | $29.0^{***}$ |
| Synthetic control | $4.2^{***}$ | $10.9^{***}$ | $22.4^{**}$ | $16.8^{**}$ |
| Callaway–Sant'Anna | $5.2^{***}$ | 4.1 | $15.5^{***}$ | $11.6^{***}$ |
| Stacked DiD | $5.4^{***}$ | — | $18.7^{***}$ | — |

*Notes:* Each cell reports the implied percentage effect at $t = +6$ or the average post-treatment effect, depending on the estimator. TWFE reports the single post-treatment coefficient; SCM reports the average gap at event time +6; Callaway–Sant'Anna reports the ATT averaged over post-treatment periods (±5 year window); stacked DiD reports the post-treatment coefficient. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

**Reverse causality in multi-facility counties.** The concentration of information-sector agglomeration in multi-facility counties raises a reverse-causality concern: do growing information sectors attract additional data centers? The data offer little support for that view. Pre-second-opening information growth does not predict a shorter gap to the second opening; in the subset with four or more years between openings, information employment was decelerating before the second facility arrived. Single-facility and multi-facility counties also show similar information-sector trajectories immediately after first entry, more consistent with planned campus expansion than with follow-on siting responding to rising local IT employment.

**Population and migration.** A final question is whether the employment effects reflect in-migration of workers drawn by data center activity rather than job creation for existing residents. Using county-level population estimates from the Census Bureau (2003–2019), we estimate a first-differenced DiD on population growth rates. The data center treatment has no statistically significant effect on population growth (a point estimate of 0.002 percentage points per year), while the employment growth rate effect is large and statistically significant at the 1 percent level (0.64 percentage points per year). The employment effect is unchanged when population growth is included as a control. These results suggest that the employment gains are largely absorbed by existing residents and commuters rather than in-migrants, though annual Census population estimates may miss some short-distance moves.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Data Availability

The data and code used in this paper will be made available in a public repository upon publication. The primary data sources are the IM3 Open Source Data Center Atlas (Pacific Northwest National Laboratory), the Bureau of Labor Statistics Quarterly Census of Employment and Wages (QCEW), the Census Bureau's Quarterly Workforce Indicators (QWI) from the LEHD program, the FHFA All-Transactions House Price Index, and the OpenStreetMap historical edit API. Author-constructed facility opening dates and the harmonized county-year panel will be deposited at a public repository (Harvard Dataverse) at the time of publication.

## Declaration of Generative AI and AI-Assisted Technologies in the Manuscript Preparation Process

During the preparation of this work, the authors used several large-language-model tools to assist with code debugging, robustness checks, and editorial revision. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the article.

# References


Alberto Abadie. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425, 2021.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010.

Juan Alcácer and Wilbur Chung. Location strategies for agglomeration economies. *Strategic Management Journal*, 35(12):1749–1761, 2014.

Juan Alcácer and Mercedes Delgado. Spatial organization of firms and location choices through the value chain. *Management Science*, 62(11):3213–3234, 2016.

Alexander W. Bartik, Janet Currie, Michael Greenstone, and Christopher R. Knittel. The local economic and welfare consequences of hydraulic fracturing. *American Economic Journal: Applied Economics*, 11(4):105–155, 2019.

Timothy J. Bartik. *Making Sense of Incentives: Taming Business Incentives to Promote Prosperity*. W.E. Upjohn Institute for Employment Research, 2020.

Kirill Borusyak, Peter Hull, and Xavier Jaravel. Quasi-experimental shift-share research designs. *Review of Economic Studies*, 89(1):181–213, 2022.

Matias Busso, Jesse Gregory, and Patrick Kline. Assessing the incidence and efficiency of a prominent place based policy. *American Economic Review*, 103(2):897–947, 2013.

Brantly Callaway and Pedro H.C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

Cardinal News. Tax abatement for data centers is now $1.6 billion a year. `https://cardinalnews.org/2026/01/22/tax-abatement-for-data-centers-is-now-1-6-billion-a-year/`, 2026.

Matias D. Cattaneo, Yingjie Feng, and Rocío Titiunik. Prediction intervals for synthetic control methods. *Journal of the American Statistical Association*, 116(536):1865–1880, 2021.

Matias D. Cattaneo, Yingjie Feng, Filippo Palomba, and Rocío Titiunik. scpi: Uncertainty quantification for synthetic control methods. *Journal of Statistical Software*, 113(1):1–38, 2024.

Doruk Cengiz, Arindrajit Dube, Attila Lindner, and Ben Zipperer. The effect of minimum wages on low-wage jobs. *Quarterly Journal of Economics*, 134(3):1405–1454, 2019.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

David Dorn. *Essays on Inequality, Spatial Interaction, and the Demand for Skills.* PhD thesis, University of St. Gallen, 2009.

Tommy Pan Fang and Shane Greenstein. Where the cloud rests: The economic geography of data centers. *Strategy Science*, 10(4):404–420, 2025.

Chris Forman, Avi Goldfarb, and Shane Greenstein. The Internet and local wages: A puzzle. *American Economic Review*, 102(1):556–575, 2012.

Daniel Goetzel, Mark Muro, and Shriya Methkupally. Turning the data center boom into long-term local prosperity. Brookings Institution, February 2026. URL `https://www.brookings.edu/articles/turning-the-data-center-boom-into-long-term-local-prosperity/`.

Paul Goldsmith-Pinkham, Isaac Sorkin, and Henry Swift. Bartik instruments: What, when, why, and how. *American Economic Review*, 110(8):2586–2624, 2020.

Good Jobs First. Data center subsidies: A growing cost for communities. `https://goodjobsfirst.org/data-center-subsidies/`, 2024.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

Michael Greenstone, Richard Hornbeck, and Enrico Moretti. Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598, 2010.

Michael J. Hicks. Data centers and local job creation. *Center for Business and Economic Research, Ball State University*, 2024.

Jonas Hjort and Jonas Poulsen. The arrival of fast internet and employment in Africa. *American Economic Review*, 109(3):1032–1079, 2019.

Patrick Kline and Enrico Moretti. Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee Valley Authority. *Quarterly Journal of Economics*, 129(1):275–331, 2014.

Enrico Moretti. Local multipliers. *American Economic Review: Papers & Proceedings*, 100 (2):373–377, 2010.

Enrico Moretti and Per Thulin. Local multipliers and human capital in the United States and Sweden. *Industrial and Corporate Change*, 22(1):339–362, 2013.

David Neumark and Helen Simpson. Place-based policies. *Handbook of Regional and Urban Economics*, 5:1197–1287, 2015.

Ohio River Valley Institute. Why data centers will be economic development duds. `https://ohiorivervalleyinstitute.org/why-data-centers-will-be-economic-development-duds/`, 2025.

Vikram Pathania and Serguei Netessine. The impact of Amazon facilities on local economies. *Journal of Policy Analysis and Management*, 2026. doi: 10.1002/pam.70065.

Cailin Slattery and Owen Zidar. Evaluating state and local business incentives. *Journal of Economic Perspectives*, 34(2):90–118, 2020.

Juan Carlos Suárez Serrato and Owen Zidar. Who benefits from state corporate tax cuts? A local labor markets approach with heterogeneous firms. *American Economic Review*, 106(9):2582–2624, 2016.