Using Multiple Imputation and Matching to Improve Network Effects on Corporate Tax Policy Interdependence

Azusa Uji
azusa.uji@gmail.com

In Young Park
inyoungpark@fas.harvard.edu

Zach Perez
zdperez@gmail.com

This Draft: May 4, 2015

Abstract

By using a spatial lag model, Cao (2010) tests the effects of economic and cultural similarities among countries on international diffusion of corporate tax policy and by extension. Our analysis attempts to improve upon the methods set forth by Cao, and further his investigation of cross-national policy interdependence. First, using recently developed statistical methods, we impute missing data to present a more thorough regression analysis. Second, we attempt to estimate the causal effects of structural equivalence with matching methods, and we show that strong covariate correlations confound matching techniques.

1 Introduction

By using a spatial lag model, Cao (1993) tests the effects of economic and cultural similarities among countries on international diffusion of corporate tax policy and by extension. While the effects of belonging geographical proximity, and cultural affinity are found to be significant determinants of capital tax rate convergence among nations in portfolio investment networks and similar intergovernmental organizations (IGO), Cao's analysis is significantly constrained by extremely high missingness in data (59 percent missingness). Moreover, while Cao considers the relevance of domestic political institutions, these variables are not controlled for in his analysis. In addition, using `matchIt` (Ho, Imai, King, and Stuart, 2005) or coarsened exact matching (Iacus, King, and Porro, 2011) on Cao's data, as recommended by the author, calls for further examination.

Our analysis improves upon Cao's methods by applying multiple imputation and matching, and furthers his investigation of cross-national policy interdependence by including domestic political variables. First, using `Amelia`, we impute missing data to present a more thorough regression analysis. Second, we identify model dependence by using Cao's spatial lag model to forecast the change in corporate tax rates for selected countries. Third, we replace Cao's k-nearest neighbors approach, which estimates the causal effects of structural equivalence, with matching methods to reduce model dependence.

In what follows, we describe in detail our replication strategy and its theoretical bases, a step-by-step walk-through of our replication, and analysis of its implications. The paper is organized as follows: section 2 outlines the blueprint of our replication, which is largely divided into three parts: imputation, matching, and including domestic political variables. Section 3 then compares the effects of these applications by using Cao's models on imputed data. Section 4 provides an in-depth discussion on the strengths and limitations of these new techniques. Section 5 concludes with recommendations for future research.

2 Description of Original Data

Each variable is stored as object data frame in panel format, except sorted by year, not country. For instance, each variable is formatted as "[num 1:1197]", from year 1 (containing all countries), year 2 (all countries), year 3 (all countries), and so on.

Figure 1: Descriptive Statistics of Cao's Variables

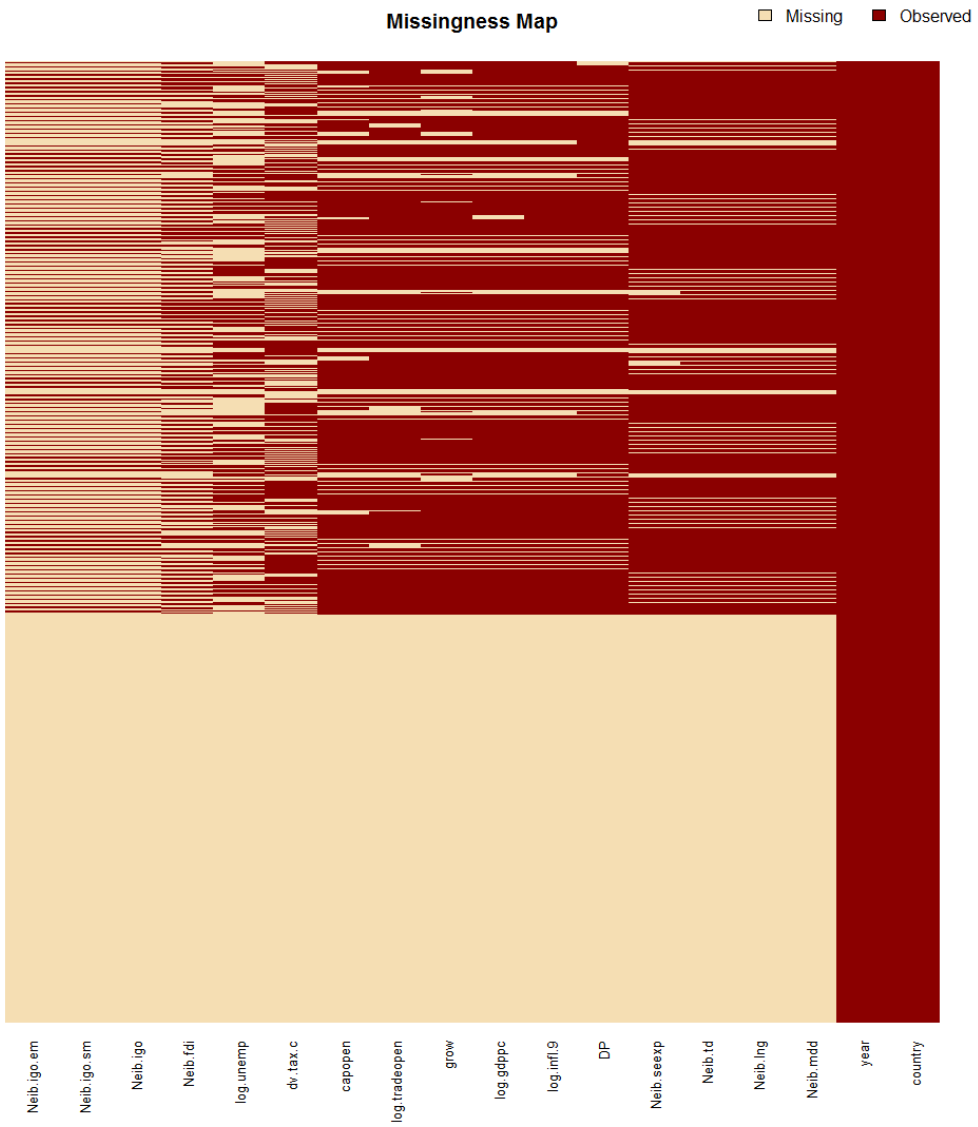
Tax rate	Tax rate (lag1)	Inflation	Growth	Unemployment	Population	GDP	Capitalopen	Tradeopen
Min. : 0.00	Min. :-1.00	Min. :-0.6931	Min. :-18.015	Min. :-0.3567	Min. :23.01	Min. :-2.0464	Min. :-1.7528	Min. :2.412
1st Qu.:25.00	1st Qu.:24.00	1st Qu.: 2.3795	1st Qu.: 2.032	1st Qu.: 1.6487	1st Qu.:32.19	1st Qu.:-0.0202	1st Qu.:-1.0948	1st Qu.:3.792
Median :30.00	Median :29.00	Median : 2.5337	Median : 3.854	Median : 2.1163	Median :34.92	Median : 1.2480	Median : 1.2431	Median :4.103
Mean :27.68	Mean :26.68	Mean : 2.6702	Mean : 3.666	Mean : 2.0381	Mean :36.66	Mean : 1.2494	Mean : 0.7442	Mean :4.117
3rd Qu.:35.00	3rd Qu.:34.00	3rd Qu.: 2.8332	3rd Qu.: 5.500	3rd Qu.: 2.4681	3rd Qu.:40.40	3rd Qu.: 2.6170	3rd Qu.: 2.6233	3rd Qu.:4.447
Max. :54.00	Max. :53.00	Max. : 6.3261	Max. : 26.200	Max. : 3.6163	Max. :52.95	Max. : 4.3856	Max. : 2.6233	Max. :5.740
NA's :1312	NA's :1312	NA's :1095	NA's :1117	NA's :1475	NA's :1071	NA's :1103	NA's :1133	NA's :1121
Distance 1	**Distance 2**	**Language**	**Trade**	**Neib.igo**	**Neib.cap**	**Neib.seexp**	**Neib.portf**	**Neib.fdi**
Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. : 0.00	Min. :13.66	Min. :13.66	Min. :13.66
1st Qu.:19.34	1st Qu.:17.94	1st Qu.: 0.00	1st Qu.:26.59	1st Qu.:26.13	1st Qu.: 79.15	1st Qu.:21.67	1st Qu.:20.70	1st Qu.:21.48
Median :24.30	Median :25.25	Median :16.00	Median :29.12	Median :27.12	Median : 190.50	Median :24.34	Median :23.24	Median :24.54
Mean :21.63	Mean :22.73	Mean :14.96	Mean :27.03	Mean :24.94	Mean : 296.32	Mean :24.17	Mean :23.03	Mean :24.11
3rd Qu.:27.50	3rd Qu.:30.00	3rd Qu.:28.28	3rd Qu.:31.23	3rd Qu.:27.63	3rd Qu.: 483.38	3rd Qu.:26.73	3rd Qu.:25.82	3rd Qu.:27.13
Max. :40.00	Max. :40.00	Max. :40.00	Max. :35.52	Max. :29.47	Max. :1572.00	Max. :33.53	Max. :30.37	Max. :32.67
NA's :1015	NA's :1015	NA's :1015	NA's :1015	NA's :1680	NA's :1015	NA's :1480	NA's :1714	NA's :1499

