Counterfactual Reconciliation: Incorporating Aggregation Constraints For More Accurate Causal Effect Estimates

Cengiz, Doruk and Tekgüç, Hasan

The Home Depot, Kadir Has University

June 2022

Counterfactual Reconciliation: Incorporating Aggregation Constraints For More Accurate Causal Effect Estimates

Doruk Cengiz and Hasan Tekgüç*

Abstract: We extend the scope of the forecast reconciliation literature and use its tools in the context of causal inference. Researchers are interested in both the average treatment effect on the treated and treatment effect heterogeneity. We show that ex post correction of the counterfactual estimates using the aggregation constraints that stem from the hierarchical or grouped structure of the data is likely to yield more accurate estimates. Building on the geometric interpretation of forecast reconciliation, we provide additional insights into the exact factors determining the size of the accuracy improvement due to the reconciliation. We experiment with U.S. GDP and employment data. We find that the reconciled treatment effect estimates tend to be closer to the truth than the original (base) counterfactual estimates even in cases where the aggregation constraints are non-linear. Consistent with our theoretical expectations, improvement is greater when machine learning methods are used.

Keywords: Forecast Reconciliation; Non-linear Constraints; Causal Machine Learning Methods; Counterfactual Estimation; Difference-in-Differences

Declarations of interest: None.

*Doruk Cengiz is a Lead Data Scientist at The Home Depot, Hasan Tekgüç is an Associate Professor of Economics at Kadir Has University (e-mail: hasan.tekguc@khas.edu.tr). Please address all correspondence to Doruk Cengiz (dcdorukcengiz@gmail.com).

1. Introduction

The burgeoning forecast reconciliation literature offers systematic approaches which combine optimum forecasts from different levels of hierarchy to ensure that the aggregation constraints arising from the grouping structure are satisfied. For example, although it is a priori knowledge that regional sales sum up to national sales, discrepancies between the sum of regional sales forecasts and the national sales forecasts may occur because the best performing models often vary at different levels of hierarchy. The methods proposed in the literature ex post correct the discrepancies while utilizing all the forecasts.

To date, the focus of the literature has been mostly limited to forecasting multivariate time series even though hierarchical or grouped data is ubiquitous in many research areas. In particular, in causal effect analyses, to assess the heterogeneous impact of a treatment as well as the average treatment effect on the treated, it is common to create hierarchical or grouped data by decomposing the overall population into subgroups. With this data, researchers separately estimate the causal effect for each of the groups by comparing what happens after implementation of the treatment (the actual universe) with what would happen in the absence of the treatment (the counterfactual universe). The aggregation constraints are satisfied in the former, whereas they may be violated in the latter, leaving us with an odd counterfactual universe where the basic rules of mathematics do not apply.

In this paper, we extend the scope of the existing forecast reconciliation literature and demonstrate the value of incorporating those identities that arise from the grouping structure of data in answering causal questions. We make four main contributions. First, we show that the

methods in the forecast reconciliation literature can be used when treatment effect heterogeneity as well as the average treatment effect on the treated are of interest. Relying on the potential outcomes framework (Rubin 1974), we reformulate the causal effect question as a question of predicting the unobserved counterfactual. This reformulation allows the use of methods developed in the forecast reconciliation literature in constructing coherent group-specific counterfactual estimates that correctly add up across the hierarchy. Second, influenced by the "optimal combination" approach of Hyndman et al. (2011) and Wickramasuriya et al. (2019), we propose a reconciliation method that allows outcome variables and the treatment effects estimate to be non-linearly transformed. In these cases, the researcher is interested in having more accurate causal effect estimates in the transformed space, so the linear aggregation constraints appear as non-linear. Therefore, we (indirectly) allow non-linear constraints. Third, building on Panagiotelis et al.'s (2021) geometric interpretation of reconciliation, we also document the factors that precisely determine the size of the overall improvement in accuracy due to the proposed reconciliation method, whether it is used for counterfactual estimation or forecasting. In addition, we show under which conditions reconciliation is redundant when the least squares estimators, the most commonly used estimators in causal inference analyses, are used. The implication of the proof is that when one of these conditions is violated, the counterfactual estimates at different levels of the hierarchy will not be coherent. Fourth, with the U.S. state-by-industry employment and GDP data, we explore whether and to what extent reconciliation is beneficial in terms of counterfactual estimate accuracy. We demonstrate that, even when the underlying theory requires the outcome variables to be non-linearly transformed, reconciliation, on average, improves the overall accuracy in the transformed space. We observe that

improvement becomes particularly greater when causal machine learning tools are used or when different models are employed for different groups in the hierarchy.

The remainder of the paper is organized as follows. In the next section, we briefly discuss forecast reconciliation and the potential outcomes framework literature. Section 3 details counterfactual reconciliation and provides new insights on reconciliation using geometric interpretation. Section 4 explains our design of experiments to measure the benefits of reconciliation. Section 5 describes the data, and Section 6 presents the findings. Section 7 concludes.

2. Literature Review

In this section, we briefly review forecast reconciliation and the potential outcomes framework literature.

2.1 Forecast Reconciliation

Many time series data are naturally hierarchical. Sales of individual firms aggregate to total industry sales. Electricity consumption of individual cities aggregate to regions; regions, to national total. Their accurate forecasts have obvious economic value, such as avoiding accumulation of unsold stocks or avoiding power outages.

Forecasts at each level can be valuable for different stakeholders. While a general manager might be interested in knowing the total worldwide sales of a particular product, a local manager might be responsible for meeting regional demand. While, ideally, best forecasts for each level add up

appropriately in the hierarchy (i.e., there is coherency among forecasts at different levels), this is usually not the case for various reasons, among which include using different models and transformations that tend to perform better at different levels.

To make forecasts, researchers used to rely on bottom up, top down, or middle out methods. The bottom up approach exploits the rich and detailed information that can be extracted from the bottom level and builds bottom level forecasts and aggregates up. The top down approach is useful when the signal to noise ratio at the most disaggregated level is very low, yet at the top level the noise is mitigated and the signal, though less rich in detail, is quite strong. Put differently, it is possible that aggregated data may reveal trends that are quite difficult, if not impossible, to capture at the bottom level (Hollyman et al., 2021). Then, the researcher predicts the total sales or demand for electricity, and sub-groups' forecasts are disaggregated according to some rules. The middle out approach aims to find the sweet spot in this rich-in-detail vs. noise tradeoff, where accuracy is maximized according to some error metric. It builds the forecasts between the most bottom and top levels. Then, the forecasts are parceled out proportionally downwards and are aggregated upwards (Hyndman and Athanasopoulos 2018).

Athanasopoulos et al. (2009) show that no single method, neither top down nor bottom up, outperforms others consistently. Almeida et al. (2016) point out that arguments for both approaches are sensible, and as specific problems and/or forecast horizons change, different approaches can perform better. In their seminal paper, Hyndman et al. (2011) introduce the optimal combination approach, which sparked the literature on forecast reconciliation with linear constraints. Hollyman et al. (2021) demonstrate that forecast reconciliation is a type of

combination of forecasts in which the original forecasts (also called base forecasts) optimized for different levels are combined in a principled, systemic, and scalable approach. The reconciled forecasts, as opposed to the original ones, conform to the *natural aggregations constraint* of hierarchical data. The paper offers a dose of humility to over-confident modelers who view a particular model as superior to the alternatives.

The key element of forecast reconciliation is coherence; i.e., bottom series aggregate to the top, while using all available forecasts. Wickramasuriya et al. (2019) provide alternative reconciliation methods. Panagiotelis et al. (2021) provide a geometric interpretation and document when the reconciled forecasts are, overall, guaranteed to be better than the original (or base) forecasts. Empirical results in Wickramasuriya et al. (2019) and Panagiotelis et al. (2021) show that reconciled forecasts, overall, tend to be, on average, more accurate than the original (base) forecasts. Spiliotis et al. (2021) introduce machine learning models to the reconciliation procedure and allow non-linear combinations of the original (base) forecasts.[1]

2.2 Potential Outcomes Framework

The second strand of literature we build on is the Rubin causal model, also known as the potential outcomes framework (Rubin 1974; Imbens and Rubin 2015; Imbens 2020). According to this framework, researchers examining causal effects of a treatment (policy or event) on treated units construct a counterfactual universe to explore "what would happen if the treatment were never implemented." In this universe, everything up until the treatment is the same as the

[1] Although beyond the scope of the current study, the nascent probabilistic forecast reconciliation literature goes beyond reconciling point forecasts and reconciles forecast densities (see Gamakumara (2020) for a more detailed analysis).

actual universe. However, the counterfactual universe is not subjected to the treatment. As a consequence, the outcome variables of the actual and counterfactual universes might diverge and the divergence is interpreted as the causal treatment effect on the treated.

In their 2017 review of recent developments in econometrics for applied researchers, Athey and Imbens (2017) highlight research in three areas: (i) new research on identification strategies such as synthetic controls and regression discontinuity, (ii) sensitivity and robustness analyses for more credible identification strategies, and (iii) recent advances in machine learning for causal effects. The intention of all these recent advances is reducing bias in the coefficient estimates while providing policy evaluation. In policy evaluation, linear regressions with period and group fixed effects (two-way fixed effects, TWFE) are the work horse of applied researchers (see Angrist and Pischke (2008) for the details and assumptions of TWFE and other commonly used linear models). de Chaisemartin and D'Haultfoeuille (2022b: 1) note that "26 of the 100 most cited papers published by the American Economic Review from 2015 to 2019 estimate such regressions"[2]. Moreover, de Chaisemartin and D'Haultfoeuille's (2022b) review shows that a rapidly growing econometrics literature is concerned with estimating the heterogeneous treatment effects of policy between groups and over-time. This recent focus on estimating heterogeneous treatment effects also seeks to reduce the bias in coefficient estimates.

On the other hand, forecast reconciliation is a post-estimation procedure within the forecasting literature and is designed to reduce forecast error even when the original estimates are unbiased. It utilizes existing linear aggregation constraints in data to improve accuracy. Hence, it is a

[2] See de Chaisemartin and D'Haultfoeuille (2022a) Table 2 for the list of papers.

procedure that complements those recent developments in applied econometrics focused on eliminating potential sources of bias in coefficient estimates.

Mathematically, the true average treatment effect on the treated (β^T) is as follows[3]:

$$\beta^T = \frac{1}{M}\sum_{j=1}^{M}\left(Y_j^1 - Y_j^0\right) \quad (1)$$

where Y indicates the outcome variable, the superscripts are the treatment indicator (i.e., superscript 1 refers to a world with treatment and superscript 0 refers to a world without treatment), and M stands for the number of trials (i.e. the number of unit-time pairs that are treated). The first term on the right-hand side of equation (1), (Y_j^1), is what we observe in the actual universe where the treatment is implemented. The second term (Y_j^0) is from the counterfactual universe which is not directly observed.

Both researchers and policy makers are commonly interested in the overall effect of treatment on the treated population as well as its differential effects on certain subgroups such as by gender, by race and ethnicity, by region or by industry. This analysis of subgroups can also illuminate how and through which channels treatment affects the population. To assess treatment effect heterogeneity, researchers construct counterfactual universes specific to each of the subgroups in addition to the one for the overall population. Then, we can generalize equation (1) which becomes:

[3] We use a notation similar to that in Rubin (1974).

$$\beta_g^T = \frac{1}{M}\sum_{j=1}^{M}(Y_{j,g}^1 - Y_{j,g}^0) \quad (2)$$

where subscript *g* indicates the group of interest (subgroup or overall population).

In reality, once the policy or treatment is implemented, it is not possible to observe both the actuals and the counterfactuals, so the latter are estimated. Thus, the causality question implicitly becomes a prediction question in the sense that we are trying to predict/estimate the unobserved counterfactuals (Athey and Imbens 2017; Athey et al. 2021). Then, equation (2) becomes:

$$\widehat{\beta_g^T} = \frac{1}{M}\sum_{j=1}^{M}(Y_{j,g}^1 - \widehat{Y_{j,g}^0}) \quad (3)$$

where hat (^) indicates the term is estimated.

3. Counterfactual Reconciliation for hierarchical and grouped data

3.1 Room for Reconciliation in Counterfactual Estimation

Converting the causal question into a prediction question allows us to use advances in the forecast reconciliation literature. When researchers analyze treatment effect heterogeneity by subgroups as well as the average treatment effects on treated units, certain linear dependence relationships or constraints naturally arise. For instance, the sum of white and non-white employees adds up to total employment. Also, the total number of individuals employed in both the manufacturing and non-manufacturing industries adds up to total employment as well. These statements are true whether the treatment is implemented or not. However, counterfactual estimates are very unlikely to satisfy these linear constraints even if they are all unbiased and consistent because the counterfactual estimates are random variables. In the theorem below we

identify the specific conditions that need to hold to satisfy linear constraints for counterfactual estimates.

Theorem: *If linear least squares estimators are employed to construct counterfactuals, and outcome variables are transformed by a linear bijective function, and the same regressor matrices with n observations and k variables are used, then linear dependencies among the outcome variables that hold in the actual universe hold in the counterfactual universe as well.*

Proof: For a sequence of n-dimensional column vectors $\{y_1, y_2, \ldots, y_G\}$, if there exists $\{a_g\}$ such that for some y_1, we have $y_1 = \sum_{g=2}^{G} a_g * y_g$, then we call $\{y_1, y_2, \ldots, y_G\}$ linearly dependent.

Using linear least squares estimators with column vectors in n-by-k matrix Z as the exogenous variables, we have $\widehat{\beta_g} = (Z'Z)^{-1}Z'y_g$ for all g. This implies that the counterfactuals of the treated units for the treated periods can be written as $\widehat{y_g^0} = y_g - \widehat{\beta_g^T} * T_g$, where the subscript g indicates the groups (subgroups or overall population), $\widehat{y_g^0}$ is the estimated counterfactual, and T represents the treatment indicator column(s) in Z. This can be simplified as $\widehat{y_g^0} = y_g - \widehat{\beta_g^T}$ since all T_g are 1 for the treated units for the treatment periods.

