Recentered Exposure to Network Designs: Connectivity, Redundancy, and Planner Choice in China's National Trunk Highway System

Submission to NBER Workshop on Development Economics and Networks for Early-Career Researchers
December 5, 2026

applicant information:
Satej S. Soman (satej@berkeley.edu)
University of California, Berkeley - School of Information
102 South Hall #4600
Berkeley, CA 94720-4600

faculty recommender:
Joshua E. Blumenstock (jblumenstock@berkeley.edu)
University of California, Berkeley - School of Information
102 South Hall #4600
Berkeley, CA 94720-4600

Motivation & Research Statement

Dear Workshop Organizers,

I am excited to enclose the following extended abstract, entitled "Recentered Exposure to Network Designs: Connectivity, Redundancy, and Planner Choice in China's National Trunk Highway System". This work examines a setting in which a policymaker's intervention itself is a network – China's national-scale highway system. I apply a design-based framework to estimate the causal impact of transportation network implementation by reconstructing the planner's objectives from archival documents. Exposure to the published network plan is recentered by expected exposure to an ensemble of counterfactual networks that meet the planner's stated goals. This approach removes exposure in units whose connectivity is mechanically implied by the planner's goals and lets counties whose variation in connectivity stems from genuine planner discretion drive identification of causal effects. The result is a spatially re-centered instrument that isolates the component of exposure driven by alternative planner choices, reducing endogeneity from guaranteed connections and clarifying the causal interpretation of the planning decisions.

This work, while in its early stages, investigates several advances:

1. A broadly-applicable method for generating counterfactual networks for causal inference, which distinguishes between network links that show true exogenous variation over space, and ones that are mechanically formed by the network generation process.
2. An orthogonal decomposition of overall graph exposure into backbone and loop effects, which captures distinct dynamics in networked systems.
3. A replication of a known paper in networks and development with modern remote sensing data (i.e. a high-resolution digital elevation model (DEM), as well as nightlight luminosity and built-up area timeseries), and a discussion of using tree-based graph instrumentasl variables to approximate the effects of networks with richer graph structure.

My research as a doctoral candidate at UC Berkeley's School of Information has focused on the intersection of network systems, geospatial machine learning, and empirical measurement in development settings. I have used graph-theoretic tools to study slum formation, model public health responses, and devise data collection strategies on networks featurized by cellular data records. Other forthcoming work in development-centric measurement uses high-resolution satellite imagery and state-of-the-art Transformer neural networks to estimate living standards of individual households in Togo, Benin, Malawi, and Bangladesh, with the goal of improving the targeting of social protection relative to approaches that assign all households in a given geographic region a single estimated consumption or wealth prediction. Projects I intend to work on in the near future use similar techniques to detect informal transit networks in developing countries. Overall, my work is about empirical measurement and identification in the spatial and network systems relevant to our understanding of economic development, and this paper tackles a network-design identification problem, with geospatial measurement providing the infrastructure reconstruction.

I believe this project is a strong candidate for the NBER conference on development and networks. I also hope to interact with other early-stage researchers whose research focus spans network systems and development, and provide perspectives on how the use of machine learning and remote sensing may complement ongoing work in this exciting field.

Thank you very much for your time and thoughtfulness in considering this submission.

Regards,
Satej S. Soman

Ph.D. Candidate
School of Information
University of California, Berkeley

satej s soman

suh-THAYJ SOH-munn

satej@berkeley.edu [website] [google scholar]

EDUCATION

PhD	Information Systems & Management	University of California, Berkeley School of Information advisor: Joshua Blumenstock	2021-2027 (expected)
MS	Computational Analysis & Public Policy	University of Chicago Harris School of Public Policy	2018-2020
BS	Electrical Engineering & Computer Science Materials Science & Engineering	University of California, Berkeley College of Engineering	2010-2014

TEACHING

UC Berkeley - INFO 206A: Introduction to Programming & Computation (Fall 2024, head TA)
UC Berkeley - INFO 206B: Introduction to Data Structures & Analytics (Fall 2024, head TA)
UC Berkeley - INFO 251: Applied Machine Learning (Spring 2025, co-TA); ⋆ Outstanding TA Award

EXPERIENCE

Visiting Scientist - Geospatial Machine Learning	Microsoft AI for Good	Jun 2025-Mar 2026
Graduate Student Researcher	University of California, Berkeley	Aug 2021-
Head of Data Science	COVID International Working Group	Jun 2020-Aug 2021
Principal Research Engineer	Mansueto Institute for Urban Innovation	Jan 2019-Jun 2020
Forward-Deployed Engineering Lead	Palantir Technologies	Aug 2014-Aug 2018