3 Replication Strategy

3.1 Multiple Imputation

As we examined Cao's data, we plotted a missingness map to conclude that imputation will contribute to Cao's analysis, which predicates on extremely high missingness in original data. As shown in the map below, nearly half of his the values in his data set are missing:

Figure 2: Missingness Map: Cao's Data



Prior to performing multiple imputation, we checked to ensure that our data meets the assumptions underlying `Amelia`. By trial and error, we learned that some of our $W * y$

matrices were perfectly collinear with one another. To pick out the offenders, we ran linear regression on tax rates by incrementally adding each independent variable to the model, in order to examine which variable was causing collinearity. We found that adding the variable NEIB.PORTF returned NA's as regression coefficients. This indicates that Neib.portf is linearly related to Neib.fdi, a variable we added to the model immediately prior to adding Neib.portf. Because Amelia cannot take as input linearly correlated or linear combinations of variables in the imputation model, we decided to drop Neib.portf in imputing the data.

Moreover, we took natural logs of major macroeconomic variables so that the distributions of tax rate (outcome variable), inflation, unemployment, GDP per capita, and trade openness are assumed to follow normal.

Figure 3: Imputation Diagnostics: Predictions for Variables

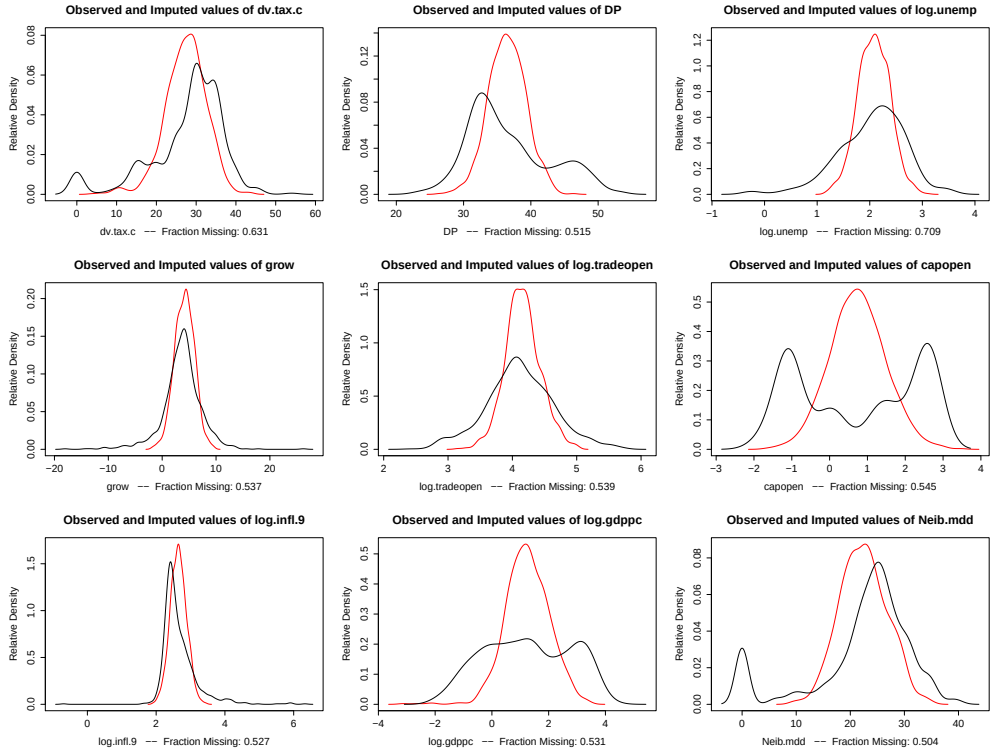
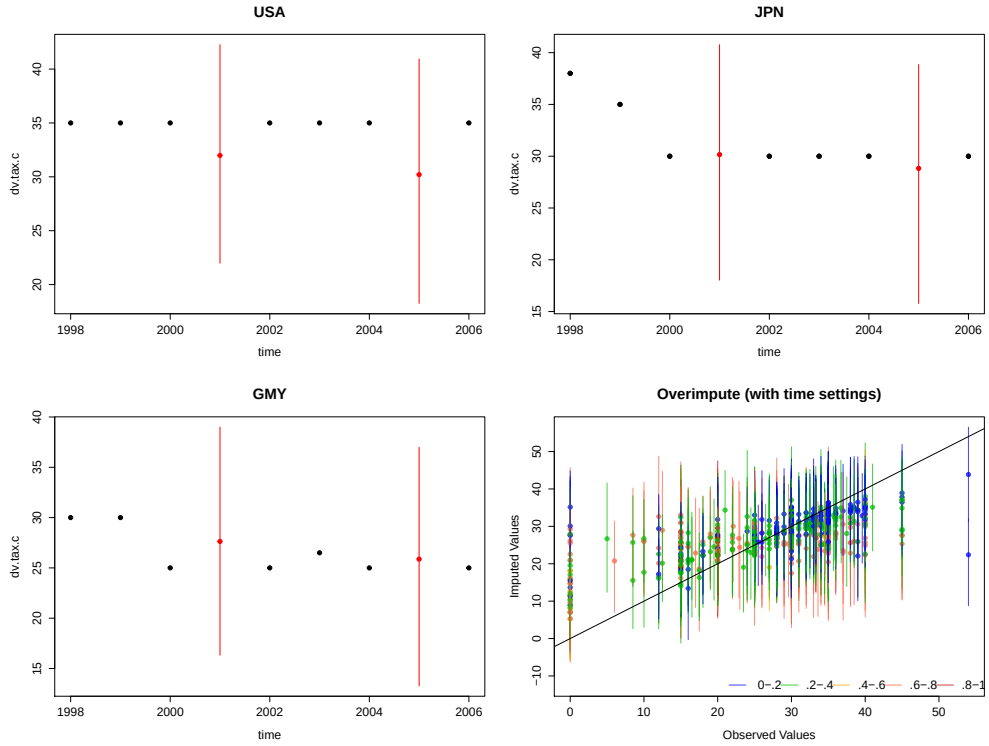


