Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Metro Infrastructure and Regional Housing Markets: Causal Inference under Spatial and Network Dependence

Ahmad Yousefi
University of Isfahan
Saeed Samadi
s.samadi@ase.ui.ac.ir
University of Isfahan
Babak Saffari
University of Isfahan

Research Article

Keywords: Metro infrastructure, Housing prices, Housing capitalization, Spatial-network spillovers, Difference-in-differences, Double Machine Learning

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11017324/v1

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

Additional Declarations: No competing interests reported.

Abstract

Regional economic outcomes are often shaped by interregional dependence, yet causal evaluations of regional policies typically model this dependence through a single geographical structure. We develop a causal inference framework for settings where regional units are linked through multiple, conceptually distinct networks and treatment adoption is staggered over time, using the construction of metro infrastructure across 31 Iranian provinces (2006–2023). The framework combines staggered-adoption difference-in-differences with Double Machine Learning (DML) and a graph-based representation of spatial and economic-network dependence. Two graph branches separately encode geographical proximity and economic linkages within DML's nuisance-learning stage, allowing the treatment effect to be estimated net of dependence transmitted through both networks.The estimates indicate a positive effect of metro construction on regional housing prices: the Callaway–Sant'Anna estimator yields an ATT of 0.2196 ($p = 0.002$), while DML yields 0.3849 (robust SE = 0.043). The two approaches agree on direction but differ in magnitude, reflecting different identifying assumptions. Three CS-DiD specifications with alternative spillover adjustments produce estimates between 0.2036 and 0.2196. Given the small number of treated regions, we complement these estimates with cluster-based inference, placebo and permutation tests, Monte Carlo experiments, and diagnostics of spatial and network identifiability. These show the two components carry distinct information but are only moderately identifiable in this sample, so we interpret the graph-based decomposition cautiously. The results illustrate how causal inference for regional interventions can accommodate staggered adoption and multiple forms of interregional dependence.

JEL Classification: R31; R42; C21; C31

1. Introduction

Assessing the causal effect of large-scale infrastructure investments—particularly public transportation development—on housing markets is a central question in regional science and regional economics. A substantial hedonic and spatial difference-in-differences literature has documented sizable price premia near new transit stations—for example, Diao, Leonard, and Sing (2017) find that homes near new stations on Singapore's Circle Line appreciated by roughly 7.8% relative to control areas, and Bardaka, Delgado, and Florax (2019) trace direct, indirect, and total capitalization effects of Denver's phased light-rail expansion. This literature, however, has largely developed around two identification challenges that remain incompletely resolved. First, transit lines are typically not constructed simultaneously but gradually across different years in different regions; recent econometric literature has shown that conventional two-way fixed effects estimators can be severely biased under such conditions, as earlier-treated units are mistakenly used as controls for later-treated units (Callaway & Sant'Anna, 2021; Goodman-Bacon, 2021; Sun & Abraham, 2021). Second, the effect of subway construction in one area does not necessarily remain confined to that area: geographical proximity (spatial spillovers) and socioeconomic linkages between regions (network spillovers) can transmit the effect to neighboring or connected areas—a phenomenon that standard Diff-in-Diff frameworks typically ignore, and that the existing spatial-DiD real estate literature has so far modeled through a single geographical channel only.

1. **1.1. Methodological Gap**
2. 1. Classical DiD-based studies typically ignore staggered adoption or model it with simple two-period designs, which can be severely biased according to recent literature.
3. 2. The DML framework is used for causal effect estimation with flexible control of confounders but does not explicitly model spatial/network dependencies between units.
4. 3. Spatial econometric studies typically consider only one adjacency matrix (geographical) and ignore economic/network spillovers beyond physical proximity.
5. 4. The application of GNNs to causal effect estimation is still nascent, and most existing studies do not report under what data conditions such complexity is justified.

1.2. Innovation and Position in the Literature

First innovation relative to the multiple-period DiD literature. The recent staggered DiD literature focuses primarily on correcting bias from “bad comparisons” in standard Two-Way Fixed Effects estimators and typically assumes cross-sectional units are independent, ignoring spatial or network spillovers between units (Goodman-Bacon, 2021; Callaway & Sant’Anna, 2021; de Chaisemartin & D’Haultfoeuille, 2020; Sun & Abraham, 2021). The innovation of this study in this dimension is not introducing a new DiD estimator, but rather the explicit combination of the Callaway-Sant’Anna framework with a separate spatial-network spillover control layer via a GNN encoder within the DML framework, such that the treatment effect is simultaneously purified from both sources of bias—staggered adoption and cross-unit dependence. Neither component alone is novel; the innovation lies in their compatible combination and in the diagnostic tools introduced to assess the validity of each component.

Second innovation relative to the Double Machine Learning literature. (a) The DML framework (Chernozhukov et al., 2018) is widely used for causal effect estimation with flexible control of confounders, but most applications use standard learners such as Random Forest, Gradient Boosting, or feedforward neural networks for nuisance functions, which do not explicitly model spatial/network dependence structures between units. The innovation of this study is replacing these standard learners with a two-branch graph encoder that allows nuisance functions μ and π to simultaneously draw information from two separate sources of interdependence (geographical proximity and economic/network linkages), without merging these sources into a single combined weight matrix (which would make it impossible to disentangle each source’s contribution).

(b) Unlike the standard DML procedure that constructs cross-fitting folds randomly, we use spectral clustering based on the adjacency matrix with a Hungarian balancing algorithm to construct block-spatial folds. This choice prevents spatial information leakage between training and test folds—an issue that is overlooked in most DML implementations on network/spatial data and can lead to falsely optimistic cross-validation performance estimates.

Third innovation relative to the spatial econometrics’ literature. The classical spatial econometrics literature uses almost exclusively a single adjacency weight matrix (typically geographical) (Anselin, 1988; LeSage & Pace, 2009). The few studies that employ multiple weight matrices simultaneously typically estimate and report spatial dependence parameters for each matrix but rarely assess their identifiability separately and

quantitatively. The innovation of this study in this dimension is twofold: first, the simultaneous use of two independent matrices (spatial and network) within a deep learning on graph framework (not merely in classical linear models) to disentangle the contribution of each source; second, and perhaps more importantly, the introduction of a formal diagnostic tool for assessing the extent to which spatial dependence parameter estimates arise from actual data signal versus the purely geometric structure of the adjacency matrix—decomposition of the log-likelihood function into data-driven and geometric components, accompanied by independent validation with continuous optimization. To the authors' knowledge, such explicit decomposition is not standard in applied spatial econometrics reporting, even though parameter identifiability issues are well-known in the theoretical literature.

Fourth innovation relative to the emerging literature on GNN applications in causal inference. The application of Graph Neural Networks to causal effect estimation is still nascent, and most existing studies focus on introducing new architectures or demonstrating performance advantages on specific datasets without systematic reporting of under what data conditions such complexity is justified (Job et al., 2025). The innovation of this study in this dimension is not the specific GNN architecture employed (which uses known components such as GIN with skip connections and multi-head attention), but rather the reporting approach: this study explicitly provides Super Learner combination mechanisms, Cinelli-Hazlett sensitivity analysis, and Monte Carlo validation for permutation attribution, allowing conclusions about the justifiability of model complexity to be derived directly from the data rather than from the researcher's prior assumptions. This approach can be proposed as a methodological template for transparent reporting in this emerging field.

