

Testing Governance Mediation with Multi-Edition City Indices: A Simulation-Validated Pipeline with a Confounding Diagnostic and Sensitivity Analysis

Kamal Abdelgowad

Independent researcher

Baher ElShaarawy

baher@creatrox.info

Independent researcher

Muhammad Essa

King Saud University

Asra Mehar

University of Nizwa

Kiran Kumar

University of Chicago

Research Article

Keywords: Smart cities, Digital governance, Mediation analysis, Monte Carlo simulation, Panel data, Composite indicators, Reproducibility

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11304542/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Testing Governance Mediation with Multi-Edition City Indices: A Simulation-Validated Pipeline with a Confounding Diagnostic and Sensitivity Analysis

Kamal Abdelgowad¹, Cairo University, Egypt.

Baher ElShaarawy², Independent Researcher, Email: baher@creatrox.info

Muhammad Essa³, King Saud University, KSA.

Asra Mehar⁴, University of Nizwa, Oman.

Dr. Kiran Kumar⁵, University of Chicago, USA

Abstract

Composite city indices such as the IESE Cities in Motion Index (CIMI) are increasingly used to test whether institutional governance mediates the association between urban technology and sustainability. These multi-edition panels contain repeated observations of the same cities, persistent unobserved city characteristics, missing scores and changing indicator definitions. We specify an eight-stage open pipeline for mediation analysis on such panels. We then evaluate it in a Monte Carlo study on the CIMI schema (183 cities, five editions, 38,500 simulated panels). When residuals clustered by city, Sobel tests with HC3 standard errors rejected a true null indirect effect in up to 11.2% of replications. The city-cluster bootstrap kept Type I error near 5%. An unobserved city characteristic affecting governance and environment inflated pooled indirect effects by 126–330%. A two-way fixed-effects (FE) mediation model stayed within 1.2% of the truth under the same confounding. A paired bootstrap contrast between pooled and FE estimates flagged confounding in 63% of replications at 36% pooled bias and in at least 97% at 69% bias. Its false-positive rate was 4.5–6.2%. Time-varying confounding biased the FE estimate by 26–87%. A closed-form residual-correlation sensitivity function recovered the true effect when the confounding correlation was supplied (bias within 1.4%, coverage 0.94–0.95). It also reports the correlation needed to explain an effect away. Single deterministic imputation inflated the indirect effect by up to 13%, and within-edition standardisation corrected rescaling but not changes in indicator content. The resulting protocol states which estimators, diagnostics and sensitivity analyses should precede mediation claims from city indices. No result concerns real cities.

Keywords: Smart cities; Digital governance; Mediation analysis; Monte Carlo simulation; Panel data; Composite indicators; Reproducibility

1. Introduction

Cities are now routinely ranked on composite indicators that combine technological, institutional and environmental scores (United Nations Department of & Social, 2024). The IESE Cities in Motion Index (CIMI) is a prominent example. Its 2022 edition ranks 183 cities across nine dimensions, including technology, governance and environment (School, 2022). The concept these indices measure remains contested (Kitchin, 2014), and indicator frameworks differ in which dimensions they include and how they weight them (Ahvenniemi et al., 2017; Huovila et al., 2019; Sharifi, 2019). Even so, because these indices are public, repeated and broad in coverage, they invite a question that case studies cannot answer: does institutional governance carry part of the association between urban technology and sustainability performance?

The conceptual case for such a pathway is well developed. Smart-city frameworks treat technology, institutions and people as interdependent rather than separable components (Albino et al., 2015; Nam & Pardo, 2011). Digital-government research argues that institutional capacity determines whether digital investment produces public value (Bannister & Connolly, 2014). Meijer and Bolívar (2016) concluded from their review that governance arrangements shape, rather than simply follow, technology adoption in cities. Case evidence points the same way: smart-city plans that neglect institutional and participatory processes tend to produce fragmented results (Angelidou, 2017; Appio et al., 2019; Castelnovo et al., 2016). Bibri and Krogstie (2017) describe governance as a structural link between technological capacity and sustainability, and Gil-Garcia et al. (2018) note that governance is rarely modelled as an intervening variable in empirical work.

Turning this proposition into a statistical test on index data is less straightforward than it appears. A multi-edition index such as CIMI is a panel, and three of its properties interfere with the standard mediation test. First, the same cities appear in every edition, so residuals are likely to be correlated within cities. Inference that treats 915 city-years as independent observations can then overstate precision (Bertrand et al., 2004; Cameron & Miller, 2015). Second, cities differ in persistent characteristics, such as administrative tradition or wealth, that are not captured by the index. If such a characteristic raises both governance and environmental scores, a pooled regression attributes its influence to the governance-to-environment path. Multilevel mediation research has shown that conflating within-unit and between-unit variation can bias indirect effects substantially (Krull & MacKinnon, 2001; Zhang et al., 2009). Third, the indicators and weights behind each CIMI

dimension change between editions, and coverage varies. This is a recognised difficulty for longitudinal use of composite indicators (Organisation for Economic & Development, 2008; Saisana et al., 2005).

Each of these problems has a known remedy in isolation. Missing values can be imputed by chained equations (van Buuren, 2018); dependence can be handled by resampling whole clusters (Field & Welsh, 2007); and persistent unit-level confounding can be removed by within-unit estimation (Wooldridge, 2010). What is missing is evidence on how these remedies perform together, at the sample size and edition structure of a real city index, and on which of the three problems matters most. Simulation studies of mediation tests exist (MacKinnon et al., 2002), but they generally assume independent observations and complete data. We are not aware of a study that evaluates mediation inference under the clustering, confounding, missingness and edition changes that characterise city-index panels.

An earlier version of this work validated a mediation pipeline on a single synthetic CIMI-schema panel with a planted pathway. That exercise confirmed that the code ran correctly, but it could not show whether the pipeline would produce valid inference on real data. The generator contained no city-level dependence or confounding, so almost any estimator would have recovered the planted effect. The present study replaces the single draw with five Monte Carlo experiments that deliberately introduce the conditions under which mediation tests are expected to fail. It also adds two tools that the earlier pipeline lacked: a test for persistent confounding and a sensitivity analysis for time-varying confounding (Imai, Keele, & Yamamoto, 2010).

The study makes five contributions. (1) It specifies an eight-stage, open-source pipeline for mediation analysis in multi-edition index panels. (2) It reports the bias, Type I error, power and interval coverage of six inference strategies under independent, clustered and confounded city structures. (3) It proposes and evaluates a paired-bootstrap contrast between pooled and within-city indirect effects, which turns a common informal comparison into a test with known error rates. (4) It derives a closed-form sensitivity function for time-varying confounding of the mediator-outcome path under the within-city estimator, and validates it by simulation. (5) It quantifies the effects of missing scores and changing indicator definitions. From these results, we derive a decision protocol for analysts who wish to test governance mediation on real CIMI extracts.

Section 2 reviews related work. Section 3 describes the pipeline and the simulation design. Section 4 reports the illustrative single-panel analysis and the Monte Carlo results. Section 5 discusses implications, Section 6 sets out limitations, and Section 7 concludes.

2. Related Work

2.1 Governance as an intervening variable in smart-city research

Early definitional work established that smart-city performance is multidimensional. Caragliu et al. (2011) and Albino et al. (2015) treated technology, governance and sustainability as co-dependent dimensions, using cross-sectional comparisons. Their designs could describe co-occurrence but could not separate a direct technology-sustainability association from one routed through governance. Later work examined how these dimensions appear in city strategies. Angelidou (2017) found that most of fifteen municipal plans emphasised technology and gave little attention to the institutional processes needed to translate it into outcomes, a pattern also noted in the innovation-ecosystem review of Appio et al. (2019).

Governance-focused reviews shifted attention from the technology itself to the institutional arrangements around it. Meijer and Bolívar (2016) argued that governance configurations condition whether digital investment yields sustainability outcomes, and Castelnovo et al. (2016) reached a similar conclusion from a multi-case assessment of European programmes. Bibri and Krogstie (2017) conceptualised governance as a structuring mediator in smart and sustainable cities. Gil-Garcia et al. (2018) called for empirical tests that link digital government and public management research. Across this literature, the mediating role of governance is argued from qualitative or small-N evidence; it is rarely estimated, and its uncertainty is rarely quantified (Table 1).

2.2 Statistical problems in panel mediation with index data

The product-of-coefficients approach estimates an indirect effect as the product of the technology-to-governance path (a) and the governance-to-outcome path (b) (Baron & Kenny, 1986). Its sampling distribution is not normal, which motivated resampling-based intervals (Preacher & Hayes, 2008). MacKinnon et al. (2002) compared fourteen tests by simulation and found that the Sobel test is conservative for small effects, whereas resampling methods have better power. The sample size needed to detect an indirect effect also rises steeply as the a and b paths shrink (Fritz & MacKinnon, 2007). These studies assumed independent observations.

City-index panels violate that assumption. When observations are grouped, conventional and heteroskedasticity-consistent standard errors understate sampling variability, sometimes severely (Bertrand et al., 2004; Cameron & Miller, 2015). The usual solution is resampling whole clusters, which preserves the dependence structure (Cameron et al., 2008; Field & Welsh, 2007). Grouping also presents a second, more serious, problem. Multilevel mediation in which the observed

association between mediator and outcome incorporates the within-group and the between-group component(s). If unobserved unit-level characteristics drive both, the pooled estimate of b is biased (Krull & MacKinnon, 2001; Preacher et al., 2010; Zhang et al., 2009). Econometric panel methods address this by removing time-invariant unit effects (Bell & Jones, 2015; Mundlak, 1978; Wooldridge, 2010). Causal mediation theory formally outlines the same criteria: The indirect effect is shown if the mediator-outcome relationship is free from unmeasured mediating confounding (Imai, Keele, & Tingley, 2010). The same authors suggested sensitivity analysis using the correlation of the disturbances of the mediator and outcome models. Panel estimators and sensitivity analysis, therefore, address different threats: the first removes time-invariant confounding, while the second asks how much time-varying confounding an estimate can withstand.