Then, for the counterfactuals to satisfy the same linear constraints as the actuals, it is sufficient and necessary to show:

$$\widehat{y_1^0} = \sum_{g=2}^{G} a_g * \widehat{y_g^0}$$

We know

$$y_1 = \sum_{g=2}^{G} a_g * y_g$$

Multiplying both sides by $(Z'Z)^{-1}Z'$

$$(Z'Z)^{-1}Z'y_1 = (Z'Z)^{-1}Z' \sum_{g=2}^{G} a_g * y_g = \sum_{g=2}^{G} a_g * (Z'Z)^{-1}Z'y_g$$

$$\widehat{\beta_1} = \sum_{g=2}^{G} a_g * \widehat{\beta_g}$$

Following the last line, for the average treatment effect on the treated, we have

$$\Rightarrow \widehat{\beta_1^T} = \sum_{g=2}^{G} a_g * \widehat{\beta_g^T}$$

Multiplying both sides by -1 and adding y_1 yields

$$y_1 - \widehat{\beta_1^T} = y_1 - \sum_{g=2}^{G} a_g * \widehat{\beta_g^T} = \sum_{g=2}^{G} a_g * y_g - \sum_{g=2}^{G} a_g * \widehat{\beta_g^T} = \sum_{g=2}^{G} a_g * (y_g - \widehat{\beta_g^T})$$

$$\widehat{y_1^0} = \sum_{g=2}^{G} a_g \widehat{y_g^0}$$

■

The theorem indicates that (i) if the regressor matrices are not identical across groups or (ii) if the outcome variables are transformed so that the linear constraints are no longer satisfied ($h(y_1) \neq \sum_{g=2}^{G} a_g * h(y_g)$ *when* $y_1 = \sum_{g=2}^{G} a_g * y_g$), then the linear dependencies are likely not to hold in the counterfactual universe.[4,5] These cases include using (i) group-specific control

[4] One interesting exception here is that if the control variables in all Z_g's and the treatment indicators are all uncorrelated, the treatment indicators do not vary between groups, and the outcome variables are not transformed, then the linear constraints would hold in the counterfactual universe as well even if different group-specific control variables are used. This result comes from the Frisch-Waugh-Lovell theorem, as adding controls that are uncorrelated with the treatment variables would have no impact on the estimated treatment effects on treated. Nonetheless, it is important to emphasize that the correlation coefficients between all control variables and the treatment indicators must be exactly 0 in the observed sample for this to happen.

[5] If there are endogenous variables in Z, researchers commonly use the two-stage least squares estimator (2SLS). The first stage of 2SLS, the endogenous components of Z are purged and $\hat{Z}$ is produced. In the second stage, $\hat{Z}$ is used. The theorem holds in this case as well, except Z is replaced by $\hat{Z}$. In Appendix Table A.3, we show that our main findings and arguments can be extrapolated to the case where 2SLS is employed.

or instrumental variables, (ii) group-specific weights for observations, (iii) data-driven methods such as machine learning to specify the exact functional forms in which the predictors enter the model, or (iv) all kinds of transformations of the outcome variable that invalidate the linear constraint.[6]

Because we know a priori that the true counterfactuals have to be on the constraint hyperplane (see Figure 1 and the accompanying discussion for visual example), there is room for improving the counterfactuals' accuracy by reconciling them when the constraints do not hold in the counterfactual universe.

3.2 Reconciliation as a Post-Estimation Procedure

Before we discuss the proposed counterfactual reconciliation method, we rewrite equation (3) to allow transformations of the outcome variable:

$$\widehat{\beta_g^T} = \frac{1}{M}\sum_{j=1}^{M}(h(Y_{j,g}^1) - h(\widehat{Y_{j,g}^0})) \quad (4)$$

where bijective function h is allowed since the underlying theory may require the outcome variable and the treatment effect estimate to be transformed. In other words, we allow cases where there are linear dependence relationships between outcome variables in grouping G, yet the outcome variables appear transformed in the theory.[7]

[6] The examples of the latter include log-transformations and normalizations by group-specific populations.

[7] For instance, empirical macroeconomists usually log-transform the Cobb-Douglas production function (its simplest form is $Y = AK^{\alpha}L^{\beta}$ where Y indicates the output (GDP); A, total factor productivity; K, capital; L, labor; α and β, the output elasticities of capital and labor, respectively) to obtain a linear model (e.g. Gechert et al., 2021; Mankiw et al., 1992). After the log-transformation, the coefficients of the multiple regression estimate A, α, and β.

Influenced by the "optimal combination" approach of Hyndman et al. (2011), for all treatments j, we propose the following constrained minimization for reconciliation:

$$\widehat{\Lambda_{g,J}} = argmin_{\Lambda_{g,j}} L_j\left(h(\widehat{Y^0_{g,J}}), h(\Lambda_{g,j})\right)$$
$$subject\ to\ A\Lambda_{g,j} = 0 \tag{5}$$

where L_j indicates that the loss function is calculated per treatment, $\widehat{Y^0_{g,J}}$ are the original counterfactual estimates, h is the transformation, and $\Lambda_{g,j}$ are the reconciled counterfactuals for treatment j and group g, respectively. A matrix indicates the linear constraints. Note that if L_j's are the mean squared error and h is the identity function, equation (5) becomes the treatment-specific orthogonal projections of the counterfactual estimates on the hyperplane where the linear constraints are satisfied.[8] When h is non-linear, numerical methods can be used to solve equation (5).[9]

There are four main benefits of the proposed reconciliation. First, according to our experiments with real-world data, the reconciled counterfactual estimates, overall, (i) are guaranteed to be closer to the true counterfactuals than the original estimates when h is a bijective linear function

Similarly, it is common to use the log of employment in the labor economics literature as well (e.g. Allegretto et al., 2017).

[8] This is called the OLS estimator in Hyndman et al. (2011) where the outcome variable is composed of the original forecasts at different aggregation levels and the right-hand side variables are columns of the "summing matrix" which orthogonally projects the original forecasts onto the hyperplane.

[9] In the experiments below, when h is non-linear, we use numerical methods to obtain the reconciled counterfactual estimates. In particular, we use the "alabama" library by Varadhan (2015) in R programming language and we employ the "nlminb" (nonlinear minimization subject to box constraints) algorithm to minimize the loss function in equation (5) (Gay 1990). For robustness, we also try Sequential Least Squares Programming (SLSQP) in the scipy library in Python programming language (Virtanen et al., 2020). Both routines produced the same reconciled counterfactual estimates (up to the tolerance).

and (ii) they tend to be closer to the truth when h is non-linear. Second, the reconciliation method alters the original treatment effect estimates as little as possible to satisfy the constraints. This is, arguably, a desirable property in particular if the research design is credible and the original treatment effect estimates are unbiased and/or consistent. In these cases, the researcher may justifiably be reluctant to substantially alter the original estimates. Third, it does not utilize any of the in-sample errors which might not be meaningful due to the use of models that tend to overfit.[10] It also does not require out-of-sample errors which might not be feasible due to the small sample size. Fourth, it is quite straightforward to incorporate weights in different subgroups. If the researcher is willing to allow the reconciliation procedure to make bigger changes for counterfactuals of a certain group, this can be included in the loss function in the form of weights.

3.3 The Magnitude of Accuracy Improvement

For the purposes of geometric exposition, in this section, we define $h(\widehat{Y^0_{g,J}}) = \widehat{y^0_{g,j}}$ and $h(\Lambda_{g,j}) = \lambda_{g,j}$ and re-write the optimization function as follows:

$$\begin{aligned} &\widehat{\lambda_{g,J}} = argmin_{\lambda_{g,j}} L_j(\widehat{y^0_{g,J}}, \lambda_{g,j}) \\ &subject\ to\ Ah^{-1}(\lambda_{g,j}) = 0 \end{aligned} \tag{6}$$

Geometrically, the reconciliation method finds the shortest path from the original counterfactual estimates to the constraint hyperplane. If h is the identity function, the loss function is the mean squared error, and we have a three-dimensional hierarchy with $Y_{tot} = Y_1 + Y_2$. What the reconciliation procedure does in this case is depicted in Figure 1 panel (a).

[10] Horizontal regressions in Doudchenko and Imbens (2016) and Athey et al. (2021) may fit the training data perfectly. Not using in-sample errors is one of the major advantages of our proposed reconciliation method because models that provide the best predictions under certain regimes can be those that perfectly interpolate the training data due to the double-descent phenomenon (Hastie et al., 2019).

Figure 1 is here

Due to the constraint, the three-dimensional space in the graph collapses into a two-dimensional plane (dotted plane). The original counterfactual estimates (blue circles), while still can be written as a three-dimensional array ($\widehat{Y_{tot}}$, $\widehat{Y_1}$, $\widehat{Y_2}$), have only two degrees of freedom: They must satisfy the constraint. When they do not, the reconciliation finds the shortest path between the original counterfactual estimates and the constraint, and the reconciled estimates (green circles) are on the plane.

Depicting a similar figure in four or more dimensional hierarchies is not feasible. Nonetheless, no matter the number of dimensions, there always appears a triangle with original counterfactual estimates, reconciled estimates, and the truth as the vertices, with the latter two on the constraint hyperplane (Figure 1 panel (b)).

Since the truth is on the plane, the shortest path from the original counterfactual estimates to the plane will form a right-angled triangle with the side from "Original" to "Truth" as the hypotenuse, unless the original counterfactual estimates are already on the plane. In other words, the reconciled counterfactuals which are orthogonal projections to the plane are overall always going to be closer to the truth unless reconciliation is unnecessary.[11] In addition, if we know the

[11] Note that it is not possible to claim that reconciliation will pull *all* the group-specific counterfactual estimates to the truth. Even when h is linear, some groups' counterfactual estimates might become less accurate after reconciliation. In practice we cannot predict, a priori, lucky instances when the original counterfactual estimates for certain groups are accurate, so reconciliation would only harm their accuracies. We do not know the truth, except in simulation exercises. Hence, we always reconcile the counterfactuals as long as reconciliation is warranted. What is guaranteed is that the reconciled counterfactual estimates will be, *overall*, closer to the truth than the original estimates. If the researcher is more confident about the counterfactual estimates of certain groups, this can be incorporated by using weights in equation (5).

exact locations of the points, using the law of sines, we can calculate the distance between the reconciled counterfactuals and the truth relative to that between the originals and the truth:

$$\frac{c}{a} = \frac{sin(C)}{sin(A)}.$$

$sin(A)$ is 1 when the triangle is right-angled, and $sin(C)$ is smaller than 1.

Figure 2 is here

In fact, there emerges another relationship among the edges of right-angled triangles that allows us to calculate the magnitude of accuracy improvement ($\frac{c}{a}$). According to the Pythagorean theorem, $a^2 = b^2 + c^2$. Dividing both sides by a^2 yields $\frac{b^2}{a^2} + \frac{c^2}{a^2} = 1$, which is the equation of a circle with a radius of 1. Since neither $\frac{b^2}{a^2}$ nor $\frac{c^2}{a^2}$ takes negative values, we observe only the first quadrant of a unit circle. In other words, when h is a bijective linear function, there is a mechanical relationship between the accuracy improvement ($\frac{c}{a}$) and the distance from the reconciled to the original counterfactual estimates relative to the prediction error of the original estimates ($\frac{b}{a}$). Then, as illustrated in Figure 2, for a given value of $\frac{b}{a}$, we can straightforwardly obtain the rate of the overall improvement in accuracy since the point ($\frac{b}{a}$, $\frac{c}{a}$) is on the circle.

Figure 3 is here

When h in equation (6) is not a linear function, the hyperplane is not flat. Figure 3 shows a case in three-dimensional space where the outcome variables are log-transformed while the constraint is linear. In other words, we still have $Y_{tot} = Y_1 + Y_2$, but the axes are shown in log scale.[12]

Here, the right-angled triangle does not necessarily appear because the constraint has a curvature. However, we can still draw the triangle and use the law of sines. Unless the original estimates are already on the plane, the angle A can be right, wide (Figure 3 panel (b)), or acute (Figure 3 panel (c)). In the former two cases, accuracy improvement is still guaranteed. In the latter, however, reconciliation might be harmful.

In cases where h^{-1} is non-linear, an important piece of information can be obtained if the graph of the constraint function in equation (6) is concave or convex. If it is strictly convex (concave), then the graph of the constraint function is above (below) any plane tangent to it due to the properties of convex (concave) functions. This implies that if our original estimates correspond to a point below (above) the graph of the convex (concave) constraint, then angle A will be greater than 90°. As shown in Figure 4 panel (a) (Figure 4 panel (c)), because the angle between the shortest line segment from the original counterfactuals to the constraint and the corresponding tangent line on the constraint is 90° and the constraint curves away from the original estimates, a wide-angled triangle appears with its vertices as original estimates, reconciled estimates, and the truth. More concretely, if, say, h^{-1} is the exponential function (a convex function), and the constraint is $\exp(y_{Tot}) = exp(y_1) + \exp(y_2)$, and $\exp(\widehat{y_{tot}})$ is

[12] The constraints appear non-linear because the axes are log-transformed.

smaller than $exp(\widehat{y_1}) + \exp(\widehat{y_2})$ (i.e. the original estimates correspond to a point below the graph of the constraint function), then angle A will be larger than 90°.