2. Literature Review

2.1. Robust Estimators for Staggered Adoption

Recent econometric literature has extensively demonstrated that conventional Two-Way Fixed Effects estimators can be severely biased in the presence of staggered treatment adoption and heterogeneous treatment effects over time. Goodman-Bacon (2021) showed that the TWFE estimator under such conditions can be decomposed as a weighted average of pairwise comparisons (including "already-treated vs. later-treated" comparisons that receive negative weights). Sun and Abraham (2021) similarly showed that event-study specifications with leads/lags can suffer from systematic contamination in the presence of treatment effect heterogeneity. Callaway and Sant'Anna (2021)—the main framework used in Section 4.2 of this paper—provided a solution that avoids this bias by disaggregating estimates to the group-time level (ATT(g,t)), enabling flexible aggregation of results to highlight treatment effect heterogeneity.

2.2. Spatial Spillovers in Diff-in-Differences Estimators

Most advances in Section 2.1 rely on the SUTVA (Stable Unit Treatment Value Assumption) assumption, which considers each unit's treatment status as having no effect on other units. This assumption is highly restrictive in the context of transportation infrastructure investments, which are inherently network-based and spatial. Butts (2023) provided a formal framework for identifying treatment effects in the presence of

spatial spillovers, disentangling the replacement effect from the combined effect of treatment and exposure to neighbors' treatment.

In the applied domain of transportation and housing prices, this challenge has previously been addressed through the "Spatial DiD" approach. Delgado and Florax (2015) first proposed the Spatial DiD model, incorporating spatial error correlation into the standard DiD specification. Diao, Leonard, and Sing (2017) applied this approach to estimate the effect of the Singapore Circle Line metro opening on housing prices, finding that homes near new stations experienced approximately 7.8% value appreciation relative to control areas. Bardaka, Delgado, and Florax (2019) extended this framework to multiple and sequential treatments (such as the gradual development of Denver's light rail line), successfully identifying direct, indirect, and total effects of each intervention separately—an approach conceptually close to the CS-DiD followed by spatial modeling in Section 4.5 of this paper, though Bardaka et al. focus on simultaneous or limited-event treatments, not staggered adoption with long time gaps (up to 16 years) as in this study's sample. Keeler and Stephens (2023) and Zhou, Chen, Han, and Zhang (2021), for Los Angeles and Shanghai respectively, provided further evidence of metro access capitalization in housing prices using Hedonic-DiD methods, without explicitly incorporating spatial spillovers.

A methodological feature of this study that has received less attention in the existing literature is the simultaneous combination of multi-group staggered adoption (not just binary or limited treatments) with two distinct spillover channels (spatial and non-geographical network) within a single framework.

2.3. Classical Spatial Econometrics

The family of Spatial Autoregressive (SAR), Spatial Error (SEM), and Spatial Durbin (SDM) models used in Sections 4.5 and 5.5 of this research are formulated based on the standard framework of LeSage and Pace (2009). This reference text formally introduces the distinction between direct and indirect (spillover) effects, which forms the basis for interpreting the coefficients in Table 7 of this study.

2.4. Causal Machine Learning: Double Machine Learning and Super Learner

Chernozhukov et al. (2018) introduced the Double/Debiased Machine Learning framework, which uses Cross-Fitting and Neyman orthogonality to enable the use of flexible machine learners for nuisance function estimation without losing the root-n convergence rate for the target causal parameter. This framework forms the basis of Section 4.3 of this study. For combining multiple heterogeneous learners in a data-driven manner, this study uses the Super Learner algorithm of van der Laan, Polley, and Hubbard (2007), who use cross-validation to show that a weighted combination of multiple learners asymptotically performs at least as well as the best single learner.

2.5. Graph Neural Networks in Causal Inference and Spatial Econometrics

The neural graph encoder architecture of this study (Section 4.3) draws from four standard graph message-passing families: Graph Convolutional Networks (Kipf & Welling, 2017), GraphSAGE (Hamilton, Ying &

Leskovec, 2017), Graph Attention Networks (Velickovic et al., 2018), and Graph Isomorphism Networks (Xu et al., 2019), which is theoretically the most expressive of these four architectures.

The application of Graph Neural Networks to causal inference is an emerging field. A recent review (Job et al., 2025) categorizes the spectrum of “graph-based causal learning” methods and highlights remaining challenges—including the need to account for spatial heterogeneity and spatial spillovers for accurate causal estimation.

2.6. Bad Controls and Sensitivity Analysis

Cinelli, Forney, and Pearl (2022) provided a graphical framework for diagnosing “bad control”—situations were controlling for a variable that is itself affected by treatment increases rather than reduces bias or absorbs the true causal effect; a discussion rooted in the analysis of Angrist and Pischke (2009). For quantitatively assessing the robustness of results to unobserved confounders, this study uses the Cinelli and Hazlett (2020) framework, which introduces the Robustness Value as a functional-form-independent measure of confounder strength. For statistical inference robust to a small number of clusters—a situation that applies precisely to this study’s sample (five valid treated provinces)—the Wild Cluster Bootstrap method of Cameron, Gelbach, and Miller (2008) is employed; these authors explicitly showed that standard asymptotic tests with fewer than 30 clusters can severely over-reject the null hypothesis.

3. Data and Institutional Framework

The data for this study comprises a balanced panel of 31 provinces over eighteen years (2006–2023), totaling 558 province-year observations. For each province-year unit, five categories of information were collected: (a) dependent variable (log housing price), (b) treatment variable (metro line operational status), (c) a set of socioeconomic explanatory variables, (d) geographical adjacency weight matrix, and (e) socioeconomic linkage weight matrix. The panel structure is completely balanced; no missing observations exist in either the time or cross-sectional dimension.

4. Methodology

4.1. Overall Framework and Rationale for Combining Three Methods

The aim of this study is to estimate the Average Treatment Effect on the Treated (ATT) of subway construction on log housing prices, in the presence of three simultaneous challenges: (a) staggered treatment adoption over time, which exposes the standard Two-Way Fixed Effects difference-in-differences approach to known biases (Goodman-Bacon, 2021; de Chaisemartin & D’Haultfoeuille, 2020), (b) the presence of observed confounding variables with potentially nonlinear relationships with both outcome and treatment, and (c) the existence of two independent sources of cross-sectional unit interdependence—geographical proximity and economic/network linkages—each of which can serve as a channel for spillover effects.

To address these three challenges simultaneously, three complementary approaches were employed: (1) the Callaway and Sant'Anna multiple-period difference-in-differences framework as the reference estimator robust to staggered adoption; (2) Double Machine Learning with a cross-fitting procedure for flexible, nonparametric control of confounders; and (3) a Graph Neural Network-based encoder with two independent branches within the DML framework, for simultaneously disentangling spatial from network spillovers. These three approaches are fitted independently and their results are compared as cross-cutting evidence, rather than treating one as a substitute for another.