Missing data and changing indicator definitions add further uncertainty. Chained-equations imputation is widely used (van Buuren, 2018), but single imputation ignores imputation uncertainty, and combining multiple imputation with bootstrap inference requires care (Rubin, 1987; Schomaker & Heumann, 2018). For time-series cross-section data, imputation models that exploit unit-level information perform better than those that ignore panel structure (Honaker & King, 2010). Changes in indicator sets and weights between editions are a known source of non-comparability in composite indices (Organisation for Economic & Development, 2008; Saisana et al., 2005), but their consequences for mediation estimates have not been examined.

Simulation studies are the accepted way to establish how a statistical procedure behaves when the truth is known (Morris et al., 2019). Table 1 positions the present study against the substantive literature: it does not test the governance pathway in real cities, but it establishes the conditions under which such a test would be informative.

Table 1. Prior research on governance in smart cities and the scope of the present study.

Study	Method	Data	Key output	Limitation for testing mediation
Caragliu et al. (2011)	Cross-sectional descriptive/OLS typology	European cities, single year	Urban performance ranking	Cross-sectional; no governance pathway tested
Meijer and Bolívar (2016)	Systematic literature review	Published smart-governance studies	Qualitative synthesis	No quantitative mediation test
Castelnovo et al. (2016)	Multi-case assessment	Three European cities	Holistic governance framework	Small-N; no statistical inference
Angelidou (2017)	Content analysis of city plans	15 municipal smart-city plans	Presence/absence of characteristics	No outcome or mediation test

Study	Method	Data	Key output	Limitation for testing mediation
Bibri and Krogstie (2017)	Interdisciplinary literature review	Cross-disciplinary corpus	Narrative synthesis	Conceptual; no index tested
MacKinnon et al. (2002)	Monte Carlo comparison of mediation tests	Simulated independent observations	Type I error, power	No clustering, confounding or missingness
This study	Eight-stage mediation pipeline with confounding diagnostic and sensitivity analysis, evaluated in five Monte Carlo experiments	Synthetic CIMI-schema panels, 183 cities $\times$ 5 editions	Bias, Type I error, power, CI coverage of ab; diagnostic power; robustness value	Synthetic only; real CIMI extract not yet analysed

3. Materials and Methods

3.1 The CIMI schema

CIMI scores cities on nine dimensions: technology, governance, environment, economy, human capital, social cohesion, mobility and transportation, urban planning, and international outreach (School, 2022). The editions used as the template were 2014, 2017, 2019, 2020 and 2022. With 183 cities, a balanced panel therefore contains 915 city-years. A real extract had not been assembled at the time of analysis. All data in this study are synthetic panels that follow this schema; none of the numbers refer to real cities.

3.2 The analysis pipeline

The pipeline has eight stages. Stages 1–5 revise the original design; Stage 5 has been promoted from a robustness check to a co-primary estimator. Stages 6–8 are new.

Stage 1: Imputation. Missing dimension scores are imputed with a round-robin chained-equations procedure (scikit-learn IterativeImputer, Bayesian ridge regression, 15 iterations), which is functionally equivalent to MICE (van Buuren, 2018). The original pipeline used a single deterministic imputation. The revised pipeline uses $m = 5$ stochastic imputations. Its “panel-aware” variant adds leave-one-out city means of the observed scores and the edition index as auxiliary predictors (Honaker & King, 2010).

Stage 2: Within-edition standardisation. Each dimension is centred and scaled by its mean and standard deviation within each edition. Scores then express a city’s position relative to other cities in the same edition, which removes edition-specific shifts in location and scale caused by methodology revisions.

Stage 3: Mediation regression. With technology as the exposure (X), governance as the mediator (M), environment as the outcome (Y), and economy and human capital as controls (C), the model for city i in edition t is

$$M_{it} = \alpha_1 + aX_{it} + \boldsymbol{\gamma}'\mathbf{C}_{it} + \varepsilon_{1,it} \quad (1)$$

$$Y_{it} = \alpha_2 + c'X_{it} + bM_{it} + \boldsymbol{\delta}'\mathbf{C}_{it} + \varepsilon_{2,it} \quad (2)$$

The indirect effect is ab , and the direct effect is c' . For reference, the Sobel standard error is

$$SE(ab) = \sqrt{b^2 SE(a)^2 + a^2 SE(b)^2} \quad (3)$$

Stage 4: City-cluster bootstrap. Whole cities, with all their editions, are resampled with replacement ($B = 999$ in the simulations; 5,000 in the illustrative analysis). The 2.5th and 97.5th percentiles of the bootstrapped ab form the 95% confidence interval. Mediation is declared when the interval excludes zero.

Stage 5: Two-way fixed-effects (FE) mediation. Equations (1) and (2) are re-estimated with city and edition fixed effects. Implemented by demeaning all variables within cities and adding edition dummies, this estimator uses only within-city changes across editions. Time-invariant city characteristics, observed or not, therefore cannot bias a or b . Inference again uses the city-cluster bootstrap.

Stage 6: Pooled-within contrast. The difference $D = ab_{\text{pooled}} - ab_{\text{FE}}$. It is bootstrapped by applying the same city resampling weights to both models in every replicate. An interval that excludes zero indicates that between-city differences, rather than within-city change, drive the pooled estimate. This formalises the logic of the Mundlak (1978) and Hausman-type comparisons discussed by Bell and Jones (2015) for the indirect effect.

Stage 7: Edition-heterogeneity diagnostic. The indirect effect is estimated separately for the latest edition and for the earlier editions pooled. The difference is bootstrapped with the same city resampling weights applied to both subsets. An interval that excludes zero flags a possible change in what an index dimension measures.

Stage 8: Sensitivity to time-varying confounding. The FE estimator cannot remove a factor that varies across editions and affects both governance and environment. Such a factor correlates the disturbances of Equations (1) and (2); let ρ denote that correlation and assume the technology-governance path remains unconfounded. Within-city least squares then estimates $b + \rho\sigma_2/\sigma_1$, where

σ_1 and σ_2 are the disturbance standard deviations. The estimated outcome residual standard deviation, s_2 , equals $\sigma_2\sqrt{(1 - \rho^2)}$, and s_1 estimates σ_1 . A bias-adjusted indirect effect is therefore

$$ab(\rho) = \hat{a} \left[\hat{b} - \frac{\rho s_2}{s_1 \sqrt{1 - \rho^2}} \right] \quad (4)$$

and the correlation at which the adjusted effect reaches zero is

$$\rho_0 = \frac{k}{\sqrt{1 + k^2}}, \quad k = \frac{\hat{b} s_1}{s_2} \quad (5)$$

Intervals for $ab(\rho)$ are obtained by applying Equation (4) within each city-cluster bootstrap replicate. The robustness value ρ_0 states how strongly unobserved time-varying factors would need to correlate the two disturbances to explain the indirect effect away. This follows the residual-correlation sensitivity analysis developed for linear structural equation models (Imai, Keele, & Tingley, 2010), written here for the within-city estimator. Reporting the value at which a conclusion would be overturned mirrors the robustness value that Cinelli and Hazlett (2020) proposed for omitted-variable bias in regression.

The indirect effect estimated in Stages 3–5 is a statistical quantity. It acquires a causal interpretation only under the assumption of no unmeasured confounding of the X – M , X – Y , and M – Y relationships (Imai, Keele, & Tingley, 2010). The FE estimator relaxes this assumption for time-invariant confounders. Stage 8 quantifies, but does not remove, the threat from confounders that vary across editions.

3.3 Illustrative single-panel analysis

The original validation exercise is retained as an illustration of the pipeline’s output. One synthetic panel was generated with a fixed random seed (42). City baselines were drawn from a normal distribution (mean 50, SD 15). Technology scores equalled the baseline plus Gaussian noise (SD 8) and a linear drift of 0.6 points per year. Governance was generated as $0.50 \times \text{technology}$ plus Gaussian noise (mean 30, SD 10). Environment was generated as $0.35 \times \text{governance} + 0.15 \times \text{technology}$ plus Gaussian noise (mean 25, SD 12). The remaining dimensions were noisy linear functions of the city baseline or of the governance and economy dimensions, bounded to the 0–100 range. Five per cent of technology scores were deleted completely at random. Pooled estimation used HC3 standard errors (MacKinnon & White, 1985), the Sobel test (Sobel, 1982) and 5,000 city-cluster bootstrap resamples. Diagnostics comprised variance inflation factors (VIF), the

Breusch-Pagan test, the Durbin-Watson statistic, a city-and-year FE specification with city-clustered standard errors, and a median-split subgroup analysis on the economy dimension.

In this generator, all city-level dependence passes through technology. Governance and environment receive independent noise in every edition. The Monte Carlo design below refers to this structure as scenario S1.

3.4 Monte Carlo design

The simulation study was planned and reported following the ADEMP structure of Morris et al. (2019): aims, data-generating mechanisms, estimands, methods and performance measures.

Aims. To estimate the bias, Type I error, power and confidence-interval coverage of mediation inference in CIMI-schema panels under (E1) different forms of city-level dependence, (E2) different amounts and mechanisms of missing data, and (E3) a change in indicator definition in the final edition. Two further experiments evaluate the new stages: (E4) the power and false-positive rate of the pooled-within contrast across a range of confounding strengths, and (E5) the bias of the FE estimator under time-varying confounding and the performance of the Stage 8 sensitivity function.

