When the toolkit runs out: single-host event studies and the limits of causal inference

Hyunhak Kim* Yun Seol†

Abstract

Tourism event evaluations increasingly observe a single host city at monthly frequency. Such designs yield one treated unit, and cluster-robust inference is asymptotic in the number of *treated* clusters, so enlarging the comparison group improves the counterfactual but not the standard error. The recent difference-in-differences literature does not help: it corrects staggered adoption, which single-host designs do not have. Taking the 2025 APEC summit in Gyeongju, Korea, we hold the data fixed and vary design choices one at a time. Correcting only the variance estimator moves the headline p-value from 0.000 to 0.200, leaving the point estimate untouched. Enlarging the donor pool and repairing a convex-hull failure moves the synthetic control estimate from $+313\%$ to $+7.5\%$. No estimate survives a post-treatment trend violation half the size of the one already visible beforehand, and a power calculation shows that further months would not change this. Across sixteen comparable studies we find pre-trend sensitivity analysis reported once. This design cannot establish an effect, which is not to say none exists.

Keywords: mega-event; synthetic control; difference-in-differences; placebo inference; Honest DiD; tourism legacy

*Kookmin University. Corresponding author.
†Kyungpook National University.

1 Introduction

Evaluating what a mega-event does for tourism used to mean comparing a projection with an outturn. It now means estimating a counterfactual, and the data available for doing so have become steadily finer. Mobile-network location records, card settlements and navigation queries are published monthly at municipal level in a growing number of countries, and the natural response has been to move the analysis down to the scale at which the event actually happened: one city, one event, high-frequency outcomes.

That move has a structural consequence which the literature has not confronted. As the unit of analysis shrinks, the treated group shrinks with it, until it contains exactly one unit. The pool of potential comparisons grows at the same time, which makes the counterfactual look better and the inference problem worse. Cluster-robust standard errors are consistent in the number of clusters, but the asymptotics that matter for a treatment assigned to a single cluster run in the number of *treated* clusters (Conley and Taber, 2011; MacKinnon and Webb, 2018). Adding comparison municipalities improves the counterfactual and does nothing whatever for the standard error.

The recent difference-in-differences literature does not help here, and it does not claim to. Callaway and Sant'Anna (2021), Sun and Abraham (2021) and the estimators surveyed by Roth et al. (2023) exist to remove the negative weighting that arises when many units adopt treatment at different times. With one unit adopting at one time there is no staggered adoption, no forbidden comparison, and no negative weight; these estimators reduce to two-way fixed effects or do not apply. The field's most visible recent advance solves a problem that single-host studies do not have, while leaving untouched the one they do.

To our knowledge, the difficulty has been stated explicitly only once in this literature. In a study of a single museum opening, Wieland (2026) reports that a placebo test "could not be conducted here, as there is only one control unit." The observation is correct and consequential, but it is made in passing and not pursued.

This article pursues it. We evaluate the tourism effect of the 2025 APEC summit in Gyeongju, Korea, a case with a single host city, a single event and municipal monthly data. The question is not what the effect was, but how far the estimated answer moves as design

choices vary while the data are held fixed. The exercise is organised as two specification ladders. The first varies only how uncertainty is quantified; the second varies only the comparison group and the functional form of the counterfactual. Each rung changes exactly one element of the design.

Two patterns emerge. Along the inference ladder the point estimates are invariant: non-local domestic visitors are estimated at $+9.1\%$ whether the standard error is heteroskedasticity-robust, cluster-robust, bootstrapped or permuted, while the associated p-value moves from 0.000 to 0.200. The movement therefore reflects the variance estimator rather than the data. Along the second ladder the announcement-regime effect falls from $+313\%$ to $+7.5\%$, driven by two separate defects: a donor pool too small to place the treated unit, and a functional form that cannot represent it, because Gyeongju's admissions exceed those of every donor and simplex weights cannot reach its level. Either defect alone is sufficient to overturn the original estimate.

Three further diagnostics point the same way. Under the sensitivity framework of Rambachan and Roth (2023), no estimate survives a post-treatment trend violation as large as the one already visible before treatment. Moving the treatment date to dates on which nothing happened produces effects *larger* than the real one, with comparable pre-treatment fit. Refitting the estimator for every municipality at every feasible date — following the specification-search logic of Ferman et al. (2020) — places the actual estimate near the middle of the resulting joint distribution. Finally, a power calculation shows that extending the panel would not have changed this, because the noise is persistent unit-level drift rather than month-to-month variation, and averaging more months does not average away a drift.

The interpretation requires care. The point estimates are positive in almost every specification and are robust to dropping any single comparison unit. We do not conclude that the summit had no effect on tourism in Gyeongju, only that this design cannot establish one, and that the same holds for the design as it is commonly implemented.

To place the case in context we coded sixteen empirical studies of event effects on tourism and regional outcomes. Multi-unit designs dominate — fourteen of sixteen — and where there are many treated units the inference is largely sound, with several studies reporting null or

heterogeneous results rather than suppressing them. A search for a specific arithmetic error, placebo p-values below the floor the donor pool permits, returned no instances. The gap that does emerge is common to designs of every size: eleven of sixteen studies test for pre-trends, and one reports any formal analysis of what follows if the test fails.

The paper therefore makes three contributions. First, it quantifies how much of a single-host event study's conclusion is carried by inference and comparison-group choices rather than by data, using a case where the original estimates are available for direct comparison. Second, it demonstrates pre-trend sensitivity analysis in a setting where the literature does not yet report it, and shows that a conservative approximation and the official procedure agree on the verdict while the official one puts the breakdown value five to thirteen times higher. Third, it shows that the estimated "legacy" ranges from $+1.5\%$ to $+52.8\%$ across five candidate outcome variables measured on the same cities over the same months, which makes the choice of legacy metric a modelling decision rather than a reporting one.

The remainder of the article is structured as follows. Section 2 reviews the related literature and characterises current practice. Section 3 describes the case, Section 4 the data and Section 5 the estimators and inference procedures. Section 6 reports the specification ladders and the accompanying diagnostics, Section 7 discusses the implications for evaluation practice, and Section 8 concludes and acknowledges limitations.

2 Literature Survey

The economic case for hosting a mega-event has long rested on projections, and the gap between projection and realisation is the oldest finding in this literature. Preuss (2007) formalised the legacy concept partly in response to it, distinguishing tangible legacies — venues, transport, accommodation stock — from intangible ones such as image, networks and know-how, and arguing that the second category is both where the durable value is claimed to lie and where measurement is hardest. Getz (2008) placed event tourism within a research agenda that has since moved steadily from input–output accounting toward quasi-experimental estimation.

The substantive findings of that literature are mixed in a specific way. Effects during the

event are consistently detected and consistently modest relative to projections. Effects afterwards are contested. Some studies find durable increases in arrivals attributable to improved destination image; others find that the event-period increase is partly displacement, with regular visitors avoiding a host city during congestion and higher prices, so that gross arrivals overstate the net gain. The direction of the post-event effect appears to depend on event type, host characteristics and how long after the event one looks. This matters for our purposes less as a substantive dispute than as an indication of how much rests on design: a literature whose central quantity changes sign across studies is one in which the estimator's properties deserve scrutiny.

A second regularity is that anticipation begins at selection rather than at the event. Hosts commit to infrastructure spending, marketing and venue construction in the years between winning a bid and staging the event, and international media attention starts at the announcement. Studies that date treatment from the event itself therefore assign several years of genuinely post-treatment data to the control period. This is now recognised: Boto-García and Santana-Gallego (2026) address it directly by treating the announcement year as the treatment date. We adopt the same logic and, because our case gives us both dates cleanly, estimate the two regimes separately rather than pooling them.

The dominant empirical design in this transition has been the gravity model with a hosting indicator. Fourie and Santana-Gallego (2011) estimated an average tourist arrival increase of 8.1% across six mega-sport events, using bilateral flows between roughly two hundred countries; Vierhaus (2019) applied the same framework to a longer panel. A decade later Fourie and Santana-Gallego (2022) revisited their own design, switching from OLS to Poisson pseudo-maximum-likelihood estimation and stating explicitly that the earlier fixed-effects specification was likely biased. That self-correction is worth noting: this is a literature that does revise itself.

More recently the field has adopted the modern difference-in-differences toolkit. Boto-García and Santana-Gallego (2026) apply stacked difference-in-differences and Wooldridge's extended two-way fixed effects estimator to 147 countries over 1991–2023, addressing the negative-weighting problem that arises when treatment timing is staggered (Goodman-Bacon, 2021; de Chaisemartin and D'Haultfœuille, 2020). They also date treatment from the announce-

ment year rather than the event year, in order to capture anticipation — a distinction we adopt at municipal scale.

To characterise practice rather than assert it, we coded sixteen studies that estimate a causal effect on a tourism or regional outcome using difference-in-differences, event-study or synthetic-control designs. For each we recorded the number of treated units, the comparison group, the variance estimator, whether a pre-trend test was reported, whether any formal sensitivity analysis for parallel-trends violations was reported, and whether a placebo test was conducted. Table 1 presents the coding.

The selection procedure was as follows. Candidates were assembled from the reference lists of the two most recent methodological treatments of event evaluation in tourism, from forward citations to the canonical mega-event studies, and from journal searches combining event terms with design terms; studies were retained if they were available in full text, quantified a dated intervention on a bounded set of places, and reported enough detail to code the inference procedure. That is a purposive audit of visible practice, not a systematic review: there is no registered protocol, no exhaustive database sweep, and no screening of grey literature. Sixteen studies read in full text can indicate what a design of this kind typically does and does not report, and that is the use made of them here. They do not support inference about the field's distribution, and none is drawn.

The sample is deliberately broader than mega-events alone. Alongside the Olympic, World Cup and European Capital of Culture studies it includes evaluations of a concert tour (Boto-García et al., 2025), of television filming (Contu and Pau, 2022), of an attraction-designation policy (Zhang and Zhang, 2023) and of two national events assessed with a panel-data counterfactual (Wan and Song, 2019). These share the design problem we are interested in — a discrete, dated intervention on a small number of places — and excluding them would have made the sample look more homogeneous than the practice is.

Three regularities emerge from the coding.

Multi-unit designs dominate. Fourteen of sixteen studies have more than one treated unit, often many more: twenty-five host countries in the gravity studies, fifteen host regions in Firgo (2021), thirty-four host cities in Falk and Hagsten (2017). This reflects the structure of the

events themselves: the Olympics and the World Cup have been held repeatedly, so a panel of hosts is available by construction. Designs structurally similar to the present case are therefore the exception rather than the rule.

Where there are many treated units, inference is largely sound. Firgo (2021) uses a wild cluster bootstrap with the null imposed and 9,999 replications, and validates the design by treating runner-up bidding regions as placebo hosts. Gomes and Librero-Cano (2018) uses runner-up cities as the comparison group, a genuinely strong identification choice. Hu and Wang (2025) reports that only nine of sixteen treated countries show significant declines, stating the heterogeneity plainly rather than averaging it away. Kobierecki and Pierzgalski (2022) reports bootstrapped p-values between 0.505 and 0.864 for three of its four cases and concludes that the effects are insignificant. Wallimann and Mehr (2026), applying synthetic difference-in-differences to eight host cities, concludes that the findings "do not support strong claims of large tourism impacts." Systematic suppression of null results is therefore not apparent in this literature.

Formal sensitivity to pre-trend violations is essentially absent. Eleven of the sixteen studies report a pre-trend test. Exactly one reports anything resembling a formal sensitivity analysis for what happens if that assumption fails, and even there the treatment is a dynamic lead–lag specification rather than the bounds of Rambachan and Roth (2023). This is a gap of diffusion rather than of care — the method dates from 2023 — but it is a gap, it is common to designs of every size, and it is fixable now.

We also looked for a specific arithmetic error: a reported placebo p-value smaller than the floor that the donor pool permits, which is $1/(J+1)$ for J donors, since the treated assignment is a member of its own reference set. We found none among the five studies where the floor is computable. Lahura and Sabrera (2023) is the instructive case. It derives a floor formula explicitly and reports $p = 0.048$ with twenty-one donors, disclosing rather than obscuring that its estimate sits at the smallest attainable value. Its floor is stated as $1/J$ rather than the $1/(J+1)$ we use — a difference of three thousandths at that pool size, and of no consequence to its conclusion. The relevant point is that the floor was computed and reported at all.