4.2. Callaway-Sant'Anna Framework for Multiple-Period Difference-in-Differences

4.2.1. Definition of Target Parameter

For each adoption cohort g (the year in which member provinces constructed a metro line) and each time period t, the average treatment effect is defined as:

$$ATT(g,t) = E\left[Y_t\left(g\right) - Y_t\left(0\right) \mid \ G = g\right]$$

1

where Yt(g) is the potential outcome in period t under treatment in period g, and Yt (0) is the potential outcome without treatment. This quantity is estimated by comparing the mean change in housing prices for provinces in cohort g relative to the reference group (in this study, never-treated provinces) between the base period and period t:

$$\widehat{ATT}\left(g\right) = \left(\overset{'}{Y}_{post,g} - \overset{'}{Y}_{pre,g}\right) - \left(\overset{'}{Y}_{post,\infty} - \overset{'}{Y}_{pre,\infty}\right)$$

2

where the subscript ∞ refers to the never-treated group.

4.2.2. Managing Cohorts Without a Pre-Period

A critical methodological consideration explicitly managed in this study is the existence of cohorts whose treatment began in the first observed period of the panel and therefore lack a usable pre-period for differencing. In the sample under study, the Tehran province cohort (adoption in the first year of data) falls into this category. For such a cohort, only a raw level comparison between the post-treatment mean of the treatment group and the reference group can be computed:

$$\widehat{ATT}_{raw}\left(g\right) = \overset{'}{Y}_{post,g} - \overset{'}{Y}_{post,\infty}$$

3

which lacks the property of eliminating time-invariant level differences between groups and therefore cannot be interpreted as a valid estimate of the causal effect. In this study, such cohorts are explicitly

identified, excluded from the final ATT aggregation process, and their raw ATT is reported only for transparency, not for causal interpretation. The importance of this procedure is demonstrated empirically by comparing scales: in this study's sample, the raw ATT for the group without a pre-period (1.1858) exceeds the mean ATT of valid cohorts (approximately 0.22) by more than five times; careless inclusion of this group in the final aggregation could have significantly biased the overall estimate upward.

4.2.3. ATT Aggregation and Statistical Inference

The final ATT is calculated as the weighted average of ATT(g) for valid cohorts (those with a pre-period), with weights based on the number of units in each cohort, where Gvalid is the set of cohorts with a pre-period (Eq. 4).

$$\widehat{ATT} = \frac{\sum_{g \in \mathcal{G}_{valid}} n_g \cdot \widehat{ATT}(g)}{\sum_{g \in \mathcal{G}_{valid}} n_g}$$

4

Inference (Bootstrap standard errors): To estimate the standard error of the ATT, a cluster-bootstrap method at the province level was used: in each bootstrap iteration (B = 999), a set of provinces (with their entire time series, to preserve within-unit dependence structure) was resampled with replacement from the original sample, and the ATT was recalculated. The standard error is derived from the standard deviation of the bootstrap distribution, and the 95% confidence interval from the 2.5th and 97.5th percentiles of this distribution. The p-value was calculated using the add-one smoothing correction formula (Eq. 5) to avoid the known problem of reaching exactly zero in completely one-sided bootstrap samples.

$$p = \min\left(1,\ 2 \cdot \min\left(\frac{1 + \#\{b : \widehat{ATT}_b \leq 0\}}{1+B},\ \frac{1 + \#\{b : \widehat{ATT}_b \geq 0\}}{1+B}\right)\right)$$

5

It should be emphasized that this procedure is a simplified approximation of panel clustering and is not equivalent to the formal multiplier bootstrap procedure based on the influence function introduced in the original Callaway and Sant'Anna (2021) paper; it was chosen for computational simplicity and its adequacy for this study's purposes with a small sample.

4.3. Double Machine Learning Framework with Graph Encoder

4.3.1. Partially Linear Model Specification

The DML framework (Chernozhukov et al., 2018) models the treatment effect with the following partially linear specification:

$$Y_{it} = \theta \cdot T_{it} + g_0(X_{it}) + \epsilon_{it} \qquad , \qquad E[\epsilon_{it} \mid X_{it}, T_{it}] = 0$$

6

$$T_{it} = m_0\left(X_{it}\right) + \nu_{it} \qquad , \qquad E[\nu_{it} \mid X_{it}] = 0$$

7

where g0 and m0 are nonparametric nuisance functions for predicting the outcome and predicting treatment (generalized propensity score). The target parameter θ (treatment effect) is estimated using the Neyman orthogonal score method after removing the effects of nuisance functions (Eq. 8), where the residuals are obtained from cross-fitted nuisance functions (i.e., for each unit, predictions are taken from a model not trained on that unit's data, to avoid overfitting bias).

$$\widehat{\theta} = \frac{\sum_{i,t} \widehat{\nu}_{it} \cdot \widehat{\epsilon}_{it}}{\sum_{i,t} \widehat{\nu}_{it}^2}$$

8

4.3.2. Two-Branch Graph Encoder Architecture

The nuisance functions g0 and m0 are fitted using a shared neural network incorporating a two-branch encoder. Each branch applies an independent Graph Isomorphism Network (GIN) with skip connections to one of two adjacency matrices: the W branch for the spatial-geographical adjacency matrix, and the A branch for the network/economic linkage matrix. For each time period t, hidden representations from both branches are computed and summed (Eq. 9), using the GIN layer update rule of Xu et al. (2019) (Eq. 10), where M ∈ {W, A} and ε(k) is a learned parameter.

$$H_t = \mathrm{Norm}\left(\mathrm{GIN}_W(X_t, W) + \mathrm{GIN}_A(X_t, A)\right)$$

9

$$H^{(k+1)} = \mathrm{MLP}^{(k)}\left(\left(1 + \epsilon^{(k)}\right)\hspace{0.17em}H^{(k)} + M\hspace{0.17em}H^{(k)}\right),\ M \in \{W, A\}$$

10

A multi-head attention layer is then applied across the time axis to the sequence Ht for all t to capture long-term temporal dependencies in the final representation. The output of this encoder is fed as input to two separate networks for estimating μ (H, T) (outcome function) and π (H, T) (treatment function), with year fixed effects added as one-hot encodings to both networks to control for common time shocks.

4.3.3. Spatial Cross-Fitting Procedure

Unlike the standard random k-fold procedure, which in spatially structured data can cause information leakage between folds (because neighboring units may be placed in different training and test folds while remaining connected through the adjacency matrix), this study uses balanced spectral clustering (with capacity-constrained allocation via the Hungarian algorithm) to construct block-spatial folds, such that

neighboring units are placed in the same fold as much as possible and boundaries between folds align with weaker spatial connections.