Data-generating mechanisms. Each replicate contained 183 cities observed in five editions (2014, 2017, 2019, 2020, 2022). For city i with persistent baseline $u_i \sim N(0,1)$, edition t and year y_t :

$$\text{econ}_{it} = 0.6u_i + 0.8e, \quad \text{hc}_{it} = 0.5u_i + 0.3 \text{econ}_{it} + 0.8e$$

$$X_{it} = 0.7u_i + 0.06(y_t - 2014) + \sqrt{0.51} e$$

$$M_{it} = aX_{it} + 0.10 \text{econ}_{it} + 0.10 \text{hc}_{it} + \tau v_i^M + \kappa \eta_i + e$$

$$Y_{it} = 0.15X_{it} + bM_{it} + 0.10 \text{econ}_{it} + 0.05 \text{hc}_{it} + \tau v_i^Y + \kappa \eta_i + e$$

where each e is an independent standard normal draw. Four further dimensions (social cohesion, mobility, urban planning, and international outreach) were generated as $0.5u_i + 0.3 \text{econ}_{it} + 0.8e$ and served as auxiliary variables for imputation. Three city-structure scenarios were crossed with four effect conditions:

- **S1 Independent** ($\tau = 0, \kappa = 0$): city dependence enters only through technology, as in the original generator.
- **S2 Clustered** ($\tau = 0.8, \kappa = 0$): independent city-level random intercepts $v_i^M, v_i^Y \sim N(0,1)$ in the governance and environment equations. Residuals are correlated within cities, but there is no confounding.

- **S3 Confounded** ($\tau = 0, \kappa = 0.6$): a shared, unobserved city characteristic η_i , correlated 0.5 with the baseline u_i . It raises both governance and environment.

The effect conditions were a null a path ($a = 0, b = 0.30$), a null b path ($a = 0.50, b = 0$), a small effect ($a = 0.20, b = 0.15$) and a medium effect ($a = 0.50, b = 0.30$).

Experiment E2 used scenario S2 with the medium effect. Technology scores were deleted at rates of 5%, 15% and 30% under three mechanisms. Under MCAR, deletion was completely random. Under MAR, the log-odds of deletion decreased by 1.5 per standard deviation of the economy score. Under MNAR, it decreased by 1.5 per standard deviation of the technology score itself, so that technologically weaker cities were more often unreported. The intercept of the logistic deletion model was calibrated to each target rate.

Experiment E3 used scenario S2 with the medium effect and no missing data. In the 2022 edition, a share $w \in \{0, 0.25, 0.50, 0.75\}$ of the variance of the environment score was replaced by an independent component of equal variance. This mimics a revision of the indicator set that changes what the dimension measures. In addition, technology and environment scores in every edition were rescaled by arbitrary edition-specific factors (1.0, 1.2, 0.8, 1.1, 1.4) and shifted by (0, 5, -3, 2, 8) points, to represent changes in scaling.

Experiment E4 crossed the strength of the shared city confounder, $\kappa \in \{0, 0.15, 0.30, 0.45, 0.60\}$, with city random intercepts of $\tau = 0.5$ and two effect conditions (null b path and medium effect). The confounder was correlated 0.5 with the city baseline. A sixth cell with $\tau = 0$ and $\kappa = 0$ reproduced scenario S1. Experiment E5 started from scenario S2 and added an independent time-varying factor. $\psi_{it} \sim N(0,1)$ with loading $\lambda \in \{0, 0.3, 0.6\}$ to the disturbances of both the governance and the environment equations. The resulting disturbance correlation is $\rho = \lambda^2 / (1 + \lambda^2)$, that is, 0, 0.083 and 0.265.

Estimand. The target was the indirect effect in within-edition standardised units, $ab \cdot \sigma_X / \sigma_Y$. The within-edition standard deviations were computed from a single population draw of 80,000 cities for each condition. Table 6 lists the resulting true values for E1. In E3, the estimand was the indirect effect before the indicator change. In E5, the population standard deviations included the variance contributed by the time-varying factor.

Methods. E1 compared six inference strategies. Four were applied to the pooled model: the Sobel test with classical OLS standard errors, the Sobel test with HC3 standard errors (the original specification), the Sobel test with city-clustered (CR1) standard errors, and a percentile bootstrap that resampled individual city-years. Two used the city-cluster percentile bootstrap: one on the

pooled model (the original Stage 4) and one on the two-way FE model (Stage 5). E2 compared four missing-data strategies, all followed by the pooled model with the city-cluster bootstrap. These were complete-case analysis, single deterministic imputation (the original Stage 1), multiple imputation with $m = 5$, and panel-aware multiple imputation with $m = 5$. For multiple imputation, 200 bootstrap draws per imputed data set were pooled into one distribution of 1,000 draws (Schomaker & Heumann, 2018). E3 applied the pooled cluster-bootstrap estimator and the Stage 7 diagnostic. E4 applied the pooled and FE estimators and the Stage 6 contrast. E5 applied the FE estimator without adjustment and with Equation (4) evaluated at the true ρ , and recorded the robustness value ρ_0 from Equation (5).

Performance measures. Bias was the mean estimate minus the true value, also expressed relative to the true value. Under null conditions, the rejection rate is the empirical Type I error; under non-null conditions, it is power. Coverage was the proportion of 95% intervals containing the true value. E1, E3, E4 and E5 used 1,000 replications per condition, giving a Monte Carlo standard error of about 0.007 for a rejection rate of 0.05 and for coverage of 0.95. E2 used 500 replications per condition (Monte Carlo standard error about 0.010 for coverage of 0.95). Rejection rates between 0.037 and 0.064, and coverage between 0.936 and 0.964, are within two Monte Carlo standard errors of nominal for E1.

3.5 Implementation

The illustrative analysis was run in Python 3.12 with pandas 3.0.2, NumPy 2.4.4, SciPy 1.17.1, statsmodels 0.15.0, scikit-learn 1.8.0, matplotlib 3.10.8 and seaborn 0.13.2, with the random seed fixed at 42 throughout. The Monte Carlo study was run in Python 3.13.16 with NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1 and scikit-learn 1.9.1. Every replicate used its own seed derived deterministically from the experiment, condition and replicate indices. Cluster bootstraps were computed from city-level cross-product matrices, which gives results identical to refitting each resample but is much faster. Estimators were checked against statsmodels on test panels. The full study (38,500 simulated panels) ran in about 90 minutes on a two-core CPU without a GPU.

4. Results

4.1 Illustrative single-panel analysis

The illustrative panel contained 915 city-years. Table 2 reports descriptive statistics in raw units, before imputation and standardisation. Shapiro-Wilk tests rejected normality for technology, human capital, social cohesion and mobility ($p < .05$) but not for the other five dimensions.

Table 2. Descriptive statistics and Shapiro-Wilk normality tests for the nine CIMI dimensions in the illustrative synthetic panel (raw score units).

Dimension	Mean	SD	Min	Median	Max	Shapiro-Wilk p	Normal ($\alpha = .05$)
Technology	51.92	15.50	9.56	51.24	100.00	0.002	No
Governance	56.12	12.36	23.60	56.47	94.29	0.319	Yes
Environment	51.99	13.21	5.91	52.40	93.96	0.333	Yes
Economy	49.09	16.70	0.00	48.29	99.43	0.102	Yes
Human capital	39.71	14.89	0.00	38.92	93.63	0.048	No
Social cohesion	29.57	14.19	0.00	29.04	81.13	0.001	No
Mobility	25.05	13.90	0.00	25.06	69.01	0.000	No
Urban planning	42.16	11.23	4.58	41.79	85.87	0.600	Yes
International outreach	34.21	12.74	0.00	34.29	75.88	0.549	Yes

Figure 1 shows the within-edition correlation structure. Technology, governance and environment were moderately to strongly correlated, as the generator implies. Economy and human capital correlated only weakly with these three dimensions, which supports their use as controls rather than as alternative mediators.

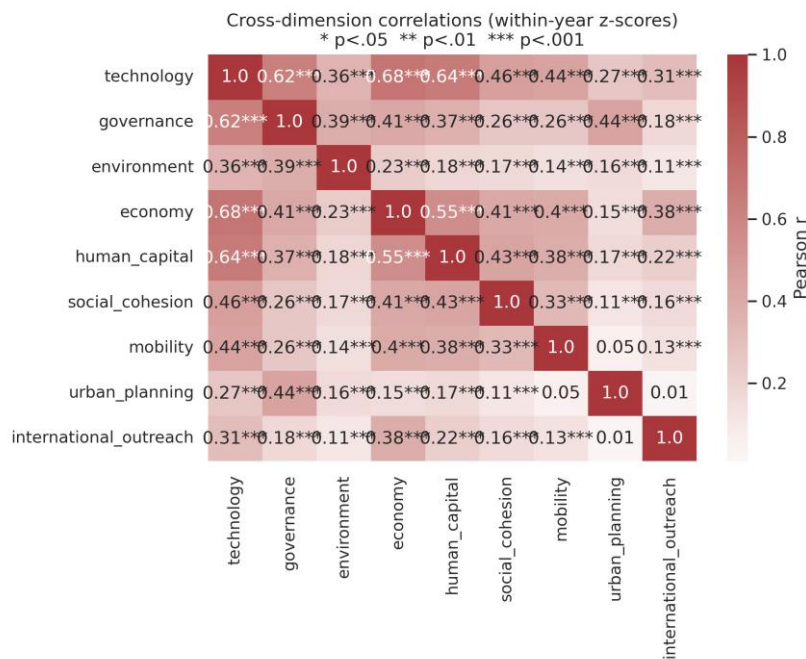


Figure 1. Within-edition standardised Pearson correlations among the nine CIMI dimensions across all synthetic city-years of the illustrative panel.