WORKS IN SUBMISSION

Scaling Croptype Mapping for Data-Sparse Settings with Geospatial Embeddings and Weak Labels
Soman, S., Zaytar, A., Tadesse, G.A., Robinson, C., Becker-Reshef, I., & Ferres, J.L. (2026)

Anaximander: A System for Interactive Orchestration of Geospatial Deep Learning Models
Soman, S., Zaytar, A., Tadesse, G.A., Hacheme, G., Danish, M.S., Becker-Reshef, I., & Ferres, J.L. (2026)

GeoFaham: A Multi-Agent Framework for Geospatial Decision Support
Danish, M.S., Hacheme, G., Zaytar, A., **Soman, S.**, Tadesse, G.A., Becker-Reshef, I., & Ferres, J.L. (2026)

PEER-REVIEWED PUBLICATIONS

A Hierarchical Bayesian Model for Estimating Age-specific COVID-19 Infection Fatality Rates in Developing Countries
Pugh, S., Levin, A., Meyerowitz-Katz G., **Soman, S.**, Owusu-Boaitey, N., Zwi, B., Malani, A., & Fosdick, B. (2024).
Statistics in Medicine 43, no. 30 (2024): 5667-5680 10.1002/sim.10259

Assessing the Burden of COVID-19 in Developing Countries: Systematic Review, Meta-Analysis, & Public Policy Implications
Levin, A., Owusu-Boaitey, N., Pugh, S., Fosdick, B., Zwi, A., Malani, A., **Soman, S.**, et al. (2022).
BMJ Global Health 7, no. 5 (2022): e008477 10.1136/bmjgh-2022-008477

Epimargin: A Toolkit for Epidemiological Estimation, Prediction, and Policy Evaluation
Soman, S., Loftus, C., Buschbach, S., Phadnis, M., & Bettencourt, L. (2021).
Journal of Open Source Software, 6(65), 3464 10.21105/joss.03464

Worldwide Detection of Informal Settlements via Topological Analysis of Crowdsourced Digital Maps
Soman, S., Beukes, A., Nederhood, C., Marchio, N., & Bettencourt, L. (2020).
ISPRS Int. J. Geo-Inf., 9 (10) 10.3390/ijgi9110685

WORKING PAPERS

Vaccine Allocation Priorities Using Disease Surveillance and Economic Data
Malani, A., **Soman, S.**, Ramachandran, S., Chen, A., Lakdawalla, D. (2022).
NBER Working Paper, (No. w29682) 10.3386/w29682

Systems Architecture for Real-Time Epidemiological Prediction and Control
Bettencourt, L. & **Soman, S.** (2020).
Mansueto Institute for Urban Innovation Research Paper, (24) 10.2139/ssrn.3748704

Adaptive Control of COVID-19 Outbreaks in India: Local, Gradual, and Trigger-Based Exit Paths from Lockdown
Malani, A., **Soman, S.**, Asher, S., Novosad, P., Imbert, C., Tandel, V., et al. (2020).
NBER Working Paper, (No. w27532) 10.3386/w27532

CONFERENCE PROCEEDINGS

Model-Driven Data Collection for Cost-Effective Poverty Prediction
Soman, S., & Blumenstock, J.E. (2024).
Thirteenth Int'l Conference on Information & Communication Technologies and Development (ICTD), Nairobi, Kenya.

Estimating Household-Level Economic Characteristics from High-Resolution Satellite Imagery
Soman, S., Constenla-Villoslada, S., Aiken, E., & Blumenstock, J.E. (2024).
World Bank Land Conference - Research Track. Washington, D.C., USA.

Can Strategic Data Collection Improve the Performance of Poverty Prediction Models?
⋆ spotlight presentation (top 10% of submissions)
Soman, S., Aiken, E., Rolf, E., & Blumenstock, J.E. (2023).
Eleventh Int'l Conference on Learning Representations (ICLR),
4th Workshop on Practical Machine Learning for Developing Countries. Kigali, Rwanda. 10.48550/arXiv.2211.08735

PATENTS

Systems and Methods for Annotating Datasets
Javitt, G., Szuflita, S., **Soman, S.**, Pandey, H., Dhulipalla, S., Shekhawat, V.
US Patent #11,176,116. Granted: 16 November, 2021.

Data Item Aggregate Probability Analysis System
Soman, S., Hoffman, D., Al Khafaji, S., Kowalik, J., Sanzovo, P., Punukollu G.
US Patent #10,691,756. Granted: 6 February, 2018.

Extended Abstract

Overview

The locations of large-scale infrastructure projects are often assigned through a network plan, but both planner objectives and realized construction reflect engineering constraints, pre-existing infrastructure, political considerations, and endogenous economic conditions. Naive estimation of the effects of infrastructure projects would be biased by this endogeneity. This paper develops a design-based framework for estimating the causal effects of planned networks when the planner's mandate admits multiple feasible realizations. I construct an ensemble of planner-consistent graphs and use expected exposure over this ensemble to recenter the planned treatment, separating exposure that is mechanically implied by the planning mandate from exposure generated by the planner's discretion.