To reduce variance from random fold selection, the entire cross-fitting procedure was repeated three times with independent random seeds, and the final residuals, and hidden representations (H) were averaged across the three repetitions. The variance between these three repetitions is reported as an indicator of cross-fitting procedure stability.

4.4. Super Learner Combination and Linear Correction

4.4.1. Super Learner Procedure

To increase the robustness of nuisance function estimation against potential failure of any single method, a Super Learner combination (van der Laan et al., 2007) of four components was employed: (1) the GNN encoder described in Section 4.3, (2) ElasticNet regression with year fixed effects, (3) Gradient Boosting, and (4) a base model (simple mean). Optimal combination weights were selected by grid search on the simplex to minimize the in-sample error of the combined model—MSE for μ and binary cross-entropy for π (Eq. 11).

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\ \Delta^{K-1}} \mathcal{L}\left(Y, \sum_{k=1}^{K} w_k \widehat{f}_k\right)$$

11

This procedure is entirely data-driven and imposes no minimum weight or manual intervention for any component—including the GNN component. The final weight of each component, including the possibility of a zero weight, is directly obtained from this optimization and reported without adjustment.

4.4.2. Residual Correction and Robust Standard Errors

After Super Learner combination, an optional linear correction layer is applied to the resulting residuals: the T and Y residuals obtained from the out-of-fold procedure are regressed on control variables and fixed effects, and these corrected residuals replace the original ones only if the variance explained by this linear regression is positive (i.e., the linear regression actually explains some residual variance not captured by the Super Learner). Otherwise (negative variance, indicating worse performance than the simple mean on out-of-fold data), this layer is deactivated and the original residuals are preserved. This decision criterion is designed to prevent degradation of residual quality by unstable correction in small samples.

Robust Standard Errors: The final standard error of the ATT is computed using two complementary methods: (1) Heteroskedasticity-Consistent (HC1/White) robust standard errors, and (2) Wild Cluster Bootstrap at the province level, which explicitly accounts for cluster dependence between observations from the same province and generally yields more conservative estimates than HC1.

4.5. Diagnostic Tests and Validation

4.5.1. Foundation Robustness Tests: Placebo, Parallel Pre-trend, and Monte Carlo Simulations

Placebo Test: To assess whether the estimated effect is merely a product of the specific treatment assignment structure in the sample, a placebo test with 200 repetitions was conducted: in each repetition, cohort labels were randomly permuted across provinces and the placebo ATT was computed using the same main estimation function. The p-value was calculated using the standard permutation test formula (add-one correction, Eq. 12) to ensure it never reaches exactly zero.

$$p = \frac{1 + \#\{b : |\ \widehat{ATT}_b\ | \geq |\ \widehat{ATT}_{obs}\ |\ \}}{1 + n_{perm}}$$

12

Parallel Pre-trend Test: For each cohort with a pre-period, the linear slope of the housing price trend in the pre-treatment period was compared between the treatment group and the reference group; a slope difference of less than 0.1 was interpreted as non-rejection of the parallel trends assumption.

Monte Carlo Simulations: To assess bias and RMSE of the CS-DiD estimate under controlled conditions, three data generating processes (DGPs) were designed: (1) simple linear relationship without strong spatial spillovers, (2) nonlinear (sinusoidal) relationship, and (3) linear relationship with strong spatial spillovers (ρ = 0.8). In each scenario, 200 artificial samples with known true treatment effect (θtrue = 0.15) were generated and CS-DiD estimation was performed on each sample; mean bias (mean estimate minus true value) and RMSE are reported as performance indicators.

4.5.2. Sensitivity Analysis to Omitted Variables (Cinelli-Hazlett)

The Robustness Value of each estimate against unobserved confounders was computed using the Cinelli and Hazlett (2020) criterion (Eq. 13):

$$RV = \frac{f^2}{1 + f^2} \quad , \quad f = \frac{|\ t|}{\sqrt{dof}}$$

13

where t is the t-statistic of the estimate and dof is the approximate degrees of freedom of the model. The Minimum Detectable Effect with 80% statistical power at 5% significance level was also computed as a complement to this analysis.

4.5.3. Supplementary Identifiability and Spatial-Network Disentanglement Tests

To assess the sensitivity of the main CS-DiD estimate to the implicit SUTVA assumption (absence of spillovers between units), the CS-DiD framework was repeated once with spatial error correlation control on the geographical adjacency matrix (W) and once on the network-economic linkage matrix (A)–based on the Delgado and Florax (2015) approach and following the cohort-based structure of Bardaka, Delgado, and Florax (2019). Unlike spillover modeling at the outcome variable level (which faces spatial lag endogeneity issues), this test models spillovers in the error structure after controlling for treatment and covariates (Eq. 14):

$$Y_{it} = \sum_{g,t} \theta_{g,t} D_{it}^{(g)} + X_{it}\beta + \alpha_i + \lambda_t + u_{it} \qquad , \qquad u_{it} = \rho \, \hspace{0.17em} M u_{it} + \epsilon_{it}$$

14

where M is either W or A. The parameter ρ is estimated by maximizing the concentrated likelihood on the residual of the model, and cohort coefficients θg,t are estimated by least squares after spatial transformation of the data. The aggregated ATT, following the procedure in Section 4.2.3, is computed as the weighted average of cohort coefficients with pre-periods.

Identifiability of Spatial Parameters: To assess whether estimated ρ, λ parameters arise primarily from actual data signal or from the geometric structure of the adjacency matrix, the concentrated log-likelihood function was decomposed into two components: the data-driven component (− n/2·log s²(ρ)) and the data-independent geometric component (log|det(I − ρW) |). The relative contribution of each component's range of variation to the total log-likelihood variation over a ρ grid was computed and reported as an identifiability index. To definitively rule out the possibility that the found optimum is merely an artifact of limited grid resolution, optimization of each parameter was independently repeated with a continuous optimization algorithm (Brent's method, implemented in scipy.optimize.minimize_scalar), and the difference between the discrete grid optimum and the continuous optimum is reported.

Disentangling Branch Contributions (Monte Carlo): The relative contribution of each branch (W and A) to the final GNN model prediction was estimated using permutation importance: the hidden representation of each branch is independently permuted within each time period, and the resulting increase in prediction error is used as the importance measure for that branch. To assess the reliability of this estimate with the limited sample size, an independent Monte Carlo validation was designed: in an artificial data generating process with known true contribution of each branch to constructing the outcome (wtrue = 0.7, atrue = 0.3) as a linear combination coefficient, the performance of the permutation importance estimate was assessed at different sample sizes (n = 300, 600, 1200). A crucial methodological point is that the theoretical target of this measure is not the raw linear combination coefficient, but the ratio of squared contributions to prediction variance (Eq. 15):

$$\text{target}_W = \frac{w_{true}^2}{w_{true}^2 + a_{true}^2}$$

15

because permutation importance is proportional to the squared contribution of each branch to output variance, not to the linear coefficient itself.

