Online Resource 1

Supplementary tables, figure and technical appendix

One decision, several rows: a modifiable unit effect in project databases, with evidence on hydrogen investment timing

Submitted to *The Annals of Regional Science*

Author details withheld for double-anonymous peer review

This file collects the material cited from the manuscript. Section 1 holds the construction diagnostics and the secondary result tables; Section 2 holds the definitions and derivations that the manuscript states without reproducing. Every quantity is produced by the replication code accompanying the submission, which reconstructs each table and figure from the raw workbooks when run in ascending numerical order.

1 Tables and figure

See Tables S1 to S9 and Fig. S1.

Table S1 Descriptive statistics, investment-unit risk panel, 2023 to 2026

Variable	Mean	Median	SD	*N*
Reaches FID by $t+1$	0.04	0.00	0.20	3211
Committed or operating units within 50 km	0.72	0.00	1.39	3211
Log committed or operating units, 50 km (primary regressor)	0.35	0.00	0.54	3211
Operating units within 50 km	0.36	0.00	0.77	3211
Committed, not operating, within 50 km	0.36	0.00	0.87	3211
Pre-FID units within 50 km	2.05	1.00	2.87	3211
Projects in the unit	1.18	1.00	0.48	3211
Announced capacity, as recorded	198368.73	23821.52	629756.62	3202
Log capacity, regressor with imputation	10.23	10.06	2.19	3211
Capacity imputed (indicator)	0.13	0.00	0.33	3211
Share electrolytic	0.88	1.00	0.32	3211
Share derivative	0.30	0.00	0.45	3211
Coastal	0.54	1.00	0.50	3211
Age of oldest announcement (years)	1.42	1.00	1.21	3211
Observations				3211
Investment units				1428
Countries				90
Events				128

Cited in Sect. 3.4 of the manuscript. The estimation sample comprises 3,211 unit-periods, 128 events and 90 countries. Capacities are in Nm^3/h. The regression uses the log capacity with median imputation and its indicator, not the recorded variable. Source: prepared by the author.

Table S2 Flow from the raw vintages to the estimation sample

Vintage	Raw rows	Admissible	Geolocated	Coverage	Units at risk	Observations	Events
2021	991	983	534	54%	224	224	11
2022	1478	1468	934	64%	479	479	30
2023	1998	1990	1452	73%	803	803	60
2024	2438	2430	1945	80%	1071	1071	38
2025	2625	2618	2330	89%	1337	1337	30
2026	2942	2934	2426	83%	–	–	–

Cited in Sect. 3.1. Admissible rows exclude aggregate placeholders and header artefacts. June 2026 supplies the outcome of the 2025 transition and is not itself followed by a further vintage. Source: prepared by the author.

Table S3 Matching log across the six vintages

Matching rule	2021	2022	2023	2024	2025	2026	Total
Seed vintage	983	0	0	0	0	0	983
IEA reference and country	0	966	1363	1883	2223	2759	9194
IEA reference, country corrected	0	1	13	2	6	3	25
Reference reassigned, rejected	0	0	1	0	0	0	1
Normalised name and country	0	0	0	1	1	2	4
Normalised name only	0	0	1	0	0	0	1
Corroborated fuzzy name	0	0	2	4	1	4	11
New project	0	501	611	540	387	166	2205
Rows in vintage	983	1468	1990	2430	2618	2934	12423

Cited in Sect. 3.1. Each row of each vintage is assigned exactly one rule, in the hierarchy set out in Sect. 2.1 below. Rejected reference reassignments are logged separately and are not counted as assignments. Source: prepared by the author.