I apply this framework to China's National Trunk Highway System (NTHS), using the corridor plan issued by the Ministry of Communications (MOC) in 1993 and the planning documents behind it. These documents call for 5 horizontal and 7 vertical named corridors totaling roughly 35,000 km, designed to link every city above 1 million people and 93% of cities above 500,000 people, as well as border crossings and major coastal ports. It was the largest road-building program in history at the time: starting from China's existing 6,600-km network in 1993 to covering 60,000 km by 2007, doubling every four years, sustained for fifteen years. The system was declared complete in 2007, well ahead of the original planning horizon, and the build-out coincided almost exactly with China's WTO-era industrialization surge. The later building phase was also key to China's fiscal stimulus response to the 1997 Asian financial crisis. Because of the difficulties in separating returns to infrastructure investment from changes in economic conditions in areas likely to grow through the 1990s, careful econometric design is required to identify the effects of the NTHS.

Prior Work & Contributions

Existing work, notably Faber (2014), instruments realized highway connections with a minimum-cost spanning tree (MST) — a network plan the MOC itself computed, printed, and rejected. A MST is a network consisting of edges representing the least-cost paths over a grid of construction costs that connects all of the targeted cities. The sample frame consists of those units "incidentally" connected to the network at a suitable distance from targeted nodes compared to units that did not receive a connection.

I identify three issues with the MST instrument strategy:

- *Non-uniqueness*: A MST network is not guaranteed to be unique, especially on empirical cost surfaces. The choice of MST a practitioner uses depends on how the underlying algorithmic implementations ((Dijkstra, 1959; Kruskal, 1956)) choose to break ties. I demonstrate that changing the exact set of equivalent-cost edges that constitute the instrument can cause the estimated effects to swing and even flip sign, as shown in Figure 1. Table 1 also documents that the strength of the instrument is tree-dependent.
- *Incidental versus mechanical connection*: Some of the non-targeted counties that are considered incidentally connected are actually on near-unique cost-minimizing paths. If the original units A and B are targeted, and a connection through unit C is the only cost-feasible way to link them, then C's connection is mechanical rather than incidental: it is implied by the decision to connect A and B and survives in every counterfactual plan. Examples of such paths include China's "Three Famous Corridors" and the Golmud-Lhasa road built by the People's Liberation Army in the 1950s. Comparing such units to unconnected counties recovers the effect of sitting on a mandated path, not the effect of a quasi-random connection.
- *Loop representation*: A spanning tree is loop-free by construction: it has zero *cycle rank* and exactly one path between any pair of nodes. The realized NTHS, like most large transportation systems, contains loops (ring roads, parallel routings, and redundant links between corridors) in service of political and military objectives rather than in response to pure economic concerns. Exposure generated by redundancy is therefore invisible to a tree instrument, which understates connectivity precisely for counties whose access depends on alternative routes. The corridor mandate itself guarantees the mismatch: a plan of five north–south and seven east–west trunk corridors is a crossing grid whose intersections necessarily create loops, so no tree can represent the planned network, let alone the realized one.

Empirical Strategy & Data

I instead recover treatment from the published plan and define the set of planner-consistent graphs as those satisfying the published corridor mandate at no greater cost than the realized plan. The ensemble is generated by Markov chain Monte Carlo sampling and consists of networks with realistic loop structure, a property MSTs cannot reflect.

The design draws on two instrumental variable (IV) approaches: the strategy of Herzog (2021), who uses the published plan for the United States Interstate Highway System as an instrument, combined with Borusyak and Hull (2023), which formalizes recentering exposure by its expectation across counterfactual draws of treatment and realization. The transportation network application in the latter paper uses counterfactual shocks in *time* – this work instead uses variation in *space* characterized by alternative network realizations consistent with the historical mandate. The planning mandate makes some exposure inevitable: counties near targeted cities or on corridors every admissible network must serve are connected in nearly all counterfactual plans, so their realized connections carry no identifying variation. I therefore measure exposure to the published plan and to the realized network as deviations from the expected exposure over the ensemble of planner-consistent networks:

$$\mu_i = \mathbb{E}_{G \in \mathcal{G}}\left[X_i(G)\right]$$
$$Z_i = X_i(G_P) - \mu_i$$
$$D_i = X_i(G_R) - \mu_i$$