All variance inflation factors were below 3.0 (technology 2.998, governance 1.625, economy 1.918, human capital 1.785; Table 3), so multicollinearity is unlikely to have distorted the path estimates.

Table 3. Variance inflation factors for predictors in the mediator and outcome models (illustrative panel).

Variable	VIF
Constant	1.000
Technology	2.998
Governance	1.625

Variable	VIF
Economy	1.918
Human capital	1.785

Technology was positively associated with governance ($a = 0.657$, $SE = 0.039$, $p < .001$; $R^2 = 0.385$). In the outcome model, both governance ($b = 0.267$, $SE = 0.036$, $p < .001$) and technology ($c' = 0.238$, $SE = 0.053$, $p < .001$) were associated with environment ($R^2 = 0.176$). The estimated indirect effect was 0.175 ($SE = 0.026$; Sobel $z = 6.74$, $p < .00001$). The city-cluster bootstrap with 5,000 resamples gave a 95% interval of [0.123, 0.229] (Table 4; Figure 2).

Table 4. Path coefficients, standard errors, Sobel test and city-cluster bootstrap interval for the indirect effect (illustrative panel; standardised units).

Path	Estimate	SE	Sobel z	Sobel p	Bootstrap 95% CI
a : technology $\rightarrow$ governance	0.657	0.039	–	–	–
b : governance $\rightarrow$ environment	0.267	0.036	–	–	–
c' : technology $\rightarrow$ environment (direct)	0.238	0.053	–	–	–
Indirect effect ($a \times b$)	0.175	0.026	6.74	< .00001	[0.123, 0.229]

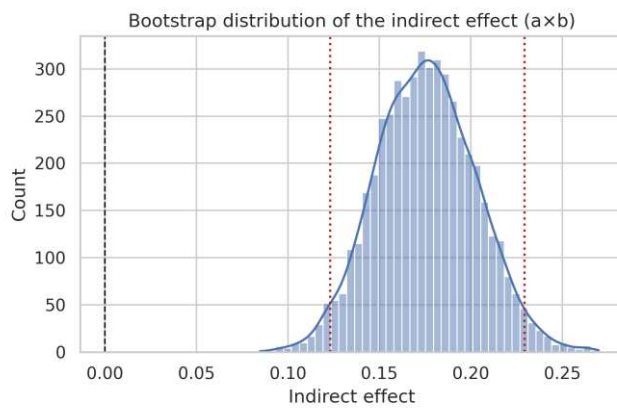


Figure 2. Distribution of the indirect effect across 5,000 city-cluster bootstrap resamples of the illustrative panel; vertical lines mark the 95% percentile interval.

The Breusch-Pagan test did not reject homoskedasticity ($LM = 4.43$, $p = .351$), and the Durbin-Watson statistic was 1.93 (Figure 3). HC3 standard errors were retained because they had been specified in advance, not because of this diagnostic. With city and year fixed effects, the technology-to-governance coefficient was 0.694 ($p < .001$; $R^2 = 0.516$), slightly larger than the pooled estimate of 0.657.

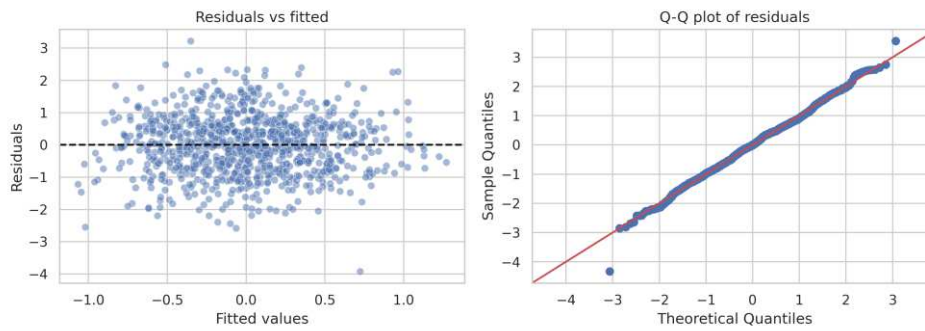


Figure 3. Residuals-versus-fitted and normal quantile-quantile plots for the outcome model of the illustrative panel.

In the median-split analysis (Table 5), technology and governance were associated with environment in both higher-economy cities ($n = 458$; technology $\beta = 0.247$, $p = .001$; governance $\beta = 0.223$, $p < .001$; $R^2 = 0.156$) and lower-economy cities ($n = 457$; technology $\beta = 0.224$, $p = .004$; governance $\beta = 0.316$, $p < .001$; $R^2 = 0.136$). The generator contained no subgroup difference, so the larger governance coefficient in lower-economy cities reflects sampling variation in one draw. Figure 4 summarises the fitted model.

Table 5. Outcome-model coefficients in higher- and lower-economy subgroups (median split on the economy dimension; illustrative panel).

Subgroup	n	Technology β	p	Governance β	p	R^2
Higher-economy cities	458	0.247	0.001	0.223	< .001	0.156
Lower-economy cities	457	0.224	0.004	0.316	< .001	0.136

Mediation model: indirect effect = 0.175, bootstrap 95% CI [0.123, 0.229]

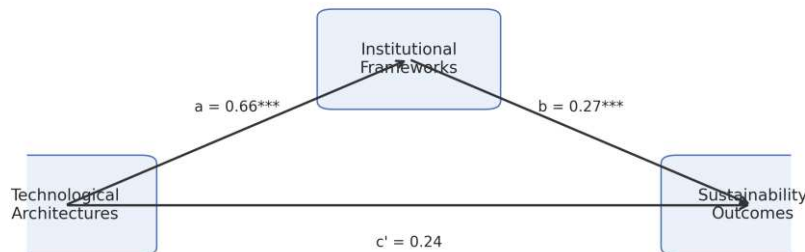


Figure 4. Fitted mediation model for the illustrative panel, showing standardised path coefficients and the city-cluster bootstrap 95% interval for the indirect effect.

This single panel shows what the pipeline outputs, but not whether those outputs can be trusted. Its generator corresponds to scenario S1, the most benign condition in the Monte Carlo study. The following subsections address that question.

4.2 E1: Inference under clustering and confounding

Table 6 reports the true indirect effects, and Table 7 and Figure 5 summarise the performance of the six inference strategies.

Table 6. True indirect effects in within-edition standardised units, by scenario and effect condition.

Effect condition	S1 Independent	S2 Clustered	S3 Confounded
Null a ($a = 0, b = 0.30$)	0.000	0.000	0.000
Null b ($a = 0.50, b = 0$)	0.000	0.000	0.000
Small ($a = 0.20, b = 0.15$)	0.028	0.023	0.023
Medium ($a = 0.50, b = 0.30$)	0.133	0.107	0.102

Table 7. Performance of six inference strategies for the indirect effect (E1; 1,000 replications per cell). Type I error is the rejection rate under a null hypothesis; power is the rejection rate under a small effect. † Original pipeline specification. ‡ Proposed co-primary estimator.

Scenario	Strategy	Type I ($a = 0$)	Type I ($b = 0$)	Power (small)	Coverage (small)	Coverage (medium)	Relative bias, medium (%)	Mean CI width (medium)
S1 Independent	Pooled, Sobel (OLS SE)	0.046	0.058	0.998	0.949	0.952	+0.1	0.070
S1 Independent	Pooled, Sobel (HC3 SE)	0.047	0.056	0.996	0.948	0.955	+0.1	0.070
S1 Independent	Pooled, Sobel (CR1 SE)	0.048	0.061	0.994	0.949	0.956	+0.1	0.070
S1 Independent	Pooled, city-year bootstrap	0.053	0.065	1.000	0.955	0.953	+0.1	0.070
S1 Independent	Pooled, city-cluster bootstrap	0.053	0.063	0.999	0.949	0.960	+0.1	0.070
S1 Independent	Two-way FE, city-cluster bootstrap	0.060	0.065	0.945	0.940	0.952	-0.3	0.084
S2 Clustered	Pooled, Sobel (OLS SE)	0.087	0.114	0.907	0.872	0.902	-0.1	0.062
S2 Clustered	Pooled, Sobel (HC3 SE)	0.089	0.112	0.901	0.873	0.899	-0.1	0.062
S2 Clustered	Pooled, Sobel (CR1 SE)	0.050	0.044	0.788	0.928	0.953	-0.1	0.074
S2 Clustered	Pooled, city-year bootstrap	0.099	0.126	0.951	0.889	0.899	-0.1	0.062
S2 Clustered	Pooled, city-cluster bootstrap	0.058	0.049	0.888	0.935	0.957	-0.1	0.074
S2 Clustered	Two-way FE, city-cluster bootstrap	0.054	0.055	0.948	0.941	0.944	-0.6	0.067
S3 Confounded	Pooled, Sobel (OLS SE)	0.875	1.000	1.000	0.000	0.000	+126.4	0.083
S3 Confounded	Pooled, Sobel (HC3 SE)	0.870	1.000	1.000	0.000	0.000	+126.4	0.083
S3 Confounded	Pooled, Sobel (CR1 SE)	0.848	1.000	1.000	0.000	0.000	+126.4	0.090
S3 Confounded	Pooled, city-year bootstrap	0.877	1.000	1.000	0.000	0.000	+126.4	0.082