Table S4 Measurement diagnostics: coordinate stability and unit geometry

Stability of coordinates, September 2025 to June 2026	
Projects geolocated in both vintages	2254
Coordinates identical to the metre	2224 (98.7%)
Displacement greater than 0.5 km	30
Displacement greater than 2 km	30
Displacement greater than 10 km	30
Displacement greater than 50 km	28
Displacement greater than 100 km	26
Geometry of the strict investment unit	
Units containing more than one project	311 of 2085
Largest unit (projects)	6
Mean size of multi-project units	2.24
Median diameter (km)	0.000
Largest diameter (km)	0.374
Units whose diameter exceeds 0.5 km	0

Cited in Sect. 3.2. Chaining through connected components could in principle produce a investment unit wider than the 0.5 km threshold; none occurs. Source: prepared by the author.

Table S5 Excess spatial concentration of cancellations, by unit of analysis

Unit of analysis	Units	Moran I	z	p
Project	1684	+0.0372	5.00	0.0005
Strict investment unit	1414	-0.0069	1.25	0.1055
Identical coordinates	1331	-0.0133	0.78	0.2200
Geographic, 0.5 km	1301	-0.0133	0.83	0.1920
Geographic, 2 km	1241	-0.0211	0.27	0.3715
Geographic, 10 km	1089	-0.0232	0.28	0.3680

Cited in Sect. 4.1, and the numerical counterpart of Fig. 1 of the manuscript. Moran statistics of the residual of a linear probability model of cancellation on project covariates and country fixed effects, on a ten-nearest-neighbour graph, for the 1,684 projects that were pre-FID and geolocated in September 2025. Inference is by within-country randomization with event counts preserved over 1,999 draws. Each row applies a different rule for individuating investment units; the criterion defining what is counted is held fixed. Source: prepared by the author.

Table S6 The family of twelve exposure estimates

Exposure measure	Coef.	(s.e.)	1 s.d. (p.p.)	p	Holm	R–W
Radius 50 km						
Operating units	-0.0003	(0.0111)	-0.01	0.978	1.000	1.000
Committed, not operating	-0.0026	(0.0113)	-0.11	0.816	1.000	1.000
Committed or operating	-0.0008	(0.0087)	-0.04	0.927	1.000	1.000
Operating capacity	-0.0000	(0.0013)	-0.01	0.983	1.000	1.000
Committed or operating capacity	0.0007	(0.0011)	+0.28	0.536	1.000	0.989
Share committed or operating	0.0036	(0.0164)	+0.10	0.828	1.000	1.000
Radius 100 km						
Operating units	-0.0102	(0.0108)	-0.61	0.348	1.000	0.833
Committed, not operating	-0.0198	(0.0107)	-1.12	0.065	0.774	0.717
Committed or operating	-0.0163	(0.0091)	-1.23	0.074	0.812	0.608
Operating capacity	-0.0013	(0.0012)	-0.51	0.280	1.000	0.738
Committed or operating capacity	-0.0013	(0.0010)	-0.60	0.230	1.000	0.885
Share committed or operating	0.0099	(0.0200)	+0.25	0.620	1.000	0.989
Joint test of the twelve coefficients						$p = 0.608$
Observations						3211
Events						128

Cited in Sect. 4.2. Twelve separate estimates of Eq. (3) of the manuscript, one per exposure measure, each on the same 3,211 observations with the full control vector, vintage fixed effects and country fixed effects. Coefficients, standard errors and raw p values use the spatio-temporal Conley estimator of Eq. S1 below, to which Holm adjustment applies; the Romano-Wolf column and the joint test come from a spatial-block bootstrap of 4,999 replications over 140 k-means blocks. Source: prepared by the author.

Table S7 Robustness of the primary estimand: estimation window, variance estimator, rare-event estimator and selection

Specification	Obs.	Events	Coefficient (s.e.)
Estimation window			
2023 to 2026 (main)	3211	128	-0.0008 (0.0087)
2021 to 2026 (extension)	3914	169	-0.0042 (0.0083)
Variance estimator			
Conley spatio-temporal (200 km, one lag)	3211	128	-0.0008 (0.0087)
Clustered on the investment unit	3211	128	-0.0008 (0.0084)
Two-way: unit and country-vintage	3211	128	-0.0008 (0.0087)
Homoskedastic	3211	128	-0.0008 (0.0085)
Rare-event estimator (log odds)			
Firth penalized logit	3211	128	0.0082 (0.2067)
Selection			
Unweighted	3211	128	-0.0008 (0.0087)
Inverse probability of observation	3211	128	-0.0010 (0.0088)