4.5.4. Control Variable Importance: Combining Two Complementary Approaches

Rationale for combining two methods: Assessing the importance of explanatory variables in nuisance functions (μ and π) is a known challenge in deep learning model interpretability: no single measure alone provides a complete picture of a variable's role in model prediction, as each method has different assumptions and structural sensitivities. To increase the robustness of findings against this methodological sensitivity, this study employs two complementary approaches based on different theoretical foundations: Integrated Gradients (IG; Sundararajan, Taly, & Yan, 2017) as a gradient-based method with the Completeness axiom property, and Permutation Importance as a model-agnostic performance-based method.

- IG measures the local, path-wise sensitivity of the model output to each feature by integrating gradients along the linear path from a baseline to the observed value. This method depends on the internal, differentiable structure of the model and measures importance encoded within the architecture.
- Permutation Importance measures each variable's importance through the increase in out-of-sample prediction error after randomly permuting that variable's values. This method is agnostic to model architecture and directly measures the generalization performance drop, not merely internal sensitivity.

This analysis was applied only to the outcome function (μ) with treatment status fixed at zero (full control scenario); its aim is to identify variables that play the greatest role in explaining housing price levels, independent of the treatment effect itself. See Table 6 in the Empirical Results section.

4.6. Software and Reproducibility

All analyses were implemented in Python using PyTorch for neural architectures, scikit-learn for Super Learner components and linear models, statsmodels for unit root tests, and SciPy for continuous optimization. For complete reproducibility, the main random seed was set to 42, and the cross-fitting procedure was repeated in three iterations with independent, specified seeds (42, 1042, 2042).

5. Empirical Results

5.1. Sample Descriptive Statistics and Treatment Structure

The final sample comprises a balanced panel of 31 provinces over 18 time periods (2006 to 2023), totaling 558 province-year observations. Of these 31 provinces, 6 (19.4% of the sample) constructed metro lines during the study period, while 25 provinces (80.6%) never constructed a metro line throughout the observation period and form the never-treated reference group. Table 1 shows the complete cohort structure.

Of the five identified cohorts, only Tehran lacks a pre-period (because its metro construction coincided with the first available year of data) and was therefore excluded from the final aggregated estimate according to the procedure described in Section 4.2.2. The final ATT estimate is based on 4 valid cohorts and 5 treated province observations.

Table 1
Treatment Adoption Cohort Structure

Cohort	Year	Provinces	Count	Pre	Post	Status
g = 1	2006	Tehran	1	0	18	No pre-period
g = 5	2010	Khorasan Razavi	1	4	14	Valid
g = 9	2014	Fars	1	8	10	Valid
g = 10	2015	Isfahan, East Azerbaijan	2	9	9	Valid
g = 17	2022	Alborz	1	16	2	Valid

5.2. Main Result: Average Treatment Effect

Table 2 shows the final ATT estimates from the two independent methods.

Table 2
Average Treatment Effect on the Treated (ATT) Estimates

Method	ATT	% Change	SE	SE Type	95% CI	p-value
CS-DiD	0.2196	+ 31.6%	0.0296	Bootstrap (cluster)	[0.162, 0.283]	0.002
DML (GNN + SL)	0.3849	+ 61.9%	0.0431	HC1 / White	—	—
DML (same)	0.3849	+ 61.9%	0.1161	Wild Cluster	[0.1856, 0.5842]	—

Both methods, independent of each other and with different identification assumptions, show the same positive direction and relatively similar scale of effect. However, the Wild Cluster standard error, which explicitly accounts for within-province cluster dependence, is approximately 2.7 times the HC1 standard error, indicating significant cluster dependence in the data, and the Wild Cluster should be considered the primary measure.

5.3. Pre-Correction Estimates: The Role of Super Learner and Linear Correction Layers

Tracing the evolution of the ATT estimate through different pipeline stages (Table 3) shows that the Super Learner combination played a decisive role in the final estimate.

Table 3
Evolution of ATT Estimate Through Pipeline Stages

Stage	ATT	ATE	ATU
After Cross-Fitting (GNN only)	1.4196	1.2352	0.5504
After Super Learner	0.3849	0.3768	0.3420
After linear correction	0.3849	0.3768	0.3420

The initial estimate based solely on the GNN encoder (before Super Learner combination) is substantially larger (1.42) than the final estimate. Examination of the optimal Super Learner weights (Table 4) shows that this difference mainly arises from the GNN component receiving zero weight in estimating the outcome function (μ).

Table 4
Optimal Super Learner Weights

Component	Weight in μ	Weight in π
GNN Encoder	0.00	0.50
ElasticNet + FE	0.20	0.20
Gradient Boosting	0.80	0.30
Base Model (mean)	0.00	0.00

These weights were computed entirely data-driven, without any manual intervention. In other words, in this specific dataset and with this sample size, the data-driven optimization procedure determined that the combination of Gradient Boosting and ElasticNet, without any contribution from the GNN model, provides the best out-of-sample prediction for the outcome function. Unlike the μ estimation where GNN received zero weight, in the π estimation (treatment probability) the GNN component received the highest weight (0.50) among the four learners—direct evidence that the usefulness of an architecture must be assessed separately for each estimation target.

5.4. Robustness and Validation Tests

5.4.1. Placebo Test

The placebo test with 200 random permutation repetitions yielded a p-value of 0.005, which is significant at the conventional 5% level. This result indicates that the estimated effect is unusually larger than the distribution of placebo effects from random treatment assignment.

5.4.2. Parallel Pre-trend Test

For all four cohorts with a pre-period (g = 5, 9, 10, 17), the difference in pre-treatment housing price trend slopes between the treatment group and the reference group was below the 0.1 threshold in all cases (0.0495, 0.0267, 0.0469, 0.0233, respectively), meaning that the parallel trends assumption—the

fundamental precondition for identification validity in the difference-in-differences framework—is not rejected for any cohort used in the final estimate.

5.4.3. Monte Carlo Simulations for Method Validation

Table 5 shows the results of Monte Carlo simulations (200 repetitions per scenario) for assessing bias and precision of the CS-DiD estimate under three controlled data generating processes.

Table 5
CS-DiD Performance in Monte Carlo Simulations

DGP Scenario	Bias	RMSE
Linear, no strong spillovers	+ 0.0029	0.1505
Nonlinear (sinusoidal)	+ 0.0008	0.1001
Linear, strong spatial spillovers (ρ = 0.8)	−0.0342	0.1537

In all three scenarios, the bias is near zero and substantially smaller than the RMSE. Even in the third scenario, where spatial spillovers were intentionally designed to be strong and thus the CS-DiD unit independence assumption is violated, the relative bias (− 0.0342) relative to the true effect (0.15) remains about 23% but is by no means misleading. These results indicate that the CS-DiD method does not produce large systematic bias even in the presence of ignored spatial spillovers.

5.4.4. Relative Importance of Control Variables

Table 6 shows the final ranking of the ten highest-ranked control variables based on both measures and a composite index. The rankings from the two methods, despite fundamental differences in computational foundations, show strong rank correlation: the unemployment rate ranks first or near-first in both methods, and the housing loan interest rate received exactly zero weight in both. This convergence, independent of the different structural assumptions of the two methods, provides evidence for the robustness of the ranking.