Scenario	Strategy	Type I ($a = 0$)	Type I ($b = 0$)	Power (small)	Coverage (small)	Coverage (medium)	Relative bias, medium (%)	Mean CI width (medium)
S3 Confounded	Pooled, city- cluster bootstrap†	0.844	1.000	1.000	0.000	0.000	+126.4	0.091
S3 Confounded	Two-way FE, city-cluster bootstrap‡	0.056	0.071	0.941	0.946	0.941	+0.2	0.065

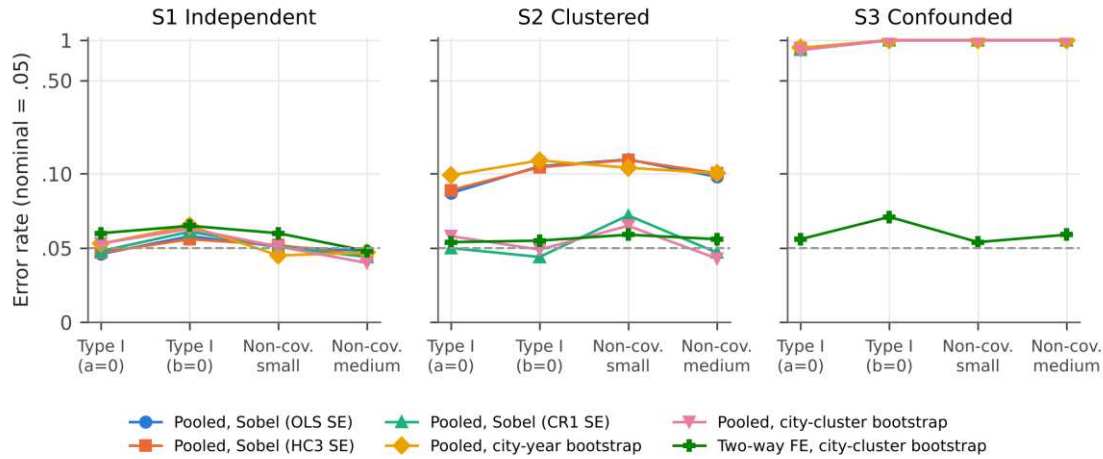


Figure 5. Error rates of six inference strategies by city-structure scenario (1,000 replications per point). The first two points in each panel are Type I error rates under a null a or b path; the last two are non-coverage rates ($1 - \text{coverage}$) of the 95% interval under small and medium effects. The dashed line marks the nominal 0.05. The vertical axis is linear up to 0.10 and logarithmic above it.

S1 Independent. When city-level dependence entered only through technology, all six strategies performed close to nominal, type I error ranged from 0.046 to 0.065, and coverage for the medium effect ranged from 0.952 to 0.960. Relative bias of the pooled estimate was below 0.3% in absolute value. Because the original illustrative generator has this structure, its successful recovery of the planted pathway (Section 4.1) would have been obtained with almost any estimator. The FE estimator was also unbiased here, but it discards between-city variation. Its intervals were 20% wider for the medium effect, and its power for the small effect fell to 0.945, against 0.999 for the pooled cluster bootstrap.

S2 Clustered. Adding independent city-level intercepts to the governance and environment equations left the pooled point estimates unbiased but invalidated inference that ignores clustering. Under a null b path, the Sobel test was rejected in 0.114 of replications with classical standard errors and 0.112 with HC3 standard errors, the specification used in the original analysis. The observation-level bootstrap was the worst performer at 0.126. For the medium effect, the intervals of these three methods covered the true value in only 0.899–0.902 of replications. Their intervals were about 16% narrower than the city-cluster bootstrap intervals, which explains the excess rejections. Both cluster-aware strategies restored nominal behaviour. The cluster bootstrap had

Type I error of 0.058 and 0.049 and medium-effect coverage of 0.957; the Sobel test with CR1 standard errors performed similarly (0.050, 0.044; coverage 0.953). The CR1 Sobel test, however, had less power than the cluster bootstrap for the small effect (0.788 versus 0.888), consistent with the known conservatism of the Sobel test for small indirect effects (MacKinnon et al., 2002). The FE estimator was the most powerful method in this scenario (0.948 for the small effect), because demeaning removes the city intercepts from the residual variance.

S3 Confounded. A shared unobserved city characteristic changed the picture entirely. All five pooled strategies share the same point estimate, and that estimate was strongly biased. For the medium effect, the mean pooled estimate was 0.232 against a true value of 0.102 (relative bias +126%). For the small effect, it was 0.097 against 0.023 (+330%). Coverage was 0.000 for every pooled method in both non-null conditions. When the true b path was zero, the pooled methods declared mediation in 100% of replications. When the true a path was zero, they did so in 84.4–87.7% of replications, because the confounder correlates with technology through the city baseline. Cluster-robust inference did not help: it corrects the standard error, not the bias. The FE estimator was nearly unbiased (relative bias +0.2% for the medium effect and -1.0% for the small effect), covered the true value in 0.941–0.946 of replications, and had Type I error of 0.056 and 0.071. The latter value is slightly above the upper Monte Carlo bound of 0.064, so the FE percentile interval may be mildly liberal when the b path is exactly zero.

4.3 E2: Missing technology scores

Table 8 and Figure 6 report bias and coverage under missing technology scores. The single deterministic imputation used in the original pipeline inflated the indirect effect, and the inflation grew with the missing rate. Under MCAR, relative bias was +2.5%, +6.6% and +12.7% at 5%, 15% and 30% missingness. Under MAR, it was +2.5%, +4.8% and +12.2%. At 30% missingness, coverage fell to 0.898 (MCAR) and 0.878 (MAR). The direction of the bias is expected. Deterministic imputation replaces each missing technology score with its predicted value from governance, environment and the other dimensions. The imputed values, therefore, lie exactly on the regression surface, which strengthens the observed technology-governance association and inflates a .

Complete-case analysis and both forms of multiple imputation were close to unbiased under MCAR and MAR (relative bias within $\pm 3\%$ at all rates). Multiple imputation gave narrower intervals than complete-case analysis at 30% missingness (mean width 0.075 versus 0.084 under MCAR), because it retains the observed governance and environment scores of incomplete city-years.

Coverage of all three methods dropped slightly below nominal at 30% missingness (0.922–0.930 under MCAR and MAR).

When weaker-technology cities were more likely to be unreported (MNAR), complete-case analysis and both multiple-imputation methods underestimated the indirect effect. At 30% missingness, relative bias was -11.4% for complete-case analysis, -6.9% for multiple imputation and -5.4% for panel-aware multiple imputation. Panel-aware imputation, which adds each city's leave-one-out mean scores from other editions, reduced this bias modestly and restored coverage to 0.950; under MCAR and MAR, it performed almost identically to standard multiple imputation. Single imputation showed small biases under MNAR at 5% and 15% missingness (-1.2% and +1.4%) and +7.0% at 30%. This apparent advantage arises only because its upward bias partly offsets the downward bias caused by the missingness mechanism, and it should not be read as robustness.

Table 8. Relative bias (%) and 95% interval coverage of the indirect effect under missing technology scores (E2; scenario S2, medium effect, true $ab = 0.107$; 500 replications per cell). All methods use the pooled model with city-cluster bootstrap. † Original pipeline specification.

Mechanism	Method	Bias 5%	Bias 15%	Bias 30%	Coverage 5%	Coverage 15%	Coverage 30%
MCAR	Complete case	+0.4	+0.5	-0.3	0.936	0.952	0.930
MCAR	Single imputation†	+2.5	+6.6	+12.7	0.940	0.936	0.898
MCAR	Multiple imputation	+0.4	0.0	-1.1	0.948	0.952	0.930
MCAR	Panel-aware MI	+0.5	+0.1	-0.9	0.950	0.954	0.930
MAR	Complete case	-0.1	-2.3	-2.9	0.944	0.958	0.924
MAR	Single imputation†	+2.5	+4.8	+12.2	0.948	0.942	0.878
MAR	Multiple imputation	+0.4	-1.6	-1.9	0.940	0.938	0.922
MAR	Panel-aware MI	+0.5	-1.4	-1.5	0.942	0.942	0.924
MNAR	Complete case	-4.8	-7.9	-11.4	0.928	0.926	0.916
MNAR	Single imputation†	-1.2	+1.4	+7.0	0.944	0.940	0.924
MNAR	Multiple imputation	-3.3	-5.2	-6.9	0.928	0.932	0.928
MNAR	Panel-aware MI	-2.6	-4.0	-5.4	0.930	0.942	0.950