Cited in Sect. 4.2. These are the blocks of the robustness table that do not appear in Table 2 of the manuscript, which retains the decision unit and fixed-effects blocks. Exposure is the log count of committed or operating units within 50 km; every row includes the full control vector, and rows outside the fixed-effects block use vintage and country fixed effects. The Firth coefficient is a log odds and is not comparable in level with the linear probability coefficients. Source: prepared by the author.

Table S8 Heterogeneity tested by interaction

Dimension	In group / outside	Difference in the 1 s.d. effect (p.p.)	p	Holm
Electrolytic	2836 / 375	-1.89 [-4.22, 0.45]	0.113	0.681
Derivative product	979 / 2232	+0.50 [-1.06, 2.06]	0.531	1.000
Coastal	1743 / 1468	-0.63 [-2.10, 0.84]	0.403	1.000
OECD	2327 / 884	-1.15 [-3.91, 1.60]	0.412	1.000
Many pre-FID neighbours	1250 / 1961	+0.83 [-0.87, 2.53]	0.339	1.000
Above-median capacity	1602 / 1609	+0.42 [-0.98, 1.83]	0.555	1.000
Joint test of the six interactions		Wald = 5.44, 6 d.f., $p = 0.489$		

Cited in Sect. 4.2. Each row is one estimate of Eq. (3) of the manuscript on the full sample, augmented by the group indicator and its interaction with exposure. The reported quantity is the interaction coefficient per standard deviation of exposure. Standard errors are the spatio-temporal Conley estimator and Holm adjustment is applied across the six tests. Source: prepared by the author.

Table S9 Status revisions between the September 2025 and June 2026 vintages

Status revision, September 2025 to June 2026	Projects	Share (%)
Reached FID, construction or operation	32	1.90
Still concept or feasibility study	1325	78.68
Recorded as dormant or on hold	191	11.34
Recorded as cancelled	105	6.24
Not matched in the June 2026 vintage	31	1.84
Total	1684	100.00

Cited in Sect. 4.2. The cohort is the 1,684 projects that were pre-FID and geolocated in September 2025. Adverse revision outweighs favourable revision by more than nine to one. Source: prepared by the author.