The two single-treated-unit studies in our sample are the informative cases, and neither

is a mega-event study. Lahura and Sabrera (2023) evaluates an infrastructure investment at one Peruvian archaeological site against twenty-one donors, and supplements the permutation test with the conformal and t-test procedures of Chernozhukov et al. (2021). Wieland (2026) evaluates a single museum opening in one German town, and states the problem directly: a placebo test "could not be conducted here, as there is only one control unit."

That statement is, to our knowledge, the only explicit acknowledgement of the difficulty in this literature. The reason it arises is structural. As the unit of analysis moves from country to region to municipality to individual site, the treated group shrinks toward one, while the pool of potential comparisons grows. The modern difference-in-differences literature has solved the problems that arise when many units adopt treatment at different times (Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; Roth et al., 2023). It has not solved, and cannot solve, the problem that arises when exactly one unit adopts treatment at one time, because the relevant asymptotics run in the number of treated clusters (Conley and Taber, 2011; MacKinnon and Webb, 2018). Every estimator in that toolkit either reduces to two-way fixed effects in this setting or is simply inapplicable.

The pressure toward this design is not perverse. Country-level estimates answer a question that is often not the one being asked. A national tourism authority deciding whether to bid, or a provincial government justifying expenditure, wants to know what happened to the host city, and averaging over a country conflates the host with regions that may have lost visitors to it. Municipal data now make the sharper question answerable descriptively, at monthly frequency, with outcomes — card spending, navigation queries — that country statistics do not carry. The move down in scale buys real precision about what occurred. What it costs is the treated variation on which inference depends, and that cost is not usually stated.

This is not an argument that small-sample designs are uninformative. Abadie et al. (2010) developed the synthetic control precisely for comparative case studies with one treated unit, and supplied permutation inference suited to it. The difficulty arises when such a design is estimated with tools built for the many-unit case — cluster-robust standard errors, conventional significance thresholds — and reported as though those tools applied. The estimator and the inference procedure have to be chosen together.

The econometric literature supplies more than the field currently uses. Cameron et al. (2008) for few clusters; Abadie et al. (2010) and Abadie (2021) for the synthetic control and its placebo inference; Doudchenko and Imbens (2016) and Ferman and Pinto (2021) for the case where the treated unit sits outside the donors' convex hull; Arkhangelsky et al. (2021) for an estimator that needs neither a level match nor unconditional parallel trends; Rambachan and Roth (2023) for sensitivity to pre-trend violations; Chernozhukov et al. (2021) for inference that does not borrow strength across units; Ferman et al. (2020) for what specification searching does to synthetic control results.

Tourism research has begun to organise this. Tokarchuk and Gabriele (2026) adapt the placebo taxonomy of Eggers et al. (2024) — treatment, outcome and population placebos — into a six-step checklist for tourism applications, and document that tourism studies typically implement only one placebo type. Their illustration is a case in which the placebo tests support the original estimate. Ours is the complementary case.

This article addresses that gap. We take a single-host, municipal-scale event study constructed according to standard practice and subject it to the full set of diagnostics described above. The contribution is not a new estimator; every method employed is established. It is, first, a quantification of how far the conclusions of such a study move as design choices vary while the data are held fixed; second, an application of pre-trend sensitivity analysis in a setting where Table 1 indicates it is not yet reported; and third, evidence that the choice of outcome variable materially determines the estimated magnitude of the tourism legacy.

3 Study context

Gyeongju was the capital of the Silla kingdom for most of a millennium and is now a mid-sized city in North Gyeongsang province, on Korea's south-eastern coast. Its historic core, temple complexes and royal tombs make it the country's principal domestic heritage destination: over our pre-treatment window it draws an average of 711,000 monthly admissions at surveyed attractions, the highest of the 70 municipalities observed throughout that window in our four-province panel. That ranking matters later, because a treated unit at the top of its comparison

Table 1: Design practice in mega-event and event tourism evaluations

Study	Method	Treated	Standard errors	PT	Sens.	Plac.
Fourie & Santana-Gallego (2011)	gravity + DID	25	clustered (pairs)	N	N	N
Falk & Hagsten (2017)	DID + PSM	34	clustered (city)	N	N	N
Gomes & Librero-Cano (2018)	DID	multiple	clustered (city)	Y	N	N
Vierhaus (2018/2019)	gravity + DID	25	not reported	N	N	N
Wan & Song (2019)	panel data approach	2	analytical, AR(p)	N	N	N
Firgo (2021)	dynamic DID	15	wild cluster	Y	Y	Y
Fourie & Santana-Gallego (2022)	gravity (PPML)	multiple	robust	N	N	N
Contu & Pau (2022)	event study (Sun &	9	clustered (county)	Y	N	Y
Kobierecki & Pierzgalski (2022)	generalised SCM	4	bootstrap	Y	N	Y
Lahura & Sabrera (2023)	SCM	1	permutation + t	Y	N	Y
Zhang & Zhang (2023)	staggered DID	multiple	not reported	Y	N	Y
Hu & Wang (2025)	SCM + fsQCA	16	placebo + t	Y	N	Y
Boto-Garcia et al. (2025)	dynamic DID	20	clustered (market)	Y	N	Y
Wieland (2026)	DID + SCM	1	not reported	Y	N	Y
Boto-Garcia & Santana-Gallego (2026)	stacked DID, ETWFE	multiple	clustered (country)	Y	N	N
Wallimann & Mehr (2026)	Synthetic DID	8	placebo inference	Y	N	Y

Coding of 16 empirical studies that estimate a causal effect on a tourism or regional-economic outcome using difference-in-differences, event-study or synthetic-control designs. "Treated" is the number of treated units. PT = a pre-trend test is reported; Sens. = a formal sensitivity analysis for violations of parallel trends is reported (e.g. Rambachan and Roth, 2023); Plac. = any placebo test is reported.

Multi-unit designs dominate (14 of 16); only 2 study uses a single treated unit. 11 of 16 report a pre-trend test, but only 1 report a formal sensitivity analysis for its violation. 9 of 16 report a placebo test.

The full coding sheet, including per-study notes, will be deposited with the replication materials on acceptance.

distribution is precisely the case in which a convex combination of donors cannot reproduce it.

In June 2024 Gyeongju was selected to host the 2025 APEC Economic Leaders' Meeting, beating two larger competitors. Leaders' Week followed in late October 2025. The sixteen months between the two dates were occupied by the visible preparation that mega-events generate — venue construction, transport upgrades, national and international press — and the expectation of a tourism legacy was explicit in the policy discussion throughout. We therefore treat confirmation and hosting as two distinct interventions rather than one, and Figure 1 shows how they sit against the coverage of our two panels.

The comparison cities are chosen to span the profile Gyeongju competes with rather than to match it on size. Andong and Pohang are the other two major destinations in the same province, the former a Confucian heritage town and the latter an industrial port city with a coastal tourism sector; they capture any effect that diffuses within the province. Jeonju and Gangneung lie outside it. Jeonju is a heritage and culinary destination, Gangneung a coastal and festival one, and between them they represent the two national alternatives a domestic visitor weighs against a trip to Gyeongju.

The five differ substantially in scale. Over the pre-treatment window monthly non-local domestic visitors average 5.6 million in Jeonju, 4.2 million in Pohang, 3.5 million in Gyeongju, 2.7 million in Gangneung and 1.4 million in Andong. Gyeongju is thus in the middle of this group on visitor volume while being first on attraction admissions — a reminder that "how large is this destination" has no single answer, and that the choice of outcome is itself a modelling decision. We return to that point in Section 7.

APEC is not a sporting mega-event, and the channels the mega-event literature was built around do not transfer intact. An Olympics or a World Cup brings spectators — hundreds of thousands of them, staying for days, buying tickets and hotel nights. A leaders' meeting brings delegations. The direct visitor count is small by tourism standards and heavily skewed towards officials, security personnel and press on per-diem itineraries rather than leisure travellers. Whatever effect such an event has on tourism is therefore mostly *indirect*, and the plausible channels differ in sign as well as in size.

Four channels appear relevant. *Anticipation and public investment* operates through the

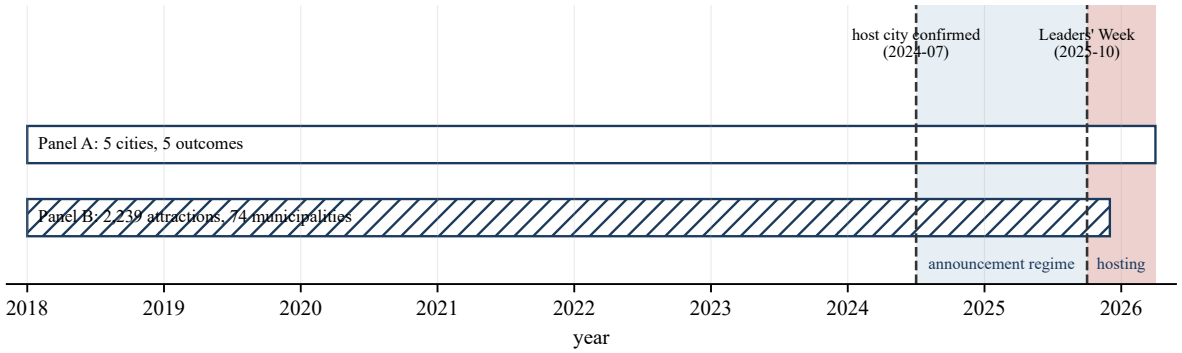

Figure 1: Study timeline. The host city was confirmed in June 2024 and the announcement regime is dated from the following month; Leaders' Week ran from 27 October to 1 November 2025, so the hosting regime is dated from October. Bars show the coverage of the two panels described in Section 4. Panel B ends in December 2025 for three of its four provinces, which limits the hosting window in that panel to three months.

announcement regime: host-city designation triggers venue and transport spending, and the resulting construction, signage and national press coverage can raise a destination's salience long before the event. *Destination image* works on a similar horizon but a longer tail — the point of hosting, in the policy discussion around Gyeongju, was that the city would be seen. Against these, *security and crowding-out* runs the other way and is specific to summits: road closures, controlled perimeters, restricted access to exactly the heritage sites that draw ordinary visitors, and advice to avoid the area can suppress leisure tourism in the very weeks the event occupies. *Displacement* is the fourth: visitors deterred from Gyeongju may go elsewhere in the province rather than not travel, which moves activity between municipalities without creating any.

Two implications follow for the research design. First, the announcement regime is where the clearer positive signal would be expected, while the hosting regime is where positive and negative channels are most likely to offset; a small hosting estimate is therefore not in itself evidence of no effect. Second, displacement complicates the role of the same-province comparison units: Andong and Pohang are the closest matches on observable characteristics, and are also the destinations most likely to absorb visitors diverted from Gyeongju. They are therefore modelled separately, and Section 6 reports the host and its neighbours as distinct treatment roles rather than pooling them.

4 Data

We assemble two monthly panels. They are complements rather than substitutes: the first carries the range of outcomes commonly used in the tourism literature but contains few units, while the second contains enough units to support inference but observes a single outcome. Table 2 summarises both.

4.1 Panel A: five cities, five outcomes

The first panel comes from the Korea Tourism Data Lab (KTDL), a platform operated by the Korea Tourism Organization that publishes municipal-level tourism indicators built from mobile-network location records, card-transaction settlements and in-car navigation queries. We collect five monthly series for each city from January 2018 to April 2026: non-local domestic visitors, domestic tourism expenditure, navigation search queries, foreign visitors, and foreign tourism expenditure. "Non-local" excludes residents of the city itself, which matters because a large share of recorded movement in a small city is commuting.

The five cities are Gyeongju, the host; Andong and Pohang, the two other major tourism destinations in North Gyeongsang province; and Jeonju and Gangneung, two comparably sized destinations outside the province. Jeonju is a heritage and food destination and Gangneung a coastal one, chosen because between them they span the two profiles that Gyeongju competes with nationally.

Four of the five series are complete for all 500 city-months. Foreign visitor counts are the exception: KTDL began publishing them at different dates for different cities, in 2020 for Gyeongju, Andong and Gangneung but only in 2022 for Jeonju and Pohang. Specifications involving foreign visitors therefore run on the common window from January 2022, and we flag them as such throughout. Foreign expenditure, by contrast, is available for the full period.