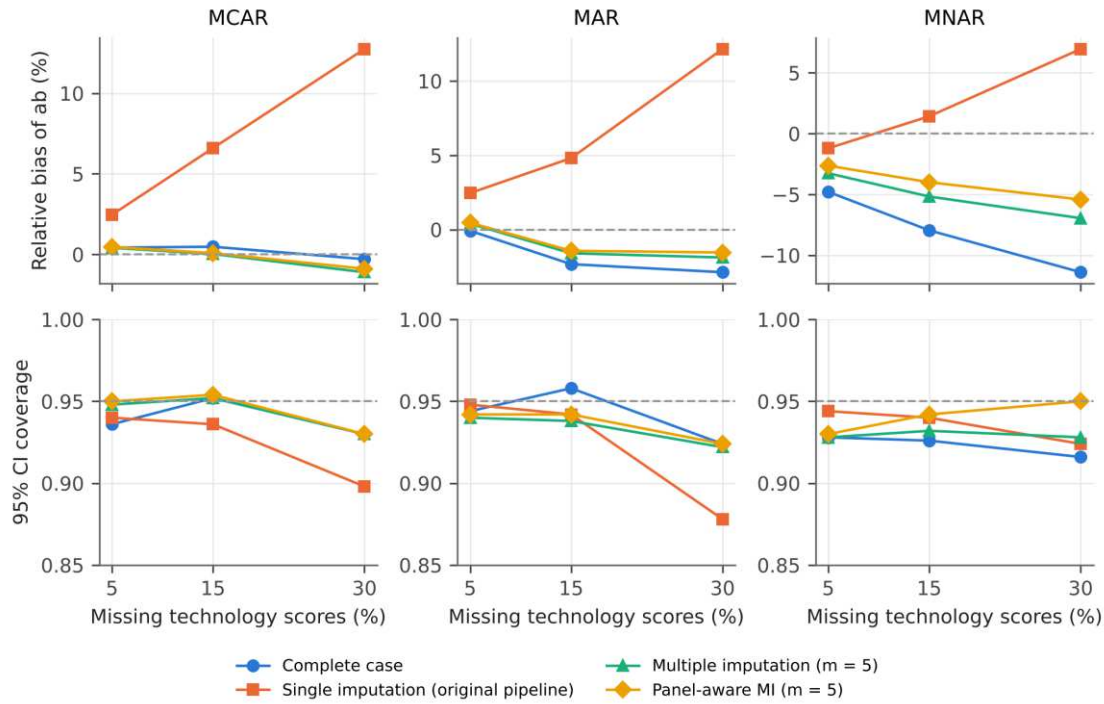


Figure 6. Relative bias (top) and 95% interval coverage (bottom) of the indirect effect under missing technology scores, by missingness mechanism and handling method (scenario S2, medium effect; 500 replications per point). Dashed lines mark zero bias and nominal 0.95 coverage.

4.4 E3: Change in indicator definition

Table 9 and Figure 7 report the effect of changing the 2022 environment indicator. With no change ($w = 0$), the pooled estimate was unbiased (-0.3% relative bias), and coverage was 0.960, even though every edition had been arbitrarily rescaled and shifted. This confirms that within-edition standardisation fully absorbs changes in the location and scale of a dimension.

A change in content is different. As the share of the 2022 environment score replaced by an unrelated component rose to 75%, the edition-specific estimate for 2022 fell from 0.107 to 0.053, while the estimate for 2014–2020 stayed at 0.106. Because 2022 contributes one-fifth of the panel, the pooled estimate was attenuated more modestly, to 0.096 (-10.4% relative bias), and coverage fell to 0.889. The edition-heterogeneity diagnostic flagged the change in 0.067, 0.155 and 0.307 of replications at 25%, 50% and 75% replacement. With $w = 0$, it flagged a change in 0.050 of replications, its nominal false-positive rate. A contrast based on one edition of 183 cities, therefore, has low power. A non-significant diagnostic cannot be taken as evidence that an indicator revision was harmless.

Table 9. Effect of replacing a share w of the 2022 environment score (E3; scenario S2, medium effect, true pre-change $ab = 0.107$; 1,000 replications per row). All editions were also arbitrarily rescaled and shifted.

w	Mean ab, 2014–2020	Mean ab, 2022	Mean ab, pooled	Relative bias, pooled (%)	Coverage, pooled	Diagnostic flags change
0.00	0.106	0.107	0.106	-0.3	0.960	0.050
0.25	0.107	0.094	0.104	-2.2	0.956	0.067
0.50	0.106	0.075	0.100	-6.4	0.921	0.155
0.75	0.106	0.053	0.096	-10.4	0.889	0.307

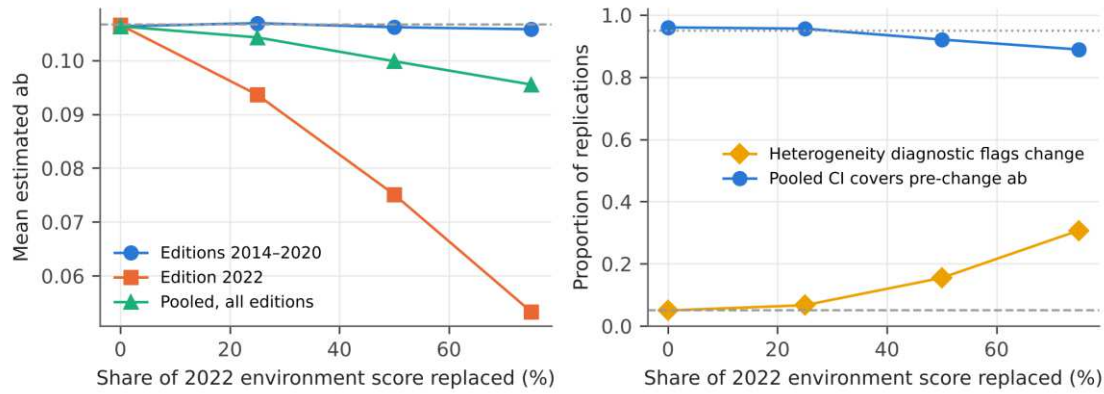


Figure 7. Effect of replacing part of the 2022 environment indicator (scenario S2, medium effect; 1,000 replications per point). Left: mean estimated indirect effect for 2014–2020, for 2022 alone and for the pooled panel; the dashed line marks the true pre-change value. Right: proportion of replications in which the edition-heterogeneity diagnostic flagged a change, and coverage of the pooled 95% interval; reference lines mark 0.05 and 0.95.

4.5 E4: Detecting persistent city confounding

Table 10 and Figure 8 summarise E4. Without a city confounder, the pooled-within contrast flagged a difference in 0.045–0.062 of replications, close to its nominal 5% level. As κ increased, the pooled indirect effect drifted upward: its relative bias for the medium effect was +12%, +36%, +69% and +111% at $\kappa = 0.15, 0.30, 0.45$ and 0.60 . The FE estimate stayed within 1.2% of the true value at every level. The contrast detected the confounding in 0.153, 0.632, 0.973 and 1.000 of replications at the same four levels. Detection under a null b path was slightly lower (0.449 at $\kappa = 0.30$), because a smaller indirect effect leaves less room for divergence.

The diagnostic, therefore, has a clear operating range. It reliably detects confounders large enough to inflate the pooled indirect effect by roughly two-thirds or more. It often misses confounding that inflates the estimate by 10–15%. A non-significant contrast should not be read as evidence that the pooled estimate is unbiased.

Table 10. Pooled and two-way FE indirect effects and the pooled-within contrast under increasing persistent city confounding (E4; 1,000 replications per row). τ is the SD of independent city intercepts, and κ is the loading of the shared city confounder.

τ	κ	True ab (medium)	Relative bias, pooled (%)	Relative bias, FE (%)	Contrast flags, medium	Contrast flags, null b
0.0	0.00	0.133	+0.1	+0.3	0.062	0.045
0.5	0.00	0.121	-0.9	-1.2	0.049	0.057
0.5	0.15	0.117	+11.9	-1.1	0.153	0.078
0.5	0.30	0.111	+35.6	+0.2	0.632	0.449
0.5	0.45	0.104	+69.3	-0.4	0.973	0.943
0.5	0.60	0.097	+110.6	+0.3	1.000	1.000

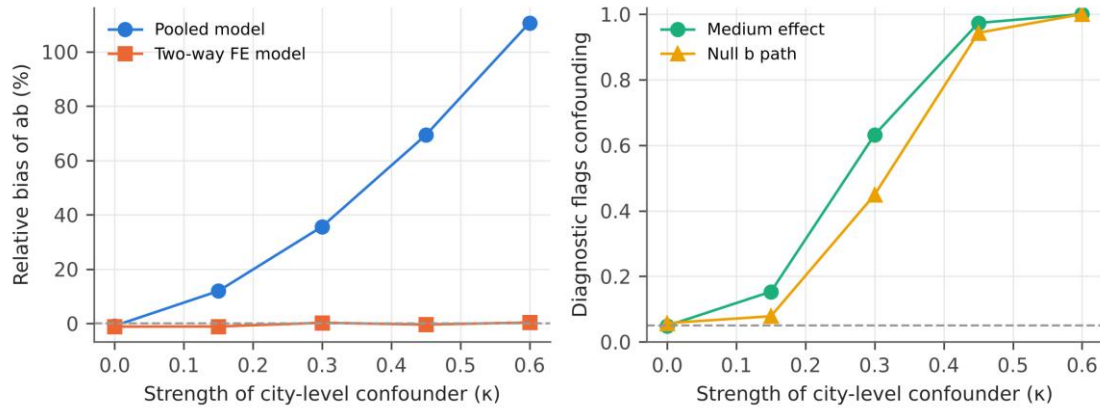


Figure 8. Pooled-within contrast as a diagnostic for persistent city confounding (E4; $\tau = 0.5$; 1,000 replications per point). Left: relative bias of the pooled and two-way FE indirect effects for the medium effect. Right: proportion of replications in which the paired city-cluster bootstrap interval for the pooled-minus-FE difference excluded zero; the dashed line marks 0.05.

4.6 E5: Time-varying confounding and sensitivity analysis

Table 11 reports E5. Without time-varying confounding ($\lambda = 0$), the FE estimator was unbiased (-0.9%) with coverage of 0.943. A time-varying factor that correlated the two disturbances by only $\rho = 0.083$ ($\lambda = 0.3$) biased the FE indirect effect by +25.9% and reduced coverage to 0.699. Under a null b path, the FE estimator then declared mediation in 0.619 of replications. At $\rho = 0.265$ ($\lambda = 0.6$), bias rose to +87.1%, coverage fell to 0.042, and the null was rejected in 1.000 of replications. Persistent-confounding control thus offers no protection against confounding that changes between editions.