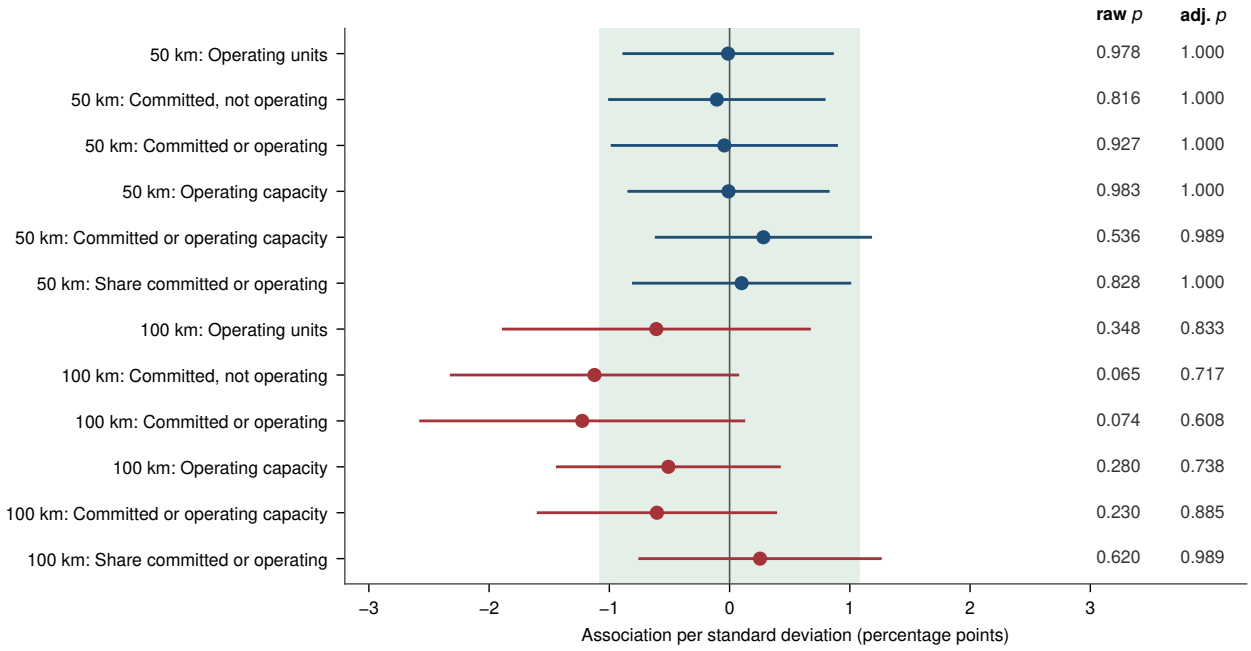


Fig. S1 The twelve exposure estimates, with raw and familywise-adjusted p values. Points are associations per standard deviation of exposure in percentage points and horizontal lines are 95% intervals; the upper six rows use a 50 km radius and the lower six a 100 km radius. The column headed raw p reports Conley spatio-temporal p values and the column headed adj. p the Romano-Wolf stepdown values from the spatial-block bootstrap. The shaded band is the 1.08-point equivalence benchmark

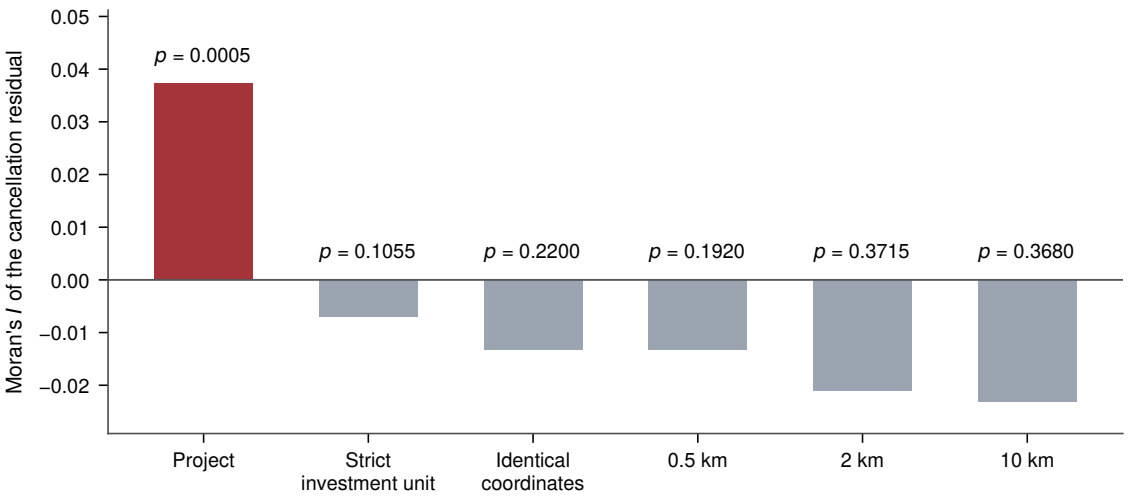


Fig. S2 Excess spatial concentration of cancellations, by unit of analysis. Bars give Moran's I of the residual of a linear probability model of cancellation on project covariates and country fixed effects, on a ten-nearest-neighbour graph, for the 1,684 projects that were pre-decision and geolocated in September 2025. The p value above each bar comes from within-country randomization with event counts preserved over 1,999 draws. The leftmost bar is the project level, the remaining five alternative rules for individuating investment units

2 Technical appendix

2.1 Matching hierarchy and corroboration rule

Vintages are linked on the project reference and the country first; then on the reference alone where the country changes and the normalized names still agree at 70 or above; then on normalized name and country; then on normalized name alone. Failing those, a fuzzy name match within country at 85 or above is accepted when the name similarity reaches 93, or when a corroboration score of at least two is reached. Coordinates closer than 10 km or capacities within a quarter of each other each score two; coordinates within 50 km or capacities within a factor of three each score one; a shared technology scores one; and coordinates more than 50 km apart score minus two. Failing all of these a new project is recorded. Corroborated fuzzy matching assigns eleven rows across the six vintages, each logged with its score. One reference reassignment, from an Indian memorandum to a Japanese pilot, was detected, rejected and recorded.

2.2 Control vector

X_{it} in Eq. (3) of the manuscript contains the log unit capacity and its missingness indicator, the electrolytic and derivative shares of the unit, a coastal indicator, an irradiation proxy, the log distance in kilometres to the nearest other investment unit in the same vintage, the log number of projects in the unit, and the age of the oldest announcement. All controls enter every specification, except that the electrolytic, derivative and coastal interaction regressions drop the control defining their group indicator. The full coefficient vectors are in the replication package.

2.3 Spatio-temporal variance estimator

Let Z_{it} denote the full regression design vector, including exposure, controls and fixed-effect indicators, and $\hat{u}_{it} = Z_{it}\hat{\varepsilon}_{it}$ the regression scores. The default variance estimator is the spatio-temporal

analogue of Conley (1999),

$$\widehat{V} = (Z'Z)^{-1}\Big[\sum_i \sum_j \sum_t \sum_u \omega(d_{ij})\,\tau(|t-u|)\,\hat{u}_{it}\hat{u}'_{ju}\Big](Z'Z)^{-1}, \tag{S1}$$

with $\omega(d) = \max\{0, 1 - d/\bar{d}\}$ at $\bar{d} = 200$ km and $\tau(\ell) = \max\{0, 1 - \ell/(L+1)\}$ at $L = 1$ vintage.

2.4 Equivalence benchmarks and tests

The first margin is calibrated on the committed-to-achievable 2030 capacity gap. The International Energy Agency puts announced 2030 capacity at 27 Mt and committed capacity at 4.3 Mt, with more than 6 Mt achievable if the projects with strong potential reach a final investment decision in 2026 or 2027, which leaves a pre-FID announced stock of 22.7 Mt and a gap of 1.7 Mt. With a baseline hazard $\bar{h}$, T remaining transitions, a pre-FID stock M and a gap Δ to be halved, the required increment δ solves

$$\big[1 - (1 - \bar{h} - \delta)^T\big] - \big[1 - (1 - \bar{h})^T\big] = \frac{\Delta/2}{M}. \tag{S2}$$

At $\bar{h} = 3.99\%$ and $T = 4$, 15.0% of the stock would reach a final investment decision unaided and reaching 18.8% requires $\delta = 1.08$ percentage points. The calibration assumes that capacity reaching a decision is proportional to units reaching one, and that the increment applies uniformly to the at-risk stock; neither is established. The second margin is set conventionally at a quarter of the baseline hazard, 1.00 percentage points.

Writing $\hat{\theta} = 100\,\hat{\beta}\,\mathrm{sd}(E)$ for the association per standard deviation and $\hat{\sigma}$ for its standard error, non-equivalence at a benchmark δ is rejected at level α when

$$\max\Big\{1 - \Phi\big((\hat{\theta} + \delta)/\hat{\sigma}\big),\; 1 - \Phi\big((\delta - \hat{\theta})/\hat{\sigma}\big)\Big\} < \alpha. \tag{S3}$$