Here G_P is the published 1993 corridor plan, digitized from archival map sources and matched to a national grid of candidate routes (Figure 2, left); G_R is the network as built by 2007; $X_i(G)$ is a generic exposure mapping as a function of the graph G. The results in this note use Faber's mapping: $X_i(G)$ indicates that county i lies within 10 km of network G. $\mathcal{G}$ is the set of planner-consistent networks (those serving the mandated corridors and targeted cities at no greater construction cost than the realized plan) and μ_i is county i's expected exposure across $\mathcal{G}$. W_i collects baseline county controls (distance to the nearest targeted city, 1990 urban registered population, prefecture-capital and city administrative status, agricultural employment share, and schooling), and α denotes province fixed effects. Estimation is just-identified two-stage least squares on peripheral counties at least 50 km from plan-targeted cities, with standard errors clustered by province throughout.

$$D_i = \pi Z_i + W_i'\Gamma + \alpha + \varepsilon_i \tag{1}$$

$$\Delta y_i = \beta \widehat{D}_i + W_i'\Gamma + \alpha + \nu_i \tag{2}$$

The ensemble is sampled by Markov chain Monte Carlo over networks built from the same candidate-route grid as the plan itself. A member network must serve the mandated corridors and targeted cities and may cost no more to build than the realized plan, with construction costs measured on a terrain- and land-cover-based cost surface. The sampler starts with a maximally-overconnected network: the Delaunay triangulation of the targeted nodes, with all minimum cost edges between adjacent nodes admitted as candidate edges. The sampler then removes nodes in order of descending edge construction costs while maintaining connectivity and coverage mandates. Proposal moves rewire routes locally, and acceptance rules enforce feasibility and the cost cap, so the chains mix over networks with realistic loop structure rather than trees. Sampling is the only practical access to this set: the space of planner-consistent networks is combinatorially large, so I follow the ensemble approach developed for legislative redistricting, where MCMC samples of the space of lawful maps serve as the reference distribution against which an enacted plan is judged (Fifield et al., 2020). The sampler reproduces exact edge probabilities on a small enumerable test problem. The set $\mathcal{G}$ consists of approximately fourteen million distinct networks, and μ_i is the share of sampled networks passing within 10 km of county i (Figure 2, right, shows the per-route construction probabilities).

The plan comes from the 1993 Ministry of Communications corridor program and surrounding planning-era documents (Jiaotong Bu [Ministry of Communications], 1993; World Bank, 1994; Zhang, 1995). County outcomes are 1997–2006 log differences in county yearbook aggregates (National Bureau of Statistics of China, 2026), complemented by DMSP-OLS nighttime luminosity (NOAA National Geophysical Data Center and Earth Observation Group, Colorado School of Mines, n.d.) and GAIA built-up area (Gong et al., 2020). The realized network comes from a GIS database of China's road system (Davis et al., 2025; Baum-Snow et al., 2012). The construction-cost surface is built from the Copernicus GLO-30 digital elevation model (European Space Agency, 2021) and Digital Chart of the World-era land cover (Defense Mapping Agency and Environmental Systems Research Institute, 1993), following the cost specification of Faber (2014).