Figure 4: Imputation Diagnostics: Tax Rates and Overimputation



3.2 Matching

The spatial lag model is designed to account for autocorrelation among the covariates. The key assumption of the model is that dependence among observations is known and is fully represented by the connectivity matrix, denoted W.

$$y = X\beta + \rho Wy + \epsilon \tag{1}$$

Due to the nature of the time-series, cross-sectional data we use, we also introduce a measure of time dependence, which we do by lagging the dependent variable and regressing on it. In order to use this lag, we must assume that changes in the dependent variable for one observation do not affect other observations until the next year. For an economic factor such as corporate tax rates, this is a reasonable assumption to make.

$$y_{i,t} = x'_{i,t}\beta + \phi y_{i,t-1} + \rho w_i y_{t-1} + \epsilon_{i,t} \tag{2}$$

We use the above model to predict the corporate tax rates of a given country from a number of domestic economic factors, X, and a connectivity network, W, accounting for spatial connectivity. We consider the spatial connectivity between countries on seven factors: geography, foreign direct investment, foreign portfolio investment, export markets, trade, and membership in three categories of intergovernmental organizations. The coefficient ρ, which we attempt to estimate will tell us the effect a certain type of connectivity has on the dependent variable.

As mentioned above, the purpose of the Wy connectivity matrix is to reduce the dependence between observations. It is usually row standardized, such that the sum of all of the connections across a row sums to one (Beck et al., 2006). We attempted to match on underlying connections, before standardizing the matrix. We believe that relevant information is lost in this standardization, and that by retrieving it, a causal claim is possible as recommended by Cao (1993).

4 Linear Regression with Spatial Lags: Effects of Imputation, Matching, and Domestic Variables

4.1 Regression Analysis

We use Cao's models to test the effects of trade, structural equivalence in exports, FDI flow, geographical and cultural proximity on national corporate tax rates. We find that all of the coefficients are statistically more significant after missing values have been filled in using imputation.

4.2 Results

The results are described in the table that follows. SE, standard error; within 950 km is used as the threshold to define connectivities; fixed country and year effects are estimated but not reported in the table. As seen in the results below, applying multiple imputation has significantly improved the effects of trade and FDI networks as well as geographical and cultural affinities on the interdependence of national corporate tax rates. Interestingly, however, *all* of the coefficients are found to be highly significant (p-values under 0.00), suggesting mishaps in the performance of imputation in `Amelia`.

Table 1: Temporally Lagged Spatial Lag Model: Geography (Minimum Distance) and Common Language, 1999–2006

	Geography		**Common Language**		**Geography & Language**	
	Cao	Imputed	Cao	Imputed	Cao	Imputed
(Intercept)	−59.87	34.93	−53.80	34.34	−60.04	34.64
	(19.50)***	(3.84)***	(19.51)***	(4.22)***	(19.62)***	(3.96)***
Tax lag at $t-1$	0.31	−0.02	0.30	−0.02	0.31	(−0.02)
	(0.05)***	(0.03)***	(0.05)***	(0.03)***	(0.05)***	(0.03)***
Inflation rate	−0.01	−0.01	−0.00	−0.01	−0.01	−0.01
	(0.03)	(0.01)***	(0.03)	(0.01)***	(0.03)	(0.01)***
GDP Growth	−0.17	−0.12	−0.14	−0.12	−0.17	−0.11
	(0.10)	(0.05)***	(0.10)	(0.06)***	(0.10)	(0.06)***
Unemployment rate	−0.11	0.12	0.09	0.11	−0.11	0.11
	(0.17)	(0.05)***	(0.17)	(0.06)***	(0.17)	(0.06)***
Age dependency	2.20	0.15	2.14	0.18	2.20	0.16
	(0.50)***	(0.07)***	(0.50)***	(0.06)***	(0.50)***	(0.07)***
GDP per capita	−0.28	−0.01	−0.27	−0.01	-0.28	−0.01
	(0.12)***	(0.01)***	(0.12)***	(0.01)***	(0.12)***	(0.01)***
Capital market openness	0.62	−0.11	−0.70	−0.13	-0.62	−0.12
	(0.47)	(0.31)***	(0.48)	(0.28)***	(0.47)	(0.30)***
Trade openness	0.03	−0.01	0.02	−0.01	0.03	−0.01
	(0.03)	(0.01)***	(0.03)	(0.01)***	(0.03)	(0.01)***
$\rho_{geography}$	0.20	0.15			0.20	0.15
	(0.10)***	(0.04)***			(0.10)***	(0.03)***
$\rho_{language}$			0.03	0.05	0.01	0.04
			(0.08)	(0.03)***	(0.08)	(0.03)***
Number of observations	336	2079	336	2079	336	2079
Residual SE	3.63		3.66		3.64	
Adjusted R^2	0.77		0.77		0.77	

Table 2: Policy Interdependence Induced by Network Position Similarity in FDI and Export Networks

	FDI		**Exports**	
	Cao	Imputed	Cao	Imputed
(Intercept)	−88.58	34.69	−62.97	30.90
	(24.21)***	(4.06)***	(19.78)***	(3.90)***
Tax lag at $t-1$	0.27	−0.02	0.31	−0.02
	(0.06)***	(0.03)***	(0.05)***	(0.03)***
Inflation rate	−0.03	−0.01	−0.01	−0.01
	(0.04)	(0.01)***	(0.03)	(0.01)***
GDP Growth	−0.18	−0.12	−0.17	−0.09
	(0.12)	(0.05)***	(0.10)	(0.06)***
Unemployment rate	−0.17	0.11	−0.13	0.12
	(0.21)	(0.05)***	(0.17)	(0.06)***
Age dependency	2.95	0.15	2.06	0.16
	(0.62)***	(0.07)***	(0.50)***	(0.07)***
GDP per capita	−0.28	−0.01	−0.24	−0.01
	(0.18)***	(0.01)***	(0.12)**	(0.01)***
Capital market openness	−0.67	−0.11	−0.56	−0.13
	(0.61)	(0.31)***	(0.47)	(0.31)***
Trade openness	0.01	−0.01	0.03	−0.01
	(0.05)	(0.01)***	(0.03)	(0.01)***
$\rho_{geography}$	0.30	0.15	0.16	0.14
	(0.12)***	(0.03)***	(0.10)**	(0.04)***
ρ_{FDI}	0.00	0.01		
	(0.10)	(0.07)***		
$\rho_{exports}$			0.33	0.16
			(0.16)***	(0.05)***
Number of observations	272	2079	336	2079
Residual SE	3.80		3.66	
Adjusted R^2	0.76		0.77	

Table 3: Temporally Lagged Spatial Lag Model: Bilateral Trade, 1999–2006

	Geography & Trade	
	Cao	Imputed
(Intercept)	−56.60	27.71
	(20.30)***	(4.09)***
Tax lag at $t-1$	0.30	−0.02
	(0.05)***	(0.03)***
Inflation rate	−0.01	−0.01
	(0.03)	(0.01)***
GDP Growth	−0.17	−0.11
	(0.10)**	(0.05)***
Unemployment rate	−0.11	0.11
	(0.17)	(0.05)***
Age dependency	2.20	0.17
	(0.50)***	(0.06)***
GDP per capita	−0.28	−0.01
	(0.12)***	(0.01)***
Capital market openness	−0.60	−0.11
	(0.48)	(0.31)***
Trade openness	0.02	−0.01
	(0.03)	(0.01)***
$\rho_{geography}$	0.21	0.12
	(0.10)***	(0.03)***
ρ_{Trade}	−0.12	0.23
	(0.20)	(0.07)***
Number of observations	272	2079
Residual SE	3.80	
Adjusted R^2	0.76	

5 Limitations of Our Approach

5.1 Multiple Imputation

We learned that `Amelia` picks up any high-frequency values and forms a normal distribution around these values. For instance, with some tax rates are stacked at zero, the distribution of tax rates after imputation shows that the predicted values of missing data follow a curve surrounding small normal distributions of high-frequency tax rates, as shown below: This was likewise an issue with other variables with high-frequency values as well.

A larger challenge, however, was to compare and contrast across multiple imputation models. The available measures being plots, our choice of an optimal imputation model was heavily reliant on graphical conjecture. In future versions of the package, we hope to see numerical measures of optimality, such as standard error or R-squared, that indicate the predictability of the models. This is particularly concerning when imputation is used to determine which group of variables best suit the analysis. In our case, upon finding that inclusion of all $W*y$ matrices does not necessarily render imputed estimates close to observed values, we had to rely on a iterative process of imputing and plotting the imputed values to check inclusion of which $W * y$ would contribute most to predicting unobserved data. With numerical criteria (other than variation in chain lengths), this process could have been less arbitrary.

Relatedly, we observed that there exists a trade-off between numerical instability underlying the imputation process and the predictability of the imputation models. For example, knowing that our continuous variable, `CapOpen`, has distributed bi-modal, and recognizing that no transformation will make this distribution appear normal, we decided that setting it as a nominal variable might help us obtain better predicted values for the missing cases. Indeed, when we ran multiple imputation, the chain lengths were 410, 458, 488, 729, and 261, numerically unstable and consuming much more time. Nevertheless, the imputation model now predicts much close to the observed values. Particularly, `CapOpen` is notably better predicted as a nominal variable: Again, in the absence of quantifiable criteria of model fit, we rely on graphical representations of the goodness of fit in our imputation outputs. As a result of setting `CapOpen` nominal, other variables are predicted better by `Amelia`.

In addition, missing values in some of the Wy matrices were replaced with zeros, before cross-multiplication, producing bimodal distributions—one peak at zero and the other at a high-frequency value. Values in the other matrices were calculated using a threshold. Existing documentation does not provide sufficient explanation for how to deal with these special cases. We could either designate these variables as "nominal," which can also bias our estimation in unexpected ways. Further examination of special conditions such as the spatial weights used in Cao will benefit future applications of multiple imputation. If country A was below this threshold, it was coded zero; otherwise, its true value was recorded. This likewise produced bimodal distributions. Despite Amelia's assumption of multivariate normality, and usefulness of log transformations for certain generic macroeconomic variables, transforming these matrices could introduce even more bias in our estimation. Indeed, outputs from Amelia (density plots as previously presented) demonstrated that Amelia assumes that most missing values are in the middle of the two peaks. In the case of matrices based on thresholds, this will bias where the country is located vis-á-vis the threshold.

5.2 Matching

In our analysis, we match observations on a spatial lag model in an attempt to retrieve lost information from a connectivity matrix in hopes of revealing a causal relationship. For instance, a unit with many connections to other units that have low-valued dependent variables could have the exact same $W_i y_{t-1}$ value as a unit with only one connection to a unit with a high-valued dependent variable. In such a situation, $w_i y_{t-1}$ does not provide much utility for measuring the effect of connections on dependent variables. Therefore, we sum the connections across entries in the connectivity matrix before standardizing and use the sums as a treatment variable. We define a treated observation as one with a large number of connections and a control observation as one with with a low number of connections. Since the number of connections is continuous, we will establish a threshold above which an observation will be considered treated.

We utilize many methods for choosing a treatment threshold. We plot histograms of connections for a W matrix and choose a natural breaking point in the number of connections. If the connections appear normally distributed, we test several different treatment thresholds,

such as mean, median, and first and third quartile, and attempt to obtain a balanced, matched set. We use both Mahalanobis and coarsened exact matching to make our matches. We match on all of the covariates of X, but not on y_{t-1} or on Wy, which are both confounding variables. For coarsened exact matching, we have chosen coarsening levels through trial and error with the goal of decreasing both univariate and multivariate imbalance.

In general, we can find good univariate balance on both Mahalanobis and coarsened exact matching methods. However, the multivariate imbalance measure, L1, rarely drops below 1.0. The results of regressing on our matched data sets for all seven connectivity matrices show that none of the coefficients are significant. We fail to find a matched set for IGO Membership matrices due to a large amount of missing data that could not be imputed. We conclude that matching techniques, when used with a spatial lag model, could be counterproductive. The first SUTVA assumption of non-interference is violated by the very nature of a spatial lag model, and the second assumption of no variation in treatment is violated without a clear separation between levels of connections in the treatment variable. Nonetheless, the door is open to future efforts regarding matching on spatial lag models.

Table 4: Coefficients for variables suspected of influencing corporate tax rates. Regressions run on matched data sets with all aforementioned independent X variables and fixed effects for country and year.

	Mahalaobis Matching			**Coarsened Exact Matching**		
	Coefficient	Standard Error	p-value	Coefficient	Standard Error	p-value
Geography	0.13	0.10	0.22	-0.12	0.33	0.72
Foreign Direct Investment	0.06	0.21	0.75	0.12	0.32	0.71
Structural Equivalence in Exports	0.13	0.09	0.14	0.39	0.69	0.57
Total Trade	0.09	0.22	0.67	-0.13	0.75	0.86
IGO Membership	0.13	0.54	0.81			
Minalist and Social-Cultural IGO Membership	0.20	0.39	0.61			
Minalist and Economic IGO Membership	0.14	0.57	0.81			

6 Conclusion

Our analysis of the spatial lag model of national corporate tax rates suggests areas for further research. First, imputation techniques can be refined to fit specific data sets. In our case, missing values in some of the Wy matrices were replaced with zeros, before cross-multiplication, producing bimodal distributions—one peak at zero and the other at a high-frequency value. Values in the other matrices were calculated using a threshold. Existing documentation does not provide sufficient explanation for how to deal with these special cases. We could either designate these variables as "nominal," which can also bias our estimation in unexpected ways. Further examination of special conditions such as the spatial weights used in Cao will benefit future applications of multiple imputation.

Second, matching on connectivity measures could be tried in combination with other methods that control for correlation among TSCS observations. Strong temporal and spatial dependence between covariates obscured the effects of spatial dependence on outcome variables. It is perhaps necessary to explore or develop different regression models for use in situations of spatial dependence before attempting to show causal relationships with matching.

Third, we might be able to incorporate domestic variables from theoretical and empirical perspectives. Policy diffusion studies such as this one have solely focused on horizontal aspect and have largely ignored how domestic factors work bottom-up in the horizontal process of diffusion. We could see some scenarios where spatial policy diffusion is conditional on specific domestic variables such as oppositions or political stability.

References

N. Beck, K. S. Gleditsch, and K. Beardsley. Space is more than geography: Using spatial econometrics in the study of political economy. *International Studies Quarterly*, 50:27–44, 2006.

X. Cao. Networks as channels of policy diffusion: Explaining worldwide changes in capital taxation, 1998-2006. *International Studies Quarterly*, 54:823–854, 1993.

D. E. Ho, K. Imai, G. King, and E. A. Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. 2005.

S. M. Iacus, G. King, and G. Porro. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 2011.