Table 6
Final Ranking of Control Variables

Rank	Variable	IG	Permutation	Mean
1	Unemployment rate	24.3%	14.8%	19.6%
2	Land price	16.4%	10.0%	13.2%
3	Rent index	6.6%	15.5%	11.0%
4	Inflation rate	7.5%	10.3%	8.9%
5	Housing investment	6.4%	9.8%	8.1%
6	Provincial GDP	7.0%	7.6%	7.3%
7	Household income	6.3%	7.6%	7.0%
8	Housing loan volume	10.9%	2.6%	6.7%
9	Construction cost	5.8%	7.3%	6.6%
10	Construction permits	5.0%	7.9%	6.5%

5.5. Comparison with Classical Panel Spatial Econometric Methods

Table 7 shows the ATT estimates from six classical panel specifications alongside CS-DiD. The range of estimates from this family of models [− 1.0398, 0.273] is considerably wide, and some specifications—particularly indirect effects (SDM)—show negative signs and large magnitudes that are difficult to explain economically. Four of the eight estimated spatial dependence parameters in this table ($\rho = 0.922$) reached exactly the boundary of the 30-point grid search across four different specifications, a clear indication of instability in estimating these parameters using classical panel methods with the existing data structure.

Table 7
Comparison of ATT Estimates from Classical Panel Methods

Method	Direct	Indirect	Spatial Parameter
CS-DiD (reference)	0.2196	—	—
OLS with Fixed Effects	0.1857	—	—
SAR (spatial W)	0.2603	—	ρ = 0.922 (boundary)
SAR (network A)	0.1340	—	ρ = 0.922 (boundary)
SEM (spatial W)	0.1924	—	λ = 0.512
SEM (network A)	0.1716	—	λ = 0.717
SDM (spatial W)	0.2726	−0.1922	ρ = 0.922 (boundary)
SDM (network A)	0.1877	−0.6959	ρ = 0.922 (boundary)
SDM (both W and A)	0.1557	−1.0398	—

5.6. Supplementary Inference Tests

5.6.1. Cinelli-Hazlett Sensitivity Analysis

Table 8 shows the sensitivity analysis results. The Robustness Value of the CS-DiD estimate (0.100) indicates that an unobserved confounder would need to explain at least 10% of the variance in both treatment and outcome to reduce the estimate to zero—a moderate level of robustness. In contrast, the DML estimate with RV = 0.139 shows slightly stronger robustness than CS-DiD; however, both estimates are in the moderate robustness range and neither decisively dominates the other.

Both estimates exceed their own MDE thresholds (0.3849 > 0.1207 for DML and 0.2196 > 0.0829 for CS-DiD), meaning that with 80% statistical power at 5% significance level, the estimated effects are statistically detectable.

Table 8
Sensitivity Analysis and Minimum Detectable Effect (MDE)

Estimate	RV	Label	MDE (80% power)
CS-DiD	0.100	Average	0.0829
DML (GNN)	0.139	Average	0.1207

5.6.2. Sample Size Limitation Detection

Automatic sample size detection explicitly flagged two limitations: the time period length (nT = 18) is below the conventional threshold of 30 periods, and the number of treated units (6) is below the conventional threshold of 10 units. These two limitations are the direct source of most cautionary considerations raised in subsequent sections.

5.7. Identifiability of Spatial Dependence Parameters: Quantitative Analysis on Residuals Controlled for Staggered Treatment

After removing the treatment effect using the CS-DiD ATT estimates for each cohort from the Y series, spatial dependence parameters were re-estimated on the residuals to avoid contamination of this estimate with the treatment effect. Table 9 shows the estimation results along with quantitative identifiability diagnostics.

Table 9
Spatial Dependence Parameter Estimates and Identifiability Diagnostics

Model	Grid (199-pt)	Continuous	Geometric Share
SAR_W	0.0600	0.0573	27.0%
SAR_A	−0.0600	−0.0611	62.7%
SEM_W	0.0600	0.0565	27.8%
SEM_A	−0.0800	−0.0767	58.1%
SDM_W	−0.1200	−0.1209	29.2%
SDM_A	−0.3600	−0.3607	40.2%
Mean	—	—	40.8%

Two complementary findings can be extracted from this table. First, the difference between the discrete grid estimate and the continuous optimization estimate for all six parameters is less than 0.004, definitively demonstrating that the reported optima are real and stable, not artifacts of limited grid resolution. Second, the share of the data-independent geometric term in the log-likelihood function ranges from 27.0% to 62.7% (mean 40.8%). Although this share is not high enough to make identification of these parameters completely impossible, it indicates moderate identifiability of these parameters with the available sample size, particularly for models using the network matrix A (SAR_A and SEM_A, with geometric shares above 58%).

5.8. Corrected Spatial-Network Spillover Estimates with Simultaneity Correction

Table 10 shows the spillover coefficient estimates before and after correcting the simultaneity problem using the spatial 2SLS method (Kelejian & Prucha, 1998).

Table 10
Comparison of Spillover Estimates With and Without Simultaneity Correction

Method	Spatial Coef.	Network Coef.
Spatial 2SLS (Kelejian–Prucha)	0.7386	0.2295
Simple Ridge (no correction)	0.6467	0.1153

The significant difference between the two methods (0.7386 vs. 0.6467 for the spatial coefficient) provides empirical confirmation that the simultaneity problem arising from the interdependence of Y with its neighbors indeed leads to measurable bias in the simple estimate. Consequently, the estimated spatial coefficient (0.7386) is at a reasonable distance from the theoretical stability boundary of the SAR model ($|\rho| < 1$), consistent with the findings in Section 5.7.

5.9. Disentangling Spatial-Geographical Branch from Network-Economic Branch Contributions

5.9.1. Main Estimation Result

Post-hoc attribution analysis on the final model showed that after prior removal of the treatment effect using cross-fitted residuals, the GNN encoder model with $R^2 = 0.9193$ explained 92% of the outcome variance. The permutation importance decomposition showed the following ratio:

Table 11
Relative Contribution of Spatial and Network Branches

Branch	Error Increase	% of Baseline	Relative Weight
Spatial (W)	0.095700	285.77%	53.8%
Network (A)	0.082035	244.97%	46.2%

The W to A ratio is 1.16, meaning that in this point estimate, the spatial branch has a slightly larger contribution (53.8% vs. 46.2%) to model prediction than the network branch.

5.9.2. Monte Carlo Validation of Estimate Reliability

To assess the reliability of the above estimate, an independent Monte Carlo simulation with known true contributions ($w_{true} = 0.7$ and $a_{true} = 0.3$ as linear coefficients) was designed at three different sample sizes. Table 12 shows the results.