Figure 4 is here

On the other hand, Figure 4 panel (b) (Figure 4 panel (d)) shows that when the graph of the constraint function is convex (concave), and the original estimates correspond to a point above (below) it, then angle A will be smaller than 90° because the constraint curves towards the original estimates. [13]

Notably, using weights in equation (5) also leads to projections where the right-angled triangle does not appear. The overall improvement in accuracy is not guaranteed in these cases either. However, some guidance has been provided in the literature regarding how these weights can be constructed to boost expected improvement in accuracy. More specifically, Wickramasuriya et al. (2019) propose using in-sample (one-step ahead) errors to construct weights. Briefly, their argument is that series with smaller errors are expected to have more accurate forecasts, so they should have larger weights, and they will be changed little during reconciliation. On the other hand, Jeon et al. (2019) and Spiliotis et al. (2021) recommend using the cross-validation procedure and the resulting out-of-sample errors in obtaining the weights. They note that in-sample errors do not reliably proxy out-of-sample errors. They also use the objective function that is used to assess the accuracy in constructing the weights. As we note in footnote 10, it is not always possible to obtain in- or out-of-sample errors. In these cases, the weights can be constructed according to intuitive judgement and opinions. Van Erven and Cugliari (2015) provide some insights as to how these weights can be selected.

[13] In our experiments, even when $\exp(\widehat{y_{tot}})$ is larger than $exp(\widehat{y_1}) + \exp(\widehat{y_2})$ (thus angle A is smaller than 90°), we observe, on average, that the overall accuracy has improved due to the reconciliation.

4. Design of Experiments

In this section, we describe the design of our experiments with the U.S. GDP and employment data to assess whether and to what extent the proposed reconciliation method provides more accurate counterfactuals.

To present evidence showing the differential impact of reconciliation under different modelling approaches, we utilize traditional two-way fixed effects (TWFE) regressions as described in Angrist and Pischke (2008) and causal machine learning methods. In the main paper, for the reasons explained in Section 4.2 and footnote 15, we focus on the generalized synthetic control (GSC) developed by Xu (2017). In the Appendix Table A.2, we present the results of our experiments for a battery of causal machine learning tools. We reach similar conclusions with the other tools as well.

4.1 Two-way Fixed Effects (TWFE) Regressions

Lalonde (1986: 604) remarks that "Econometricians intend their empirical studies to reproduce the results of experiments that use random assignment without incurring their costs." Rubin (1974) points out that in most social sciences only available data are observational (nonrandomized) because (i) the cost of randomized experiments can be prohibitive, (ii) there can be ethical concerns about randomly assigned treatments (i.e., effect of drug addiction on health), and (iii) estimates based on experiments would be delayed (i.e., effect of diet on longevity). As a result, social scientists most commonly employ difference-in-differences (DiD) research design to mimic experimental research design (Cunningham, 2021).

DiD compares changes in the outcome of the treatment group from the pre-treatment period to the post-treatment period with the same outcome of the control group over the same time horizon. In its simplest form with two time periods (post-treatment and pre-treatment), two cross-sectional units (treated and untreated), and no additional controls, DiD estimates the causal treatment effect as follows:

$$\widehat{\beta^T} = Y_{1,1} - \widehat{Y_{1,1}^0} = Y_{1,1} - Y_{0,1} - (Y_{1,0} - Y_{0,0}) \quad (7)$$

where $Y_{1,1}$ is the outcome of the treated unit in the post-treatment period and $\widehat{Y_{1,1}^0}$ is the counterfactual estimate. The left-hand side equation is equivalent to equation (3). The counterfactual estimate, $\widehat{Y_{1,1}^0}$, is constructed using the outcome of the untreated unit in the post-treatment period ($Y_{0,1}$) after correcting it for the level differences between the treated unit and the untreated unit in the pre-treatment period ($(Y_{1,0} - Y_{0,0})$). For the case where there are many untreated units, pre-treatment periods, and additional control variables, the same logic applies. After having accounted for the control variables, DiD constructs the counterfactual estimates by correcting the average observed post-treatment outcome of the untreated units according to the level differentials.[14] When there are multiple post-treatment periods or treated units, the treatment effect can be separately estimated for each of them unless the researcher believes the treatment has no dynamic effect and its effect is constant across units.

A mathematically equivalent way to obtain the same causal effect with the use of a regression equation is:

[14] See Angrist and Pischke (2008) for more details on how DiD estimates the treatment effect, when there are many treated and control units, pre- and post-treatment periods, and additional control variables.

$$Y_{i,t} = \alpha_i + \gamma_t + \beta^T T_{i,t} + \varepsilon_{i,t} \quad (8)$$

where i and t indicate cross-sectional and time dimensions, respectively. $T_{i,t}$ is the binary variable for treatment, and it takes on the value of 1 when the unit is treated and 0 otherwise. β^T is the coefficient of interest, and its estimate, $\hat{\beta}^T$, is the average treatment effect on the treated. α_i and γ_t are a set of indicator variables which capture cross-sectional and time-specific effects, respectively. Comparing equation (8) with equation (7), in the basic DiD case, $\alpha_i = (Y_{1,0} - Y_{0,0})$ corrects the level difference between treated and untreated units, and $\gamma_t = Y_{0,t}$ corrects the difference between pre- and post-treatment observation of the outcome variable to the extent of any change in the untreated unit. Thus, β^T is the remaining difference of the outcome variable in the post-treatment period of the treated unit, and it is numerically identical in equations (7) and (8). The main benefit of the mathematical equivalence between equations (7) and (8) is its allowing easily generalization of the basic DiD setup to many treated/untreated units and pre-/post-treatment periods and additional controls.

In addition to the standard OLS assumptions, the primary requirement of a causal DiD estimate to be valid is the parallel trends assumption. According to this assumption, after accounting for all the control variables and fixed effects, there should be no systematic differences between the outcomes of interest in the control sample (composed of untreated units) and the treated sample in the pre-treatment period. In other words, we ascribe all the observed differences between the counterfactual estimate and the actual in the post-treatment period to the treatment. Therefore, the validity of the counterfactual estimate relies on the parallel trends assumption.

DiD design and TWFE regressions are quite common and facilitate estimation of counterfactuals. We employ the following to construct group-specific counterfactuals:

$$Y_{i,t,g} = \alpha_{i,g} + \gamma_{t,g} + \beta_g^T T_{i,t,g} + \varepsilon_{i,t,g} \quad (9)$$

where i and t indicate the cross-sectional (state or industry) and time (quarter) dimensions, respectively. g indicates the group we focus on, so we run separate regressions for each of the groups. $\alpha_{i,g}$ and $\gamma_{t,g}$ capture the cross-sectional- and time-specific effects of the group. $T_{i,t,g}$ is the treatment indicator, and, as noted in Section 2.2, the counterfactuals are constructed using the estimate of β_g^T. $\varepsilon_{i,t,g}$ is the error term.

4.2 Generalized Synthetic Control Method (GSC)

When working with real-world data, equation (9) might be too simplistic, and the parallel trends assumption might be violated. Confounding variables that correlate with both treatment and outcome variables may bias the estimates. Researchers address this omitted variable bias by including additional control variables. However, oftentimes, they cannot or do not directly observe the required control variables. With recent advances in machine learning tools, researchers can derive these confounder variables using data-driven methods.

Xu's (2017) generalized synthetic control (GSC) method captures these time-varying systematic differences among cross-sectional units in a quite intuitive way: These confounders tend to cause

the error term of equation (9) to follow certain patterns.[15] These patterns reveal themselves among the principal components of the error term. Then, GSC determines which of these principal components are "important" based on the cross-validation procedure, and it purges them.[16]

GSC augments equation (9) as follows:

$$Y_{i,t,g} = \alpha_{i,g} + \gamma_{t,g} + \beta_g^T T_{i,t,g} + F_{t,g}\,\lambda_{i,g} + \varepsilon_{i,t,g} \qquad (10)$$

The two additional sets of terms in equation (10), $F_{t,g}$ and $\lambda_{i,g}$, are defined as group-specific latent factors (i.e., time varying coefficients) and factor loadings (i.e., state specific intercepts). They are time- and cross-section-varying confounders such as group-specific regional trends which are not explicitly specified by the researcher.

We note that GSC is quite different from the synthetic control method developed by Abadie et al. (2010). The GSC equation primarily adds the factors and the loadings to equation (9). Therefore, equation (10) can also be interpreted as having different moderately "important" group-specific control variables in traditional TWFE regressions. As noted in Section 3.1, using different right-hand side variables opens up room for the reconciliation procedure to improve counterfactual accuracy. Therefore, GSC results in our experiments can be read as improvements in

[15] In the main text, we focus on the GSC primarily because it is quite straightforward to show the direct relationship between the GSC and the two-way fixed effects, as the former augments the latter by including additional control variables derived from the patterns detected in the latter's error term. It has also been shown to perform well in policy evaluation analyses (Gobillon and Magnac 2016). For a literature review on causal machine learning tools, see Liu et al. (2021).

[16] This implies that if GSC deems none of the principal components as important, equation (10) collapses to equation (9).

counterfactual accuracy resulting from reconciliation in the presence of group-specific covariates.

5. Data

We demonstrate the benefits of the proposed reconciliation method using the 1990-2019 quarterly state-by-industry employment statistics from the Quarterly Census of Employment and Wages (QCEW) and the 2005Q1-2021Q1 quarterly state-by-industry GDP statistics from the U.S. Bureau of Economic Analysis (U.S. BEA). The final vintage of QCEW quarterly data is published in the third quarter of the following year (U.S. Bureau of Labor Statistics, 2021b). Both datasets were obtained in June 2021. Since QCEW data ends in 2019, we use the final vintage. BEA GDP by industry estimates are initially announced the following quarter; however, these estimates are subject to annual updates in the month of September during subsequent years (U.S. BEA, 2021a; 2021b). Hence the state-by-industry GDP data used in this analysis should be regarded as June 2021 vintage.[17]

Figure 5 is here

QCEW data are a publicly available data set that reports the quarterly count of employment at multiple geographic areas and industry levels. The data are constructed based on employer reports and cover more than 95% of all jobs in the U.S. The QCEW is commonly considered to be very high quality employment data since it lacks sampling error. However, due to the Bureau

[17] Researchers who use TWFE regressions (such as equation (9)) or more complex models (such as causal machine learning as in equation (10)) tend not to be concerned about the stationarity of their data due to the demeaning along time and panel dimensions. Nonetheless, especially if the time dimension is long, non-stationarity of the panel data can be particularly harmful. Following Acemoglu et al. (2019), we confirm that the panel data we use in our experiments are stationary by regressing the outcome variable on its lagged value after the demeaning. For all outcome variables and methods used, the 95% confidence intervals produced using standard errors clustered at the state level rarely contain 1 or -1. The details of this exercise are provided in Appendix and in Figure A.1.

of Labor Statistics' (BLS) confidentiality policy, the reported values at more disaggregated levels are suppressed to protect the identity of cooperating employers. This is problematic in our case since these methods tend to break the linear identities that we exploit to obtain more accurate counterfactuals. More concretely, due to data suppression, total employment in state A is sometimes greater than the sum of individuals employed in each of the 5-digit NAICS industries in state A. As a result, we utilize only total employment counts using two broad categories: goods producing (NAICS 101) and service producing (NAICS 102). Figure 5 Panel (b) shows the hierarchical structure of the employment data.

We use the current GDP data which are also publicly available at the BEA website. Specifically, we obtained GDP in current dollars (SQGDP2) at 19-Industry detail. The industry detail is based on the North American Industry Classification System (NAICS). Similar to QCEW, BEA suppresses lower-level GDP data to avoid disclosures of confidential data (such as GDP of the construction sector in D.C. in 2005). In other words, the reported gross state product (GSP) of Georgia can be more than the sum of individual industry products in the same state in some quarters. Hence, we calculate new state-level totals by aggregating available data in order to obtain linear identities. Figure 5 Panel (a) shows the hierarchical structure of GDP data.

Table 1 is here

Table 1 Panel (a) presents summary statistics for GDP data.[18] We have 65 quarters of data (2005Q1-2021Q1). At the state level, the total number of observations is 3,315 (65*51 including

[18] For more details, see Appendix Table A.1 where we report summary statistics by industry.

DC); the state-by-industry level observation number is 62,985.[19] Average quarterly current GDP by state is $290 billion with a standard deviation of $377 billion. At the industry-state level, average quarterly GDP is $15 billion (standard deviation $29 billion). The table also presents the average and standard deviations of natural logarithms of quarterly GDP. As expected, taking the natural logarithms of the data reduces variation significantly in the data as evidenced by the large differences in the coefficient of variation between the levels and natural logarithms (by 10 to 15 times). Table 1 Panel (b) presents summary statistics for employment data. Mean employment by state exceeds 2 million. Roughly, a quarter of total employment is in goods-producing sectors; the remainder, in service producing sectors (see notes to Table 1 for exact classification). Again, taking natural logarithms of data reduces the coefficient of variation by 12 to 15 times.

6. Simulations and Results

6.1 Simulations

To show the benefits of the proposed counterfactual reconciliation procedure, we randomly generate placebo laws that are enacted in some states but not in others. Given that the placebo laws have no effect whatsoever, any treatment effect estimate that is not zero indicates a divergence between the true counterfactuals and the estimated counterfactuals. Our expectation is that the reconciliation procedure will exploit this divergence and bring the estimated counterfactuals closer to the truth.

[19] For simplicity we present only one of the possible hierarchies. We could arrange the data with total industry gross product or employment at the top and state-level industry gross product or employment at the bottom. Then we would have 65*19=1,235 observations at the top level and 65*19*51=62,985 at the bottom. Table A.1 presents summary statistics for quarterly GDP by industry. Real estate and rental and leasing represent the largest industry on average.

Briefly, our simulations involve the following steps: First, we vary the number of treated states (where placebo laws are enacted) between 1 and 5. Second, we randomly choose the exact date when the treatment is implemented in the treated states and we vary the length of the pre-treatment periods between 24 and 36. Third, we vary the length of the post-treatment periods between 1, 4, and 12. This means that, in total, we have 12 (=2*2*3) different scenarios.[20] Fourth, using this simulated data, we estimate the group-specific treatment effects (β_g^T) using equations (9) and (10). This produces the original counterfactual estimates. Then, using equation (5), we also produce the reconciled estimates.[21] In order to reduce the impact of outliers, we repeat each scenario 250 times to measure how well the counterfactual reconciliation performs on average. Fifth, we calculate the average root mean square prediction error (RMSPE) separately for the reconciled and the original counterfactual estimates. In the following tables, for each of these 12 scenarios and the two data sets (GDP and employment), we report the ratios of the average root mean squared prediction error of the reconciled counterfactual estimates to that of the original counterfactual estimates:

$$\frac{\overline{RMSPE_{reconciled}}}{\overline{RMSPE_{original}}} = \frac{\sum_{\#iteration}\sqrt{\sum_j \sum_g (h(\widehat{\Lambda_{g,j,\text{iteration}}}) - h(Y_{g,j,\text{iteration}}))^2}}{\sum_{\#iteration}\sqrt{\sum_j \sum_g (h(\widehat{Y^0_{g,j,\text{iteration}}}) - h(Y_{g,j,\text{iteration}}))^2}}, \quad (11)$$

[20] This simulation is highly similar to the one employed in Bertrand et al. (2004).

[21] We estimate a single group-specific treatment effect coefficient even when we allow multiple treated units and horizons. This implies that to obtain the original counterfactual estimates, we subtract the same number ($\hat{\beta}_g^T$) from the groups' outcome variables in the treated units in the post-treatment periods. Ideally, different group-specific treatment effect estimates are calculated for each treated unit and post-treatment period to capture heterogeneity across cross-sections and dynamic effects. However, for various reasons, researchers may regularize $\hat{\beta}_g^T$ and report the average group-specific treatment effect on the treated. Note that the reconciliation procedure is always separately performed for each treated unit and time (the treatment index j in equation (5) corresponds to the combination of i and t in treated states in post-treatment periods in equations (9) and (10)).

where $iteration$ represents the iteration number ($1 \leq iteration \leq 250$). Following the notation in Section 2.2, j is the index of treatment (for instance, if the horizon is 4 and there are 5 treated units, $1 \leq j \leq 20$) and g indicates groups. A ratio less than one indicates average improvement in accuracy due to the reconciliation compared to the original estimates.

In the tables below, when "overall" is indicated, all the groups in all hierarchy levels are considered in calculating RMSPE ratios. When "top" or "bottom" is indicated, we report RMSPE ratios for the groups that correspond to the "top" or "bottom" level. Note that for each time period and cross-sectional unit, we produce reconciled counterfactuals for all the groups simultaneously using equation (5). Since RMSPE is calculated only after the original and reconciled estimates have been constructed, the way it is calculated ("overall", "top", or "bottom") does not impact on the estimates.

6.2 Results

Panels (a) in tables 2, 3, and 4 present results for GDP data, and panels (b) present results for employment data. Tables 2, 3, and 4 differ from each other by whether the outcome variables are transformed or not and whether two-way fixed effects (equation (9)) or machine learning tools (equation (10)) are used. Table 2 presents the results for the selected machine learning method, GSC (equation 10), where the outcome variables are untransformed (equation 3). By design, GSC may construct different control variables ($F_{t,g}\ \lambda_{i,g}$ in equation 10) for different groups. Table 3 presents linear TWFE estimation results where the right-hand side controls are the same

in each case (equation 9) but the outcome variables are transformed (equation 4).[22] Finally, Table 4 presents GSC (equation 10) with log-transformed outcome variables (equation 4).

Table 2 is here

Table 2 reports that for every alternative scenario, the reconciled counterfactuals are, overall, more accurate. The decline in average root mean square prediction error (RMSPE) ranges from 1.5% to 9.5%. Given that the reconciliation procedure uses only the original counterfactual estimates and known linear constraints, the improvement can be considered free lunch if the overall improvement in accuracy is of interest. The average magnitude of the RMSPE ratios is in line with our expectations. Considering that GSC captures different confounders for each group, the more dissimilar the group-specific control variables across the hierarchy, the greater the potential improvement by reconciliation. Our observations also indicate that there is no clear pattern between the magnitude of the improvement in accuracy and the horizon, number of states that are treated, or the number of pre-treatment quarters in the data.

When we focus on specific levels in the hierarchy, we observe RMSPE ratios greater than 1. This is also in line with our expectations. As noted in footnote 11, reconciliation does not necessarily decrease the RMSPE of every group. Therefore, it is possible that the accuracies of certain groups in the hierarchy deteriorate while the overall accuracy improves.

Table 3 is here

Table 3 employs TWFE (equation (9)) to estimate the counterfactuals; the outcome variables are log-transformed (h function is the logarithmic function). Even though improvement in accuracy is not guaranteed due to the transformation, the ratios are mostly less than one. This suggests

[22] We do not report TWFE results (equation 9) when the outcome variable is untransformed (equation 3), since this corresponds to the case discussed in the theorem in Section 3.1.

that, on average, the reconciliation procedure has produced more accurate counterfactuals, albeit the magnitude of the improvement is quite small. Small improvements are expected since the true treatment effect is null and we employ a relatively basic regression equation with no controls other than fixed effects. Thus, reconciliation only marginally alters the original estimates. As explored below, when the distances between the original and the reconciled counterfactuals (b) are small compared to those between the original counterfactual estimates and the truth (a), the improvement due to the reconciliation tends to be very small, yet still present.

Table 4 is here

Table 4, on the other hand, depicts a different story when the regressor matrices are not identical for different groups ($F_{t,g}\,\lambda_{i,g} \neq F_{t,g'}\,\lambda_{i,g'}\ \ when\ g \neq g'$). In this case, the outcome variables are log-transformed and the right-hand side variables are different. The distances between the original and the reconciled counterfactuals (b) are sizable compared to those between the original counterfactual estimates and the truth (a). Overall, this has led the improvements in accuracy to range from 1% to 7%.

Consistent with our previous observations in footnote 11, Table 4 clearly shows that, even when the overall accuracy is improved due to the reconciliation, individual group's accuracies may deteriorate. In panel (A), we note that reconciliation, on average, has improved the original GDP estimates at the state-by-industry level (bottom level) at the expense of the corresponding estimate at the state level (top level). While the overall accuracies, on average, have improved in every case, the relatively poor performances of the reconciled estimates for GSP may not be acceptable. In such cases, using group-specific weights in equation (5) would be appropriate.

Figure 6 is here

In Figure 6, we further explore two relationships related to accuracy improvements when the outcome variable is not transformed (panel (a)) and when it is log-transformed (panel (b)): The first of these relationships is between the magnitude of the A angle from figures 1 and 3 and the accuracy improvement ($\frac{c}{a}$). The second is between the distance from the original to the reconciled counterfactual estimates relative to the error of the original counterfactual estimate ($\frac{b}{a}$), and the accuracy improvement ($\frac{c}{a}$).

The left-hand side graph in panel (a) shows that angle A is always $90°$ when h is the identity function, so reconciliation is guaranteed to improve the accuracy of the counterfactuals. In this situation, for a given distance between the original counterfactual estimates and the truth, the magnitude of the improvement depends only on the distance from the original to the reconciled counterfactual estimates relative to the original counterfactual error ($\frac{b}{a}$). The right-hand side of panel (a) shows that when we plot $\frac{b}{a}$ against $\frac{c}{a}$, all the results from the experiments correspond to a point on the quadrant (the dotted curve).

In Figure 6 panel (b), we focus on the cases in which the outcome variable is log-transformed. The graph on the bottom-left shows that angle A is not always exactly $90°$. The median angle is $90.07°$, the mean is $90.35°$, and more than half of the angles are between $89.30°$ and $90.95°$. This suggests that in the majority of the cases, the triangles that will appear due to reconciliation will be very similar to right-angled triangles. However, when the angle is quite small, we can expect to observe deteriorations in counterfactual accuracy.

The right-hand side graph in panel (b) indicates that even though the strict mathematical relationship between $\frac{b}{a}$ and $\frac{c}{a}$ that we have observed in panel (a) breaks down, the results are still expected to be around the quadrant (the dotted curve). The graph indicates that the greater the relative difference between the original and the reconciled counterfactual estimates, the greater the accuracy improvements tend to be, albeit there are certain cases where accuracy deteriorated. The LOESS (locally estimated scatterplot smoothing) fit further corroborates that there are no regions in panel (b) where we should expect, on average, accuracy to deteriorate.[23]

7. Conclusion

The data in causal inference analyses can be hierarchical or grouped in nature which allows the use of forecast reconciliation methods. We show that when researchers are concerned with the heterogeneous effects of treatments (policy or event) on subgroups as well as the average causal effect on the overall population, there is room for reconciliation to obtain more accurate treatment effect estimates via more accurate counterfactuals. We propose a reconciliation method that is in the same spirit as Hyndman et al.'s (2011) "optimal combination", but which also allows non-linear constraints.

We further develop the geometric interpretation of the forecast reconciliation proposed in Panagiotelis et al. (2021). We provide additional insights into the exact determinants of the magnitude of the accuracy improvement due to reconciliation. In our experiments with real-world data, we demonstrate that the proposed reconciliation method tends to improve

[23] We use the default values of the parameters of the function loess in the stats package in R.

counterfactual predictions even when the outcome variables and the treatment effect estimates are non-linearly transformed (or the linear aggregation constraint appears non-linear in the transformed space). This can be particularly crucial in causal inference analyses, as it is common that the theoretical models require non-linear transformations of variables; thus, the accuracy of the causal effect estimate is evaluated in the transformed space. In addition, we also show that the benefits of reconciliation can be quite pronounced when causal machine learning tools are employed.

Finally, while the current study provides tools for obtaining coherent point estimates for treatment effects, it does not discuss how to obtain coherent confidence intervals. This topic merits future examination.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105(490), 493-505.

Acemoglu, D., Naidu, S., Restrepo, P., & Robinson, J. A. (2019). Democracy does cause growth. *Journal of political economy*, *127*(1), 47-100.

Allegretto, S., Dube, A., Reich, M., & Zipperer, B. (2017). Credible research designs for minimum wage studies: A response to Neumark, Salas, and Wascher. ILR Review, 70(3), 559-592.

Almeida, V., Ribeiro, R., & Gama, J. (2016). Hierarchical time series forecast in electrical grids. In K.J. Kim and N. Joukov (eds.), Information Science and Applications (ICISA) 2016. Singapore: Springer, pp. 995-1005.

Angrist, Joshua D., and Jörn-Steffen Pischke. (2008). Mostly Harmless Econometrics: An Empiricist's Companion, Princeton, NJ: Princeton University Press.

Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. International Journal of Forecasting, 25(1), 146-166.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 31(2), 3-32.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2021). Matrix completion methods for causal panel data models. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2021.1891924

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics, 119(1), 249-275.

Cunningham, S. (2021). Causal Inference. The Mixtape. New Haven: Yale University Press.

De Chaisemartin, C., & D'Haultfoeuille, X. (2022a). *Difference-in-differences estimators of intertemporal treatment effects* (No. w29873). National Bureau of Economic Research.

De Chaisemartin, C., & D'Haultfoeuille, X. (2022b). *Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey* (No. w29691). National Bureau of Economic Research.

Doudchenko, N., & Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis (No. w22791). National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w22791/w22791.pdf

van Erven T., Cugliari J. (2015). Game-Theoretically Optimal Reconciliation of Contemporaneous Hierarchical Time Series Forecasts. In: Antoniadis A., Poggi JM., Brossat X. (eds) Modeling and Stochastic Learning for Forecasting in High Dimensions. Lecture Notes in Statistics, vol 217. Springer, Cham. https://doi.org/10.1007/978-3-319-18732-7_15

Gay, D. M. (1990). Usage summary for selected optimization routines. Computing science technical report, 153, 1-21.

Gamakumara, P. (2020). Probabilistic forecast reconciliation. Doctoral dissertation, PhD thesis. Monash University.

Gechert, S., Havranek, T., Irsova, Z., & Kolcunova, D. (2021). Measuring capital-labor substitution: The importance of method choices and publication bias. Review of Economic Dynamics. https://doi.org/10.1016/j.red.2021.05.003

Gobillon, L., & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. Review of Economics and Statistics, *98*(3), 535-551.

Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*. https://arxiv.org/abs/1903.08560

Hollyman, R., Petropoulos, F., & Tipping, M. E. (2021). Understanding forecast reconciliation. European Journal of Operational Research, *294*(1), 149-160.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. Computational Statistics & Data Analysis, *55*(9), 2579-2589.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice, second ed., OTexts, Melbourne, Australia. https://otexts.com/fpp2/index.html

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Journal of Economic Literature, 58(4), 1129-79.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 76(4) 604-620.

Lee, D. S., McCrary, J., Moreira, M. J., & Porter, J. R. (2021). *Valid t-ratio Inference for IV* (No. w29124). National Bureau of Economic Research.

Liu, L., Wang, Y., & Xu, Y. (2021). A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data. https://arxiv.org/pdf/2107.00856.pdf

Mankiw, N. G., Romer, D., & Weil, D. N. (1992). A contribution to the empirics of economic growth. The Quarterly Journal of Economics, 107(2), 407-437.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica: Journal of the econometric society*, 1417-1426.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. International Journal of Forecasting, 37(1), 343-359.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688.

Spiliotis, E., Abolghasemi, M., Hyndman, R. J., Petropoulos, F., & Assimakopoulos, V. (2021). Hierarchical forecast reconciliation with machine learning. Applied Soft Computing. https://doi.org/10.1016/j.asoc.2021.107756

U.S. Bureau of Labor Statistics (2021a). Quarterly Census of Employment and Wages. Accessed in June 2021: https://www.bls.gov/cew/

U.S. Bureau of Labor Statistics (2021b). Quarterly Census of Employment and Wages: Handbook of Methods. https://www.bls.gov/opub/hom/cew/pdf/cew.pdf

U.S. Bureau Economic Analysis (2021a). GDP by State. Accessed in June 2021: https://www.bea.gov/data/gdp/gdp-state.

U.S. Bureau Economic Analysis (2021b). SQGDP2 series are available at: https://apps.bea.gov/iTable/iTable.cfm?reqid=70&step=1&isuri=1&acrdn=1#reqid=70&step=1&isuri=1&acrdn=1

Varadhan, R. (2015). Alabama: Constrained nonlinear optimization. R package version 2015.3-1.

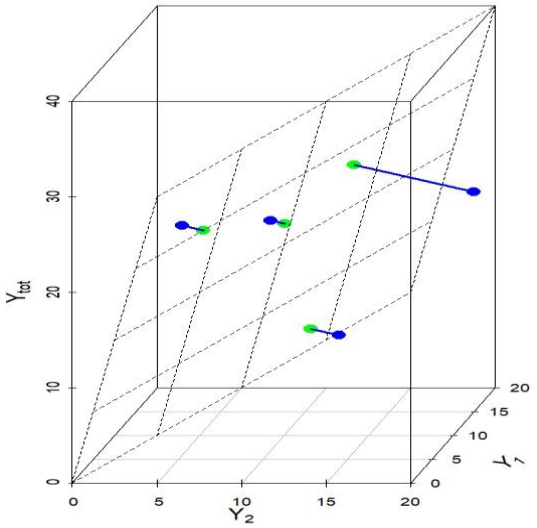
Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., et al. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17, 261–272. doi:10.1038/s41592-019-0686-2

Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. Journal of the American Statistical Association, 114(526), 804-819.

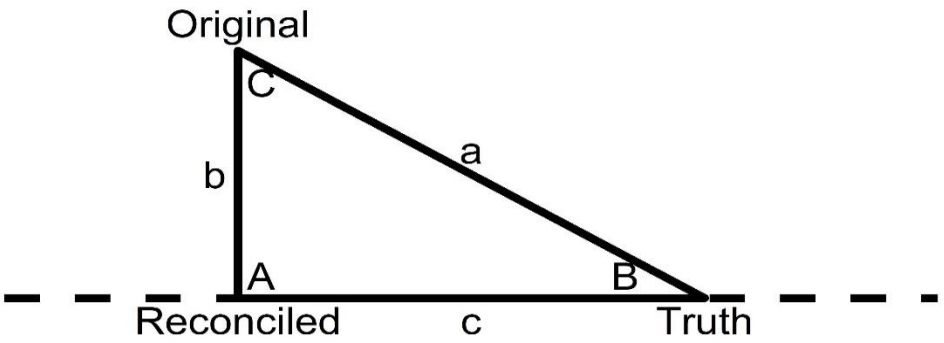
Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis* 25(1): 57–76.

Figures

Figure 1: Visual representation of reconciliation when the constraint is linear



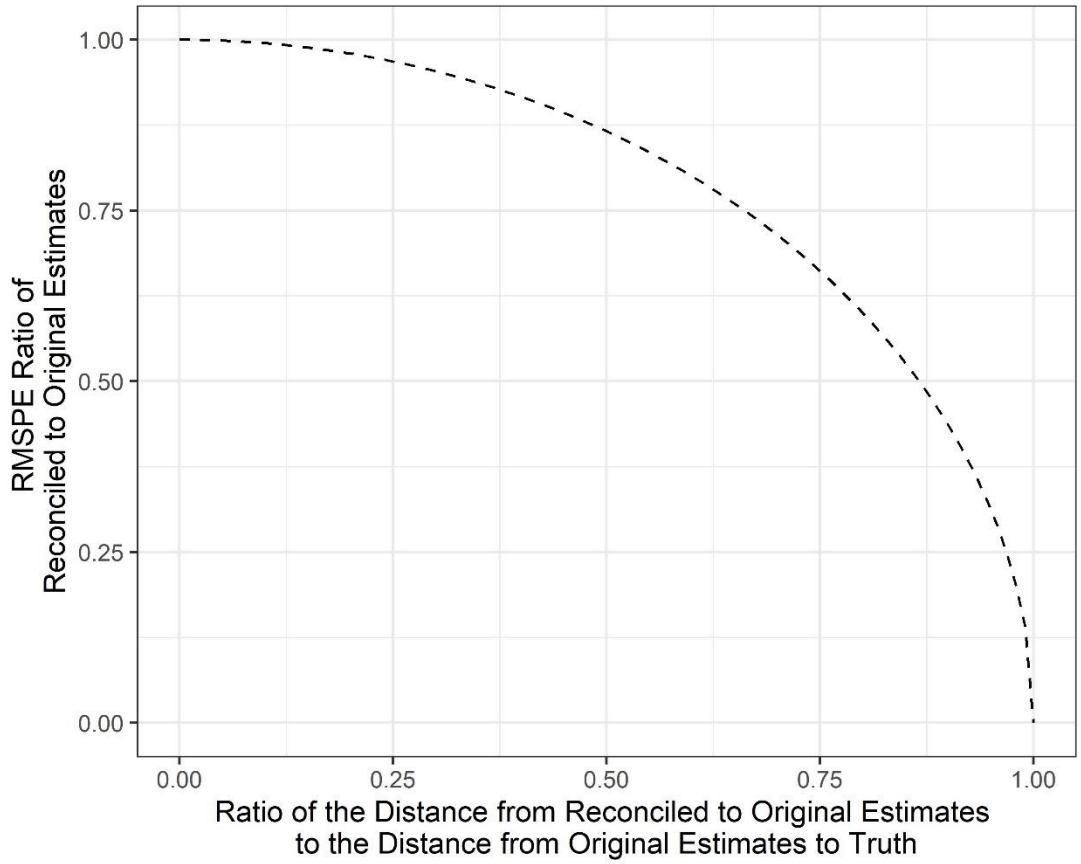
(a)



(b)

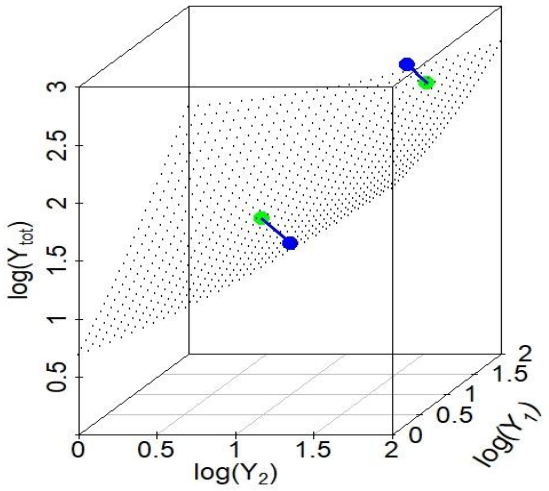
Notes: In Figure 1(a), blue (green) dots are the original (reconciled) counterfactual estimates. Figure 1(b) presents a two-dimensional slice corresponding to one of the counterfactual estimates. The dashed line in Figure 1(b) is the linear constraint plane where truth and reconciled counterfactuals reside. Side *b* corresponds to the orthogonal distance between blue and green dots from Figure 1(a). Side *a* is the original counterfactual error.

Figure 2: Relationship between accuracy improvement and distance between original and reconciled counterfactuals

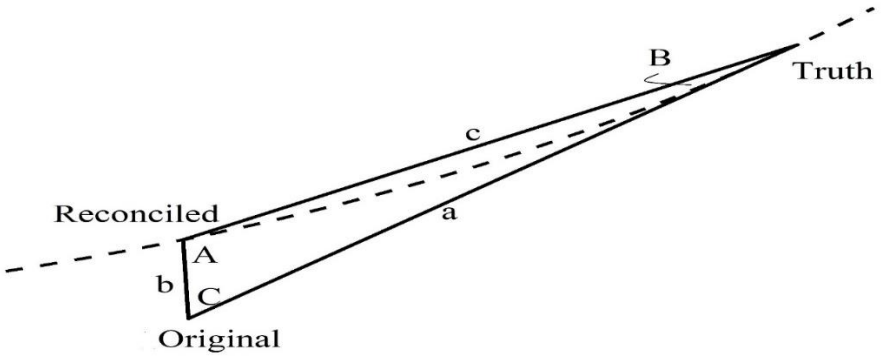


Notes: RMSPE stands for root mean square prediction errors. The x-axis plots the ratio of the distance between reconciled and original estimates to the original counterfactual error, $\frac{b}{a}$, from Figure 1(b). If the original and counterfactuals are geometrically very close (i.e. $\frac{b}{a}$, close to zero), the RMSPE ratio of reconciled and original estimates will be very close, too (i.e. ratio on y-axis will be close to one). If the triangle in Figure 1 (b) is a very elongated right angled triangle with a very short base (side c), then side b will be almost as long since hypotenuse, $\frac{b}{a}$ will be close to one, and the corresponding RMSPE ratio will be significantly less than one.

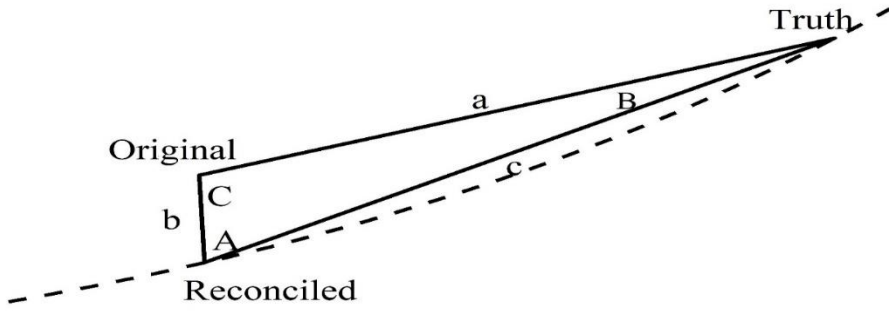
Figure 3: Visual representation of reconciliation when the constraint is non-linear



(a)



$(b)\ A > 90°$

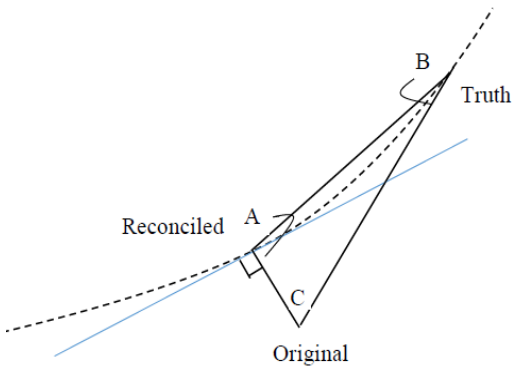


$(c)\ A < 90°$

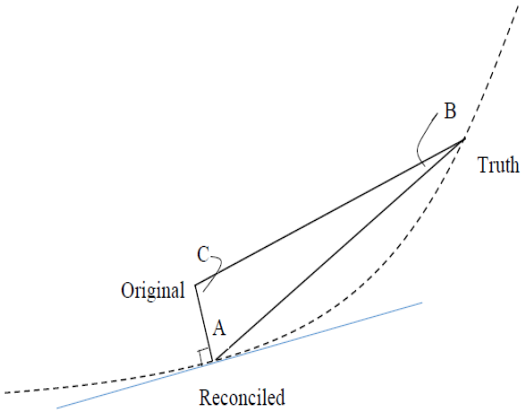
Notes: In panel (a), blue (green) dots are the original (reconciled) counterfactual estimates. Panels (b) and (c) present a two-dimensional slice corresponding to one of the counterfactual estimates. Dashed lines in panels (b) and (c) are the non-linear constraint plane where truth and reconciled counterfactuals reside. Side *b* corresponds to the vertical distance between blue and green dots from panel (a). Side *a* is the original counterfactual error. Side c is the error of the reconciled estimates.

Figure 4: Relationship between Convexity/Concavity of the Graph of the Constraint Function and Angle A

Convex Constraint

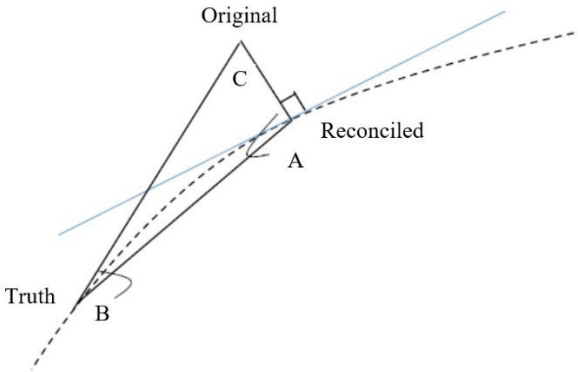


(a) $A > 90°$

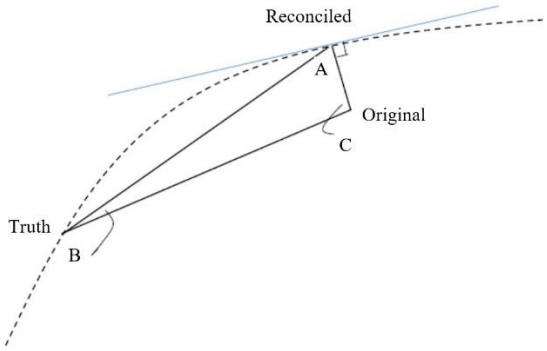


(b) $A < 90°$

Concave Constraint



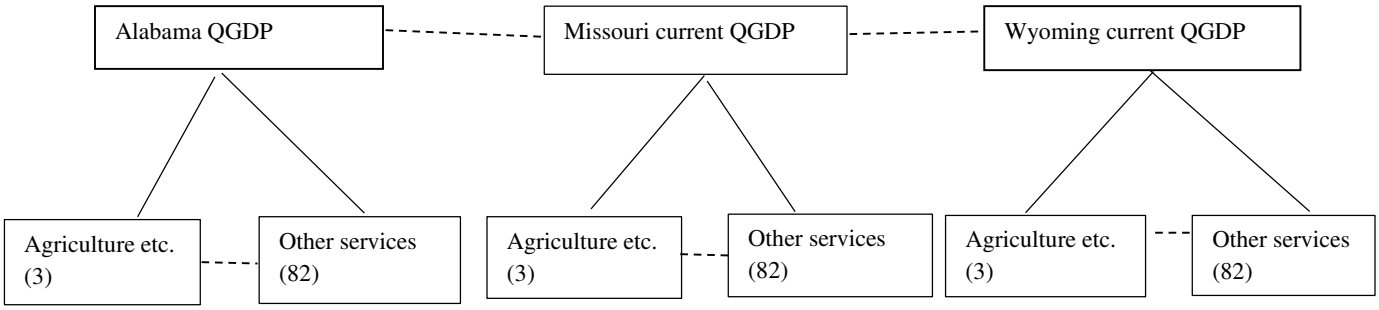
(c) $A > 90°$



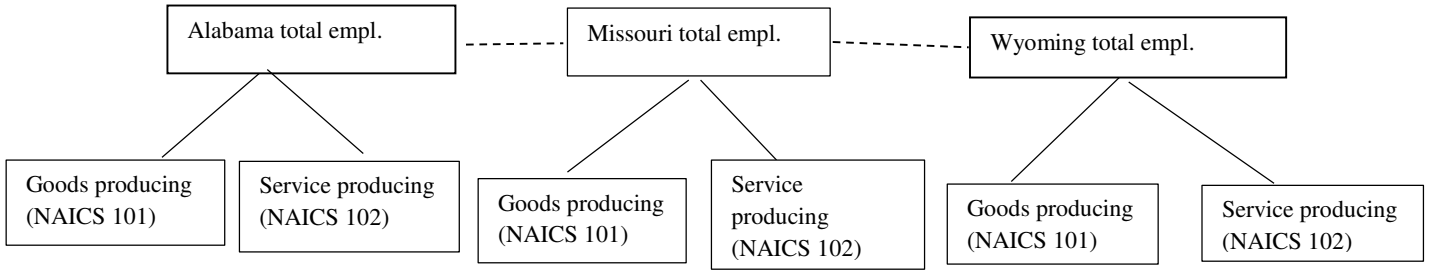
(d) $A < 90°$

Notes: See Figure 3 notes for the triangles and the dashed lines. The blue straight lines indicate the lines tangent to the constraints that pass through the points corresponding to reconciled counterfactual estimates.

Figure 5: Hierarchical data structure of quarterly GDP and employment data



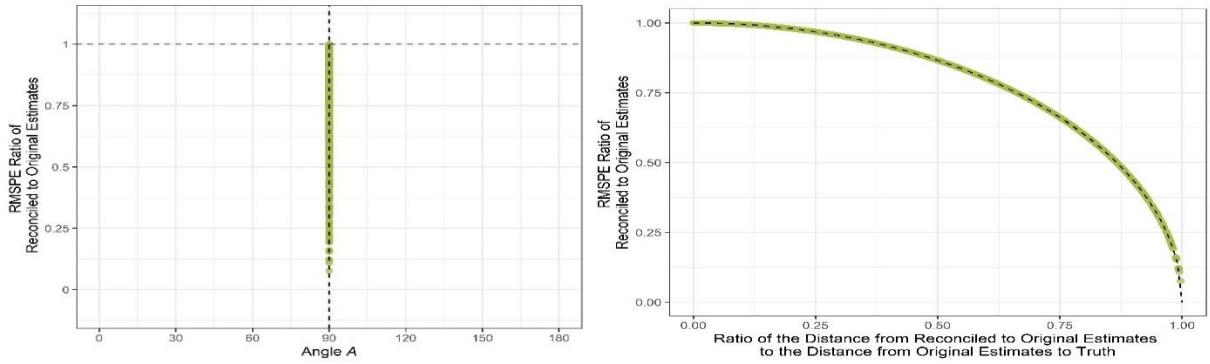
Panel (a): GDP data



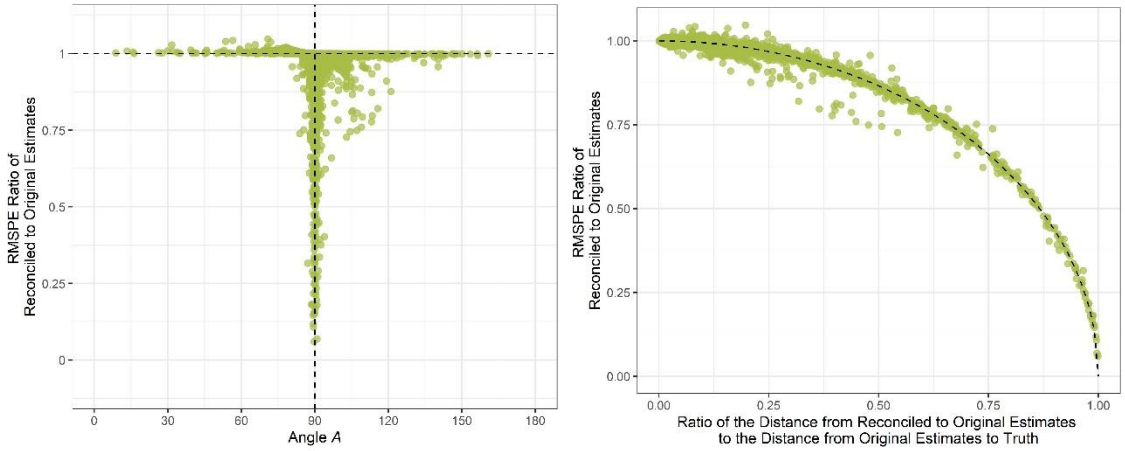
Panel (b): Employment data

Notes: In both panels we depict only three of the 51 states (including DC). Dashed lines between states represent non-depicted states. In Panel A we depict only two of the 19 industries, and dashed lines between industries represent non-depicted ones. Please see notes to Table 1 for a full list of the 19 industries. Both panels can also be depicted by switching the states and industries in the aggregation hierarchy.

Figure 6: Relationships between the magnitude of the *A* angle, distance between counterfactuals, and the accuracy improvement



Panel (a) Linear constraint



Panel (b): Non-linear constraint

Notes: Left-hand side graphs show the relationship between angle A (from figures 1 and 3) and accuracy improvement. When the constraint is linear (top left), angle A can only be a right angle (see also Figure 1b). When the constraint is non-linear (bottom left), angle A can be an acute or a wide angle (see also Figure 3b and 3c). In simulations, the majority of angle A's are clustered around 90° or above, and quite small angles are rare. Right-hand side graphs depict the relationship between $\frac{b}{a}$ and $\frac{c}{a}$ (from figures 1 and 3). In the case of linear constraint, results are always on the quadrant (top right). In the case of non-linear constraint, the results still cluster around the quadrant.

Tables

Table 1: Summary statistics

Panel (a): Current quarterly GDP

	quarterly state GDP		quarterly state by industry GDP	
	in millions $	Natural log of GDP	in millions $	Natural log of GDP
Mean	290,226	12.014	15,275	8.602
Stan. dev	376,670	1.055	29,464	1.631
CV	1.298	0.088	1.929	0.190
# Obs.	3,315	3,315	62,985	62,985

Panel (b): Total Employment

	State total		Goods producing (NAICS 101)		Service producing (NAICS 102)	
	Empl.	Natural log of empl.	Empl.	Natural log of empl.	Empl.	Natural log of empl.
mean	2,111,498	14.068	443,102	12.478	1,668,396	13.823
Std. Dev.	2,317,813	1.022	474,762	1.107	1,873,278	1.024
CV	1.098	0.073	1.071	0.089	1.123	0.074
# Obs.	6,120	6,120	6,120	6,120	6,120	6,120

Notes: CV stands for coefficient of variation (standard deviation divided by the mean). For GDP, there are 65 quarters (2005Q1-2021Q1), 51 states (including DC) and 19 industries. At the bottom level we have 65*51*19 = 62,985 observations. 19 industries are: agriculture, forestry, fishing and hunting (3), mining, quarrying, and oil and gas extraction (6), utilities (10), construction (11), manufacturing (12), wholesale trade (34), retail trade (35), transportation and warehousing (36), information (45), finance and insurance (51), real estate and rental and leasing (56), professional, scientific, and technical services (60), management of companies and enterprises (64), administrative and support and waste management and remediation services (65), educational services (69), health care and social assistance (70), arts, entertainment and recreation (76), accommodation and food services (79), and other services except government and government enterprises (82). See Table A.1 for a summary of industry statistics. For employment, there are 120 quarters (1990Q1-2019Q4), 51 states (including DC) and two broad industries. Goods producing is composed of NAICS 101 which is the total of natural resources and mining (1011), construction (1012), manufacturing (1013). Service producing is composed of trade, transportation, and utilities (1021), information (1022), financial activities (1023), professional and business services (1024), education and health services (1025), leisure and hospitality (1026), and other services (1027).

Table 2: Relative mean RMSPEs after reconciliation of the GSC counterfactuals for untransformed outcome variables

		Horizon = 1		Horizon = 4		Horizon = 12	
Panel (a): GDP							
# Treated Units = 1	overall	0.982	0.979	0.980	0.983	0.978	0.982
-Total Gross State Product (GSP)	top	0.977	0.967	0.964	0.973	0.957	0.965
-GSP by Industry	bottom	0.999	1.006	1.019	1.017	1.039	1.063
# Treated Units = 5	overall	0.985	0.985	0.981	0.980	0.979	0.978
-Total Gross State Product (GSP)	top	0.977	0.976	0.966	0.965	0.959	0.956
-GSP by Industry	bottom	1.021	1.013	1.023	1.038	1.036	1.075
Panel (b): Employment							
# Treated Units = 1	overall	0.949	0.944	0.938	0.917	0.899	0.897
-Total State Employment	top	0.954	0.950	0.947	0.890	0.882	0.928
-State Employment by Industry	bottom	1.017	1.000	0.999	0.992	0.987	0.973
# Treated Units = 5	overall	0.938	0.922	0.935	0.911	0.911	0.919
-Total State Employment	top	0.968	0.933	0.953	0.922	0.879	0.882
-State Employment by Industry	bottom	0.968	0.977	0.975	0.966	1.022	1.054
# Quarters in Training Period:		24	36	24	36	24	36

Notes: RMSPE stands for root mean squared prediction error. GSC stands for Generalized Synthetic Control. Each value in the table represents the ratio of the average RMSPE of the counterfactuals after reconciliation to the average RMSPE of the original counterfactuals. While calculating RMSPEs, "overall" indicates that all groups in all levels of the hierarchy are considered, whereas "top" and "bottom" indicate that only the top and bottom levels are considered as presented in Figure 5 is considered. We consider alternative horizons (post-treatment quarters), training periods (pre-treatment quarters), and number of treated units (states). The two columns under each horizon category display the results under 24 and 36 pre-treatment periods, respectively. The results can be interpreted as the benefit of reconciliation when there are group-specific control variables in the regression equations.

Table 3: Relative mean RMSPEs after reconciliation of the TWFE counterfactuals for log-transformed outcome variables

		Horizon = 1		Horizon = 4		Horizon = 12	
Panel (a): GDP							
# Treated Units = 1	overall	1.000	1.000	0.997	0.998	0.998	0.998
-Total Gross State Product (GSP)	top	0.916	0.898	1.135	1.055	1.116	1.169
-GSP by Industry	bottom	1.000	1.001	0.997	0.998	0.998	0.998
# Treated Units = 5	overall	0.997	0.997	0.996	0.994	0.998	0.998
-Total Gross State Product (GSP)	top	1.320	1.173	1.276	1.381	1.039	1.224
-GSP by Industry	Bottom	0.995	0.997	0.995	0.993	0.998	0.997
Panel (b): Employment							
# Treated Units = 1	overall	0.999	0.997	0.998	0.998	0.998	0.998
-Total State Employment	top	1.002	0.970	0.985	1.004	0.986	0.968
-State Employment by Industry	bottom	0.998	1.005	1.003	0.976	1.002	1.005
# Treated Units = 5	overall	0.998	0.998	0.998	0.997	0.997	0.997
-Total State Employment	top	1.002	0.981	1.001	0.986	0.999	0.979
-State Employment by Industry	bottom	0.998	1.003	0.998	1.000	0.999	1.001
# Quarters in Training Period:		24	36	24	36	24	36

Notes: RMSPE stands for root mean squared prediction error. TWFE stands for two-way fixed effects. Each value in the table represents the ratio of the average RMSPE of the counterfactuals after reconciliation to the average RMSPE of the original counterfactuals. While calculating RMSPEs, "overall" indicates that all groups in all levels of the hierarchy are considered, whereas "top" and "bottom" indicate that only the top and bottom levels are considered as presented in Figure 5 is considered. We consider alternative horizons (post-treatment quarters), training periods (pre-treatment quarters), and number of treated units (states). The two columns under each horizon category display the results under 24 and 36 pre-treatment periods, respectively.

Table 4: Relative mean RMSPEs after reconciliation of the GSC counterfactuals for log-transformed outcome variables

		Horizon = 1		Horizon = 4		Horizon = 12	
Panel (a): GDP							
# Treated Units = 1	overall	0.967	0.984	0.979	0.932	0.976	0.966
-Total Gross State Product (GSP)	top	3.230	2.088	3.568	6.094	3.134	4.047
-GSP by Industry	bottom	0.956	0.982	0.974	0.909	0.970	0.959
# Treated Units = 5	overall	0.979	0.992	0.945	0.989	0.990	0.992
-Total Gross State Product (GSP)	top	2.610	2.696	5.950	3.871	2.806	2.723
-GSP by Industry	Bottom	0.974	0.990	0.929	0.986	0.987	0.990
Panel (b): Employment							
# Treated Units = 1	overall	0.965	0.974	0.965	0.970	0.963	0.970
-Total State Employment	top	0.937	0.961	0.927	0.907	0.855	0.953
-State Employment by Industry	bottom	0.986	0.987	0.979	0.997	0.999	0.990
# Treated Units = 5	overall	0.970	0.984	0.968	0.974	0.962	0.975
-Total State Employment	top	0.944	0.935	0.963	0.936	0.845	0.897
-State Employment by Industry	bottom	0.986	1.002	0.983	0.990	0.995	1.004
# Quarters in Training Period:		24	36	24	36	24	36

Notes: RMSPE stands for root mean squared prediction error. GSC stands for Generalized Synthetic Control. Each value in the table represents the ratio of the average RMSPE of the counterfactuals after reconciliation to the average RMSPE of the original counterfactuals. While calculating RMSPEs, "overall" indicates that all groups in all levels of the hierarchy are considered, whereas "top" and "bottom" indicate that only the top and bottom levels are considered as presented in Figure 5 is considered. We consider alternative horizons (post-treatment quarters), training periods (pre-treatment quarters), and number of treated units (states). The two columns under each horizon category display the results under 24 and 36 pre-treatment periods, respectively. The results can be interpreted as the benefit of reconciliation when there are group-specific control variables in the regression equations.

Appendix

Appendix Table A.1 provides more detailed information about the GDP data sets we use.

Appendix Table A.2 shows the relative mean RMSPEs after reconciliation of the counterfactual estimates obtained using various causal machine learning (ML) tools. In addition to GSC, the following are used:

- EM: Expectation Maximization algorithm proposed by Gobillon and Magnac (2016);
- MC: Matrix Completion method proposed by Athey et al. (2021);
- SC: Synthetic Control method proposed by Abadie et al. (2010);
- EN: Proposed method by Doudchenko and Imbens (2016);
- EN-T: Transposed version of the proposed method in Doudchenko and Imbens (2016).

We report the relative mean RMSPEs for the case where the outcome variable is log-transformed, the horizon is 1, there are 24 quarters in the training period, and there is a single treated unit. Other scenarios are omitted as they yield the same conclusion.

Reconciliation and 2SLS

In Appendix Table A.3, we show the results for the case when 2SLS is employed. In the example below we focus on a very common case where we do not observe the true treatment indicator ($T_{i,t,g}$) but we observe error-prone treatment variables. More specifically, we define four error-prone treatment variables:

$$T_{i,t,g}^{*1} = T_{i,t,g} + u_{i,t}$$

$$T^{*2}_{i,t,g} = T_{i,t,g} + \epsilon_{i,t}$$

$$T^{*3}_{i,t,g} = T_{i,t,g} + u_{i,t,g}$$

$$T^{*4}_{i,t,g} = T_{i,t,g} + \epsilon_{i,t,g}$$

where $u_{i,t}$ and $\epsilon_{i,t}$, and $u_{i,t,g}$ and $\epsilon_{i,t,g}$ are independent from each other. $u_{i,t}$ and $u_{i,t,g}$ are correlated with the outcome, whereas $\epsilon_{i,t}$ and $\epsilon_{i,t,g}$ are not.[24,25]

In these simulations, we use $T^{*2}_{i,t,g}$ as an instrument for $T^{*1}_{i,t,g}$, and $T^{*4}_{i,t,g}$ as an instrument for $T^{*3}_{i,t,g}$. More specifically, following equation (9), we have the following second stages, where both are the same as equation (9) except the treatment indicator is replaced by the predicted $T^{*1}_{i,t,g}$ ($\widehat{T^{*1}_{i,t,g}}$) and $T^{*3}_{i,t,g}$ ($\widehat{T^{*3}_{i,t,g}}$).

$$Y_{i,t,g} = \alpha_{i,g} + \gamma_{t,g} + \beta_g^T \widehat{T^{*1}_{i,t,g}} + \eta_{i,t,g} \qquad \text{(A.1)}$$

$$Y_{i,t,g} = \alpha_{i,g} + \gamma_{t,g} + \beta_g^T \widehat{T^{*3}_{i,t,g}} + \eta'_{i,t,g} \qquad \text{(A.2)}$$

The corresponding first stages are as follows:

$$T^{*1}_{i,t,g} = \alpha_{i,g} + \gamma_{t,g} + \beta_g^T T^{*2}_{i,t,g} + \zeta_{i,t,g} \qquad \text{(A.1')}$$

$$T^{*3}_{i,t,g} = \alpha_{i,g} + \gamma_{t,g} + \beta_g^T T^{*4}_{i,t,g} + \zeta'_{i,t,g} \qquad \text{(A.2')}$$

[24] Note that even if none of the error terms were correlated with the outcome variables, using any of these treatment variables would yield biased coefficients due to attenuation bias and 2SLS would still be warranted.

[25] The error terms that are uncorrelated with the outcome variables are normally distributed with mean 0 and standard deviation 1/20. $u_{i,t}$ and $u_{i,t,g}$ are created by adding $0.1 * Y_{i,t,g}$ to the noise. Although not necessary, the reason for us choosing a rather small standard deviation is to ensure that the first-stage F-statistics are large enough to avoid criticisms raised by Lee et al. (2020). In all cases, the first-stage F-statistics are greater than 100.

where the last terms in equations are error terms. One important note here is that when we use $T^{*2}_{i,t,g}$ as the instrument for $T^{*1}_{i,t,g}$, neither the control variables (fixed effects) nor $T^{*2}_{i,t,g}$ vary by group. Hence, the regressor matrices in equation A.1 vary by panel unit and time, but not by group. As a result of this, when the outcome variables are linearly transformed, Theorem 1 applies. In this case, all the RMSPE ratios are exactly 1 because the reconciliation is redundant. However, when we use $T^{*4}_{i,t,g}$ as the instrument for $T^{*3}_{i,t,g}$, the regressor matrix varies by group because $\widehat{T^{*3}_{i,t,g}}$ varies by group.

Table A.3 reports the average RMSPE ratios for the case where the training data are 24 periods long, the horizon is 1, and there is only 1 treated unit although the findings can straightforwardly be generalized for longer or shorter training data or horizons, or where there are more treated units. As seen, the reconciliation, on average, has improved the overall RMSPEs in all cases except when Theorem 1 applies.[26] The improvement tends to be greater when the instrument varies by group. This further shows that the proposed reconciliation method is not specific to the cases where variables of interest are exogenous: It is warranted when 2SLS is employed as well.

Reconciliation and Stationarity

Although, the benefits of the proposed reconciliation depend entirely on angle A and the relative sizes of the edges in triangles in Section 3, one interesting question regarding the impact of reconciliation on counterfactual accuracy might be whether the stationarity of the data affects it.

[26] As before, the overall counterfactual accuracy improvement has occurred at the expense of some groups.

The impact of non-stationarity on coefficient estimates is an ever present concern of time series data. Researchers who use panel data tend not to be overly concerned with the stationarity of their data due to time and panel demeaning. Nonetheless, non-stationarity can be harmful for causal inference. To assess whether the data in our simulations are stationary, we show in Figure A.1 the coefficients estimated from group-specific regressions of the outcome variables on their first lags after fixed effects (and other factors) are purged. This method was used in a similar context in Acemoglu et al. (2019) (panel data with log of GDP as the outcome variable with, on average, 38.8 periods per country): If the estimated coefficient of the lagged term after purging fixed effects and other factors is greater or equal to 1 in absolute terms, then the panel data are non-stationary. Otherwise, as noted in Acemoglu et al. (2019), if the coefficient is statistically significantly less than 1 (even if it is quite close to 1, such as 0.973 (0.006) as reported in their Table 2 column (1)), then the data are stationary. [27]

More specifically, using the pre-treatment periods of all states and groups, we first purge (i) group-specific time and state fixed effects in TWFE regressions (equation (9)); and (ii) group-specific time and state fixed effects, and other factors deemed important by GSC (equation (10)). Then, we regress the purged outcome variable on its first lag: $\widetilde{Y_{\iota,t,g}} = \beta_0 + \gamma \widetilde{Y_{\iota,t-1,g}}$, where $\widetilde{Y_{\iota,t,g}}$ indicates the outcome variable after the fixed effects (and factors) are purged.

In our simulations, we vary the length of the pre-treatment period (24 or 36), and we use equation (9) with log(GDP) and log(employment), and equation (10) with GDP, log(GDP), employment, and log(employment) as outcome variables. To save space, we report only the regressions with GDP and log(GDP) as outcome variables and 36 pre-treatment periods, though

[27] As noted in Acemoglu et al. (2019), since the number of observations per panel unit in our setting is sufficiently large, the Nickell bias should be small (Nickell 1981).

the conclusion about the stationarity of the data is even stronger in other cases (shorter pre-treatment period or employment regressions). Therefore, we show 9 plots in three panels that show the lower limits of the 95% confidence interval of the lagged term estimates (top graphs), point estimates of the lagged term (middle graphs), and the upper limits of the 95% confidence intervals of the lagged term estimates.

All in all, while the point estimates tend to approximate close to 1 when equation (9) is used, though the 95% confidence intervals produced using standard errors clustered at the state level almost never contain 1 or -1. This is in line with Acemoglu et al.'s (2019) observations. When equation (10) is used, the point estimates tend to be positive and quite close to 0. These results suggest that the data in our simulations are stationary.

Lastly, we report the relative RMSPEs when we use the first difference estimator and update equation (9) as follows:

$$\Delta Y_{i,t,g} = \gamma_{t,g} + \beta_g^T \Delta T_{i,t,g} + \varepsilon_{i,t,g}. \qquad \text{(A.3)}$$

As expected, using the first-difference estimator with state and time fixed effects significantly reduces the correlation between the outcome variable and its lag (the average (median) value of the corresponding $\widehat{\gamma_g}$ is -0.19 (-0.22) after time demeaning). Nevertheless, the relative RMSPEs in Table A.4 lead us to conclusions very similar to the ones reached in Table 3. The overall accuracy slightly improved in all cases.

Table A.1: Quarterly current GDP by industry

		quarterly state GDP				quarterly state GDP	
		millions $	Log-transformed			millions $	Log-transformed
Total	mean	290,226	12.014	Finance and insurance	mean	24,507	9.359
	st. dev	376,670	1.055		st. dev	40,487	1.231
	CV	1.298	0.088		CV	1.652	0.132
	n	3,315	3,315		n	3,315	3,315
Agriculture, forestry, fishing and hunting	mean	3,247	7.120	Real estate and rental and leasing	mean	43,711	10.055
	st. dev	4,973	2.129		st. dev	64,638	1.104
	CV	1.531	0.299		CV	1.479	0.110
	n	3,315	3,315		n	3,315	3,315
Mining, quarrying, and oil and gas extraction	mean	5,979	6.756	Professional, scientific, and technical services	mean	24,329	9.311
	st. dev	19,688	2.432		st. dev	36,562	1.279
	CV	3.293	0.360		CV	1.503	0.137
	n	3,315	3,315		n	3,315	3,315
Utilities	mean	5,532	8.123	Management of companies and enterprises	mean	6,131	7.937
	st. dev	6,346	0.997		st. dev	7,426	1.409
	CV	1.147	0.123		CV	1.211	0.178
	n	3,315	3,315		n	3,315	3,315
Construction	mean	13,529	8.937	Administrative and support and waste management and remediation services	mean	9,951	8.561
	st. dev	17,167	1.204		st. dev	12,894	1.183
	CV	1.269	0.135		CV	1.296	0.138
	n	3,315	3,315		n	3,315	3,315
Manufacturing	mean	39,117	9.820	Educational services	mean	4,129	7.549
	st. dev	48,439	1.479		st. dev	5,925	1.308
	CV	1.238	0.151		CV	1.435	0.173
	n	3,315	3,315		n	3,315	3,315
Wholesale trade	mean	19,966	9.244	Health care and social assistance	mean	24,126	9.564
	st. dev	26,125	1.190		st. dev	28,090	1.038
	CV	1.308	0.129		CV	1.164	0.109
	n	3,315	3,315		n	3,315	3,315
Retail trade	mean	19,036	9.326	Arts, entertainment, and recreation	mean	3,382	7.347
	st. dev	23,172	1.042		st. dev	5,604	1.229
	CV	1.217	0.112		CV	1.657	0.167
	n	3,315	3,315		n	3,315	3,315
Transportation and warehousing	mean	9,960	8.649	Accommodation and food services	mean	9,528	8.635
	st. dev	11,760	1.124		st. dev	11,994	1.016
	CV	1.181	0.130		CV	1.259	0.118
	n	3,315	3,315		n	3,315	3,315
Information	mean	16,828	8.792	Other services (except government and government enterprises)	mean	7,238	8.346
	st. dev	33,233	1.335		st. dev	8,626	1.065
	CV	1.975	0.152		CV	1.192	0.128
	n	3,315	3,315		n	3,315	3,315

Notes: CV stands for coefficient of variation (standard deviation divided by mean). There are 65 quarters (2005Q1-2021Q1) and 51 states (including DC), hence 65*51=3,315 observations for each industry.

Table A.2: Relative mean RMSPEs after reconciliation of the counterfactual estimates obtained using alternative causal ML methods for log-transformed outcome variables

	GSYNTH	EM	MC	SC	EN	EN-T
Outcome: GDP	0.967	0.970	0.972	0.976	0.973	0.981
Outcome: Employment	0.965	0.949	0.968	0.981	0.956	0.983

Table A.3. Relative mean RMSPEs after reconciliation of the 2SLS counterfactuals for untransformed and log-transformed outcome variables

		Group Invariant Instrument (Eq. A.1 and A.1')		Group Varying Instrument (Eq. A.2 and A.2')	
Outcome Variable:		GDP	log(GDP)	GDP	log(GDP)
	overall	1.000	0.995	0.981	0.992
-Total Gross State Product (GSP)	top	1.000	1.499	0.967	1.451
-GSP by Industry	bottom	1.000	0.993	1.070	0.990
Outcome Variable:		Employment	Log (Employment)	Employment	log (Employment)
	overall	1.000	0.999	0.842	0.928
-Total State Employment	top	1.000	0.996	0.801	0.793
-State Employment by Industry	bottom	1.000	0.999	0.977	0.991

Notes: RMSPE stands for root mean squared prediction error. 2SLS stands for two-stage least squares. Each value in the table represents the ratio of the average RMSPE of the counterfactuals after reconciliation to the average RMSPE of the original counterfactuals. While calculating RMSPEs, "overall" indicates that all groups in all levels of the hierarchy are considered, whereas "top" and "bottom" indicate that only the top and bottom levels are considered. We report only when the horizon (post-treatment quarters) is one, training period (number of pre-treatment quarters) is 24, and number of treated units (states) is one.