Three measurement conventions apply. Expenditure is nominal won as settled, with no price deflation: the estimand is a within-month contrast across cities, so a national deflator would be absorbed by the time effects and a city-level one does not exist at this frequency. Inflation is therefore not a confounder here unless it differs across the five cities, which we cannot rule

out but have no reason to expect. All series enter in logs, and no city-month in Panel A is zero or missing except the foreign-visitor gaps described above, so no zero handling is required on this panel; the zeros that do arise are on Panel B, where pandemic closures drove some municipalities to no recorded admissions, and the two panel versions handle them differently by construction. Finally, municipal boundaries were stable across the window — there were no mergers, splits or reclassifications among the five cities or the 74 municipalities between 2018 and 2026 — so a unit means the same place in every month.

4.2 Panel B: 2,239 attractions in 74 municipalities

The second panel is the Major Tourist Attraction Admissions statistic, compiled by the Korea Culture and Tourism Institute and distributed through the national statistics portal. It records monthly admissions at individually named attractions – temples, museums, parks, historic sites – separately for domestic and foreign visitors. We use the four provinces for which we hold the full site-level series: North Gyeongsang (the host province), South Gyeongsang, Gangwon and North Chungcheong. Together these cover 2,239 attractions in 74 *sigungu* (municipalities).

Two features of this source require care. First, the published file headers declare coverage through March 2026 for three of the four provinces, but admissions are in fact populated only through December 2025 in those three; only North Chungcheong carries 2026 months, and there only for a subset of sites. We use the dates actually populated, not the declared span, which limits the post-summit window in this panel to three months.

Second, the roster of reporting attractions is not fixed. Sites enter and leave the statistic, so a municipality's total mixes genuine changes in visits with changes in which sites report. We therefore construct two versions of the panel and carry both through every specification. The *balanced* version keeps only attractions observed in every month of the estimation window and only municipalities whose balanced total is strictly positive in every month, which yields 36 municipalities; the requirement that the total never hit zero is binding because pandemic closures drove some municipalities to zero admissions. The *full* version keeps every attraction and handles closure months with a log-plus-one transform, retaining 55 municipalities. Neither version is unambiguously preferable: the balanced panel achieves a fixed composition at the

cost of roughly a third of the comparison units, and the full panel makes the opposite trade. Both are therefore reported throughout, and any disagreement between them is stated explicitly.

Balancedness carries a second and less obvious cost. Requiring observation in every month of the window allows post-treatment months to determine roster membership, so an attraction that ceased or began reporting after the summit is excluded or admitted on the basis of post-treatment information. Section 6 rebuilds the roster from pre-treatment months alone and reports what changes.

For the analysis we restrict Panel B to January 2018 through December 2025, so that its window aligns with Panel A and excludes the earliest years, when the reporting roster was least stable.

4.3 Treatment dates

Gyeongju was confirmed as the 2025 APEC host city on 27 June 2024. We date the announcement regime from July 2024, the first full month after confirmation.

Dating the hosting regime requires more care. APEC 2025 is commonly described as a November summit, and the Gyeongju Declaration was adopted on 1 November. Leaders' Week, however, ran from 27 October to 1 November and the Leaders' Meeting itself from 31 October to 1 November, so five of the six days of Leaders' Week, together with the delegate, security and media influx preceding it, fall in October. In a monthly series the event therefore falls predominantly in October, and we date the hosting regime from October 2025. Table A1 in the appendix sets out the full calendar.

The choice is consequential. On Panel B, whose coverage ends in December 2025, dating the regime from November would assign the summit month itself to the pre-period and leave two hosting months rather than three. October 2025 also records the highest attraction admissions total in the late sample.

Gyeongju additionally hosted the first Senior Officials' Meeting cluster from 24 February to 9 March 2025, bringing roughly 1,500 delegates to the city. Those months fall within the announcement regime and are retained there: they form part of the preparation phase rather than the event, and assigning them a separate indicator would use two of the fifteen announcement

Table 2: Data sources and panel structure

Source / outcome	Units	Sites	Coverage	Obs.	Freq.
Panel A. Korea Tourism Data Lab, five cities					
Non-local domestic visitors	5	100	2018–2026	500	monthly
Domestic tourism expenditure	5	100	2018–2026	500	monthly
Navigation search queries	5	100	2018–2026	500	monthly
Foreign visitors	5	76	2020–2026	332	monthly
Foreign tourism expenditure	5	100	2018–2026	500	monthly
Panel B. Major tourist attraction admissions					
North Gyeongsang	23	651	2010–2025	132,178	monthly
South Gyeongsang	21	527	2010–2025	92,119	monthly
Gangwon	18	583	2010–2025	130,492	monthly
North Chungcheong	12	478	2010–2026	82,504	monthly
Total	*74*	*2,239*			

Panel A is drawn from the Korea Tourism Data Lab, which aggregates mobile-network visit counts, card-based expenditure and navigation queries at the municipal level. Panel B is the Major Tourist Attraction Admissions statistic compiled by the Korea Culture and Tourism Institute and released through KOSIS.
"Units" counts cities in Panel A and *sigungu* (municipalities) in Panel B; "Sites" counts individual attractions.
Foreign visitor counts begin later than the other Tourism Data Lab series and start at different dates across cities, so specifications using them run on the common window from 2022 onwards.
Panel B headers declare coverage to 2026-03 for three of the four provinces, but admissions are populated only to 2025-12 in those three; the table reports the dates actually populated.

months on a distinction the data cannot support.

The dating gives 15 announcement months in both panels, and either seven post-summit months (Panel A) or three (Panel B). Because the two regimes are mutually exclusive, the hosting coefficient is an effect relative to the pre-designation baseline, not an increment over the announcement regime; where we want the increment we say so and estimate it directly.

4.4 Variables not carried forward

The original commissioned study on which this paper builds constructed a news- and search-based "attention index" and used it as an explanatory variable. We do not. Every version of that index available to us is a single national monthly series with no municipal dimension: it takes the same value for every city in a given month. In a panel with year-month fixed effects such a series is collinear with the time effects by construction, and any coefficient it appears to carry is identified purely from aggregate co-movement over the sample period – a period dominated by the pandemic collapse and recovery. We return to this in Section 6, because the original study's mechanism analysis rests on exactly that identification.

5 Method

The object of the analysis is not a single preferred estimate but the mapping from design choices to conclusions. We therefore estimate the same effect under multiple specifications and report the full set, organised as two ladders: one varying only how uncertainty is quantified, the other only the comparison group.

Confirmation as host city and hosting the summit are distinct events, sixteen months apart, and there is no reason to expect them to have the same effect. We therefore replace the usual single post-treatment indicator with two:

$$\log Y_{it} = \alpha_i + \delta_t + \beta_A \left(\mathrm{Treat}_i \times \mathrm{Ann}_t\right) + \beta_H \left(\mathrm{Treat}_i \times \mathrm{Host}_t\right) + \varepsilon_{it}, \tag{1}$$

where Ann_t marks 2024-07 to 2025-09 and Host_t marks 2025-10 onward. On the five-city panel we also split Treat_i into Gyeongju and the two other North Gyeongsang cities, to separate the host's own response from any spillover to its neighbours.

5.1 Inference with a single treated unit

Equation (1) has a single treated unit, which is the central inferential difficulty addressed in this article.

Heteroskedasticity-robust standard errors are invalid here because the disturbances are serially correlated within units. Cluster-robust standard errors address that, but their asymptotics run in the number of clusters, and more importantly the relevant asymptotics for a treatment assigned to a single cluster run in the number of *treated* clusters, which is one and cannot be increased by adding comparison units (Conley and Taber, 2011; MacKinnon and Webb, 2018). Adding control municipalities improves the counterfactual; it does not repair the standard error.

We therefore report three inference procedures alongside each other. The first is the conventional cluster-robust standard error, included only so that it can be compared with what follows. The second is a wild cluster bootstrap with the null imposed and Rademacher weights (Cameron et al., 2008); with five clusters there are only $2^5 = 32$ distinct sign vectors, so we enumerate all of them rather than sampling. That removes simulation error from the bootstrap

distribution; it does not make the test exact, since exhausting the weights does nothing about the single treated cluster underneath. The third is randomization inference: we permute the assignment of cities to roles, holding the size of each role fixed, and compare the observed coefficient with the resulting distribution. For the Gyeongbuk-versus-rest specification the thirty role assignments collapse to ten distinct treatment indicators, and for the host-versus-spillover specification all thirty are distinct, so the attainable p-values are coarse — in steps of 0.1 and 0.033 respectively. They are nonetheless attainable, and a p-value of 0.20 obtained from ten distinct assignments is more informative than one of 0.001 obtained from a variance estimator whose assumptions do not hold. These are reported as design-conditioned permutation diagnostics rather than as exact randomization tests: Gyeongju was not randomly assigned to host APEC, so exchangeability of the city labels is an assumption about the comparison design rather than a property guaranteed by a known assignment mechanism.

On the attraction panel the same logic applies with more comparison units. There we permute the treated label across all municipalities, which gives a permutation floor of $1/36$ or $1/55$ depending on the panel version.

5.2 Synthetic control and synthetic difference-in-differences

The synthetic control method (Abadie et al., 2010) constructs a counterfactual as a convex combination of donors chosen to match the treated unit's pre-treatment path. It does not require parallel trends, which makes it the natural complement to equation (1).

It does, however, require the treated unit to lie inside the donors' convex hull. Gyeongju does not. Its mean pre-treatment log admissions exceed those of every donor municipality in all four provinces, so simplex weights cannot reach its level and the fitted synthetic control is biased downward by construction, inflating the post-treatment gap that would otherwise be read as an effect. We report the standard estimator to document this failure, and then re-estimate allowing an intercept, so that weights match the *shape* of the trajectory rather than its level (Doudchenko and Imbens, 2016; Ferman and Pinto, 2021). Inference is by placebo permutation: we run the identical procedure treating each donor as if it were Gyeongju and rank the treated unit within the resulting distribution. Two statistics are reported and they

answer different questions. The post-to-pre RMSPE ratio asks whether the treated unit's post-treatment divergence is large *relative to how well it was fitted*, and we compute it both over all placebos and after discarding those whose pre-treatment fit is more than five times worse than Gyeongju's, since a placebo that never fitted can produce a large ratio for reasons unrelated to treatment. The absolute ATT addresses whether the effect itself is unusual, and carries no fit screen when the donor pool is varied, since that exercise is designed to make deterioration in fit visible rather than to condition on it. Both statistics are reported.

Synthetic difference-in-differences (Arkhangelsky et al., 2021) sits between difference-in-differences and the synthetic control, and needs neither of the assumptions that give each of them trouble here. It retains unit fixed effects, so it never needed the level match, and it weights pre-periods as well as units, so it does not lean on unconditional parallel trends. Unit weights $\hat{\omega}$ solve a ridge-penalised simplex least-squares problem matching the treated unit's pre-treatment path; time weights $\hat{\lambda}$ solve the analogous problem predicting each donor's post-period mean from its pre-period path; the estimator is then a weighted difference-in-differences using both. The time weights matter substantively here: they concentrate on 5 to 9 of the 78 available pre-periods, discarding stretches in which the donors are least comparable.

Agreement between this estimator and the demeaned synthetic control is informative because the two rely on different assumptions. As an implementation check, imposing uniform unit and time weights reproduces the unweighted difference-in-differences estimate to ten decimal places.

5.3 Sensitivity to violations of parallel trends

Pre-trend tests are rejected in both panels, so the question is not whether parallel trends holds but how large a violation would have to be to explain the estimate away. We answer it with the framework of Rambachan and Roth (2023), using the authors' own implementation. Under $\Delta^{SD}(M)$ the differential trend is extrapolated linearly from the pre-period and allowed to curve away by at most M log points per month; under $\Delta^{RM}(\bar{M})$ each post-treatment step is at most $\bar{M}$ times the largest step observed before treatment. The breakdown value is the smallest M or $\bar{M}$ at which the robust confidence set covers zero. The benchmark $\bar{M} = 1$ is the natural

reference: an estimate that breaks down below it survives only if one believes the post-treatment differential trend was smaller than the one already visible in the data.

Three implementation choices should be noted. We trim the pre-window to $\tau \geq -12$, both because the fuller window reaches into the pandemic recovery — where month-to-month differential swings reach 0.11 to 0.75 log points, making $\bar{M} = 1$ an extraordinarily permissive benchmark — and because the conditional test's cost grows steeply in the number of post-periods. For the same reason we assess the hosting regime through a separate event study re-centred on the summit itself, which reduces its post-window to three periods. This re-centring changes the estimand: the pre-periods are then the announcement regime, so the quantity being bounded is the hosting effect *relative to* the announcement regime rather than relative to the pre-designation baseline.

The third choice works against the reported conclusion. The procedure takes an estimated covariance matrix as given, and the matrix supplied here is cluster-robust by municipality — the estimator this article argues is unreliable when only one cluster is treated. The direction of that unreliability is known: with a single treated cluster such variances are understated, so the robust confidence sets we feed into the breakdown calculation are too narrow and the breakdown values too generous. Every specification already fails at $\bar{M} = 1$ with these inputs, and a correctly widened variance could only lower the breakdown values further. We therefore read this exercise as a bound in the safe direction, and we do not treat it as independent validation of the one-city design, whose inference problem it inherits rather than solves.

5.4 Additional diagnostics

Every procedure above borrows strength from the number of comparison units. With one treated unit it is worth having an interval that does not. We therefore report conformal confidence sets (Chernozhukov et al., 2021). The procedure subtracts a candidate effect from the post-treatment periods and asks whether the adjusted post-treatment block still looks unusual against every moving-block permutation of the treated unit's own residual series — the whole series, not the pre-period alone, since under the null the adjusted residuals are exchangeable throughout. The assumption is exchangeability of the treated unit's residuals over time, which

is the assumption we can actually defend here. It is also the assumption most affected by the pandemic: a series containing a collapse and a recovery is not clearly exchangeable in time, and these intervals are therefore treated as a check on the other procedures rather than as primary evidence.

The two placebo exercises usual in this literature each hold one researcher choice fixed. A donor permutation asks whether Gyeongju is unusual among municipalities *at this date*; a placebo-in-time asks whether this date is unusual *for Gyeongju*. Neither addresses whether the unit–date pair is unusual, which is the relevant question when the analyst selected both.

Following the logic of specification search in Ferman et al. (2020), and in the spirit of the specification curve (Simonsohn et al., 2020), we therefore compute the joint distribution: every municipality treated as if it were the host, at every feasible month, with the demeaned synthetic control refitted each time. This is 2,088 fits on the balanced panel and 3,190 on the full panel, spanning 58 candidate dates. We solve the simplex-constrained problems in batch on a GPU by projected gradient with the sort-based Euclidean simplex projection, having verified that the solutions reproduce the SLSQP solutions used elsewhere in the paper.

Finally we ask what the design could have detected, and here two quantities need separating. Using the donor municipalities' own gaps as the null, we compute for each post-window length H the 95th percentile of the absolute mean gap over all rolling H-month windows. That is the *critical magnitude*: below it an estimate cannot be told from ordinary drift. It is a rejection threshold, not a detectable effect, because a true effect sitting exactly on it would be rejected only about half the time. The *minimum detectable effect* adds the 80th percentile of the signed gap distribution to the critical magnitude, giving the effect size this design would find four times in five. Reporting the first alone, as is common practice, understates the effect size the design would require. The null is constructed in two ways: from a held-out stretch of the pre-period, with weights fitted earlier, and from donor gaps during the announcement window itself, which is post-pandemic and therefore the noise level that actually applies to the hosting regime. Comparing the resulting curves with the estimated effect indicates whether additional months of data would have altered the conclusion.

6 Results

6.1 Baseline estimates

Before considering how far the estimates move, we report their levels. On the five-city panel, comparing the three North Gyeongsang cities with Jeonju and Gangneung, the announcement-regime effects are -0.6% for non-local domestic visitors, $+3.0\%$ for domestic tourism expenditure, $+15.4\%$ for navigation queries, $+35.2\%$ for foreign visitors and $+69.7\%$ for foreign expenditure. Under heteroskedasticity-robust standard errors the last three are significant at conventional levels, two of them at $p < 0.001$. Under randomization inference across city-role assignments none is, with p-values between 0.2 and 1.0.

Separating the host from its neighbours, Gyeongju's own announcement-regime effects are $+5.7\%$ on visitors, $+6.0\%$ on expenditure and $+145.4\%$ on foreign expenditure, while the two neighbouring cities show -3.6%, $+1.6\%$ and $+41.2\%$. The only estimate anywhere in this set that clears the randomization floor is Gyeongju's foreign expenditure in the announcement regime, at $p = 0.033$ — rank one of thirty possible role assignments. It is reported here as the only result that survives, but two considerations limit the weight it can bear.

The first is that 0.033 is the permutation floor. The host-versus-spillover specification admits thirty distinct role assignments, so no result can come in lower: this is not a small p-value that happened to be attained but the only value below 0.05 the test can return at all. The Gyeongbuk-versus-rest specification is coarser still — its thirty assignments collapse to ten distinct treatment indicators, so its p-values are attainable only in steps of 0.1, so it cannot produce a result significant at conventional levels under any realisation of the data.

The second is multiplicity. Across the two specifications we estimate thirty coefficients — five outcomes, two regimes, and either one or two treated roles. At a nominal five percent level, such a family would be expected to yield approximately 1.5 false positives under a global null. One was obtained. A single nominal rejection out of thirty is consistent with the null rather than evidence against it. We therefore treat the set as failing to reject and report the surviving coefficient for completeness.

The hosting-regime effects are smaller than the announcement-regime effects on the five-

city panel for four of five outcomes — foreign expenditure falls from $+69.7\%$ to $+29.3\%$, foreign visitors from $+35.2\%$ to $+11.2\%$ — and larger on the attraction panel, where the hosting window is only three months and captures the summit month itself. We do not interpret this as evidence of decay, since neither difference is distinguishable from zero; it is reported because the pattern differs across the two panels.

One consequence of the dating runs counter to expectation. Including October in the hosting window adds the highest-admissions month of the late sample to the treated period, which would be expected to raise the hosting estimate. It lowers it: on the full attraction panel the synthetic control's hosting effect falls from $+27.7\%$ under the November dating to $+21.3\%$ under the October one, and synthetic difference-in-differences falls from $+10.8\%$ to $+8.2\%$. The reason is that October is not distinctively Gyeongju's peak. Averaging pre-treatment log admissions by calendar month and centring each series on its own annual mean, October is the highest of the twelve months for the donor municipalities and for Gyeongju alike, and by almost the same margin: $+0.592$ against $+0.581$ in logs on the full panel. An estimator matched on pre-treatment seasonal shape therefore absorbs the October surge rather than attributing it to the summit. The observed increase around the event is largely attributable to the season in which the event was scheduled.

6.2 Sensitivity to the inference procedure

Table 3 and Figure 2(a) report the first ladder. Rung A1 is the original design: Gyeongju against Andong and Pohang, a single post-treatment indicator, heteroskedasticity-robust standard errors. A2 adds Jeonju and Gangneung to the comparison group. A3 through A5 change nothing about the specification — same outcome, same sample, same regressors — and vary only how the standard error is computed. A6 additionally splits the post period into the two regimes.

The pattern is the central result. From A2 onward the point estimates are unchanged: non-local domestic visitors sit at $+9.1\%$ at A2, A3, A4 and A5, foreign tourism expenditure at $+111.4\%$, navigation queries at -3.8%. What moves is the p-value. For domestic visitors it travels from 0.000 under HC3, to 0.033 under cluster-robust errors, to 0.250 under the enumerated wild cluster bootstrap and 0.200 under randomization inference. For foreign expenditure

it travels from 0.000 to 0.003 to 0.188 to 0.200.

No information about tourism in Gyeongju is added between A2 and A5; what changes is the treatment of the variance estimator's assumptions. The original significance is a property of HC3 applied to five clusters with one treated unit rather than a property of the data.

Two further features of the ladder are worth noting. First, the navigation-query estimate is negative throughout and never approaches significance under any procedure, which provides an internal check: the ladder does not deflate all estimates toward zero but reports what each procedure implies. Second, moving from A1 to A2 — adding two comparison cities — changes the navigation coefficient from -12.8% to -3.8% and its p-value from 0.002 to 0.320. The choice of comparison group therefore matters as much as the variance estimator, and neither is typically reported as a design choice.

6.3 Sensitivity to the donor pool and the convex hull

Table 4 and Figure 2(b) report the second ladder, on the attraction panel. B1 reproduces the original design: a synthetic control for Gyeongju built from Andong and Pohang alone. It returns an announcement-regime effect of $+313.7\%$ with a placebo p-value of 0.333 — which, with two donors, is the smallest value the test can return at all.

This estimate is not interpretable as an effect. Its pre-treatment RMSPE is 0.871 in logs — the synthetic control misses Gyeongju's own pre-treatment path by a factor of about 2.4 on average, before any treatment has occurred — because Gyeongju's admissions exceed those of every donor and simplex weights cannot reach its level. The estimator is therefore reporting a level gap that it was never able to close.

B2 serves as a diagnostic. Adding the other nineteen North Gyeongsang municipalities changes nothing at all: identical weights (Andong 0.864, Pohang 0.136), identical fit, identical estimate to three decimal places. Andong is already the largest municipality in the province, so nineteen additional donors leave the optimiser on the same boundary solution. The only thing that does move is the p-value, from 0.333 to 0.091, and it moves purely because the reference set grew from two donors to twenty-one — an improvement in the resolution of the test rather than in the estimate being tested. Enlarging the donor pool is therefore not sufficient in itself;

it helps only when the pool contains a unit close to the treated unit's level. Only at B3, when a fourth province introduces Danyang — 0.115 log points below Gyeongju rather than 0.737 — does the fit improve, and the estimate falls to $+209.0\%$.

Rungs B4 to B6 allow an intercept. The fit improves substantially: pre-treatment RMSPE falls to 0.132, better than the median donor's. The announcement effect falls with it, to $+52.8\%$ with two donors, $+28.7\%$ with twenty-one, and $+7.5\%$ with fifty-four, where the placebo p-value is 0.745. Two distinct problems compound in B1: a donor pool too small to place the treated unit, and a functional form that cannot represent it even in principle. Each can be varied separately. Holding the standard estimator fixed and enlarging the pool alone (B1 to B3) moves the estimate by 105 percentage points; holding the two-donor pool fixed and allowing an intercept alone (B1 to B4) moves it by 261. Changing both (B1 to B6) moves it by 306, which is less than the sum of the two separate moves. The two choices interact, so the ladder should be read as a sensitivity path rather than an additive decomposition: each problem is individually sufficient to overturn the original estimate, but the total cannot be apportioned between them.

6.4 Convergence across estimators

Table 5 places the demeaned synthetic control, synthetic difference-in-differences and conformal intervals side by side. The estimators agree on imprecision: none separates its estimate from its own placebo distribution, and they fail to do so while relying on different assumptions, which is what makes the agreement informative rather than mechanical. They do not agree on sign.

Point estimates are positive in seven of the eight regime-by-estimator cells, the exception being synthetic difference-in-differences in the announcement regime on the full panel, which is -0.7%. On the balanced panel they are larger in the hosting regime than the announcement regime: $+12.3\%$ and $+38.8\%$ under the synthetic control, $+1.8\%$ and $+26.4\%$ under synthetic difference-in-differences. None is distinguishable from its placebo distribution, with p-values between 0.265 and 0.927. All four conformal intervals contain zero.

The synthetic difference-in-differences estimates are systematically smaller than the synthetic control estimates in the announcement regime ($+1.8\%$ versus $+12.3\%$; -0.7% versus

$+7.5\%$). The difference is interpretable: the time weights concentrate on the pre-periods in which donors most closely resemble Gyeongju, down-weighting stretches that the unweighted estimator treats as comparable. Part of what the synthetic control attributes to the announcement is therefore the continuation of a pre-existing divergence.

6.5 Sensitivity to pre-trend violations

Pre-trend tests reject in both panels, so Table 6 reports the quantitative sensitivity instead. Under the relative-magnitudes restriction, the breakdown value is 0.00 for all three five-city outcomes — their robust confidence sets contain zero even when no post-treatment trend violation whatever is permitted — 0.25 for the attraction panel's announcement regime, and 0.50 for the summit-centred hosting specifications.

None of the seven targets survives $\bar{M} = 1$. The strongest result in the paper would be overturned by a post-treatment differential trend half the size of the one already visible before treatment.

We report alongside these a conservative approximation computed before the official procedure was available, which returned breakdown values between 0.019 and 0.087. On the two attraction specifications where both procedures report an announcement-regime breakdown value, the official implementation raises it from 0.052 and 0.019 to 0.250 — factors of roughly five and thirteen — and still leaves every one below the benchmark. The direction of the correction is relevant: the more exact procedure moves the breakdown values against our conclusion without altering it.

6.6 Placebo evidence

Figure 4 reports the placebo exercises.

Panel (a) moves the treatment date to three dates on which nothing happened, truncating the sample each time so no genuine post-treatment month contaminates the placebo. On the balanced panel the real date yields $+12.3\%$; the three fake dates yield $+32.2\%$, $+17.7\%$ and $+18.4\%$. All three exceed the real estimate. Their placebo p-values (0.278 to 0.417) are *smaller* than the real one's (0.611), and their pre-treatment fits are comparable (0.143–0.161 against

0.164), so the result is not an artefact of poor fit at the placebo dates. The estimator is capturing Gyeongju's persistent drift relative to its donors rather than the treatment.

Panel (b) generalises this. Refitting the demeaned synthetic control for every municipality at every feasible date gives 2,088 estimates on the balanced panel and 3,190 on the full panel. The actual Gyeongju-in-2024-07 estimate sits near the middle of that joint distribution: the share of unit–date pairs producing an absolute effect at least as large is 0.572 on the balanced panel and 0.795 on the full panel, and restricting to placebo fits at least as good as the real one barely moves these (0.554 and 0.792).

This relaxes more of the analyst's discretion than either conventional placebo. A donor permutation asks whether Gyeongju is unusual at this date; a placebo-in-time asks whether this date is unusual for Gyeongju. Each holds fixed one choice the analyst made. The joint grid holds neither fixed, and the estimate is unremarkable within it.

Two qualifications apply. The cells are not draws from a common null: the windows overlap, they carry different exposure to the pandemic, and some later ones contain genuinely post-treatment months. The grid is therefore a descriptive reference distribution indicating where the observed estimate falls among comparable fits, not a test with calibrated size. In addition, it varies the designation date only and so applies to the announcement regime; the hosting regime has too few post-treatment months to permit the same sweep.

6.7 Statistical power

The hosting estimate rests on three months in the attraction panel and seven in the city panel, so a natural response is to await further data. Figure 5(b) asks whether waiting would work.

Two thresholds should be distinguished. The *critical magnitude* is the 95th percentile of the donor municipalities' own H-month average absolute gaps: an estimate below it cannot be told from ordinary drift. The *minimum detectable effect* is larger, because an effect sitting exactly at the critical magnitude would be rejected only about half the time; writing G for the signed donor gap distribution, the effect this design would detect four times in five is the critical magnitude plus the 80th percentile of G.

Using the donors' gaps during the announcement window as the null — a post-pandemic

stretch, and therefore the noise level that actually applies — the critical magnitude at the three months in hand is 0.515 log points on the balanced panel, against an observed hosting effect of 0.328. The minimum detectable effect there is 0.699. Extending the panel to nine post-summit months lowers the critical magnitude only to 0.418 and the minimum detectable effect to 0.568, both still above the estimate. On the full panel the gap is wider and never closes within twelve months.

The reason the curve falls so slowly is that the gaps are strongly serially correlated: the noise is persistent unit-level drift, not month-to-month variation, and averaging more months does not average away a drift. This is a property of the design rather than of the sample length available at the time of writing: the planned data extension would not have altered the conclusion.

6.8 Robustness checks

Four further checks assess whether the results above are artefacts of specific design choices.

No single donor carries the estimate. Dropping each comparison municipality in turn and refitting gives an announcement-regime effect ranging from $+1.6\%$ to $+11.9\%$ on the full panel, against a baseline of $+7.5\%$, and a hosting effect from $+19.2\%$ to $+24.0\%$ against $+21.3\%$. No sign reversal occurs in any of the fifty-four iterations. The most influential single donor is Cheongju, with a weight of 0.110; removing it moves the hosting estimate by 2.7 percentage points. The estimates are therefore not driven by any one comparison unit.

Selecting the roster on pre-treatment information moves the balanced panel towards the full one. Requiring an attraction to report in every month of the estimation window uses post-treatment months to decide membership, so an attraction that stopped or started reporting after the summit is admitted or excluded on the basis of what happened after treatment. The exposure is substantial: 555 attractions satisfy the requirement over the pre-period alone and 429 satisfy it over the whole window, so 126, approximately one quarter, are determined by post-treatment reporting. It is also asymmetric. Among attractions selected on pre-treatment information, every one of Gyeongju's fourteen reports in every post-summit month, against 85.8% of the donors'.

Rebuilding the roster from pre-treatment months only and leaving membership fixed there-

after halves the balanced panel's announcement estimate, from $+12.3\%$ to $+6.6\%$, and raises its placebo p-value from 0.611 to 0.757. The hosting estimate barely moves, from $+38.8\%$ to $+38.5\%$. The correction therefore brings the balanced panel's announcement estimate into line with the full panel's $+7.5\%$, which is what one would expect if part of the gap between the two panels was selection rather than composition. We retain the window roster as the reported specification for comparability with existing work, noting that the pre-selected roster is the more defensible construction and yields the same qualitative conclusion.

Heterogeneity-robust estimators change nothing, as expected. We report Callaway and Sant'Anna (2021) and Sun and Abraham (2021) estimates because they are now the default expectation for any difference-in-differences paper, and note that they cannot help here. With one treated unit adopting at one date there is no staggered adoption and no negative weighting to correct. The two estimators return identical figures to four decimal places — 0.1795 on the balanced panel, 0.1466 on the full panel for the announcement regime — confirming that they have collapsed to the same comparison.

They do, however, surface something the two-way fixed effects estimate hides. On the full panel the fixed-effects estimate is 0.3609 against their 0.1466, a factor of two and a half. The difference is the implicit baseline: the group-time estimators anchor on the period immediately before treatment, while two-way fixed effects anchors on the whole pre-period mean. When the pre-trend is as pronounced as it is here, that choice moves the estimate by more than most of the choices that are ordinarily deliberated. On the balanced panel, where the pre-trend is milder, the two agree to within 0.005.

Balanced and full panels agree in direction and differ in level. The balanced panel returns larger estimates throughout — $+12.3\%$ against $+7.5\%$ in the announcement regime, $+38.8\%$ against $+21.3\%$ in the hosting regime. Neither is unambiguously correct. The balanced panel holds the attraction roster fixed at the cost of dropping nineteen municipalities to pandemic closures; the full panel retains them at the cost of a shifting roster. Both are reported throughout, and the qualitative conclusions do not depend on the choice.

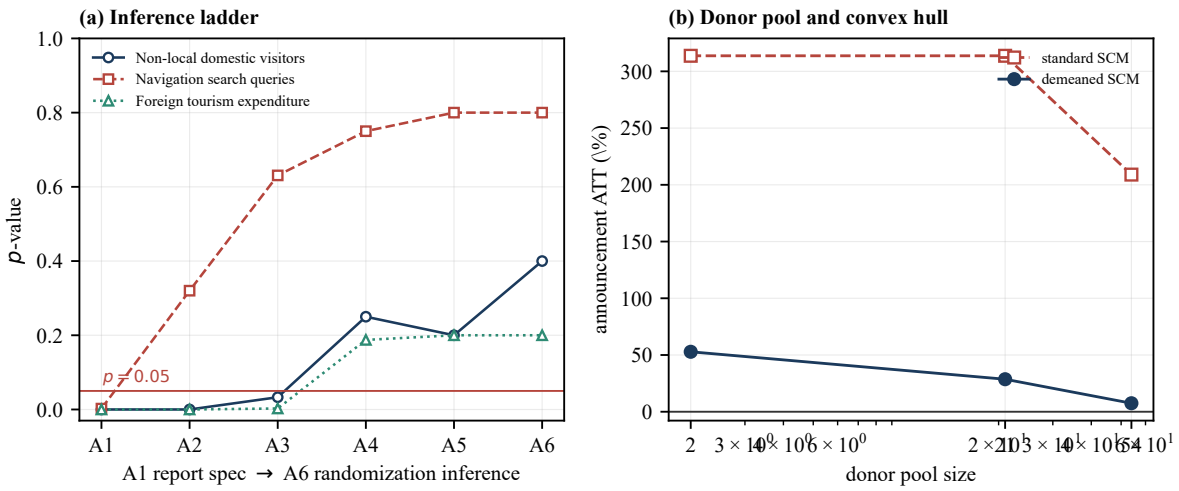

Figure 2: Specification ladders. Panel (a): p-values for the Gyeongju-versus-comparison difference-in-differences as the inference procedure changes, holding the specification fixed; the underlying point estimates are identical across rungs A2–A5. Panel (b): announcement-regime ATT from the synthetic control as the donor pool grows, with and without an intercept. Five cities and 500 city-months in panel (a); 55 municipalities and 5,280 municipality-months in panel (b).

6.9 Reassessment of the original mechanism analysis

Finally, the commissioned study reported a mediation chain running from a news-and-search attention index through navigation search to visits and spending, with 24.9% of the total effect mediated. Re-running the authors' own code on the current data reproduces the structure but not the figures: the mediated share is 18.0%, the first-stage coefficient 0.0029 rather than 0.0039. The discrepancy is a vintage problem — the panel was extended in July 2024 and that section was not re-estimated before the report was finalised.

The deeper issue is identification. As noted in Section 4, the attention index has no municipal variation. Regressing it on the other regressors in a specification with year-month fixed effects returns an R^2 of 1.000000, and the design matrix has rank 102 of 117 columns: the index is absorbed by construction. The reported coefficients exist only in specifications that omit time effects, where they are identified from aggregate co-movement over a window dominated by the pandemic. A coefficient that requires the omission of time effects reflects correlation with the national tourism cycle rather than a mechanism, and it is not carried forward.

Table 3: Ladder A: inference. The estimates hold still; the p-values do not

Rung	Design change	Inference	Dom. visitors Effect	p	Nav. queries Effect	p	Foreign spend Effect	p
A1	Gyeongju vs Andong/Pohang	HC3	+10.8%	0.000	-12.8%	0.002	+88.3%	0.000
A2	+ Jeonju/Gangneung as controls	HC3	+9.1%	0.000	-3.8%	0.320	+111.4%	0.000
A3	same spec, clustered SE	cluster	+9.1%	0.033	-3.8%	0.631	+111.4%	0.003
A4	same spec, wild cluster bootstrap	wcb	+9.1%	0.250	-3.8%	0.750	+111.4%	0.188
A5	same spec, randomization inference	ri	+9.1%	0.200	-3.8%	0.800	+111.4%	0.200
A6	+ announcement / hosting split	ri	+7.6%	0.400	-7.8%	0.800	+106.5%	0.200

City-level difference-in-differences on the five-city Tourism Data Lab panel, with Gyeongju as the treated unit throughout; the rungs vary which cities serve as comparisons, not which city is treated. Each rung changes exactly one design choice, and the outcome variable and estimation window are held fixed, so movement is attributable to the design and not to a different sample.

A1–A2 change the comparison group; A3–A5 change only how the standard error is computed; A6 splits the post period into the announcement and hosting regimes. The point estimates barely move between A2 and A5. What moves is the p-value, because HC3 standard errors are not valid with five clusters and one treated unit.

WCB = wild cluster bootstrap with the null imposed, all $2^5 = 32$ Rademacher sign vectors enumerated; enumerating every weight makes the bootstrap exhaustive, which is not the same as making the test exact. RI = permutation over which of the five cities is designated treated, refitting each time; the reference set has five members, so p cannot fall below 0.2 on this ladder. Both are design-conditioned diagnostics: city labels are not exchangeable by construction, since Gyeongju was not randomly assigned to host.

Table 4: Ladder B: donor pool and the convex hull

Rung	Variant	Donors	Level gap	Pre-RMSPE	ATT ann.	p	ATT host.	p	$p_{\min}$
B1	standard	2	+0.737	0.871	+313.7%	0.333	+349.2%	0.333	0.333
B2	standard	21	+0.737	0.871	+313.7%	0.091	+349.2%	0.091	0.045
B3	standard	54	+0.115	0.442	+209.0%	0.036	+306.0%	0.036	0.018
B4	demeaned	2	+0.737	0.331	+52.8%	0.667	+38.3%	0.667	0.333
B5	demeaned	21	+0.737	0.185	+28.7%	0.182	+47.5%	0.318	0.045
B6	demeaned	54	+0.115	0.132	+7.5%	0.745	+21.3%	0.545	0.018

Synthetic control on the attraction-admissions panel, log monthly admissions, treated unit Gyeongju. "Level gap" is Gyeongju's pre-treatment mean minus that of the highest donor, in log points; a positive value means Gyeongju lies outside the donors' convex hull and simplex weights cannot reach its level.

B1 and B2 are numerically identical. Andong is the largest sigungu in Gyeongbuk, so adding the other nineteen Gyeongbuk units leaves the optimiser on the same boundary solution and the fit does not improve at all. Only the fourth province helps, because Danyang sits 0.115 log points below Gyeongju rather than 0.737. A larger donor pool is not useful in itself; it is useful when it contains a unit near the treated one's level.

"Demeaned" allows an intercept, matching the shape of the trajectory rather than its level (Doudchenko and Imbens, 2016; Ferman and Pinto, 2021).

p-values are two-sided placebo permutations within the stated donor pool, computed over the $J + 1$ assignments as $(1 + \#\{j : |T_j| \geq |T_{\text{treated}}|\})/(J + 1)$, with ties counted as extreme. $p_{\min} = 1/(J + 1)$ is the smallest value the test can return, shown because it does most of the work at the top of the ladder: with two donors no result can fall below 0.333, so the original design cannot reject at conventional levels whatever the data say. These are design-conditioned placebo diagnostics rather than exact randomization tests; Gyeongju was not randomly assigned to host.

Table 5: Convergence across estimators

		Demeaned SCM		Synthetic DiD		Conformal
Sample	**Regime**	ATT	p	ATT	p	95% CI (log)
balanced	announcement	+12.3%	0.588	+1.8%	0.861	[-0.307, +0.243]
balanced	hosting	+38.8%	0.265	+26.4%	0.278	[-0.165, +0.635]
full	announcement	+7.5%	0.735	-0.7%	0.927	[-0.344, +0.268]
full	hosting	+21.3%	0.531	+8.2%	0.764	[-0.458, +0.617]

The three estimators fail differently, which is the point of running all of them. The event-study difference-in-differences needs parallel trends and does not have them; the synthetic control needs the treated unit inside the donors' convex hull, which fails in levels and is repaired by demeaning; synthetic difference-in-differences keeps unit fixed effects and reweights pre-periods, so it needs neither.

p-values are placebo permutations over donors. Conformal intervals follow Chernozhukov et al. (2021), using moving-block permutations of the treated unit's own residual series, which is the inference procedure with the best justification when there is one treated unit.

All four conformal intervals contain zero.

Table 6: Breakdown values for violations of parallel trends

Outcome	**Target**	**Original CS**	$\bar{M}$ approx.	$\bar{M}$ official	M (SD)	**Survives** $\bar{M} = 1$
Non-local domestic visitors	announcement 0to5	[-0.079, +0.147]	0.000	0.000	0.050	no
Navigation search queries	announcement 0to5	[-0.011, +0.238]	0.000	0.000	0.010	no
Foreign tourism expenditure	announcement 0to5	[-0.104, +0.453]	0.000	0.000	0.010	no
Attraction admissions (balanced)	announcement 0to5	[+0.036, +0.210]	0.052	0.250	0.000	no
Attraction admissions (full)	announcement 0to5	[+0.053, +0.197]	0.019	0.250	0.100	no
Attraction admissions (balanced),	hosting	[+0.187, +0.471]	0.087	0.500	0.010	no
Attraction admissions (full), summ	hosting	[+0.171, +0.401]	0.036	0.500	0.010	no

Sensitivity of the event-study estimates to violations of parallel trends (Rambachan and Roth, 2023), computed with the authors' `HonestDiD` package. $\bar{M}$ is the breakdown value under Δ^{RM}: the post-treatment violation expressed as a multiple of the largest violation actually observed before treatment. M is the breakdown value under Δ^{SD}, bounding the curvature of the differential trend in log points per month.

$\bar{M} = 1$ is the canonical benchmark: an estimate that breaks down below it requires believing the post-treatment differential trend was *smaller* than the one visible beforehand. 0 of 7 targets survive it.

The "approx." column is a conservative Python approximation computed before the official package was run; it adds a normal interval to the identified set and plugs in the estimated pre-period coefficients, so it is a lower bound. It is reported here to show that the official procedure raises the breakdown values by a factor of four to thirteen and still leaves all of them below the benchmark.

Summit-centred specifications re-date treatment to 2025-11, which makes the pre-periods the announcement regime; the estimand is then the hosting effect relative to the announcement regime.

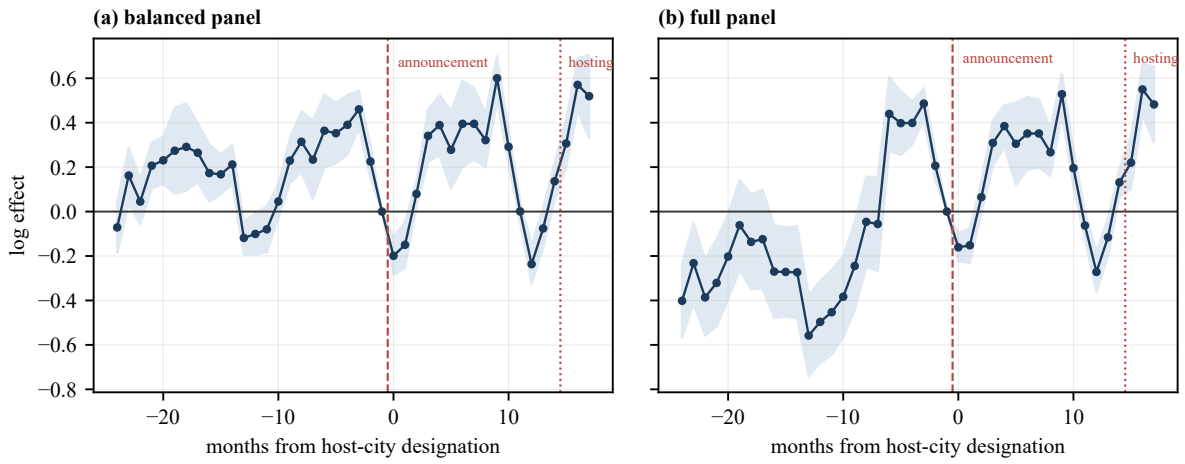

Figure 3: Event-study coefficients for Gyeongju against comparison municipalities, by months from host-city designation. Shaded bands are 95% confidence intervals from municipality-clustered standard errors, shown for comparability with standard practice; the paper's inference rests on the permutation procedures reported in Tables 5 and 6. Pre-trend tests reject in both panels. The plot is a conventional two-way fixed effects event study, so the visual heuristics for reading pre-treatment coefficients apply to it directly; this is not generally true of the plots produced by newer estimators (Roth, 2024).

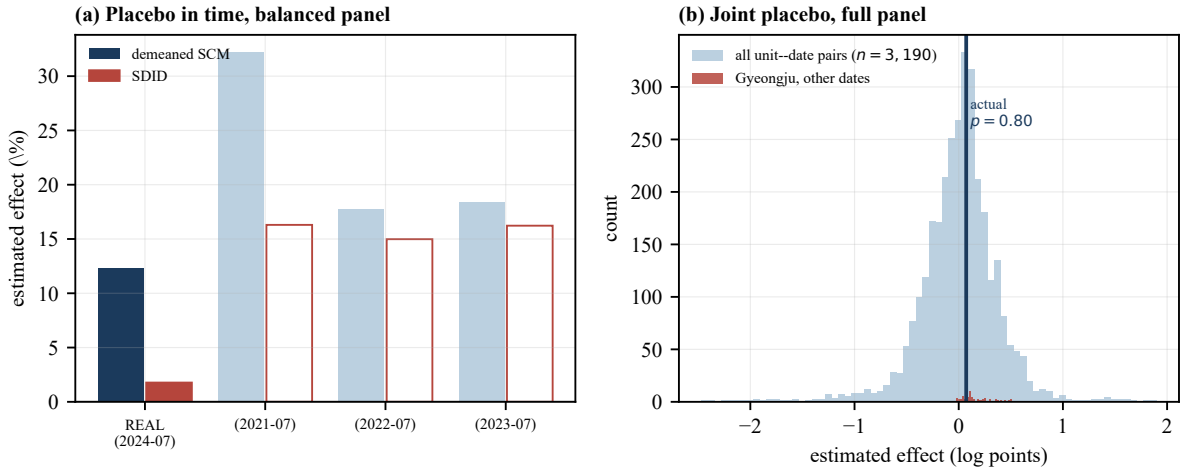

Figure 4: Placebo evidence. Panel (a): the same pipeline run with the treatment date moved to three dates on which nothing happened, balanced panel, sample truncated before the true designation in each case. Panel (b): the joint distribution of estimated effects over every municipality and every feasible treatment date, full panel ($n = 3{,}190$ fits); the vertical line is the actual Gyeongju estimate at the actual date.

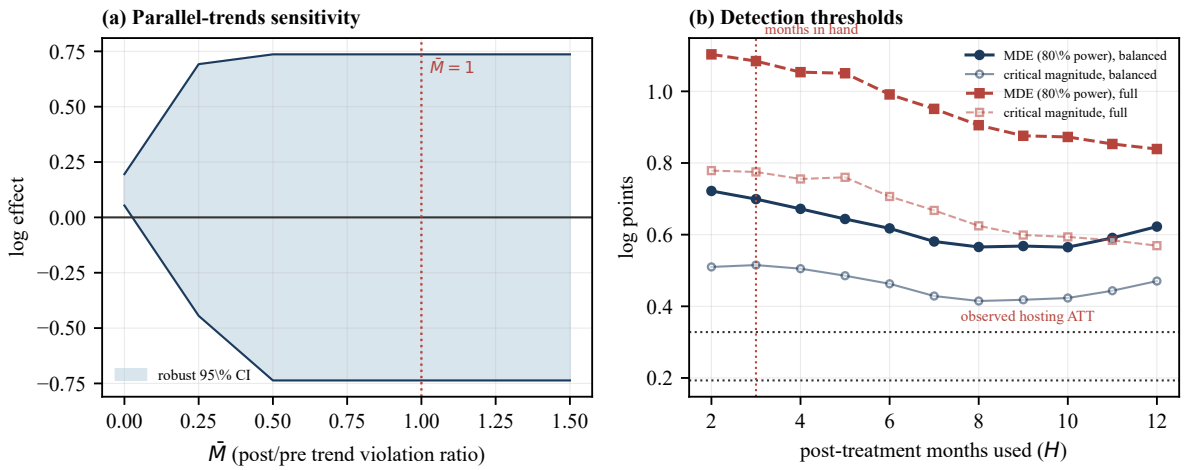

Figure 5: Panel (a): robust confidence sets for the announcement-regime effect on the full attraction panel under $\Delta^{RM}(\bar{M})$ (Rambachan and Roth, 2023); the set covers zero well before the canonical benchmark $\bar{M} = 1$. Panel (b): two detection thresholds as a function of post-treatment window length, both computed from donor municipalities' own gaps during the announcement window. The lower curve is the critical magnitude, the 95th percentile of donor gaps, below which an estimate cannot be told from drift; the upper curve is the minimum detectable effect at 80% power. Dotted horizontal lines mark the observed hosting effects.

7 Discussion

7.1 Outcome choice and the measured legacy

Section 6 estimated the announcement-regime effect on five outcomes measured on the same five cities over the same months, using one specification. On the common 2022-onward window the estimates are $+8.1\%$ for navigation search queries, $+1.7\%$ for visitors, $+1.5\%$ for tourism expenditure, $+34.3\%$ for foreign visitors and $+52.8\%$ for foreign expenditure. None is significant under randomization inference, and the pairwise differences are not significant either, so we make no claim that the funnel narrows in any causal sense.

The descriptive point nonetheless holds. A study reporting domestic visitors would report an effect close to zero; one reporting foreign expenditure would report a headline figure some thirty-five times larger. Both are defensible measures of the tourism legacy, both are drawn from the same platform, and the choice between them is made prior to estimation and is rarely discussed. The mega-event literature contains studies using arrivals, overnight stays, hotel revenue, admissions and expenditure more or less interchangeably. These results suggest that such outcomes are not interchangeable, and that comparisons of effect sizes across studies using

different outcome variables may reflect the choice of outcome rather than differences between the events.

One qualification applies to the term itself. The post-summit window is three months on the attraction panel and seven on the city panel. This is sufficient to characterise short-run and early post-event outcomes, but not to identify a legacy in the sense the term ordinarily carries, of a durable shift in a destination's position years after the event. The term is used here to denote the family of outcomes the literature labels in this way, and the estimates throughout should be interpreted as short-run.

This is a measurement problem prior to being an inference problem, and it is not addressed by improved standard errors.

7.2 Reporting recommendations for single-host designs

Tokarchuk and Gabriele (2026) propose a placebo checklist for tourism research. Our results suggest four additions specific to designs with a single treated unit. The first three are established practice not yet followed in this literature; the fourth is a proposal.

State the permutation floor. A placebo test over J donors cannot produce a p-value below $1/(J+1)$. Our own rung B1 is the illustration: with two donors the smallest attainable value is 0.333, so the treated unit ranking first establishes only that the test had nowhere lower to go. Reporting the floor alongside the p-value makes this distinction visible; Lahura and Sabrera (2023) does so explicitly.

Check the convex hull before fitting. If the treated unit's pre-treatment level exceeds every donor's, simplex-weighted synthetic control cannot match it and the resulting gap is mechanical. This is checkable in one line before any estimation, and the fix — allowing an intercept (Doudchenko and Imbens, 2016; Ferman and Pinto, 2021) — is standard. In our case, applying it while changing nothing else moves the estimate by 261 of the 306 percentage points that separate the original figure from ours. Enlarging the donor pool without checking is not a substitute: adding nineteen municipalities changed our estimate by exactly nothing, because none of them was closer to the treated unit than the donor already carrying the weight.

Report pre-trend sensitivity, not just a pre-trend test. A rejected pre-trend test tells the

reader that the identifying assumption fails but nothing about how much that matters. The breakdown value answers the second question in a single number, and Rambachan and Roth (2023) is implemented in publicly available software. Eleven of the sixteen studies coded here test for pre-trends and one goes further; this is among the least costly improvements available.

Rank the estimate against both choices at once. Conventional placebo tests hold one researcher choice fixed — the donor permutation fixes the date, the placebo-in-time fixes the unit. Computing the joint distribution over both is now computationally inexpensive: the full grid of 5,278 synthetic control fits used here runs in seconds on a consumer GPU. Where the analyst selected both the unit and the date, this is the distribution against which the estimate should be ranked. The procedure is not new — it applies the specification-search logic of Ferman et al. (2020) along an additional dimension — but it is now inexpensive enough to be adopted routinely.

7.3 Absence of evidence and implications for practice

It is important to be precise about what the analysis does and does not establish.

The point estimates in this paper are positive in almost every specification. On the attraction panel they are larger in the hosting regime than in the announcement regime, under both estimators; on the five-city panel the ordering reverses for four of five outcomes, and we do not know which pattern is real because neither difference is distinguishable from zero. The sign is consistent across panels. The estimates also survive leave-one-out donor deletion with no sign reversal in any of fifty-four iterations, and the leave-one-out range on the hosting effect is $+19.2\%$ to $+24.0\%$ on the full panel. The estimates are therefore not driven by a single influential comparison unit.

What we have shown is that the same design produces effects of comparable or larger magnitude at dates when nothing happened, that the estimate sits mid- distribution among unit–date pairs, and that a modest continuation of the pre-existing trend would account for it. These are statements about what the design can distinguish rather than about the underlying phenomenon. Concluding from this article that APEC Gyeongju had no tourism effect would overstate the evidence in the opposite direction to the claims under examination.

The overall assessment is that a single-host, municipal-scale design is underpowered against the drift that ordinarily separates comparable municipalities, and that this is a property of the design rather than a temporary shortage of data. Our power calculation makes the second half of that sentence concrete: extending the panel to nine post-summit months would leave the minimum detectable effect above the estimate, because the noise is persistent rather than transitory.

That has a consequence for how such evaluations are commissioned and read. Mega-event bids are justified with projected tourism gains, and ex-post evaluations are commissioned to verify them. If the evaluation design cannot distinguish a real effect from ordinary municipal drift, then a positive finding and a null finding carry the same information, and the exercise provides political cover rather than evidence.

This is not an argument against ex-post evaluation, but an argument for stating detectable magnitudes in advance. A power calculation of the kind in Section 6 can be run before the event, using only historical data on candidate comparison units, and would tell a commissioning body whether the evaluation it is paying for can answer the question it is asking. Where the answer is no, the options are to widen the design — more treated units, which usually means studying a class of events rather than one — or to commission a descriptive study and label it as such.

The alternative, illustrated by the case examined here, is the production of estimates that are large and conventionally significant but not robust to standard diagnostics.

8 Conclusion

This article evaluated the tourism effect of a mega-event on its host city and found that the estimated answer depends more on design choices than on the data. Holding the sample fixed, the p-value on the headline estimate travels from 0.000 to 0.200 as the variance estimator is corrected, without the point estimate moving at all; the synthetic control estimate falls from $+313\%$ to $+7.5\%$ as the donor pool is enlarged and the convex-hull failure repaired; no estimate survives a trend violation half the size of the one visible before treatment; and the estimate sits mid-distribution among placebo unit–date pairs. The design cannot establish an effect, and

more data would not change that, because the noise it must overcome is persistent drift rather than transitory variation.

Several limitations qualify these conclusions.

This is a single case. The diagnostics are general but the magnitudes are not, and a different host city — one closer to the middle of its comparison distribution, with a longer stable pre-period — might survive the same set of tests. What generalises is the structure of the problem rather than the size of the correction.

The literature coding covers sixteen studies, which is sufficient to characterise practice but not to support claims about the distribution of designs in the field. Two of its findings were more favourable to existing work than anticipated: multi-unit designs dominate, and no instance of the arithmetic error searched for was identified. A larger coding exercise might revise these findings.

The hosting regime rests on three months in the attraction panel and seven in the city panel. Published headers for three of the four provinces declare coverage we do not in fact have. The power analysis indicates that additional months would not alter the conclusion, but this is an extrapolation from the observed noise structure rather than a direct observation.

We did not carry forward the attention index used in the original commissioned study, because it has no municipal variation and is absorbed by year-month fixed effects. A genuinely city-varying attention measure — city-level search volume or geo-tagged media — would allow the mediation question to be addressed directly, and we regard this as a worthwhile extension.

Two directions for further work follow. The first is to apply this set of diagnostics to studies whose data are public, extending the single case into a replication exercise; the coding sheet records which studies permit this. The second is to ask, prospectively, what class of event evaluations can be identified at all. Where a single host is unavoidable, designs that pool events — several summits, several host cities, staggered over time — restore the treated variation on which the modern toolkit depends, at the cost of the specificity that made municipal data attractive in the first place. This trade-off is better made explicitly at the design stage than discovered after estimates are published.

Data availability

The panels are compiled from Korean public statistical sources that restrict onward redistribution, so the underlying files cannot be posted. Analysis code and the derived series required to reproduce every table and figure will be deposited in a public repository upon acceptance.

References

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

Abadie, A., Diamond, A., and Hainmueller, J. (2010). 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.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.

Boto-García, D., Anguera-Torrell, O., and Escalonilla, M. (2025). When Taylor Swift hits town: Megastar-driven effects on hotel performance. *Tourism Economics*.

Boto-García, D. and Santana-Gallego, M. (2026). Old questions, new methods: Revisiting the economic effects of hosting mega-sport events. *Tourism Management*, 113:105324.

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

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.

Chernozhukov, V., Wüthrich, K., and Zhu, Y. (2021). An exact and robust conformal inference method for counterfactual and synthetic control methods. *Journal of the American Statistical Association*, 116(536):1849–1864.

Conley, T. G. and Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, 93(1):113–125.

Contu, G. and Pau, S. (2022). The impact of TV series on tourism performance: the case of Game of Thrones. *Empirical Economics*, 63:3313–3341.

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

Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. NBER Working Paper 22791, National Bureau of Economic Research.

Eggers, A. C., Tuñón, G., and Dafoe, A. (2024). Placebo tests for causal inference. *American Journal of Political Science*, 68(3):1106–1121.

Falk, M. and Hagsten, E. (2017). Measuring the impact of the European Capital of Culture programme on overnight stays: evidence for the last two decades. *European Planning Studies*, 25(12):2175–2191.

Ferman, B. and Pinto, C. (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, 12(4):1197–1221.

Ferman, B., Pinto, C., and Possebom, V. (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management*, 39(2):510–532.

Firgo, M. (2021). The causal economic effects of Olympic Games on host regions. *Regional Science and Urban Economics*, 88:103673.

Fourie, J. and Santana-Gallego, M. (2011). The impact of mega-sport events on tourist arrivals. *Tourism Management*, 32(6):1364–1370.

Fourie, J. and Santana-Gallego, M. (2022). Mega-sport events and inbound tourism: New data, methods and evidence. *Tourism Management Perspectives*, 43:101002.

Getz, D. (2008). Event tourism: Definition, evolution, and research. *Tourism Management*, 29(3):403–428.

Gomes, P. and Librero-Cano, A. (2018). Evaluating three decades of the European Capital of Culture programme: a difference-in-differences approach. *Journal of Cultural Economics*, 42:57–73.

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

Hu, P. and Wang, D. D. (2025). The impact of war on tourism: A synthetic control and causal configuration analysis. *Annals of Tourism Research*, page 104014.

Kobierecki, M. M. and Pierzgalski, M. (2022). Sports mega-events and economic growth: A synthetic control approach. *Journal of Sports Economics*, 23(5):567–597.

Lahura, E. and Sabrera, R. (2023). The effect of infrastructure investment on tourism demand: a synthetic control approach for the case of Kuelap, Peru. *Empirical Economics*, 65:443–478.

MacKinnon, J. G. and Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135.

Preuss, H. (2007). The conceptualisation and measurement of mega sport event legacies. *Journal of Sport & Tourism*, 12(3-4):207–228.

Rambachan, A. and Roth, J. (2023). A more credible approach to parallel trends. *Review of Economic Studies*, 90(5):2555–2591.

Roth, J. (2024). Interpreting event-studies from recent difference-in-differences methods.

Roth, J., Sant'Anna, P. H. C., Bilinski, A., and Poe, J. (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, 4(11):1208–1214.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199.

Tokarchuk, O. and Gabriele, R. (2026). Placebo analysis for causal inference in tourism. *Tourism Management*, 114:105360.

Vierhaus, C. (2019). The international tourism effect of hosting the Olympic Games and the FIFA World Cup. *Tourism Economics*, 25(7):1009–1028.

Wallimann, H. and Mehr, A. (2026). The impact of the UEFA Women's EURO on hotel overnight stays: Evidence from a causal analysis. *Journal of Sports Economics*.

Wan, S. K. and Song, H. (2019). Economic impact assessment of mega-events in the United Kingdom and Brazil. *Journal of Hospitality & Tourism Research*, 43(7):1044–1067.

Wieland, T. (2026). Assessing the impact of tourist attractions through the integration of causal inference and demand-side economic analysis: A case study of the Sensoria experience museum in Holzminden, Germany. Working paper, version 2.0.0. Replication code and data: `https://github.com/geowieland/sensoria`.

Zhang, Y. and Zhang, J. (2023). Tourist attractions and economic growth in China: A difference-in-differences analysis. *Sustainability*, 15(7):5649.

Table A1: APEC 2025 event calendar and derived treatment dates

Date	Location	Event	Bears on dating
2023-11	San Francisco	Korea confirmed as 2025 APEC host economy	
2024-06-27	Seoul	Gyeongju designated host city	✓
2024-07	–	First full month of the announcement regime (T_0)	✓
2025-02-24–2025-03-09	Gyeongju	First Senior Officials' Meeting cluster, c. 1,500 delegates	✓
2025-10-27–2025-11-01	Gyeongju	APEC Economic Leaders' Week	✓
2025-10-31–2025-11-01	Gyeongju	APEC Economic Leaders' Meeting	✓
2025-10	–	First month of the hosting regime (T_1)	✓
2025-11-01	Gyeongju	Gyeongju Declaration adopted	
2025-12	–	Last month of attraction-panel coverage (Panel B)	
2026-04	–	Last month of city-panel coverage (Panel A)	

Treatment dates used throughout: announcement regime from 2024–07, hosting regime from 2025–10. The announcement regime is 15 months long.