Results

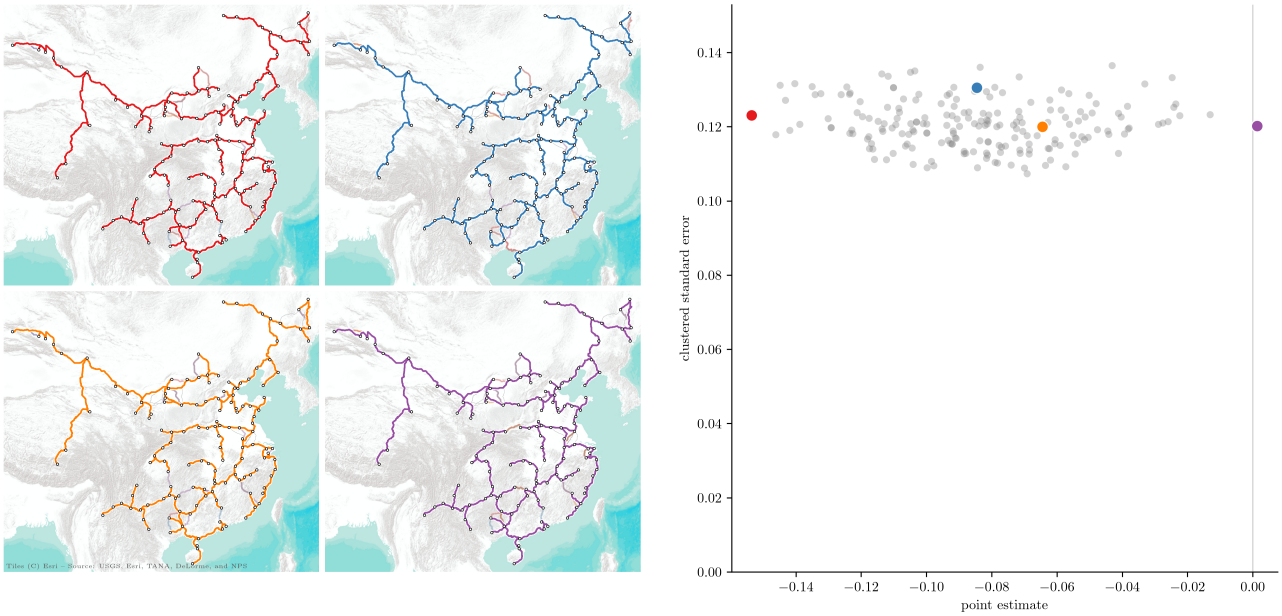


Figure 1: *Left*: Four equal-cost minimum spanning trees drawn from the ensemble. Equivalent cost edges not used in the highlighted tree construction are highlighted in low-opacity traces, showing the areas where the planner had indifference to cost. *Right*: Two-stage least squares point estimates versus standard errors across 200 identical-cost trees, with the estimates corresponding to the left panel highlighted by corresponding color. Note that while each individual point estimate is precisely estimated, the range of estimates spans from negative to positive. The same trees' F-statistics are presented in Table 1.

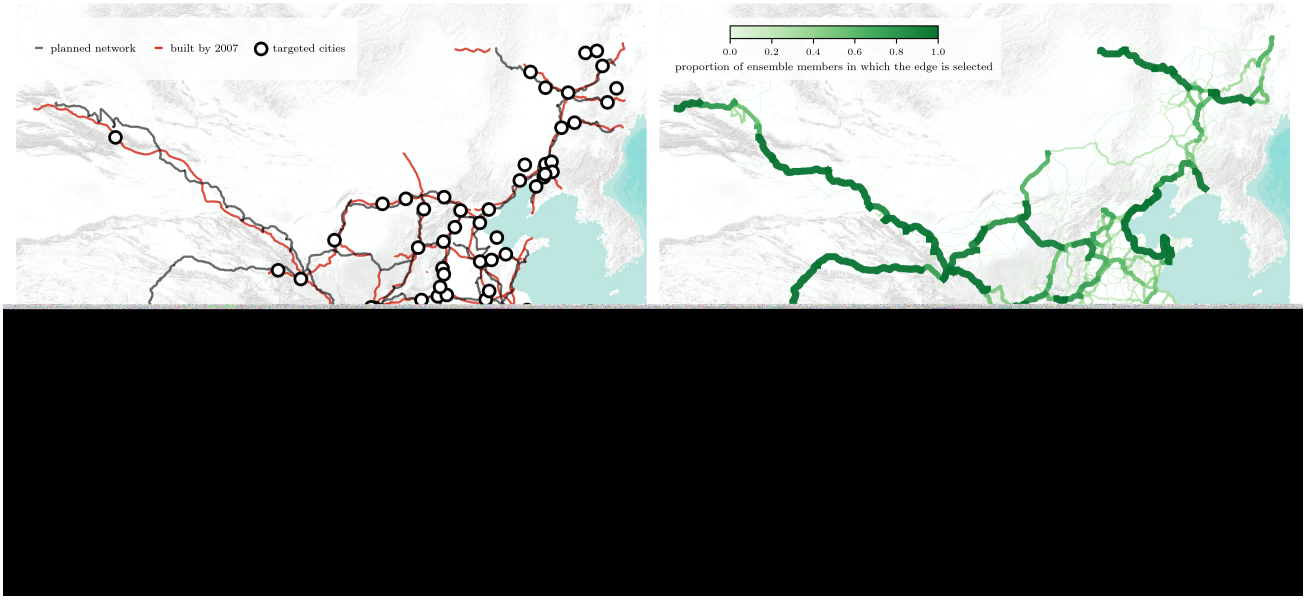


Figure 2: *Left*: the digitized 1993 corridor plan over the network built by 2007, with the plan's targeted cities marked. *Right*: candidate routes colored by the proportion of ensemble members in which each edge is selected. Routes where counties are connected as a function of global plan mandates, (like the Liaoxi corridor in northeastern China, which connects Tangshan to Shenyang by a road that must go between mountains to the northwest and the sea to the southeast) are assigned probability 1, whereas routes over flat plain areas where the planner is indifferent to their equal construction cost are assigned non-zero and non-unit probabilities.