Table A.4: Relative mean RMSPEs after reconciliation of the FD counterfactuals for log-transformed outcome variables

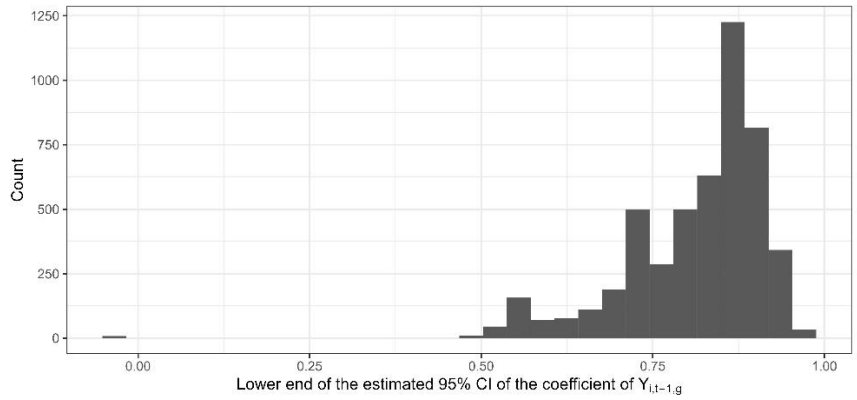
		Horizon = 1		Horizon = 4		Horizon = 12	
Panel (a): GDP							
# Treated Units = 1	overall	0.999	0.999	0.999	0.999	0.999	0.980
-Total Gross State Product (GSP)	top	1.112	1.059	1.317	1.156	1.373	2.404
-GSP by Industry	bottom	0.999	0.999	0.999	0.999	0.998	0.973
# Treated Units = 5	overall	0.996	0.997	0.994	0.993	0.997	0.992
-Total Gross State Product (GSP)	top	1.759	1.173	2.098	2.790	1.847	2.815
-GSP by Industry	bottom	0.995	0.997	0.992	0.989	0.995	0.987
# Quarters in Training Period:		24	36	24	36	24	36

Notes: RMSPE stands for root mean squared prediction error. FD stands for the first difference estimator in equation (A.3). Each value in the table represents the ratio of the average RMSPE of the counterfactuals after reconciliation to the average RMSPE of the original counterfactuals. While calculating RMSPEs, "overall" indicates that all groups in all levels of the hierarchy are considered, whereas "top" and "bottom" indicate that only the top and bottom levels are considered. We consider alternative horizons (post-treatment quarters), training periods (pre-treatment quarters), and number of treated units (states).

Figure A.1: Point and confidence interval estimates from the regressions of outcome variables on their first lags

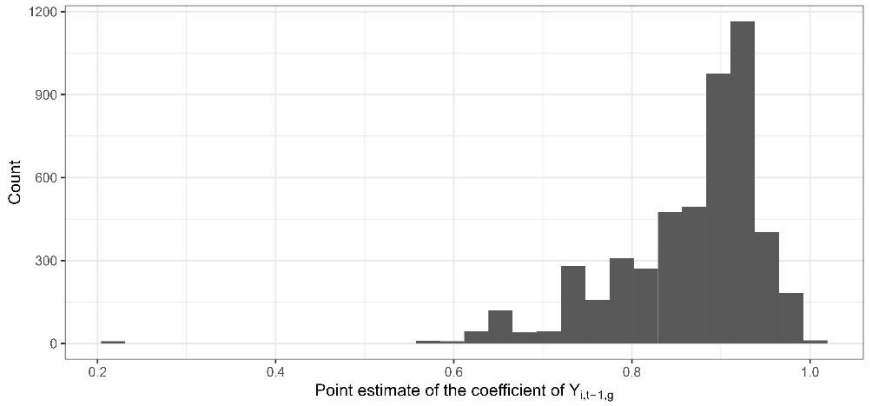
Panel A: Equation (9) with log(GDP) as the outcome variable (36 pre-treatment periods)

Lower Limits of the 95% CI in Simulations



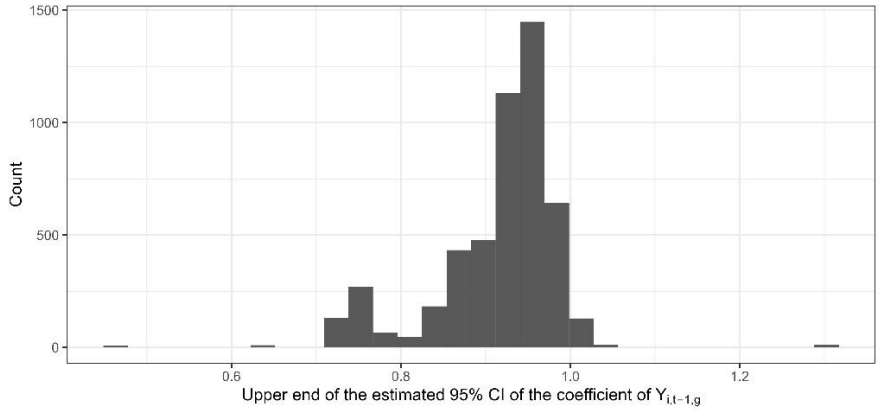
Only 0% of the lower end of the 95% CI are -1 or less.

Point Estimates



Only 0.2% of the estimated effect are 1 or greater.

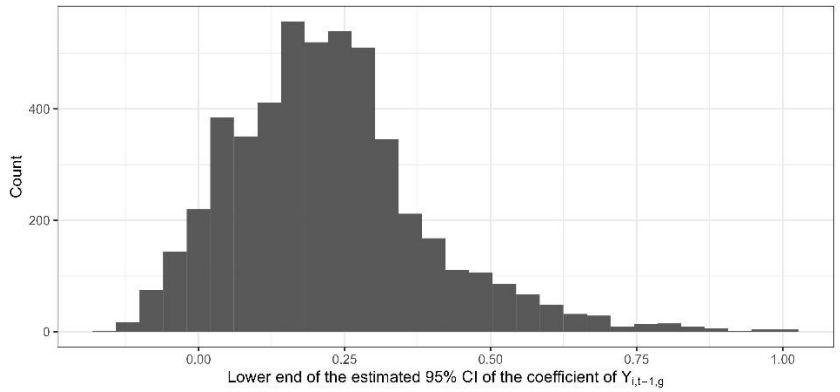
Upper Limits of the 95% CI in Simulations



Only 2.1% of the upper end of the 95% CI are 1 or greater.

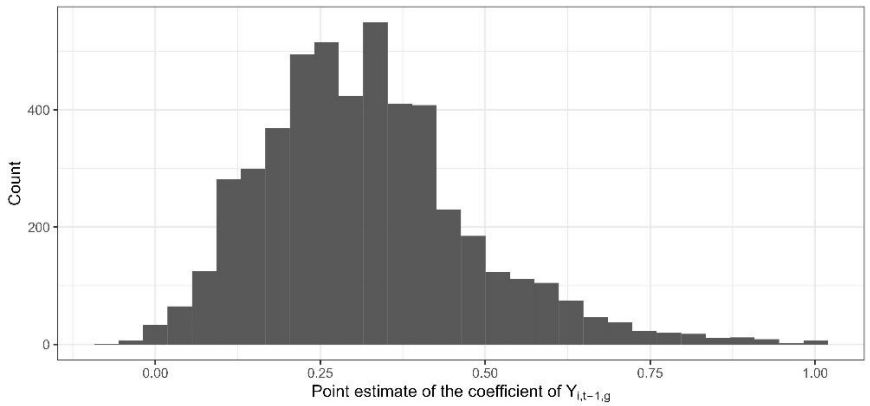
Panel B: Equation (10) with GDP as the outcome variable (36 pre-treatment periods)

Lower Limits of the 95% CI in Simulations



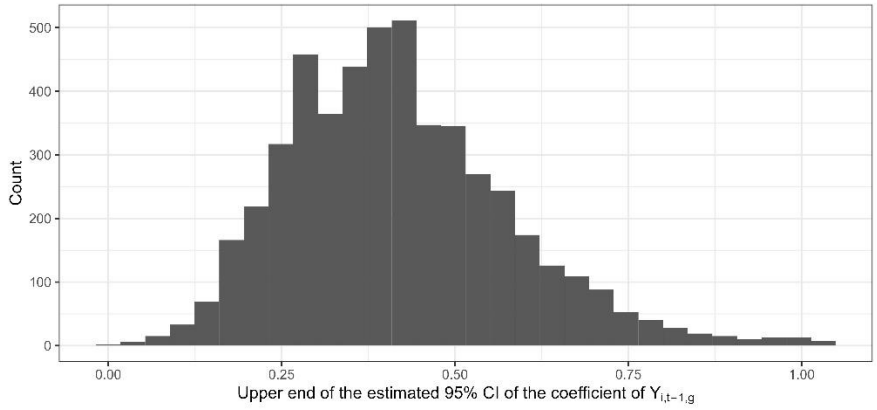
Only 0% of the upper end of the 95% CI are -1 or less.

Point Estimates



Only 0.1% of the estimated effect are 1 or greater.

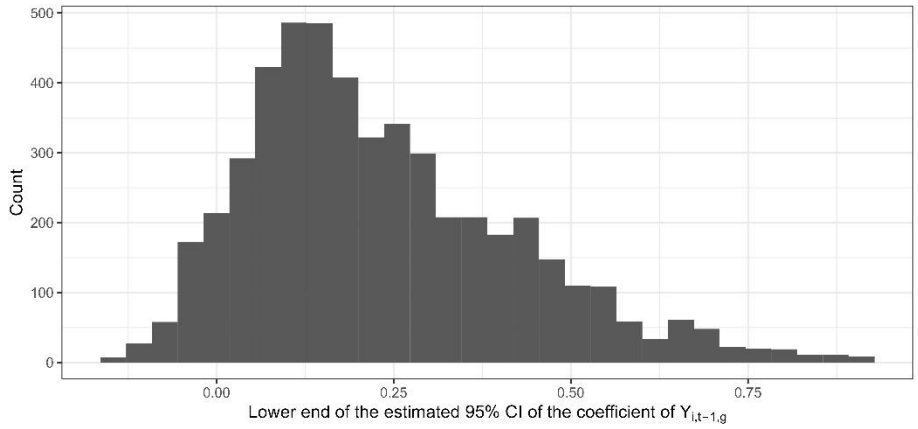
Upper Limits of the 95% CI in Simulations



Only 0.3% of the upper end of the 95% CI are 1 or greater.

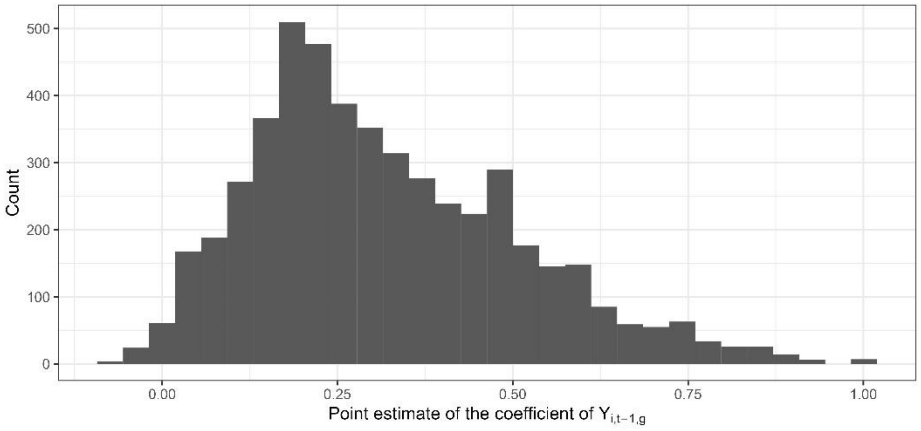
Panel C: Equation (10) with log(GDP) as the outcome variable (36 pre-treatment periods)

Lower Limits of the 95% CI in Simulations



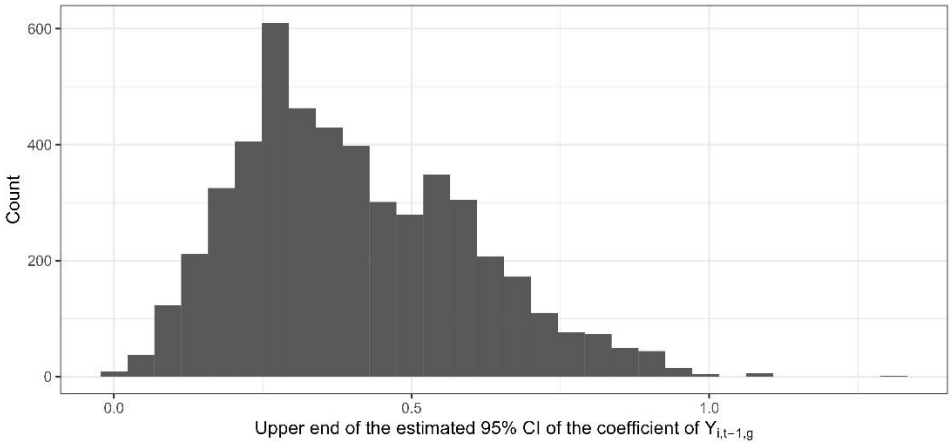
Only 0% of the lower end of the 95% CI are -1 or less.

Point Estimates



Only 0% of the estimated effect are 1 or greater.

Upper Limits of the 95% CI in Simulations



Only 0.1% of the upper end of the 95% CI are 1 or greater.