The sensitivity function in Equation (4) removed this bias when evaluated at the true ρ . Relative bias was -1.4% and -1.1%, coverage was 0.938–0.950, and Type I error under the null b path was 0.062 and 0.055. The robustness value behaved as intended. Under a null b path with $\rho = 0.265$, its median was 0.263, almost exactly the true disturbance correlation. In other words, the spurious indirect effect would be explained away by a correlation equal to the confounding that produced it. For a genuine medium effect without time-varying confounding, the median ρ_0 was 0.284. In 90%

of replications, explaining that effect away would have required a disturbance correlation of at least 0.243.

Table 11. Two-way FE indirect effect under time-varying confounding, without adjustment and with the Stage 8 sensitivity function evaluated at the true disturbance correlation ρ (E5; scenario S2 plus a time-varying factor with loading λ ; 1,000 replications per row). ρ_0 is the median robustness value from Equation (5).

λ	True ρ	Estimator	Relative bias, medium (%)	Coverage, medium	Type I error (null b)	Median ρ_0 , medium	Median ρ_0 , null b
0.0	0.000	FE, unadjusted	-0.9	0.943	0.062	0.284	0.000
0.3	0.083	FE, unadjusted	+25.9	0.699	0.619	0.358	0.083
0.3	0.083	FE, adjusted at true ρ	-1.4	0.940	0.062	0.358	0.083
0.6	0.265	FE, unadjusted	+87.1	0.042	1.000	0.504	0.263
0.6	0.265	FE, adjusted at true ρ	-1.1	0.950	0.055	0.504	0.263

5. Discussion

5.1 Pooling across cities is the main threat

The five experiments separate problems that are usually discussed together. Their magnitudes differ by an order of magnitude. Ignoring city-level clustering roughly doubled the Type I error of the Sobel test but left the point estimate unbiased. Missing technology scores biased the indirect effect by at most 13% under the original single-imputation approach, and by less under multiple imputation. A substantial change in the definition of one dimension in one edition attenuated the pooled estimate by about 10%. By contrast, a single unobserved city characteristic affecting governance and environment inflated the pooled indirect effect by 126% for a medium effect and 330% for a small one. It also produced false declarations of mediation in most null replications. Time-varying confounding was equally damaging to the FE estimator: a disturbance correlation of only 0.083 inflated the FE indirect effect by about a quarter.

It is important to note that this order is significant to the smart city literature. Case evidence or cross-sectional comparison typically backs up the proposition that governance plays a mediating role in the technology-sustainability relationship (Bibri & Krogstie, 2017; Castelnovo et al., 2016; Meijer & Bolívar, 2016). Based on index data, it does seem to support the case, since it's a pooled regression and formal inference. Our results suggest caution. Cities that score well on technology also tend to be wealthier and administratively stronger, and these characteristics plausibly raise governance and environmental scores directly. A pooled analysis cannot distinguish that pattern from mediation, and better standard errors do not solve the problem. This is the multilevel conflation problem identified by Krull and MacKinnon (2001) and Zhang et al. (2009), here demonstrated at the scale and structure of a city index.

5.2 Fixed effects, the pooled-within contrast, and their limits

The two-way FE mediation model was the only strategy that performed acceptably in all three city-structure scenarios of E1. It removed bias from time-invariant confounding and, under clustering, was also the most powerful method. Its estimate is narrower, however. It asks whether cities whose technology scores rose more than their own average across editions also showed above-average governance and, through governance, above-average environmental scores. It does not compare one city with another. When there is no city-level confounding (S1), this restriction costs precision, a concern sharpened by the fixed number of cities in an index (Fritz & MacKinnon, 2007): intervals were about 20% wider, and power for a small effect fell from 0.999 to 0.945.

The real CIMI data may result in a higher cost. There may be relatively small differences in governance scores across the city, given the absence of within-city variation over a 5-year period and 8 editions. The FE estimator will then have little power, and a null FE result should not be interpreted as a denial of mediation. One useful step would be to make estimates of both models and compare them. This is usually done informally, in line with the Mundlak (1978) type reasoning that Hausman explained in Bell and Jones (2015). E4 demonstrates that there is a well-defined operating range for a paired bootstrap version of this comparison. It had a false-positive rate of around 5%, and it was able to detect confounding reliably when the pooled bias was approximately 2/3 of the true effect. There is often a distortion of 10-15% from the pooled estimate as a result of omitting a confounder that is below this point. The contrast is thus best used to outline a divergence, rather than to excuse the pooled estimate in the absence of any divergence. When there is no city confounding (shown by the illustrative panel), the pooled and FE technology-to-governance coefficients (0.657 and 0.694) were similar, as anticipated.

5.3 Time-varying confounding and the value of a robustness statement

E5 shows the limit of within-city estimation. A factor that changes between editions and affects both governance and environment, such as a national reform or an economic shock, biases the FE indirect effect just as a persistent factor biases the pooled one. No estimator in the pipeline can remove such confounding without additional data. What Stage 8 offers is a transparent statement of how strong it would need to be. When the confounding correlation was known, Equation (4) recovered the true effect with nominal coverage. In practice, ρ is not known. The useful output is then ρ_0 , the correlation at which the effect would vanish. Readers can judge whether time-varying confounding of that size is plausible for the cities and period studied. For the medium effect in E5, ρ_0 was about 0.28, so only fairly strong time-varying confounding would explain the effect away.

For a spurious effect created by confounding, ρ_0 matched the true confounding correlation. A reported ρ_0 thus carries interpretable information in both directions.

5.4 Missing data and indicator changes

Missing data were a secondary problem in these experiments, but the original choice of single deterministic imputation was the wrong one. It produced bias that grew with the missing rate and intervals that did not reflect imputation uncertainty. Multiple imputation with bootstrap draws pooled across imputations removed most of this bias when data were missing completely at random or at random given observed scores (Rubin, 1987; Schomaker & Heumann, 2018). Under MNAR, no method recovered the true effect. Adding city-level information from other editions helped only modestly, although Honaker and King (2010) show that panel structure is valuable for imputation in time-series cross-section data. For real index data, the plausible mechanism is MNAR: cities with weaker digital infrastructure are probably less likely to report the indicators behind the technology dimension. Sensitivity analysis under explicit MNAR assumptions is therefore advisable when the missing share is not small (Imai, Keele, & Yamamoto, 2010; van Buuren, 2018).

Within-edition standardisation did what it was designed to do: arbitrary changes in the location and scale of scores between editions had no effect on the estimates. It cannot, however, correct changes in what a dimension measures. The edition-heterogeneity diagnostic detected a 75% replacement of the 2022 environment indicator in only 0.307 of replications. This low power follows from the design: one edition of 183 cities supports only an imprecise edition-specific estimate. Analysts should therefore not rely on a statistical test to detect indicator revisions. The methodological notes that accompany each CIMI edition, and sensitivity analyses that drop revised editions, are more informative (Organisation for Economic & Development, 2008; Saisana et al., 2005).

5.5 A decision protocol for real CIMI extracts

Table 12 translates the results into a sequence of decisions. The protocol is not specific to CIMI and applies to other multi-edition city or country indices with a similar structure (Huovila et al., 2019; Sharifi, 2019).

Table 12. Decision protocol for testing governance mediation on a multi-edition city index, derived from experiments E1–E5.

Step	Decision	Recommended practice	Evidence
1. Document the index	Which editions changed indicators,	Record revisions from each edition’s methodology notes before modelling; plan sensitivity analyses that drop revised editions.	E3

Step	Decision	Recommended practice	Evidence
	weights or city coverage?		
2. Handle missing scores	Single or multiple imputation?	Use multiple imputation ($m \geq 5$) with bootstrap draws pooled across imputations; avoid single deterministic imputation beyond about 5% missingness; report missingness by dimension and edition.	E2
3. Align editions	How to compare scores across editions?	Standardise each dimension within edition; this removes rescaling but not changes in content.	E3
4. Choose inference	Which standard errors or intervals?	Use the city-cluster bootstrap (or CR1 standard errors); do not rely on OLS, HC3 or city-year resampling.	E1 (S2)
5. Choose the estimator	Pooled or within-city?	Estimate both pooled and two-way FE mediation models; report the paired bootstrap contrast (Stage 6); treat FE as primary when they diverge, and do not read a non-significant contrast as proof that pooling is safe.	E1 (S1, S3), E4
6. Check stability over editions	Is the indirect effect stable?	Report edition-specific and leave-one-edition-out estimates; do not treat a non-significant heterogeneity test as proof of stability.	E3
7. Assess time-varying confounding	How robust is the FE effect?	Report the robustness value ρ_0 and $ab(\rho)$ for plausible ρ (Stage 8); argue from substantive knowledge whether confounding of that size is credible.	E5
8. Interpret	Statistical or causal mediation?	State the identifying assumptions; note that FE and the sensitivity analysis do not establish temporal ordering.	Section 6

These recommendations change the original pipeline in three ways. Multiple imputation replaces single deterministic imputation. HC3-based Sobel tests are no longer used for inference. The FE estimator is promoted from a robustness check on a path to a co-primary estimator of the indirect effect. Two stages are added: a formal pooled-within contrast and a robustness value for time-varying confounding.

6. Limitations

All results come from synthetic panels. They describe how the pipeline behaves when the truth is known; they say nothing about the relationship between technology, governance and environment in real cities.

The data-generating mechanisms were linear, and the remaining errors were Gaussian, and the effects were homogeneous across the cities and across the editions. Real CIMI scores are bounded, partly derived from rankings, and may have skewed or heavy-tailed distributions. There might also be regional or income group variations in effects. The size of the biases reported here depends on the chosen confounder strength, random-intercept size and effect sizes, although E4 shows how pooled bias grows with confounder strength. The direction of the findings should transfer to other parameter values more readily than their exact magnitudes.

The FE estimator excludes only factors that will not change during the course of the study. The Stage 8 sensitivity analysis quantifies time-varying confounding of the governance-environment

path but does not remove it. It also assumes that the technology-governance path is unconfounded, and its key parameter ρ cannot be estimated from the data. The power of it is to put that assumption on the table and up for discussion. Due to the contemporaneous nature of the mediation model, the analysis also does not allow a temporal prioritisation of the variables technology, governance and environment, or a reverse impact of environmental performance on governance. Because of the number of editions, one cannot hope to create reliable lagged models.

Only the technology dimension displayed missingness, and the panel was balanced in terms of city membership. It is possible for real CIMI editions to have different cities represented, and missingness could occur in multiple dimensions. A single replacement in one dimension and one edition were used to represent changes in indicator definitions. For real indices, there can be a gradual growth in revisions from edition to edition and dimension to dimension.

The dimensions of CIMI are composites. If two dimensions share underlying indicators, or are constructed from correlated sources, part of their association is mechanical. The simulation did not model such overlap, and it should be checked against the index documentation before any substantive analysis.

Finally, the illustrative single-panel analysis (Section 4.1) used a different generator and software versions from the Monte Carlo study. Its numbers illustrate the pipeline's output and should not be compared directly with the Monte Carlo estimands.

7. Conclusion

This study asked whether a multi-edition city index can support a test of whether governance mediates the association between urban technology and sustainability. An eight-stage pipeline was specified on the CIMI schema and evaluated in 38,500 simulated panels. Recovering a planted mediation pathway from one synthetic panel, as in the earlier version of this work, is a weak test because it succeeds under conditions in which almost any estimator works. The Monte Carlo evaluation gives a sharper answer. City-cluster bootstrap inference is necessary because heteroskedasticity-robust Sobel tests and observation-level bootstraps reject too often when cities are observed repeatedly. Multiple imputation should replace single deterministic imputation, which inflated the indirect effect as missingness increased. Within-edition standardisation handles rescaling between editions but not changes in indicator content. Most importantly, pooled mediation estimates are not credible when persistent city characteristics affect both governance and sustainability. Under such confounding, the two-way fixed-effects mediation model remained close to unbiased, while every pooled method failed. A paired contrast between pooled and fixed-effects

estimates detects such confounding when it is large, but not when it is modest. Fixed effects, in turn, fail under time-varying confounding. For that threat, the robustness value p_0 tells readers how strong the confounding would need to be to overturn the conclusion.

The next step is to apply the revised pipeline, unchanged, to a real CIMI extract assembled from the published editions. Its output should be reported with both pooled and fixed-effects estimates, edition-specific estimates, and the assumptions required for a causal reading. Further simulation work should examine unbalanced city coverage, confounding of the technology-governance path, missingness across several dimensions, and sensitivity analyses for data missing not at random.

Declarations

Data and code availability. All simulation code, the random-seed scheme and the raw replicate-level results are openly available at [repository URL and DOI to be inserted before submission]. No real CIMI data were used.

Conflict of interest. The authors declare no conflict of interest.

Funding. This research received no external funding.

CRedit author contribution statement. All authors contributed equally to conceptualisation, methodology, software, formal analysis and writing.

Ethics statement. Not applicable; the study used synthetic data only.

Nomenclature

Abbreviation	Meaning
CIMI	IESE Cities in Motion Index
X	Exposure (technology dimension)
M	Mediator (governance dimension)
Y	Outcome (environment dimension)
ab	Indirect effect
c'	Direct effect
CR1	Cluster-robust covariance estimator with small-sample correction
FE	Fixed effects
HC3	Heteroskedasticity-consistent covariance estimator, type 3
MAR / MCAR / MNAR	Missing at random / completely at random / not at random
MI	Multiple imputation
MICE	Multiple imputation by chained equations
VIF	Variance inflation factor

References

- Ahvenniemi, H., Huovila, A., Pinto-Seppä, I., & Airaksinen, M. (2017). What are the differences between sustainable and smart cities? *Cities*, 60, 234-245. <https://doi.org/10.1016/j.cities.2016.09.009>
- Albino, V., Berardi, U., & Dangelico, R. M. (2015). Smart cities: Definitions, dimensions, performance, and initiatives. *Journal of Urban Technology*, 22(1), 3-21. <https://doi.org/10.1080/10630732.2014.942092>
- Angelidou, M. (2017). The role of smart city characteristics in the plans of fifteen cities. *Journal of Urban Technology*, 24(4), 3-28. <https://doi.org/10.1080/10630732.2017.1348880>
- Appio, F. P., Lima, M., & Paroutis, S. (2019). Understanding smart cities: Innovation ecosystems, technological advancements, and societal challenges. *Technological Forecasting and Social Change*, 142, 1-14. <https://doi.org/10.1016/j.techfore.2018.12.018>
- Bannister, F., & Connolly, R. (2014). ICT, public values and transformative government: A framework and programme for research. *Government Information Quarterly*, 31(1), 119-128. <https://doi.org/10.1016/j.giq.2013.06.002>
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173-1182. <https://doi.org/10.1037/0022-3514.51.6.1173>
- Bell, A., & Jones, K. (2015). Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data. *Political Science Research and Methods*, 3(1), 133-153. <https://doi.org/10.1017/psrm.2014.7>
- Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1), 249-275. <https://doi.org/10.1162/003355304772839588>
- Bibri, S. E., & Krogstie, J. (2017). Smart sustainable cities of the future: An extensive interdisciplinary literature review. *Sustainable Cities and Society*, 31, 183-212. <https://doi.org/10.1016/j.scs.2017.02.016>
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3), 414-427. <https://doi.org/10.1162/rest.90.3.414>
- Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2), 317-372. <https://doi.org/10.3368/jhr.50.2.317>
- Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of Urban Technology*, 18(2), 65-82. <https://doi.org/10.1080/10630732.2011.601117>
- Castelnovo, W., Misuraca, G., & Savoldelli, A. (2016). Smart cities governance: The need for a holistic approach to assessing urban participatory policy making. *Social Science Computer Review*, 34(6), 724-739. <https://doi.org/10.1177/0894439315611103>
- Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1), 39-67. <https://doi.org/10.1111/rssb.12348>
- Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3), 369-390. <https://doi.org/10.1111/j.1467-9868.2007.00593.x>
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233-239. <https://doi.org/10.1111/j.1467-9280.2007.01882.x>
- Gil-Garcia, J. R., Dawes, S. S., & Pardo, T. A. (2018). Digital government and public management research: Finding the crossroads. *Public Management Review*, 20(5), 633-646. <https://doi.org/10.1080/14719037.2017.1327181>

- Honaker, J., & King, G. (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, 54(2), 561-581. <https://doi.org/10.1111/j.1540-5907.2010.00447.x>
- Huovila, A., Bosch, P., & Airaksinen, M. (2019). Comparative analysis of standardised indicators for smart sustainable cities: What indicators and standards to use and when? *Cities*, 89, 141-153. <https://doi.org/10.1016/j.cities.2019.01.029>
- Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, 15(4), 309-334. <https://doi.org/10.1037/a0020761>
- Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1), 51-71. <https://doi.org/10.1214/10-STS321>
- Kitchin, R. (2014). The real-time city? Big data and smart urbanism. *GeoJournal*, 79, 1-14. <https://doi.org/10.1007/s10708-013-9516-8>
- Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, 36(2), 249-277. https://doi.org/10.1207/S15327906MBR3602_06
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7(1), 83-104. <https://doi.org/10.1037/1082-989X.7.1.83>
- MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305-325. [https://doi.org/10.1016/0304-4076\(85\)90158-7](https://doi.org/10.1016/0304-4076(85)90158-7)
- Meijer, A., & Bolívar, M. P. R. (2016). Governing the smart city: A review of the literature on smart urban governance. *International Review of Administrative Sciences*, 82(2), 392-408. <https://doi.org/10.1177/0020852314564308>
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074-2102. <https://doi.org/10.1002/sim.8086>
- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, 46(1), 69-85. <https://doi.org/10.2307/1913646>
- Nam, T., & Pardo, T. A. (2011). *Conceptualising smart city with dimensions of technology, people, and institutions* Proceedings of the 12th Annual International Conference on Digital Government Research: Digital Government Innovation in Challenging Times, <https://doi.org/10.1145/2037556.2037602>
- Organisation for Economic, C.-o., & Development. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. OECD Publishing. <https://doi.org/10.1787/9789264043466-en>
- Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3), 879-891. <https://doi.org/10.3758/BRM.40.3.879>
- Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, 15(3), 209-233. <https://doi.org/10.1037/a0020141>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.
- Saisana, M., Saltelli, A., & Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2), 307-323. <https://doi.org/10.1111/j.1467-985X.2005.00350.x>
- Schomaker, M., & Heumann, C. (2018). Bootstrap inference when using multiple imputation. *Statistics in Medicine*, 37(14), 2252-2266. <https://doi.org/10.1002/sim.7654>

- School, I. B. (2022). *IESE Cities in Motion Index 2022*.
<https://www.iese.edu/media/research/pdfs/ST-0633-E>
- Sharifi, A. (2019). A critical review of selected smart city assessment tools and indicator sets. *Journal of Cleaner Production*, 233, 1269-1283.
<https://doi.org/10.1016/j.jclepro.2019.06.172>
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, 13, 290-312. <https://doi.org/10.2307/270723>
- United Nations Department of, E., & Social, A. (2024). *UN e-government survey 2024*.
<https://desapublications.un.org/publications/un-e-government-survey-2024>
- van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC.
- Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.
- Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organisational Research Methods*, 12(4), 695-719. <https://doi.org/10.1177/1094428108327450>