2.5 Selection calibration

Weighting by the inverse probability of appearing in the panel, from a model of observation on status, capacity, technology, country and vintage with a pseudo-R^2 of 0.21, moves the primary coefficient from -0.0008 to -0.0010; estimated probabilities range from 0.31 to 0.99 with a median of 0.92, the largest weight is 3.2 and the effective sample size is 3,167 of 3,211. Weighting cannot address selection on unobservables. As a stylized calibration, and not a formal bound, suppose the standardized conditional association combines linearly across observed and unobserved units,

$$\theta_{\mathrm{pop}} = \pi\,\theta_{\mathrm{obs}} + (1 - \pi)\,\theta_u, \tag{S4}$$

with π the coverage and θ_u the association among unobserved units, and that residualized exposure scales comparably in both groups. Neither supposition follows from the estimator, since a conditional coefficient with controls and fixed effects is not in general a convex combination of subpopulation coefficients. Under them, and requiring the one-sided 95% upper end to stay inside the benchmark, equivalence survives as long as $|\theta_u|$ stays below 1.75 percentage points per standard deviation at the least favourable coverage in the window, 73%, and below 2.11 at the average coverage of the three starting vintages, 81%, that is 44% and 53% of the baseline hazard. The exercise indicates the order of magnitude unobserved selection would need to reach; it does not bound it.

2.6 Why no endogenous spatial lag is estimated

Lacombe and LeSage (2018) show that in a spatial autoregressive discrete-choice model the coefficient on a regressor is not the quantity of interest, and that the effect must be reported as a decomposition into direct, indirect and total impacts. That decomposition requires a credibly identified spatial parameter. Instrumenting the lag on this panel, with the exogenous characteristics of second-order neighbours as instruments, yields a partial first-stage F statistic between 3.5 and 5.3 across the twelve exposure measures, far below any conventional threshold, so the resulting impacts would not be interpretable. The exposure measures used in the manuscript are accessibility variables on the right-hand side, the alternative Andersson and Gråsjö (2009) propose for representing spatial dependence when a lag cannot be identified.

2.7 Validation of the investment-unit rule

The concordance figures reported in Sect. 3.2 of the manuscript are computed by `code/57_validation_unites.py` in the replication package. All pairs of geolocated projects within 0.5 km of each other are enumerated on the great-circle distance, on the same perimeter as the unit construction of Sect. 3.2, and split according to whether the strict rule assigns them to the same investment unit. Status concordance is the share of vintages in which both are observed and carry the same status; revision concordance is the share of consecutive intervals in which both change status or neither does. Retained pairs return 47.6% and 85.8%, pairs kept apart 37.2% and 77.6%.

Two cases from the data show that neither proximity nor naming suffices alone. At Eemshaven in the Netherlands, three rows share coordinates to six decimals under two different project names, one referring to the wind farm supplying the electrolyser and the other to the energy park in two phases, and carry an identical status in all six vintages: a name-based rule would count three projects, and a nearest-neighbour graph three mutual neighbours. Two rows in Abu Dhabi share coordinates to six decimals because both carry a city-level geocode, one a green hydrogen plant under construction throughout and the other a blue ammonia project in feasibility study; proximity alone would merge two unrelated decisions. Both cases are reproduced with their coordinates and their six status readings by `code/57_validation_unites.py`.

Revision concordance so defined is dominated by inaction. Of the 1,283 intervals observed on retained pairs, 1,000, or 77.9%, are intervals in which neither row changes; the corresponding figure for the 1,389 intervals on pairs kept apart is 1,039, or 74.8%. The measure is therefore an indicator of coherence rather than of unicity, and the manuscript reports it as such. The test that conditions on movement is the informative one. Restricting attention to intervals in which at least one row changes status leaves 283 intervals on retained pairs, of which 101, or 35.7%, have both rows changing, against 39 of 350, or 11.1%, on pairs kept apart. The ratio is 3.2 and the gap 24.6 percentage points. Neither figure establishes that a retained pair is one decision; together they establish that the rule selects pairs whose statuses move together far more often than proximity alone would produce, which is the claim the manuscript makes.

The enumeration returns the 1,051 co-located pairs of Sect. 3.2. Of these, 486 are retained by the strict rule and 565 kept apart.