Table 1: First-stage estimates: spanning-tree instruments and the documentary plan instrument

Instrument	Coef.	(SE)	F
Panel A: equal-cost minimum spanning trees (ensemble draws)			
1. • Equal-cost spanning tree draw 1	0.2786	(0.0431)	41.73
2. • Equal-cost spanning tree draw 2	0.2758	(0.0459)	36.05
3. • Equal-cost spanning tree draw 3	0.2912	(0.0419)	48.37
4. • Equal-cost spanning tree draw 4	0.2931	(0.0400)	53.67
Panel B: documentary plan instrument (1993 corridor plan)			
6. Plan exposure, raw	0.3122	(0.0422)	54.82
7. Plan exposure, recentered (Z and D as deviations from μ)	0.4207	(0.0605)	48.40

Notes: The endogenous variable is an indicator for a realized NTHS connection within 10 km of the county (in row 7, its deviation from the ensemble-mean exposure μ_i). Rows 1–4: the instrument is connection within 10 km of one equal-cost minimum spanning tree of the routed plan network, drawn uniformly from the exact-tie closure with per-edge routes sampled in proportion to their multiplicity; row markers match the colors of Figure 1. The first stage is strong for every one of the 200 ensemble draws (cluster-robust F between 36.1 and 53.7); the four displayed draws span the ensemble's full range of both the point estimate and the first-stage F, including the ensemble minimum (row 2). The point-estimate instability documented in Figure 1 is a second-stage phenomenon, not a weak-instrument artifact. Rows 6–7: the instrument is connection within 10 km of the digitized 1993 corridor plan; row 7 recenters both the instrument and the endogenous variable by μ_i, the expected exposure over the counterfactual-network ensemble. All rows: peripheral sample of n = 1,583 counties at least 50 km from plan-targeted cities, identical county-level controls, province fixed effects, and CRV1 standard errors clustered by province (30 clusters).

Table 2: Second-stage (2SLS) estimates of realized connection effects: raw versus recentered plan exposure

	Raw Z		Recentered $\tilde{Z}$		
Outcome (1997–2006 log difference)	Coef.	(SE)	Coef.	(SE)	n
Industrial gross value added	−0.0219	(0.1383)	−0.0211	(0.1104)	1,499
Non-agricultural gross value added	−0.0127	(0.1111)	0.0138	(0.0807)	1,507
Government revenue	−0.1639	(0.1625)	0.0048	(0.1521)	1,501
GDP	0.0206	(0.0928)	0.0294	(0.0672)	1,574
Agricultural gross value added	−0.0362	(0.0677)	0.0043	(0.0410)	1,522
Population	0.0151	(0.0120)	0.0150	(0.0135)	1,515
Nightlight luminosity	−0.1541	(0.0873)	0.0502	(0.0787)	1,528
Built-up area	0.0869	(0.0616)	0.0777	(0.0403)	1,552
First-stage F	54.82		48.40		

Notes: Anderson–Rubin confidence sets, robust to weak instruments, are bounded and include zero in every cell. Each cell reports the 2SLS coefficient on the realized-connection indicator, with the CRV1 province-clustered standard error in parentheses (30 clusters). Left block: raw plan exposure Z instruments the raw connection indicator D. Right block: both the instrument and the endogenous variable enter as deviations from the ensemble-mean exposure μ_i ($\tilde{Z}_i = Z_i - \mu_i$, $\tilde{D}_i = D_i - \mu_i$). All columns: peripheral sample (counties at least 50 km from plan-targeted cities), identical county-level controls, and province fixed effects. The reported first-stage F is computed on the full first-stage sample (n = 1,583); per-outcome estimation-sample F values range 53.06–58.26 (raw) and 48.35–52.15 (recentered). The first six outcomes are county yearbook aggregates; nightlight luminosity and built-up area are satellite-derived.

Both instruments are strong in the first stage (Table 1), so the second-stage nulls in Table 2 are not weak-instrument artifacts. The recentered design substantially changes the estimated effect and produces null AR confidence sets across outcomes. This motivates replacing the binary treatment with continuous travel-time exposure and decomposing that exposure by network topology (cf. Section "Future Work" below). The recentered estimates are also more precise than the raw ones for seven of the eight outcomes, consistent with the recentering absorbing design noise rather than discarding signal. The tightest cell, built-up area, points to modest land development where planner discretion actually varied. These are preliminary magnitudes on a coarse treatment; the travel-time treatment described below is the natural refinement.

Future Work

Future work on this project will measure realized treatment as the change in travel time between 1993 and 2007 to reflect both road-grade-specific speeds and the pre-existing network, rather than as a binary indicator of geographic proximity.

I then plan to separate the effects of backbone connectivity from those of loop connectivity. A naive approach would take the graph G of node set V and edge set E, and choose a minimum spanning tree $T \subset E$. Exposure to the edges of T would identify the backbone effects, while exposure to the non-tree edges $E \setminus T$ identifies loop effects. This approach would run into collinearity - since T is not guaranteed to be unique, and the choice

of T naturally determines $E \setminus T$, the backbone-versus-loop decomposition's exact results are not well-defined. Figure 3 shows this in practice: each of the grey dots are the locations of the backbone effect and loop effect estimated by using a MST. Not only do the estimates trace out a nearly-linear trend (textbook collinearity), the exact sign of each component of exposure also swings with tree choice.

An alternative is using the *Hodge decomposition*, which splits the edge weights of a network into a *gradient component* (e.g. flows generated by potentials on the nodes, the tree-like backbone) and an orthogonal *cycle component* that circulates on loops. The former captures core-periphery integration—the margin emphasized by minimum spanning trees—while the latter captures local market thickening, route redundancy, and alternative access that tree-based designs cannot reflect. Writing $X_i^*(G)$ and $X_i^\circ(G)$ for county i's exposure to the gradient- and cycle-space components of network G, each component is recentered by its own ensemble mean μ_i^* or μ_i°. The decomposition is exactly additive: every unit of exposure belongs to precisely one of the two orthogonal components, so the two-channel design partitions, rather than reweights, total exposure:

$$
\begin{aligned}
X_i(G) &= X_i^*(G) + X_i^\circ(G) \\
Z_i^* = X_i^*(G_P) - \mu_i^*, &\qquad Z_i^\circ = X_i^\circ(G_P) - \mu_i^\circ \\
D_i^* = X_i^*(G_R) - \mu_i^*, &\qquad D_i^\circ = X_i^\circ(G_R) - \mu_i^\circ
\end{aligned}
$$

Each component of realized exposure is instrumented by both components of plan exposure in the first stage,

$$D_i^* = \pi^{**} Z_i^* + \pi^{*\circ} Z_i^\circ + W_i'\Gamma + \alpha + \varepsilon_i^* \tag{3}$$

$$D_i^\circ = \pi^{\circ *} Z_i^* + \pi^{\circ\circ} Z_i^\circ + W_i'\Gamma + \alpha + \varepsilon_i^\circ \tag{4}$$

with the second stage

$$\Delta y_i = \beta^* \widehat{D}_i^* + \beta^\circ \widehat{D}_i^\circ + W_i'\Gamma + \alpha + \nu_i \tag{5}$$

A Hodge decomposition estimate, which is unique and basis-free, fixes a single point in Figure 3, avoiding the basis dependence of tree choice and associated multicollinearity.

Though this project is primarily about identification in network settings, I hypothesize that a non-zero cycle-space coefficient, conditional on the relevant accessibility measures, would provide evidence against treating least-cost accessibility as a sufficient statistic for the economic effects of network topology (Allen and Arkolakis, 2022). I also suspect that political distortions to infrastructure plans (as documented by Alder and Kondo (2020) for China's NTHS and Burgess et al. (2015) in Kenya) concentrate in the cycle space. The planning documents explicitly assign some network redundancy to strategic and military objectives, providing a documented source of planner discretion distinct from economic routing. The paper aims to provide both a reanalysis of China's trunk highway system and a general framework for causal inference with planned network treatments, including a mechanism for decomposing network effects into distinct channels.

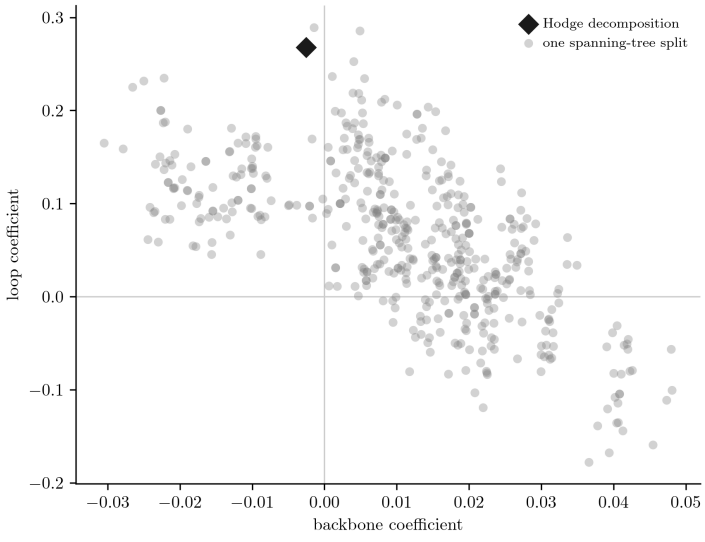


Figure 3: Backbone and loop associations of realized growth under five hundred arbitrary spanning-tree splits of the planned network (grey), against the unique Hodge decomposition of the same network (diamond). Associations are descriptive and uninstrumented.

References

Simon Alder and Illenin Kondo. Political distortions and infrastructure networks in China: A quantitative spatial equilibrium analysis. Working paper, version of January 2020. Earlier version: Society for Economic Dynamics 2018 Meeting Papers 1269, 2020.

Treb Allen and Costas Arkolakis. The welfare effects of transportation infrastructure improvements. *The Review of Economic Studies*, 89(6):2911–2957, 2022. doi: 10.1093/restud/rdac001.

Nathaniel Baum-Snow, Matthew A. Turner, Loren Brandt, J. Vernon Henderson, and Qinghua Zhang. China historical roads and rail. University of Toronto Map and Data Library, 2012. URL https://mdl.library.utoronto.ca/collections/geospatial-data/china-historical-roads-and-rail. Vector GIS, temporal extent 1962–2010.

Kirill Borusyak and Peter Hull. Nonrandom exposure to exogenous shocks. *Econometrica*, 91(6):2155–2185, 2023. doi: 10.3982/ECTA19367.

Robin Burgess, Remi Jedwab, Edward Miguel, Ameet Morjaria, and Gerard Padró i Miquel. The value of democracy: Evidence from road building in kenya. *American Economic Review*, 105(6):1817–1851, 2015. doi: 10.1257/aer.20131031.

Steven J. Davis, Meijun Qian, and Wen Zeng. A comprehensive GIS database for China's surface transport network with implications for transport and socioeconomics research. NBER Working Paper 33515, National Bureau of Economic Research, 2025.

Defense Mapping Agency and Environmental Systems Research Institute. Digital chart of the world, 1993. 1:1,000,000 vector basemap; drainage, lakes and coastline layers.

Edsger W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959. doi: 10.1007/BF01386390.

European Space Agency. Copernicus global digital elevation model (GLO-30), 2021. Produced using Copernicus WorldDEM-30 © DLR e.V. 2010–2014 and © Airbus Defence and Space GmbH 2014–2018, provided under COPERNICUS by the European Union and ESA.

Benjamin Faber. Trade integration, market size, and industrialization: Evidence from China's National Trunk Highway System. *The Review of Economic Studies*, 81(3):1046–1070, 2014. doi: 10.1093/restud/rdu010.

Benjamin Fifield, Michael Higgins, Kosuke Imai, and Alexander Tarr. Automated redistricting simulation using markov chain monte carlo. *Journal of Computational and Graphical Statistics*, 29(4):715–728, 2020. doi: 10.1080/10618600.2020.1739532.

Peng Gong, Xuecao Li, Jie Wang, Yuqi Bai, Bin Chen, Tengyun Hu, Xiaoping Liu, Bing Xu, Jun Yang, Wei Zhang, and Yuyu Zhou. Annual maps of global artificial impervious area (GAIA) between 1985 and 2018. *Remote Sensing of Environment*, 236:111510, 2020. doi: 10.1016/j.rse.2019.111510.

Ian Herzog. National transportation networks, market access, and regional economic growth. *Journal of Urban Economics*, 122:103316, 2021. doi: 10.1016/j.jue.2020.103316.

Jiaotong Bu [Ministry of Communications]. Guodao zhuganxian xitong guihua [national trunk highway system plan], June 1993. MOC document 1993600 (Jiaojifa [1993] No. 600). Full text not read; routes, spurs and 34,422 km total as reported in later official sources.

Joseph B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956. doi: 10.1090/S0002-9939-1956-0078686-7.

National Bureau of Statistics of China. County-level socioeconomic statistics. China Data Online (All China Marketing Research), 2026. URL https://www.china-data-online.com/. Accessed through the UC Berkeley library license.

NOAA National Geophysical Data Center and Earth Observation Group, Colorado School of Mines. Version 4 DMSP-OLS nighttime lights time series. Accessed through Google Earth Engine, collection NOAA/DMSP-OLS/NIGHTTIME_LIGHTS, n.d. URL https://eogdata.mines.edu/dmsp/downloadV4composites.html. Annual cloud-free composites, 1992–2013. Data collected by the US Air Force Weather Agency.

World Bank. China: Highway development and management: Issues, options and strategies. Report 13555-CHA, World Bank, Washington, DC, October 1994. URL http://documents.worldbank.org/curated/en/590541468770427393. Annex 3, "Criteria for Identifying Routes of the National Trunk Highway System," pp. 80–93; Map IBRD 24766R (April 1993).

Baosheng Zhang. Guodao zhuganxian guihua [national trunk highway plan]. In Jingtong Liu, editor, *Zhongguo jiaotong nianjian 1995 [China Transport Yearbook 1995]*, pages 143–144. Zhongguo jiaotong nianjian she [China Transport Yearbook Press], 1995. Column (Highway construction). Entry signed (MOC Highway Department, Zhang Baosheng). p. 144: attached map (Schematic map of the NTHS plan).