Leaders' Week straddles the month boundary, with five of its six days in October. We therefore treat October 2025 as the summit month. Dating the hosting regime from November instead would place the summit month in the pre-period and reduce the attraction panel's hosting window from three months to two.

The February–March 2025 Senior Officials' Meeting cluster sits inside the announcement regime and is not given a separate indicator.

Appendix

This appendix reports the material referenced in the text: the event calendar the treatment dates are derived from (Table A1); the two versions of the attraction panel and what the choice between them costs (Table A2); the weights that the synthetic control and synthetic difference-in-differences place on donors and pre-periods (Tables A3 and A4); leave-one-out donor sensitivity (Table A5); heterogeneity-robust estimators set against two-way fixed effects (Table A6); the five candidate legacy metrics (Table A7); the original mechanism analysis under progressively stronger controls (Table A8); and the placebo evidence in full (Table A9).

Table A2: Balanced and full versions of the attraction panel

					Synthetic control				DID	
Panel	**Units**	**Obs.**	**Pre-RMSPE**	**Med. donor**	**ATT ann.**	p	**ATT host.**	p	**ATT ann.**	p
balanced	36	3,456	0.164	0.228	+12.3%	0.588	+38.8%	0.265	+25.0%	0.361
full	55	5,280	0.132	0.260	+7.5%	0.735	+21.3%	0.531	+33.2%	0.655

The balanced panel keeps only attractions observed in every month of the window, and only municipalities whose balanced total is strictly positive throughout; pandemic closures make the second condition binding, and it costs nineteen municipalities. The full panel keeps every attraction and absorbs closure months with a log-plus-one transform.

Neither version is the correct one. The balanced panel buys a fixed reporting roster at the cost of comparison units; the full panel does the reverse. Both are carried through every specification in the paper and the qualitative conclusions do not depend on the choice.

The two estimators are reported in separate column blocks and should not be read across. Synthetic control p-values are donor-placebo permutations; the DID p-value is a randomization test over unit assignments. The estimands also differ: the synthetic control targets Gyeongju against a weighted donor composite, the DID against the unweighted remainder of the panel.

Table A3: Synthetic control donor weights (demeaned specification)

Donor	**Weight**
balanced panel	
Changwon (South Gyeongsang)	0.199
Mungyeong (North Gyeongsang)	0.153
Jecheon (North Chungcheong)	0.114
Cheongju (North Chungcheong)	0.105
Yeongju (North Gyeongsang)	0.103
Chungju (North Chungcheong)	0.101
Uljin (North Gyeongsang)	0.067
Yeongcheon (North Gyeongsang)	0.063
Pohang (North Gyeongsang)	0.053
Uiseong (North Gyeongsang)	0.042
sum	*1.000*
full panel	
Changwon (South Gyeongsang)	0.178
Sacheon (South Gyeongsang)	0.163
Cheongju (North Chungcheong)	0.110
Pohang (North Gyeongsang)	0.090
Yangyang (Gangwon)	0.076
Mungyeong (North Gyeongsang)	0.074
Gangneung (Gangwon)	0.068
Geochang (South Gyeongsang)	0.067
Hoengseong (Gangwon)	0.062
Cheongsong (North Gyeongsang)	0.030
Gimcheon (North Gyeongsang)	0.027
Chungju (North Chungcheong)	0.019
Yecheon (North Gyeongsang)	0.019
Goseong (South Gyeongsang)	0.009
Jeungpyeong (North Chungcheong)	0.003
Seongju (North Gyeongsang)	0.003
Eumseong (North Chungcheong)	0.002
sum	*1.000*

Weights from the demeaned synthetic control fitted to Gyeongju's pre-treatment path, 2018-01 to 2024-06. Donors receiving less than 0.001 are omitted.

No single donor dominates: the largest weight is Changwon at 0.199 in the balanced panel and 0.178 in the full panel. Table A5 reports what happens when each is removed.

Table A4: Synthetic difference-in-differences weights

Donor / period	Weight
Unit weights $\hat{\omega}$ (top 8)	
Geochang (South Gyeongsang)	0.031
Sacheon (South Gyeongsang)	0.030
Pohang (North Gyeongsang)	0.028
Cheongju (North Chungcheong)	0.028
Changwon (South Gyeongsang)	0.028
Hoengseong (Gangwon)	0.027
Yangyang (Gangwon)	0.027
Gimcheon (North Gyeongsang)	0.027
Time weights $\hat{\lambda}$ (top 8)	
2024-06	0.409
2024-02	0.274
2024-05	0.208
2024-01	0.042
2022-07	0.035
2018-10	0.028
2020-08	0.004

Full panel, announcement regime. Synthetic difference-in-differences (Arkhangelsky et al., 2021) weights pre-periods as well as units.

The time weights are the substantively interesting half: they place positive weight on only a handful of the 78 available pre-periods, discarding stretches in which the donors least resemble Gyeongju. That concentration is why the synthetic difference-in-differences estimate is smaller than the synthetic control estimate — part of what the latter reads as an effect is a pre-existing divergence.

Table A5: Leave-one-out donor sensitivity

Specification	Donors	ATT ann.	ATT host.	Sign flips	Weight
balanced panel					
baseline (all donors)	35	+12.3%	+38.8%		
leave-one-out range		[+7.1, +13.6]	[+31.8, +44.0]	0/35	
drop Changwon		+7.1%	+31.8%		0.199
drop Cheongju		+13.0%	+44.0%		0.105
drop Yeongcheon		+12.5%	+33.8%		0.063
full panel					
baseline (all donors)	54	+7.5%	+21.3%		
leave-one-out range		[+1.6, +11.9]	[+19.2, +24.0]	0/54	
drop Cheongju		+8.0%	+24.0%		0.110
drop Geochang		+10.3%	+23.9%		0.067
drop Changwon		+1.6%	+19.2%		0.178

Each donor municipality is removed in turn and the demeaned synthetic control refitted. The three rows shown per panel are the donors whose removal moves the hosting estimate most.

There is no sign reversal in any iteration. The most influential donor is Cheongju, whose removal shifts the hosting estimate by 3.5 percentage points on the full panel. The estimates are not carried by any single comparison unit.

Table A6: Heterogeneity-robust estimators against two-way fixed effects

Panel	Regime	CS	SA	TWFE
balanced	announcement	0.1795	0.1795	0.1838
balanced	hosting	0.5448	0.5448	0.5491
full	announcement	0.1466	0.1466	0.3609
full	hosting	0.5160	0.5160	0.7303

Coefficients in logs. Callaway and Sant'Anna (2021) and Sun and Abraham (2021) exist to remove the negative weighting that arises under staggered adoption. There is no staggered adoption here — one unit, one date — so they have nothing to correct, and they return identical figures to four decimal places.

They do expose one thing the two-way fixed effects estimate hides. On the full panel the fixed-effects figure is more than twice theirs, because the group-time estimators anchor on the period immediately before treatment while two-way fixed effects anchors on the whole pre-period mean. With a pronounced pre-trend that implicit choice moves the estimate more than most choices researchers deliberate over. On the balanced panel, where the pre-trend is milder, they agree.

Standard errors are omitted deliberately: with one treated unit they are not interpretable, which is the subject of Table 3.

Table A7: The same event, five candidate legacy metrics

Candidate legacy metric	Ann.	p	Host.	p
native window				
1. Navigation search	+17.0%	0.30	+13.9%	0.20
2. Visitors	+0.2%	1.00	+1.1%	0.70
3. Tourism spending	+3.5%	0.50	+2.3%	0.90
4. Foreign visitors	+34.3%	0.30	+9.5%	0.60
5. Foreign spending	+69.0%	0.20	+25.0%	0.80
common 2022-01+ window				
1. Navigation search	+8.1%	0.40	+5.3%	0.40
2. Visitors	+1.7%	0.80	+2.6%	0.90
3. Tourism spending	+1.5%	0.60	+0.2%	1.00
4. Foreign visitors	+34.3%	0.30	+9.5%	0.60
5. Foreign spending	+52.8%	0.20	+13.0%	0.80

Gyeongbuk versus non-Gyeongbuk, one specification, five outcomes measured on the same cities over the same months. p-values are randomization inference over city-role assignments; with five cities the attainable floor is 0.1.

No estimate is significant and no pairwise difference between metrics is significant either, so this is not evidence that the funnel narrows. The descriptive point is what survives: over the common window the announcement-regime estimate a researcher would report ranges from $+1.5\%$ to $+52.8\%$ depending only on which series is called the legacy.

Table A8: The original mechanism analysis under stronger controls

Rung	Specification	π	p	Direct	β	δ	Mediated
M1	report spec (city + month-of-year, cluster SE on 3 cities)	+0.0029	0.050	+0.0016	+0.1246	+0.8472	18.0%
M2	same, Newey-West HAC(12) SE	+0.0029	0.158	+0.0016	+0.1246	+0.8472	18.0%
M3	+ linear and quadratic time trend	+0.0051	0.004	+0.0008	+0.3526	+0.8529	69.1%
M4	+ year fixed effects	+0.0077	0.000	+0.0003	+0.3321	+0.8156	89.7%
M5	+ full year-month fixed effects	absorbed	–	absorbed	+0.2362	+0.8783	–
M6	report spec, COVID years dropped (2020-2022)	+0.0034	0.041	+0.0007	+0.1482	+0.7856	42.7%
M7	report spec in first differences	+0.0078	0.000	-0.0001	+0.4686	+0.7473	102.9%

Re-estimation of the mediation chain reported in the original commissioned study, using the authors' own code. M1 is their specification: city dummies, month-of-year dummies and macro controls, standard errors clustered on three cities.

M5 is the diagnostic rung. The attention index has no municipal variation, so a full set of year-month fixed effects absorbs it by construction: regressing it on the other regressors returns an R^2 of 1.000000 and the design matrix has rank 102 of 117 columns. A coefficient that exists only when time effects are omitted is a correlation with the national tourism cycle, not a mechanism.

δ, the visit-to-spending elasticity, is identified from within-city variation and is stable across every rung. It is the one quantity in this section that survives.

The original report quotes $\pi = 0.0039$ and a mediated share of 24.9%. Re-running the same code now returns 0.0029 and 18.0%: the panel was extended after that section was estimated and it was not refreshed before the report was finalised.

Table A9: Placebo evidence in full

Specification	n	Effect	p (joint)	SDID / p (unit)	Fit / p (date)
Panel A. Placebo in time					
real (2024-07) (balanced)	15	+12.3%	0.611	+1.8%	0.164
placebo (2021-07) (balanced)	15	+32.2%	0.278	+16.3%	0.143
placebo (2022-07) (balanced)	15	+17.7%	0.278	+15.0%	0.159
placebo (2023-07) (balanced)	12	+18.4%	0.417	+16.2%	0.161
real (2024-07) (full)	15	+7.5%	0.745	-0.7%	0.132
placebo (2021-07) (full)	15	+8.9%	0.691	+9.2%	0.092
placebo (2022-07) (full)	15	+0.5%	0.964	+3.2%	0.099
placebo (2023-07) (full)	12	+34.9%	0.327	+31.6%	0.095
Panel B. Joint placebo over unit and date					
balanced panel	2,088	+12.3%	0.572	0.611	0.931
full panel	3,190	+7.5%	0.795	0.745	0.741

Panel A moves the treatment date to dates on which nothing happened, truncating the sample before the true designation each time. Columns report the number of post-treatment months, the demeaned synthetic control estimate, its placebo p-value, the synthetic difference-in-differences estimate and the pre-treatment RMSPE.

All three fake dates on the balanced panel produce larger effects than the real one, with smaller p-values and comparable pre-treatment fit, so this is not an artefact of the estimator fitting the fake dates badly.

Panel B refits the estimator for every municipality at every feasible date. Columns report the number of fits, the actual estimate, and its rank expressed as a p-value against the joint distribution, against the unit dimension alone, and against the date dimension alone. The conventional placebo tests are the last two columns; each holds one researcher choice fixed.