Table 12
Monte Carlo Validation of Post-hoc Attribution Estimates

Sample Size (n)	Estimated W	Estimated A	Bias vs. Target
5,400	0.540 ± 0.254	0.460 ± 0.254	+ 0.3047
10,800	0.495 ± 0.035	0.505 ± 0.035	+ 0.3502
21,600	0.609 ± 0.232	0.391 ± 0.232	+ 0.2362

The theoretical target of this measure is not the raw linear coefficient (0.7 and 0.3), but the ratio of squared contributions to prediction variance ($w^2/(w^2+a^2)$ = 0.845 for the spatial branch), because permutation importance is proportional to the squared contribution of each branch, not to the linear coefficient itself. With this corrected reference target, even at the largest sample size tested in the simulation (n = 21,600, approximately 39 times the actual study data size of 558 observations), the bias remains 0.117 and the standard deviation of the estimate (0.232) is large.

5.10. Heterogeneous Treatment Effects (CATE)

The mean heterogeneous treatment effects estimated by the TauNet network was 0.2519—same sign as the ATT estimate (positive) in all specifications. However, the scale of this estimate (0.2519) is not identical to the ATT (0.3849), and given that estimating heterogeneous effects requires substantially more data than estimating a constant average effect, and only 6 treated units were available for training this network, this estimate should still be interpreted with caution and not as an independent, final finding alongside the ATT.

6. Discussion

6.1. Economic Interpretation of the Main Effect and Its Position in International Literature

The central finding of this study—a 31.6% increase in housing prices in provinces with metro lines, with qualitative convergence between two independent estimation methods (CS-DiD and DML)—should be placed in the context of the international literature on transportation infrastructure investment. Classical studies in developed economies typically report a range of 5 to 25% price increases within walking distance of metro stations, with considerable heterogeneity depending on urban context, business cycle stage, and identification method. This study's estimate (computed at the province level, not at the local station-distance level) falls at the upper end or slightly beyond this range.

6.2. GNN Encoder Performance in Super Learner: Two Contrasting Results for Two Estimation Targets

One notable finding of this study—reported fully and without adjustment—is that the entirely data-driven Super Learner procedure assigned substantially different weights to the GNN component in two separate competitions: in estimating the outcome function (μ), the GNN received zero weight; but in estimating the treatment function (π), the GNN received the highest weight (0.50) among all four learners (ElasticNet = 0.2,

Gradient Boosting = 0.3, base model = 0.00). This duality itself is direct, data-driven evidence that assessing the usefulness of a complex architecture should not be made as a general judgment (“GNN is good” or “GNN is bad”), but should be assessed specifically for each estimation target.

6.3. The Importance of Reporting Identifiability in Spatial Econometrics: A Common Gap in the Literature

The quantitative analysis in Section 5.7—which showed that 27 to 63 percent of the log-likelihood function variation in SAR/SEM models comes from the data-independent geometric component, not actual signal—points to a broader methodological gap in the applied spatial econometrics literature. The common practice in this literature is to report the point estimate of the spatial dependence parameter (ρ or λ) along with its standard error, without examining whether this parameter is fundamentally identified from data-driven signal or from the geometric structure of the adjacency matrix. This gap is particularly important in studies with a small number of spatial units or limited residual variation, because under such conditions the geometric term—which has no connection to the empirical question under investigation—can account for a substantial share of the maximization function.

The diagnostic tool introduced in this study—decomposition of the log-likelihood range into data-driven and geometric components, accompanied by independent validation with continuous optimization to rule out grid artifacts—provides a simple and generalizable procedure for this purpose.

6.4. The Importance of Explicit Staggered Adoption Management vs. Simplified Procedures

The comparison of the raw ATT scale for the Tehran cohort (1.1858, without a pre-period) with the mean ATT of valid cohorts (approximately 0.22)—a difference of more than five times—provides clear empirical evidence of the practical importance of recent multiple-period difference-in-differences procedures (Goodman-Bacon, 2021; Callaway & Sant’Anna, 2021; de Chaisemartin & D’Haultfoeuille, 2020). The standard, common Two-Way Fixed Effects procedure, which implicitly aggregates all cohorts with even negative weights in the final estimate, could have been significantly biased in this specific dataset precisely due to the presence of a cohort with a raw ATT several times larger than the rest, contaminated by a raw level comparison (not a difference-in-difference).

6.5. Considerations for Transparency in Machine Learning Applications to Causal Inference

Given that economic data volumes are predominantly limited and small, and given broader concerns in quantitative social sciences about the reproducibility crisis and selective reporting of results, this study argues that the value of an analytical framework should not be judged solely by the “cleanliness” of its final results, but should include an assessment of the extent to which the framework enables the diagnosis and reporting of its own limitations. The diagnostic tools introduced in this paper (Sections 4.3 and 4.4) provide a practical example of this approach and can be applied independently of this study’s specific application (metro effect on housing prices) in other research based on DML and GNN with small panel data.

7. Limitations and Methodological Considerations

7.1. Sample Size Limitation and Statistical Power

Given that economic data volumes are predominantly limited and small, the most important limitation of this study is the small effective sample size for identifying certain model parameters. Of the 31 provinces in the data, only 6 (Tehran, Khorasan Razavi, Fars, Isfahan, East Azerbaijan, Alborz) constructed metro lines during the study period. After excluding the Tehran cohort (which lacks a pre-period because its metro construction occurred in the first year of data, making its estimate based on a raw level comparison rather than a difference), the final CS-DiD estimate is based on only 4 independent cohorts and 5 treated observations. Automatic sample size detection shows that both the time period length (nT = 18 vs. conventional threshold of at least 30) and the number of treated units (6 vs. conventional threshold of 10) are below conventional values for panel analyses with sufficient statistical power.

7.2. Summary and Recommendations for Future Studies

Despite the limitations noted above, the qualitative convergence of the main results (direction and relatively similar effect scale in DML, CS-DiD, and SEM_W methods) and the absence of substantial systematic bias in controlled Monte Carlo simulations (near-zero bias in three different DGP scenarios: linear, nonlinear, and strong spatial spillovers) provide valid evidence for the validity of the overall direction of the estimated effect. However, the fundamental limitations arising from sample size (particularly in estimating spatial parameters and disentangling branch contributions) require that these findings be considered as preliminary results conditional on the available data. Increasing the number of treated units (by expanding the geographical scope of the study or adding more time periods) offers the greatest leverage for increasing the identification power of these parameters and reducing the uncertainty reported in this section, and is suggested as the primary priority for future complementary studies.

8. Conclusion

8.1. Summary of Diagnostic Framework Findings

This study introduced an integrated framework for estimating the causal effect of subway construction on housing prices, combining three complementary approaches—multiple-period Callaway-Sant'Anna difference-in-differences, Double Machine Learning with cross-fitting, and a Graph Neural Network-based encoder with two independent branches for disentangling spatial-geographical from network-economic spillovers—in a coherent structure. Using a panel of 31 Iranian provinces over 18 years, two independent estimates showed that subway construction is associated with a significant 31.6% to 61.9% increase in housing prices (CS-DiD: ATT = 0.2196 with province-level cluster bootstrap, p = 0.002; DML: ATT = 0.3849 with robust standard errors for heteroskedasticity and cluster dependence). This qualitative convergence, combined with the absence of substantial systematic bias in the CS-DiD method under controlled Monte Carlo simulations with three different data generating processes, provides valid evidence for the existence and direction of this effect.

Beyond this specific empirical finding, this study introduced a set of quantitative diagnostic tools that are generalizable independently of this research's specific application: formal decomposition of spatial econometric model log-likelihood into data-driven and data-independent geometric components, accompanied by dual validation (discrete grid vs. continuous optimization) for assessing the actual identifiability of spatial dependence parameters; and Monte Carlo validation with a corrected theoretical target (ratio of squared contributions, not raw linear coefficients) for assessing the reliability of permutation-based attribution estimates at a given sample size. Application of these tools to this study's data showed that spatial dependence parameters (with a mean geometric component share of 40.8% in the log-likelihood function) are moderately identifiable, and disentangling spatial/network branch contributions (53.8% vs. 46.2%) has considerable noise with the available sample size (558 observations).

8.2. Message for Applied Researchers

Three specific practical messages can be extracted from this study for researchers in applied econometrics and social data science. First, reporting spatial dependence estimates in SAR/SEM/SDM models should be accompanied by a quantitative identifiability index, particularly in studies with small panels or limited residual structure after controlling for other effects. Second, in any difference-in-differences study with staggered treatment adoption, explicit examination of cohort composition and identification of cohorts without a pre-period should be a standard part of the analysis procedure; in this study, the exclusion of a cohort with a raw ATT more than five times larger than the mean of the rest provided a concrete example of the practical importance of this procedure. Third, in applications of deep learning models—particularly GNNs—to causal inference, the use of data-driven combination procedures like Super Learner should be standard, so that the data itself—not the researcher's prior assumption—determines whether model complexity is justified.

Declarations

Author Contribution

A. Y.: Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Visualization. S. S: Conceptualization, Supervision, Methodology, Writing – review & editing, Project administration.B. S: Supervision, Validation, Writing – review & editing.

Acknowledgement

I sincerely thank all esteemed professors for their valuable guidance and support. We are particularly grateful to my dear wife, Sahar Abbasi Anari, whose unwavering encouragement and assistance were instrumental in the successful completion of this research.

Data Availability

The data and code that support the findings of this study. https://github.com/yousefiAI/GIN-DML

References

1. Abate, G. D. (2017). Essays on spatial and spatio-temporal modeling of housing prices. Doctoral dissertation, University of Copenhagen.
2. Ahlfeldt, G. M., Redding, S. J., Sturm, D. M., & Wolf, N. (2015). The economics of density: Evidence from the Berlin Wall. Econometrica, 83(6), 2127–2189.
3. Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics: An Empiricist's Guide. Princeton University Press.
4. Anselin, L. (1988). Spatial Econometrics: Methods and Models. Kluwer Academic Publishers.
5. Bardaka, E., Delgado, M. S., & Florax, R. J. G. M. (2019). A spatial multiple treatment/multiple outcome difference-in-differences model. Transportation Research Part A, 121, 325–345.
6. Bramoullé, Y., Djebbari, H., & Fortin, B. (2016). Peer effects in networks: A survey. Annual Review of Economics, 8, 639–666.
7. Butts, K. (2023). Difference-in-Differences with Spatial Spillovers. arXiv preprint arXiv:2105.03737.
8. Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. Journal of Econometrics, 225(2), 200–230.
9. Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. Review of Economics and Statistics, 90(3), 414–427.
10. Can, A. (1992). Specification and estimation of hedonic housing price models. Regional Science and Urban Economics, 22(3), 453–474.
11. de Chaisemartin, C., & D'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review, 110(9), 2964–2996.
12. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68.
13. Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B, 82(1), 39–67.
14. Cinelli, C., Forney, A., & Pearl, J. (2022). A crash course in good and bad controls. Sociological Methods & Research, 53(3), 1071–1104.
15. Debrezion, G., Pels, E., & Rietveld, P. (2007). The impact of railway stations on residential and commercial property value: A meta-analysis. The Journal of Real Estate Finance and Economics, 35(2), 161–180.
16. Delgado, M. S., & Florax, R. J. G. M. (2015). Difference-in-differences techniques for spatial data. Economics Letters, 137, 123–126.
17. Diao, M., Leonard, D., & Sing, T. F. (2017). Spatial-difference-in-differences models for impact of new mass rapid transit line on private housing values. Regional Science and Urban Economics, 67, 64–77.
18. Ding, J., Cen, W., Wu, S., Chen, Y., Qi, J., Huang, B., & Du, Z. (2024). A neural network model to optimize the measure of spatial proximity in geographically weighted regression approach. International Journal of Geographical Information Science, 38(7), 1315–1335.

19. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression. John Wiley & Sons.
20. Geerts, M., vanden Broucke, S., & De Weerdt, J. (2024). A benchmark study of machine learning models for house price prediction. International Journal of Data Science and Analytics.
21. Gholizadeh, A., & Samadipoour, S. (2024). Behavioral and Non-Behavioral Factors Affecting Housing Prices and Inflation in Iran. Quarterly Journal of Economic Research, 24(3).
22. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
23. Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics, 225(2), 254–277.
24. Hall, P. (1985). Resampling a coverage pattern. Stochastic Processes and Their Applications, 20(2), 231–246.
25. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. NeurIPS, 30.
26. Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press.
27. Jackson, M. O. (2010). Social and Economic Networks. Princeton University Press.
28. Job, S., Tao, X., Cai, T., Xie, H., Li, L., Li, Q., & Yong, J. (2025). Exploring causal learning through graph neural networks. WIREs Data Mining and Knowledge Discovery.
29. Keeler, Z. T., & Stephens, H. M. (2023). The capitalization of metro rail access in urban housing markets. Real Estate Economics, 51(3), 686–720.
30. Kelejian, H. H., & Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure. Journal of Real Estate Finance and Economics, 17(1), 99–121.
31. Khatami, S. B., Parikh, H., Chen, H., Roy, S., & Salimi, B. (2024). Graph machine learning based doubly robust estimator for network causal effects. AISTATS 2025, 258, 4366–4374.
32. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. ICLR 2017.
33. Lahiri, S. N. (2013). Resampling Methods for Dependent Data. Springer.
34. LeSage, J., & Pace, R. K. (2009). Introduction to Spatial Econometrics. CRC Press.
35. Neyman, J. (1959). Optimal asymptotic tests of composite statistical hypotheses. In Probability and Statistics, 213–234.
36. Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.
37. Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 90(429), 122–129.
38. Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82(1), 34–55.
39. Sun, L., & Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 225(2), 175–199.

40. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. ICML 2017, 70, 3319–3328.
41. van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical Applications in Genetics and Molecular Biology, 6(1).
42. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph attention networks. ICLR 2018.
43. Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? ICLR 2019.
44. Zhou, Z., Chen, H., Han, L., & Zhang, A. (2021). The effect of a subway on house prices: Evidence from Shanghai. Real Estate Economics, 49(S1), 199–234.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineSupplementaryMaterialARS.docx