Fiscal Capacity and Inequality: Evidence from Brazilian Municipalities*

Florian M. Hollenbach† & Thiago Silva‡

July 17, 2018

Accepted for Publication at the Journal of Politics

*Supplementary material for this article is available in the appendix in the online edition. Replication files are available in the JOP Data Archive on Dataverse (`http://thedata.harvard.edu/dvn/dv/jop`).

†Corresponding Author, Assistant Professor, Department of Political Science, Texas A&M University, 2010 Allen Building, 4348 TAMU, College Station, TX, USA, 77843-4348. Email: fhollenbach@tamu.edu. Phone: 979-845-5021. URL: `fhollenbach.org`

‡PhD Candidate, Department of Political Science, Texas A&M University, 2010 Allen Building, 4348 TAMU, College Station, TX, USA, 77843-4348. Email: nsthiago@tamu.edu.

Abstract

We argue that in young democracies, wealthy elites can limit their taxes by constraining the fiscal capacity of the state. Corrupting local officials and undermining fiscal capacity are some of the mechanisms by which high-income earners can lower their own tax liabilities, even when voters favor higher *de jure* levels of taxation. The incentive to undermine fiscal capacity is especially compelling when inequality is high, as the median voter is likely to support higher progressive taxation and redistribution. Using data from over $5,500$ Brazilian municipalities, we show that localities with higher levels of inequality accrue less revenue from local property taxes. These results are robust to estimating a number of cross-sectional models, as well as panel models with time and municipal fixed effects. Moreover, we show that municipalities with high levels of inequality are less likely to apply to a federal grant program to increase their capacity to collect taxes.

Key Words: Fiscal Capacity, Taxation, Inequality, Democracy

Scholars often presume that governments can enforce their preferred fiscal policies. This assumption has been empirically proven to be false, as governments' ability to collect taxes varies dramatically around the world. What explains these differences across countries, and who might have an interest in maintaining low levels of tax capacity that make evasion easier?

One of the key research questions in political economy is why some countries redistribute more than others (e.g., Acemoglu et al., 2015). In particular, why do many democracies with high levels of inequality redistribute far less than the Meltzer-Richard-Romer models would lead us to expect (Romer, 1975; Meltzer and Richard, 1981)?

Most of the research on redistribution starts with the assumption that states are capable of efficiently collecting taxes and redistributing income, and thus focuses on examining the timing and impact of government decisions to implement redistributive policies. More recent studies have argued that political and economic elites in formerly autocratic regimes may undermine future political processes and limit political choices through institutional designs (Ardanaz and Scartascini, 2013; Albertus and Menaldo, 2014) or low state capacity (Acemoglu et al., 2015). Therefore, even if democratic polities are firmly in favor of redistributive policies, institutions and bureaucratic legacies may undermine the political and administrative process to *de facto* block redistribution.

In this paper, we investigate the idea that economic elites in democracies can undermine the state's ability to collect revenues and that they to do so when levels of inequality are high. Specifically, we ask whether local economic and political elites can undermine efforts to increase taxation in democracies by inhibiting the ability to collect taxes.

We think of the state's capacity to enforce tax policies as endogenous and argue that when

citizens vote for higher taxes, economic elites (the wealthy) have incentives to undermine the state's ability to collect taxes. The higher the equilibrium level of redistribution would be in a world with perfect tax collection, the stronger is the incentive for economic elites to erode the state's fiscal capacity. Weakening the state's administrative and tax capacity gives economic elites a mechanism with which to constrain policy choices and *de facto* levels of taxation outside the political system.

To investigate the theoretical argument, we use data on tax revenues from over $5,500$ Brazilian municipalities. We show that, controlling for a variety of other factors, localities with higher levels of inequality raise less revenue from local property taxes. These results are robust to estimating a variety of cross-sectional models for 2000 and 2010, as well as panel models with time and municipal fixed effects. We also show that municipalities with high levels of inequality were less likely to apply to a federal grant program to increase their local tax capacity.

Fiscal Capacity & Public Spending

Research in political science and economics often starts with the premise that in democratic polities, higher economic inequality ought to be associated with political demands for redistribution. Much of this work builds on the seminal model developed by Meltzer and Richard (1981), who showed that as the difference in mean income and income of the median voter increases, levels of taxation and redistribution should rise. The idea that democracy can and would be used for redistribution when inequality exists is not new, however, and goes at least as far back as Marx. While the Meltzer-Richard model is only one specific formalization, we expect rational voters in democracies to vote for higher taxation and redistribution as

long as their marginal benefit from higher rates is positive. When taxes are linear or progressive, poorer citizens ought to prefer higher taxes than the rich. More so, if the benefit of government spending is higher for poor than rich voters, the optimal tax rate for the poor increases. Contrary to these expectations, empirically there is little evidence that inequality is associated with higher redistribution in democracies (e.g., Benabou, 1996; Perotti, 1996; Kenworthy and Pontussen, 2005).

The lack of empirical support for the Meltzer and Richard (1981) model at the cross-national level is frequently noted. Some factors that possibly condition the relationship between inequality and redistribution are differences between social insurance and redistributive policies (e.g., Moene and Wallerstein, 2001), institutional structures (e.g., Persson and Tabellini, 2003; Iversen and Soskice, 2006), religion (Scheve and Stasavage, 2006), and ethnicity (Alesina and Glaeser, 2004). More recently, scholars have argued that politics in authoritarian regimes can have lasting effects on fiscal policies, potentially long after the transition to democracy. Albertus and Menaldo (2014), for example, argue that autocratic elites can shape the institutional design of subsequent democracies to influence and shape future politics – i.e., by influencing the "rules of the game" (Albertus and Menaldo, 2014). Ardanaz and Scartascini (2013) contend that higher inequality leads to more legislative malapportionment, which makes enacting redistributive policies more difficult once the democratic regime is established.

While the design of political institutions with many veto points is one strategy to inhibit redistribution in democracies, undermining state capacity with the goal to keep the state from collecting revenue may be an equally compelling strategy. Economic elites may cripple the political process by stifling the state's ability to raise revenue. Theoretical models show

that non-democracies with higher levels of income inequality should see lower investment in state capacity (Besley and Persson, 2011).

Similarly, Acemoglu and Robinson (2008) argue that possible changes in *de jure* political institutions give economic elites reasons to invest in subverting the state, to "capture democracy" and gain influence over policy decisions. An inefficient state with corrupt ("captured") bureaucrats may be a valuable strategy for economic elites to safeguard themselves against the political power of the masses (Acemoglu, Vindigni and Ticchi, 2011).

In line with these explanations, we argue that economic elites in democracies can exploit and further weaken the state's ability to collect revenue in an effort to block taxation demanded by voters. We contend that in democratic systems, rich or wealthy citizens can keep levels of taxation low, using both democratic and undemocratic means. The wealthy have incentives to ensure that their interests are (over) represented and that taxation is limited. One way to do so is by undermining the state's ability to collect taxes, i.e., by constraining its fiscal capacity. Raising taxes is a complicated undertaking that involves collecting large amounts of data and requires a functioning and efficient bureaucracy (Besley and Persson, 2009). Yet many governments cannot enforce the tax policies chosen by their governing bodies (Bird and Zolt, 2008; Gordon and Li, 2009). In such settings, wealthy residents may have strong incentives to undermine the state and limit their personal tax payments by lowering the state's ability to collect taxes.

To illustrate our argument, consider a theoretical society with rich (r) and poor (p) citizens, in which the median voter sets the *de jure* tax rate and is a member of the poor. Both wealth and income are taxable. Assume all revenue is used to finance a public good, such as education, or used as direct transfers. Assuming the median voter is decisive, she

should vote for higher taxes until the marginal benefit from the financed public good is equal to her marginal cost of taxation. If taxes are not regressive and revenue is used for public goods or transfers, then the optimal tax rate at which the marginal benefit equals the marginal cost for the poor rises with increasing inequality.

As the tax becomes more progressive and spending benefits poor citizens more than the rich, the effect of inequality on the tax rate ought to be more pronounced. Thus, in accordance with the standard theory, if citizens vote rationally and based on income, we should see higher levels of *de jure* taxation in states with higher levels of inequality. On the other hand, the difference between pre- and post-tax income of the wealthy elite would increase with higher levels of inequality. With this standard argument in mind, one could hypothesize that higher inequality leads to higher taxation (i.e., *de jure* tax rates) in democracies.

The distinction between *de jure* and *de facto* taxation is important for our theoretical argument. As taxes have to be administered and collected, *de jure* tax rates must not translate into the same *de facto* level of taxation. For example, with a *de jure* tax rate of 15%, even the most efficient and effective tax administration does not achieve 15% realized revenue. We define the *de facto* tax rate as the actual share of the tax base that is collected in taxes. As the capacity of the tax administration decreases, the difference between *de jure* and *de facto* tax rates becomes greater.

In a democracy with weak administrative capacity and firm entrenchment of the wealthy in the political process, elites have strong incentives to undermine the state's ability to collect taxes. As outlined above, when inequality is higher, the *de jure* tax rate is likely to rise. When *de jure* tax rates increase, however, it becomes more profitable for economic elites to combat

the state's ability to assess their tax liabilities or to influence the political process through other means. Alternative avenues for influence could include bribing local tax officials who are responsible for tax assessment, placing cronies in essential positions in the local bureaucracy, or impeding the purchase of necessary tools to make tax collection more efficient. Thus, in sufficiently weak states, we contend that economic elites can undermine tax collection, and the motivation to do so increases with higher levels of inequality.

We expect these tactics to be more likely in the context of highly progressive taxes. As a given tax becomes more progressive, the rich pay a higher share of tax revenue, which increases their motivation to fight tax collection. The difference between *de jure* and *de facto* rates should thus be more significant for more progressive taxes. Similarly, as spending benefits the poor more, we expect the relationship between inequality and the *de jure* taxation to become stronger, again raising incentives for elites to fight taxation.

Based on this theoretical argument, we develop our central hypothesis. Specifically, we expect that higher inequality is associated with less fiscal capacity, and therefore less *de facto* tax revenue. Our approach contrasts with the above outlined traditional hypothesis that higher inequality is associated with more tax revenue.

Research Design: The Case of Brazil

In this paper, we use data on tax collection from over $5,500$ Brazilian municipalities to investigate the empirical argument. There are several reasons for using the case of Brazil and its municipalities as the unit of analysis.

The democratization of Brazil in the mid-1980s advanced the country socially and politically (Oliven, Ridenti and Brandão, 2008). There are now few barriers to voter registration

(Limongi, Cheibub and Figueiredo, 2015), and compulsory voting ensures a turnout close to 80% (Nicolau, 2012). Since its transition to democracy, Brazil has been known for its high levels of income inequality, making it one of the most unequal democracies in the world. Inequality has been surprisingly resilient and stable throughout the transition from the military dictatorship (1964–1985) to the new democratic regime (Barros, Henriques and Mendonça, 2000; Souza and Medeiros, 2015).

The relatively recent transition to democracy and the persistence of inequality are two reasons that make it an intriguing case with which to investigate our argument. If the *standard* arguments were correct, we would have expected a stark increase in redistribution and taxation after Brazil's democratization in the 1980s. The argument we make above is one possible explanation for why this has *not* been the case.

The Case for Studying Municipalities

The Brazilian federative union is composed of 26 states and the federal district. Brazil has $5,570$ municipalities, its lowest level of government, which have more political autonomy than localities in any other Latin American country (Nickson, 1995; Rodríguez and Velásquez, 1995). Most political responsibilities lie with the federal union or states, yet the 1988 constitution gave substantial autonomy to the municipalities (Andrade, 2007; Baiocchi, 2006; Samuels, 2004). In line with the increase in political authority, municipalities can institute and collect taxes within their jurisdiction and use the revenue to implement local policies (Arretche, 2004; Andrade, 2007).

The municipalities are largely funded by transfers from the federal and state governments. These transfers have significantly declined, however, leading to budget shortfalls and low

revenues in many municipalities. One of the most critical local tax sources is the taxation of property and land in urban areas, the *Imposto Predial e Territorial Urbano* (IPTU): the urban land and building tax. This tax is solely available to municipalities, and its importance as a local revenue source has increased significantly (De Cesare and Ruddock, 1999).

We aim to investigate whether elites use low levels of administrative capacity, as well as undermine it further, to limit their taxation. To do so, we focus on the case of the property tax in Brazilian municipalities. While the IPTU is one of the principal sources of local revenue in Brazil (property taxes represent an average of 30% of the local tax revenue) (Smolka and Furtado, 1996; De Cesare and Ruddock, 1999), comprehensive studies of this tax indicate that it is still overlooked and has unrealized potential (De Cesare and Ruddock, 1999; Afonso and Araújo, 2006; Afonso, Araújo and Nóbrega, 2013).

While property taxation is a tax on wealth, we believe our theoretical argument, which is primarily about income inequality, still applies here. The IPTU is the second most important local revenue source available to municipalities (Afonso, Araújo and Nóbrega, 2013) and has the potential to be highly progressive. Therefore, if voters observe high levels of inequality and as a result demand more taxation and spending, the IPTU is the primary local mechanism to raise these funds. Moreover, administration of the property tax requires high administrative capacity (Bahl and Martinez-Vasquez, 2008; Kelly, 2013), making it a worthwhile endeavor for elites to engage in actions to undermine the collection of these taxes.

The distributive effects of the tax and relevant spending instruments are similarly important. We have strong reason to believe that the property tax is progressive by design, and that municipal spending largely benefits the poor. First, after the new constitution was enacted in 1988, a progressive property tax system was considered a potential policy mech-

anism to overcome urban social inequalities and attain equity (De Cesare, 2012; De Cesare and Smolka, 2004; Carvalho, Jr., 2015). After a period of legal ambiguity, a constitutional amendment was passed in 2000, that explicitly allowed progressive tax rates for the IPTU (Carvalho, Jr., 2013). In reality, however, the IPTU has been found to be a regressive tax (Carvalho, Jr., 2006, 2015; Afonso, Araújo and Nóbrega, 2013).

Several causes for the regressivity of the IPTU have been suggested. Directly in line with our argument, one significant reason for its regressive nature is the poor collection of the IPTU. This is due to administrative mismanagement, administrative inefficiency, the high cost of maintaining the property register, and the discrepancy between the government's real estate evaluations and their market value (De Cesare, 2005; Carvalho, Jr., 2006, 2015). Tax exemptions for large companies and tax evasion are also responsible for the high regressivity (De Cesare and Smolka, 2004; Carvalho, Jr., 2006).

De Cesare (2005) and Afonso, Araújo and Nóbrega (2013) found that changes in IPTU rates depend on the approval of councilors in the municipal legislature. Not surprisingly, property owners in wealthier areas regularly resist higher rates, and even more so if the revenue will be invested in poorer areas of the municipality (De Cesare, 2005; Afonso, Araújo and Nóbrega, 2013). Similarly, organized groups of landowners tend to pressure public authorities to minimize their fiscal burden (Afonso, Araújo and Nóbrega, 2013). This is exacerbated by the fact that new valuations of properties have to be approved by the municipal legislatures, giving the wealthy an avenue to undermine the administrative process of tax collection (Carvalho, Jr., 2013). Thus, at least part of the regressivity of the IPTU is due to differences in the *de jure* and *de facto* tax rates.

If properly enforced, the IPTU has the potential to be redistributive and the exact

mechanisms outlined in this manuscript, i.e., elite resistance against higher taxes, are at least partially responsible for its regressivity. In addition to the potential progressivity of the tax itself, government spending at the municipal level primarily benefits the poor. In other words, the marginal benefit of additional spending is higher for the poor than the rich. For example, the most significant share of local budgets is spent on education, with health spending being second. Municipalities primarily finance preschools and primary schools as well as education infrastructure and school lunches (Gadenne, 2017).[1] While not directly redistributive transfers, we contend that spending on these goods is redistributive in nature and has greater benefits to poorer segments of society.

In line with our argument, Gadenne (2017) finds that investments allocated to modernize local tax administrations do increase tax revenue. The additional income is spent on the provision of public goods, with three-quarters of the extra revenue going towards public education. This results in an eight percent increase in locally-funded school infrastructures and six percent more children enrolled in municipal schools (Gadenne, 2017).

Measuring Fiscal Capacity Using the Property Tax

Property taxes are difficult to enforce for both administrative and political reasons (Bahl and Martinez-Vasquez, 2008; Kelly, 2013). According to Kelly (2013), we can decompose total property tax revenue into two parts. First, the total level of potential revenue, which equals the tax rate applied to the total tax base, i.e. *de jure* tax rate above. The second, equally

[1] According to data from the Brazilian Ministry of Finance (National Treasury (DFOFM), 2017), the share of public goods spending that goes to education and health grew from 25% and 11% in 1990 to 34% and 17% in 2000, and 41% and 32% in 2010, respectively.

important, determinant of total revenue is made up of "administration-related variables." These variables are the coverage ratio, i.e., the share of properties captured in the municipality's registry; the valuation ratio, i.e., the ratio of valuation in the taxpayer registry to the market valuation of properties; and the collection ratio, i.e., the percent of levied taxes that are collected. While tax rates and the base are both relevant determinants of the tax revenue collected by the state, the administrative capacity is fundamental for property taxes to raise significant revenue (Kelly, 2013; Bahl and Martinez-Vasquez, 2008).

Calculating IPTU liability (i.e., the valuation) requires several types of information, such as property size, location of the property, property use, front and backyard area, property construction standard, etc. (Carvalho, Jr., 2006). Before valuation, properties must be registered in the municipal cadaster. Carvalho, Jr. (2006) estimates that only 60% of the urban real estate in Brazil is registered. Another important aspect of property tax collection is the frequency of assessment, i.e., how often does the administration update/assess the value of properties? The Brazilian central government recommends evaluating property values every five years, with yearly adjustments. The guidelines do not seem to be regularly followed, however. For example, while Porto Alegre in the 1990s had more regular assessments than other municipalities, the assessed values of residential properties were only 19.2% of their sales prices (De Cesare, 2012).

While it is almost impossible to accurately and reliably measure fiscal capacity, we use realized property tax revenue as a proxy for local fiscal capacity. We assume that given the control variables included in the regression models below, at least some of the variation in the *policy-related variables* are held constant across our cases. For example, we include controls for local GDP, population size, and share of the rural population, which ought to explain

differences in the tax base. We add controls for revenue needs (i.e., transfers from the federal government, oil revenue) and political determinants (left-leaning mayors), which should at least partly account for differences in tax rates.[2] Lastly, we discuss some robustness checks based on smaller samples with more direct measures of administrative capacity.

Kelly (2013, 147) identifies the incompleteness of property registries (cadasters) as the most pressing administrative issue when it comes to property tax collection in developing countries, with a lack of "necessary political will to *collect and enforce the property tax*" (emphasis added) as an additional major hurdle. Anecdotal evidence suggests that municipalities in Brazil find it difficult to increase their administrative capacity. As De Cesare and Ruddock (1999) point out, wherever localities aim to increase the quality of assessment and revenue of the property tax, they are met with strong opposition. Qualitative evidence of tax fraud and incompetence in local government tax collection is easy to find. For example, in 2014, the public prosecutor's office of São Paulo was investigating companies suspected of carrying out a fraud scheme in the city's IPTU collection in partnership with tax collectors (IPTU inspectors). The inspectors calculated the correct tax, but recorded only half the area when visiting buildings. The other half of the tax was paid as a bribe to the inspectors. While the bribe was paid once, the scheme guaranteed a tax bill that was 50% of the *de jure* amount for all subsequent years (Estadão, 2014).

Similarly, a group of employees in the São Paulo City Hall was accused of fraud and irregularities concerning charges of the Service Tax and the IPTU. Members of the group defrauded the IPTU, by making changes to the cadaster, which was estimated to have cost

[2]Unfortunately, complete data on tax rates at the local level are not available.

city hall about half a billion Brazilian reais (approximately 160 million $US in today's value) (G1-Globo, 2013).

Other examples of fraud and local difficulties with tax collection include charges of public servants making improper changes to the collection system (G1-Globo, 2012), fraud schemes in the city of Campinas (collection of less than 10% of property values), and the municipality of Taboão da Serra (Folha de São Paulo, 2011). These tax evasion schemes cost at least R$ 15 million for Campinas (Folha de São Paulo, 1999) and caused a minimum loss of R$10 million to Taboão da Serra, a municipality with more than 250,000 inhabitants (Folha de São Paulo, 2011).

Some reader may question the use of property tax at the local level as the unit of analysis. The majority of taxes are levied at the federal level, which raises the question whether elites would try to undermine local capacity. We believe that the collection of local property taxes is nevertheless highly relevant for this study. First, these taxes, if properly enforced, are likely to be progressive. Based on the theoretical argument, all else equal, elites ought to prefer paying lower property taxes. Additionally, undermining the local property tax administration in the respective municipality is most likely easier and less costly than attempting to do so at the federal level. Thus, the marginal benefit of undermining tax capacity may be highest at the local level. While we lay out a general argument above, we believe that if it holds true, we should find evidence of these processes at the local level. Given the large variation in inequality and tax revenues in municipalities across Brazil, we think these represent an excellent test case for our argument.

Empirical Strategy: Data & Models

To investigate whether high-income earners use low levels of fiscal capacity to limit redistribution and taxation in high-inequality municipalities, we collected data on tax revenues, political, and socioeconomic variables for the years 1990, 2000, and 2010 from different sources. The dependent variable, our proxy for fiscal capacity at the local level, is the property tax revenue collected by municipalities. The measure of revenue collection comes from the Brazilian Ministry of Finance, released by the National Treasury Secretariat, and is made available by the Institute of Applied Economic Research (IPEA, 2016).[3]

Brazil exhibits high geographic variation in both inequality and tax collection. Our preferred measure of income inequality in the municipalities, the Gini coefficient, ranges from 0.28–0.8 in Brazil for 2010. The use of subnational data allows us to hold many variables constant across observations. For example, we do not have to worry about differences in the political system affecting our results.

We include several control variables in the regression model to account for possible confounders and partial out tax rates and tax base. First, we add a control for municipal GDP to account for the fact that higher inequality may be caused by increasing incomes, while more affluent municipalities have a larger tax base, and are more likely to be more efficient at

[3]Based on personal communication with IPEA, some ambiguity about the meaning of zeros in the IPTU revenue data exists. It is possible that some observations with a value of zero are actually missing data, while for other observations the zeros are meaningful values that indicate zero revenue. This issue mostly applies to the panel model. We use the original data in the main text but undertake additional robustness checks in the online Appendix in section F.

revenue collection. We also control for population size. Brazilian municipalities are heterogeneous regarding their size, economic condition, and capacity to tax. Studies have shown that municipal size is positively correlated with property tax revenue (Gomes, Alfinito and Albuquerque, 2013; Avellaneda and Gomes, 2014). Both of these measures were gathered from the Brazilian Institute of Geography and Statistics (IBGE, 2016).

Since municipalities are only allowed to collect property taxes from urban areas, it is pertinent for us to account for differences in urbanization. Hence, we control for the share of the population living in rural areas. We also include a measure of municipal spending on housing and urbanization. The inclusion of this variable is important, as spending on housing and urban development affects real estate evaluations and increases the base for calculating the IPTU tax. A second relevant fiscal variable included in our models is the level of transfers from both the federal and state governments to each of the municipalities (Brollo et al., 2013; Litschig and Morrison, 2013). Data on transfers and housing spending was gathered from the Institute of Applied Economic Research (IPEA, 2016). Additionally, we control for municipal revenue from oil exploration (royalties). Royalty payments made to municipalities in which oil has been discovered and explored increased from R$167 million in 1997 to R$4.7 billion in 2008 (Monteiro and Ferraz, 2012). Royalty payments are associated with an increase in the number of municipal employees (Monteiro and Ferraz, 2012) and municipal revenues (Caselli and Michaels, 2009). Similar to intergovernmental transfers, we expect that royalties from oil exploration undermine local governments' incentives to increase their own revenue capacity and may also affect inequality.

In addition, in our cross-sectional models, we include an indicator variable with a value of 1 if the mayor of the municipality is from a left party, and 0 otherwise. The inclusion

of this variable is an attempt to understand whether left-leaning parties are more likely to raise the fiscal capacity/redistributive taxation and whether they are able to achieve this goal. Given our theoretical argument, we do not expect left-leaning party governance to have a strong effect on *de facto* tax revenue. Additionally, this control may partial out some of the differences due to *de jure* tax rates. Political data were collected from the Superior Electoral Court (TSE do Brasil, 2016), and leftist parties were classified based on surveys and roll-call vote studies of Brazilian legislators (Power and Zucco Jr., 2009, 2012; Samuels and Zucco Jr., 2014; Saiegh, 2015).

We were able to collect these variables for the years 2000, 2010, and approximately 1990. We first estimate cross-sectional models for both 2000 and 2010. We estimate standard ordinary least squares (OLS) regressions for the cross-sectional models, but calculate standard errors clustered by states. The dependent variable (*IPTU revenue*) and the independent variables *housing*, *GDP*, *transfers*, *oil revenue*, and *population* were log transformed to reduce the right-skewness of their distributions.[4]

In addition to the cross-sectional models for two time periods (2000 and 2010), we also estimate a panel model for 1991, 2000, and 2010, in which we include municipal and year fixed effects. Using the unit-specific intercepts, we aim to control for unobserved confounders that do not vary over time or across units.

[4]To avoid creating missing values, prior to taking the log we add 1 to the values of *IPTU*, *housing*, *oil revenue*, and *transfers* variables.

Empirical Analysis: Results and Discussion

Figure 1 illustrates our general findings in the cross-sectional models. The plot displays the coefficient estimates for our cross-sectional model for 2010 with standard errors clustered by state.[5]

Our results consistently lend support to our hypothesis. Particularly, the coefficient for inequality (*Gini*) is estimated to be negative and is statistically significant in all models. Higher inequality is associated with lower property tax revenue, i.e, as inequality rises a municipality's ability to collect IPTU from its citizens decreases. For example, according to the results displayed in Figure 1, holding all covariates at their median value and increasing inequality from the 25th percentile value (0.45) by one standard deviation (to 0.52) is associated with a decrease in logged IPTU revenue from 10.92 to 10.49.

In line with our expectations, the coefficient for GDP is precisely estimated and positive, which indicates that richer municipalities can raise more revenue from property taxes. In contrast, the larger the share of the population living in rural areas, the lower the revenue from the IPTU.

The results for population size are somewhat surprising. Higher population size may be associated with lower revenues. The estimates for intergovernmental *transfers* are also

[5]Table A.1 in the Appendix presents the estimation results for six different models for the 2000 and 2010 data. All models were estimated using OLS. Models 2 and 4 were estimated computing robust standard errors, and Models 3 and 6 were estimated computing standard errors clustered at the state level. We also estimate all models based on data that is multiple imputed using Gaussian copulas (Hoff, 2007). The results are shown in Table A.2 in the Appendix and suport the results presented here.

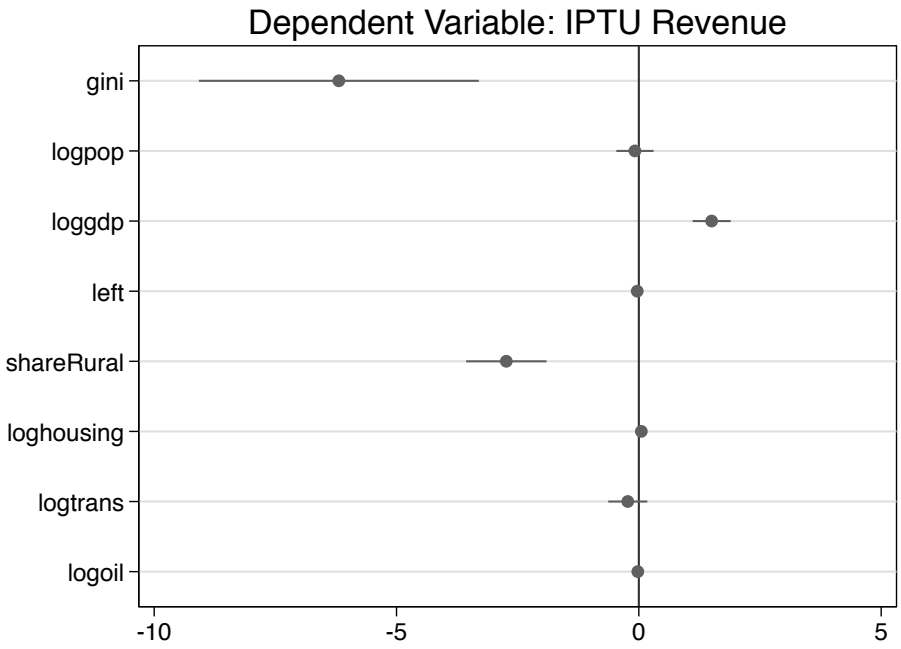


Figure 1: Coefficient Estimates from Model 6 of Table A.1 in Appendix A. Cross-Sectional Model for 2010 with non-imputed data. Standard errors clustered by state. Dependent variable: IPTU revenue in Brazilian reais (logged). The negative and significant estimate for Gini indicates that, as inequality increases the state's ability to raise revenue from citizens decreases substantially.

not precisely estimated in models with clustered standard errors. The results do indicate that municipalities that are more dependent on transfers collect lower revenues from the IPTU. These results are similar to our findings for *oil revenue*. Throughout all models, the coefficient for oil revenue is estimated to be negative, but the precision of the estimates varies across the different models. Also as expected, mayors from left-leaning political parties are not associated with higher revenues: the coefficient for leftist party mayor is very small, inconsistent, and estimated with high uncertainty.[6]

In the Supplementary Online Appendix in section B, we provide additional evidence for

[6]As an additional robustness check, Table A.3 in the Appendix displays the results from four spatial autoregressive models. Overall, the results from the spatial models are consistent with the findings presented above.

the robustness of these results by adding several potentially relevant controls and estimating bivariate models without controls. The results do not change subtantially for any of these specifications. The effect of inequality remains negative and significant when we add controls for voter turnout, competitiveness of the mayoral race, other municipal tax revenues, share of the population vulnerable to poverty[7], share of municipal GDP produced in the industrial sector, number of families that benefit from the cash transfer program (Bolsa Família), or the size of the cash benefits. The estimated effect of inequality is negative and statistically significant in all of these specifications, except when we include total logged cash benefits paid out and cluster standard errors by state. In that particular model, the coefficient on inequality is significant only at the 10% level. Lastly, we can add GDP growth over the previous decade to our cross-sectional models and the results remain substantially the same.

To provide further evidence for the robustness of our results and alleviate concerns about the dependent variable, we also estimate several models with other potential measures of fiscal capacity at the municipal level. For some of these, however, the sample size is reduced significantly. The results are presented in section C in the Appendix. First, we show that the cross-sectional results are robust to calculating our dependent variable as the ratio of IPTU revenue to municipal GDP or as a ratio to total municipal tax revenue. We also provide the results when using revenue from a different local tax source (ITBI, a tax on property transfers) as the dependent variable. The results do not change substantially.

Lastly, we also create a variable measuring the ratio of registered properties for which the property tax was paid to total registered properties (collection rate). These data are collected

[7]Variable is defined as the share of the population with incomes less than R$255.00 a month.

for 1998. While imperfect, we use this measure as an alternative dependent variable for our cross-section of 2000 (the closest year for which we have data). Again, the relationship with inequality is estimated to be negative and significant.

Panel Model Estimation

So far, we have shown that across different municipalities, higher inequality is robustly associated with less municipal revenue collected from property taxation. These findings lend support to our theoretical argument that in higher-inequality districts, wealthy elites undermine the state's ability to collect taxes. The results are robust to including many potential confounders as controls.

Nevertheless, other potential factors may affect both tax capacity and inequality. In this section, we present evidence based on a simple panel model at the municipal level for 1991, 2000, and 2010, with both municipal and year fixed effects.[8] By including both municipal and time fixed effects we can control for unobservables at the municipal level that do not vary over time, as well as shocks in time that do not change across the different municipalities.[9] Given these additional parameters, the results from the three-period panel model can serve as an additional check on the results presented above.

[8]Since several variables are not available for 1990, we use 1991 as our earliest observation. In addition, we could not find data for municipal GDP for the early 1990s. We thus have to rely on a GDP measurement from 1985 in the panel data for 1991.

[9]Since inequality within a municipality may also create incentives to redraw municipal boundaries, we conduct an analysis using a sub-sample based on municipality age. The results, presented in Appendix E, indicate that a possible split of municipalities due to high inequality does not seem to be driving our results.

We specify the following model for the three-period panel data:

$$y_{it} = \alpha_i + \gamma_t + \beta \mathbf{X_{it}} + \delta G_{it} + \epsilon_{it}, \tag{1}$$

where α_i and γ_t are municipality- and year-specific intercepts, $\mathbf{X_{it}}$ is a matrix of time-varying covariates, and β is a vector of the corresponding estimated coefficients. G_{it} is the main variable of interest, the Gini coefficient for municipality i at time t. Based on our theoretical argument, we expect its coefficient δ to be negatively signed. We present the results based on standard errors clustered at the state level.

Figure 2 displays the results from the three-period panel model. Growth in population and transfers over time are associated with higher levels of tax revenue and the 95% confidence intervals do not include zero. The coefficients for GDP, share of the rural population, and logged spending on housing are very close to zero and not significant at conventional levels. Most importantly, the coefficient for inequality is negative, and its 95% confidence interval does not cover zero. An increase in inequality over time is associated with less municipal revenue from property taxes. This finding gives additional credence to the theoretical argument.[10]

As a robustness check, we estimate the same model in a two-period panel for 2000 and 2010.[11] Surprisingly, once we add year fixed effects, the coefficient for inequality is estimated

[10]Some of the municipalities in our sample were created after 1990. We, therefore, subset the data to those municipalities created prior to 1985. The results remain the same if we do not subset.

[11]For the two-period panel model, we subset the data to municipalities created before 2000 (results shown in Table D.1 in the Appendix).

to be positive in the two-period model with controls (2000 and 2010). This suggests that a something changed in high inequality municipalities between 2000 and 2010. It is possible that the introduction of the federal cash benefits program Bolsa Família in 2003 led to these changes, though there is no clear way to test this. Since Bolsa Família was started in 2003, we can not include it as a covariate in the panel models. As we discussed above, however, the results in the cross-section for 2010 are robust even when controlling for Bolsa Família benefits. [12]

As with the cross-sectional model, we estimate the three-period panel model as a bivariate model with unit and year fixed effects. We also add a linear time trend and a quadratic time-trend to the three-period panel model. The results remain the same. Lastly, we estimate the two-period panel model using data on the collection rate (i.e., the ratio of paid to levied taxes) for 180 municipalities. These data were originally collected by Carvalho, Jr. (2017). Our general finding: a significant and negative relationship of inequality with fiscal capacity remains. On average, the greater the inequality, the smaller the IPTU collection rate. The results of these robustness checks are presented in section D of the Appendix.

[12]We thank an anonymous reviewer for alerting us to the possible effects of the Bolsa Família program. Table D.1 in the Appendix also displays the results for both panel models when the data are multiple imputed using Gaussian copulas (Hoff, 2007). The results are mostly unchanged, and in fact, the effect of inequality on property tax revenue is estimated to be stronger.

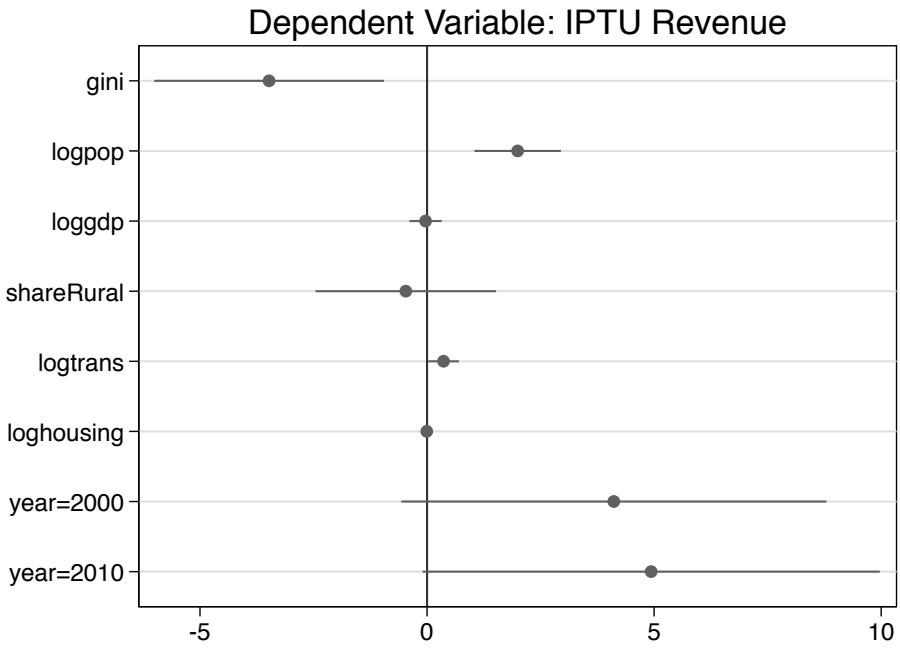


Figure 2: Coefficient Estimates from Model 1 of Table D.1 in Appendix D. Panel Model (1991, 2000, 2010) with year and municipal fixed effects, standard errors clustered at the state level. Dependent variable: IPTU revenue in Brazilian reais (logged). The results are consistent with the cross-sectional model, indicating that increases in inequality are associated with lower capacity to collect taxes.

Selection on Unobservables

In this section, we briefly discuss a sensitivity analysis of the regression results, as suggested by Oster (2017). We estimate how strong selection on unobservables compared to observables would have to be if the effect of inequality is due to bias. Two concepts are required. The first is the "relative degree of selection on observed and unobserved variables" (δ), i.e., how much more important are the variables included in the regression models compared to unobservables. Generally, Oster (2017) suggests considering results to be robust if $\delta > 1$. Secondly, R_{max} is defined as the maximum attainable R^2 for the particular regression, if all relevant variables were included. Of course, the most conservative test is with R_{max} set to

one, the highest possible R^2. Based on empirical evidence using the results of randomized experiments, Oster (2017) suggests that a R_{max} of 1.3 times the R^2 from the relevant regression might be more appropriate. We estimate δ for each of three regression models of interest using the highest possible values of R_{max}, $R_{max} = 1$.

Table 1: Selection on Unobservables

	2000	2010	Panel Model
$R_{max} = 1$	$\delta = 1.92$	$\delta = 2.62$	$\delta = 4.82$

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
Test for 2000 from Cross-Sectional Model 3 of Table A.1 in Appendix A.
Test for 2010 from Cross-Sectional Model 6 of Table A.1 in Appendix A.
Test for Panel Model from Panel Model 1 of Table D.1 in Appendix D.

The relevant values are displayed in Table 1. The results imply that it is unlikely that our results are due to selection on unobservables, as the estimated δ for all three models are above the critical value of 1, even when we use the maximum possible value of one for R_{max}.

Applications to Capacity-Building Program

The empirical analyses and the robustness checks in the previous section have provided evidence in line with our theoretical argument. Nevertheless, questions may remain with regards to our dependent variable and the identification of the theoretical mechanism. In this section, we investigate if inequality levels influenced whether municipal governments applied for grants to improve their tax administration.

In 1997, the Brazilian federal government initiated the Modernization Program of the Tax Administration (PMAT), with the goal of improving municipalities' tax administration. The foremost objective of the program was to increase municipalities' revenues by improving tax registration and collection processes, modernizing taxpayer services and enhanc-

ing municipalities' fiscal responsibility and capacity (Afonso et al., 1998; Guarneri, 2002). The program focuses on the modernization of information technology, computer equipment, training of human resources, specialized technical services, and the physical infrastructure of municipalities' public administration (Guarneri, 2002; Corrêa, 2009).

The financial funds of the program are provided to the municipalities by the Brazilian Development Bank (BNDES) through credit lines opened by BNDES financial partner institutions. The current financing amount limit is either a maximum of R$60 million per municipality or R$36 per capita (the financing accepted is based on the lower value of these criteria) (Corrêa, 2009).

Gadenne (2017) has taken advantage of the program to show that higher levels of fiscal capacity – and, ergo, local tax revenue – cause positive changes in municipal education infrastructure. If our argument is correct, we should find that municipalities with higher levels of inequality are less likely to apply to the program (even though their revenues are lower). We, therefore, estimate the probability that a municipality joins the PMAT program until 2010 as a function of its inequality level (Gini coefficient) and controls included in our previous models (all measured in 1991). We also include municipal revenue raised from IPTU collection as a control. According to our argument, the elites' constraint on the state should be stronger under higher levels of inequality. Thus, we expect that the greater the municipality's inequality, the lower the likelihood it will apply to PMAT.

As shown in Table 2, the results support this expectation. Across linear probability, logit models, and when we cluster standard errors by state (Models 2 and 4), the coefficient on inequality is negative and precisely estimated. Greater inequality appears to be associated with a lower likelihood of application to PMAT, a finding that is also reflected in the work

Table 2: Municipal Applications to the Capacity-Building Program (PMAT)

	Dependent variable: PMAT Application			
	(Model 1) OLS	(Model 2) OLS	(Model 3) Logit	(Model 4) Logit
Gini	-0.242***	-0.242***	-2.786**	-2.786**
	(0.050)	(0.070)	(1.166)	(1.230)

Notes: Dependent variable: Binary variable PMAT (1 = municipality applied to PMAT, 0 = municipality didn't apply to PMAT).
All four models include controls for IPTU revenue (logged), population (logged), GDP (logged), rural share, transfers (logged). Full Table is displayed in Table C.6 in Appendix C. Model 1 and Model 3 with robust standard errors. Model 2 and Model 4 with standard errors clustered by state.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

by Gadenne (2017). For space reasons we omit the control variables from the Table, but full results are presented in Appendix C.

These results are consistent with our expectation that more unequal municipalities will have a lower capacity to collect taxes. Although PMAT currently reaches all regions of Brazil, the program is heavily concentrated in the less unequal south and southeast regions of the country (Corrêa, 2009; Grin, 2014). While the south and southeast have received 73.4% of all established contracts in 2009, municipalities in the north and northeast regions of Brazil (more unequal) account for only 3.8% of the contracts (Grin, 2014). After 13 years, the fact that only 369 municipalities (6.63% of the Brazilian municipalities in 2011) participate in the PMAT reveals a low acceptance of the program among municipal governments in general (Grin, 2014).

Conclusion

Some of the most famous formal models in political economy make the prediction that taxation ought to increase with inequality in democracies (Romer, 1975; Meltzer and Richard, 1981). Yet in many cases, scholars do not find the stated relationship to be true. We argue that this may be explained by wealthy elites undermining the state's ability to collect taxes in highly unequal democracies, especially when the state's capacity is already limited.

To investigate this proposition, we use data on property tax revenue, inequality, and other economic variables from over $5,500$ municipalities in Brazil. Using cross-sectional, as well as panel models, and undertaking a variety of robustness checks, we show that municipalities with higher levels of inequality have lower levels of fiscal capacity/raise less revenue from the local property tax. The evidence is consistent with our theoretical argument. We do acknowledge, however, that we can not yet identify the exact causal mechanism and that other potential explanations are possible. On the other hand, our results are strengthened by the fact that municipalities with higher inequality were also significantly less likely to apply for federal programs that could aid their tax collection efforts.

If wealthy elites do actively undermine tax administration in highly unequal societies, this should have consequences for how we view democratic policy-making and the delivery of public goods. A democratic political system is no panacea: even if the will of the voters may be translated into policies, the state is not always able to properly enforce the policy choices made. On the other hand, it may be that as democracies stabilize and become further removed from their authoritarian origins, they can slowly diminish the influence of elites and increase capacity. This possibility should be further investigated in future cross-national

work. Similarly, as we argue in the paper, we think that our findings are generalizable to national level politics. Yet, subsequent studies ought to investigate whether the lack of evidence in line with the Meltzer and Richard (1981) model cross-nationally can be explained by the theoretical argument made here.

Lastly, future research should further consider the exact mechanisms by which economic elites can undermine the state's capacity to collect revenues and enforce policies. Better understanding of these processes will help us gain a better grasp of the difficulties of policy-making in (young) democracies and thus the threats to their existence. Additionally, further research ought to investigate how limited state capacity can influence the nexus between voters and politicians. For example, low levels of capacity may impact voters' preferred policies and evaluation of politicians, especially when it comes to taxation and public goods.

Acknowledgements

We thank Pedro de Carvalho Júnior and Eduardo Grin for their valuable help with the data. We also thank Pablo Beramendi, Jose Cheibub, Eunyoung Ha, Jan Pierskalla, and Ken Scheve for valuable feedback. We also thank three anonymous reviewers and the editorial team at the Journal of Politics for their feedback and help on this paper. Previous versions of this manuscript were presented at the 112th Annual Meeting of the American Political Science Association, Philadelphia, PA and the workshop in Political Economy and Political Violence at Texas A&M University.

References

Acemoglu, Daron, Andrea Vindigni and Davide Ticchi. 2011. "Emergence and Presistence of Inefficient States." *Journal of the European Economic Association* 9(2):122–208.

Acemoglu, Daron and James Robinson. 2008. "Persistence of Power, Elites, and Institutions." *American Economic Review* 98(1):267–293.

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo and James A. Robinson. 2015. Democracy, Redistribution, and Inequality. In *Handbook of Income Distribution*, ed. Anthony B. Atkinson and François Bourguignon. Vol. 2 Amsterdam: Elsevier chapter 21, pp. 1885–1966.

Afonso, José Roberto, Cristóvão Correia, Erika Amorim Araújo, Júlio Ramundo, Maurício David and Rômulo Santos. 1998. "Municípios, Arrecadação e Administração Tributária: Quebrando Tabus." *Revista do BNDES* 10:1–36.

Afonso, José Roberto and Erika Amorim Araújo. 2006. Local Government Organization and Finance: Brazil. In *Local Governance in Developing Countries*, ed. Anwar Shah. Washington: World Bank pp. 381–418.

Afonso, José Roberto, Erika Amorim Araújo and Marcos Antonio Nóbrega. 2013. *IPTU no Brasil: Um Diagnóstico Abrangente*. Vol. 4 São Paulo: Instituto Brasiliense de Direito Publico IDP Ltda; FGV Projetos.

Albertus, Michael and Victor Menaldo. 2014. "Gaming Democracy: Elite Dominance dur-

ing Transition and the Prospects for Redistribution." *British Journal of Political Science* 44(3):575–603.

Alesina, Alberto and Edward L. Glaeser. 2004. *Fighting Poverty in the US and Europe: A World of Difference*. Oxford, UK: Oxford University Press.

Andrade, Luis Aureliano de. 2007. O Município na Política Brasileira: Revisitando Coronelismo, enxada e voto. In *Sistema Político Brasileiro: Uma Introdução*, ed. Lúcia Avelar and Antônio Octávio Cintra. São Paulo: Editora Unesp pp. 243–260.

Ardanaz, Martin and Carlos Scartascini. 2013. "Inequality and Personal Income Taxation: The Origins and Effects of Legislative Malapportionment." *Comparative Political Studies* 46(12):1636—1663.

Arretche, Marta. 2004. "Federalismo e políticas sociais no Brasil: problemas de coordenação e autonomia." *São Paulo em Perspectiva* 18(2):12–26.

Avellaneda, Claudia and Ricardo Gomes. 2014. "Is Small Beautiful? Testing the Direct and Nonlinear Effects of Size of Municipal Performance." *Public Administration Review* 75(1):137–149.

Bahl, Roy and Jorge Martinez-Vasquez. 2008. The Determinants of Revenue Performance. In *Making the property tax work: Experiences in developing and transitional countries*, ed. Roy Bahl, Jorge Martinez-Vazquez and Joan Youngman. Cambridge: Lincoln Institute of Land Policy pp. 35–57.

Baiocchi, Gianpaolo. 2006. *Decentralization and Local Governance in Developing Countries: a comparative perspective*. Cambridge: MIT Press.

Barros, Ricardo, Ricardo Henriques and Rosane Mendonça. 2000. "Desigualdade e Pobreza no Brasil: Retrato de uma Estabilidade Inaceitável." *Revista Brasileira de Ciências Sociais* 15(42):123–142.

Benabou, Roland. 1996. Inequality and Growth. In *NBER Macroeconomics Annual 1996*, ed. Ben Bernake and Julio J. Rotemberg. National Bureau of Economic Research.

Besley, Timothy and Torsten Persson. 2009. "The Origins of State Capacity: Property Rights, Taxation, and Politics." *American Economic Review* 99(4):1218–1244.

Besley, Timothy and Torsten Persson. 2011. *Pillars of Prosperity: The Political Economics of Development Clusters*. Princeton: Princeton University Press.

Bird, Richard M. and Eric M. Zolt. 2008. "Tax Policy in Emerging Countries." *Environment and Planning C: Government and Policy* 26:73–86.

Brollo, Fernanda, Tommaso Nannicini, Roberto Perotti and Guido Tabellini. 2013. "The Political Resource Curse." *American Economic Review* 103(5):1759–1796.

Carvalho, Jr., Pedro H. B. 2006. "O IPTU no Brasil: Progressividade, Arrecadação e Aspectos Extra-Fiscais." IPEA Discussion Paper 1251.

Carvalho, Jr., Pedro H. B. 2013. "Property Tax Performance in Rio de Janeiro." *Journal of Property Tax Assessment & Administration* 10(4):19–36.

Carvalho, Jr., Pedro H. B. 2015. "Distributive Aspects of Real Estate Property and its Taxation Among Brazilian Families." IPEA Discussion Paper 184. `http://repositorio.`

ipea.gov.br/bitstream/11058/4956/1/DiscussionPaper_184.pdf (Accessed September 18, 2017).

Carvalho, Jr., Pedro H. B. 2017. Property Tax Performance and Potential in Brazil PhD thesis Faculty of Economic and Management Sciences at the University of Pretoria.

Caselli, Francesco and Guy Michaels. 2009. "Do Oil Windfalls Improve Living Standards? Evidence from Brazil." https://www.nber.org/papers/w15550.pdf (Accessed March 2017).

Corrêa, Letícia. 2009. "Atuação do BNDES nos investimentos na gestão do setor público: estudo do caso Pmat–Santo André (SP)." *BNDES Setorial: Setor Público* 30:211–236.

De Cesare, Claudia. 2005. O Cadastro como Instrumento de Política Fiscal. In *Cadastro Multifinalitário como Instrumento de Política Fiscal e Urbana*, ed. Diego Erba, Fabrício Oliveira and Pedro Lima Jr. Brasília: Ministério das Cidades.

De Cesare, Claudia. 2012. "Improving the Performance of Property Tax in Latin America." Cambridge: Lincoln Institute of Land Policy (Working Paper).
URL: *http://www.lincolninst.edu/sites/default/files/pubfiles/improving-performance-property-tax-latin-america-full_0.pdf (Accessed March 2017)*

De Cesare, Claudia and Les Ruddock. 1999. The Property Tax System in Brazil. In *Property Tax: An International Comparative Review*, ed. William McCluskey. United Kingdom: Ashgate Publishing Ltd chapter 13, pp. 266–282.

De Cesare, Claudia and Martim Smolka. 2004. "Diagnóstico sobre o IPTU." Cambridge: Lincoln Institute of Land Policy (Working Paper).

Estadão. 2014. "MP investiga 84 empresas beneficiárias de fraude no IPTU em SP.". http://sao-paulo.estadao.com.br/noticias/geral,mp-investiga-84-empresas-beneficiarias-de-fraude-no-iptu-em-sp,1125251?success=true (Accessed January 31, 2017).

Folha de São Paulo. 1999. "Sindicato pede abertura de inquérito.". http://www1.folha.uol.com.br/fsp/campinas/cm25029913.htm (Accessed January 31, 2017).

Folha de São Paulo. 2011. "Três vereadores são presos em Taboão.". http://www1.folha.uol.com.br/fsp/cotidian/ff0405201118.htm (Accessed January 31, 2017).

G1-Globo. 2012. "Polícia prende 13 suspeitos de fraude na arrecadação do IPTU em Cuiabá.". http://g1.globo.com/mato-grosso/noticia/2012/11/policia-prende-13-suspeitos-de-fraude-na-arrecadacao-do-iptu-em-cuiaba.html (Accessed January 31, 2017).

G1-Globo. 2013. "Escutas indicam que fiscais também fraudaram o IPTU em São Paulo.". http://g1.globo.com/sao-paulo/noticia/2013/11/escutas-mostram-que-fiscais-tambem-fraudaram-o-iptu.html (Accessed January 31, 2017).

Gadenne, Lucie. 2017. "Tax Me, But Spend Wisely? Sources of Public Finance and Government Accountability." *American Economic Journal: Applied Economics* 9(1):274–314.

Gomes, Ricardo, Solange Alfinito and Pedro Albuquerque. 2013. "Analyzing Local Government Financial Performance: Evidence from Brazilian Municipalities 2005-2008." *RAC - Revista de Administração Contemporânea* 17(6):704–719.

Gordon, Roger and Wei Li. 2009. "Tax Structures in Developing Countries: Many Puzzles and a Possible Explanation." *Journal of Public Economics* 93(7-8):855–866.

Grin, José Eduardo. 2014. "Trajetória e avaliação dos programas federais brasileiros voltados a promover a eficiência administrativa e fiscal dos municípios." *Revista de Administração Pública* 48(2):459–480.

Guarneri, Lucimar. 2002. "Modernização da gestão pública: uma avaliação de experiências inovadoras." *BNDES Social* 4:9–103.

Hoff, Peter D. 2007. "Extending the Rank Likelihood for Semiparametric Copula Estimation." *Annals of Applied Statistics* 1(1):265–283.

IBGE. 2016. "Brazilian Institute of Geography and Statistics." `http://www.ibge.gov.br/` (Accessed January 2016).

IPEA. 2016. "Institute of Applied Economic Research." `http://www.ipeadata.gov.br` (Accessed January 2017).

Iversen, Torben and David Soskice. 2006. "Electoral Institutions and the Politics of Coalitions: Why Some Democracies Redistribute More Than Others." *American Political Science Review* 100(2):165–182.

Kelly, Roy. 2013. Property Tax Collection and Enforcement. In *A Primer on Property Tax:*

Administration and Policy, ed. Williim J. McCluskey, Gary C. Cornia and Lawrence C. Walters. West Sussex, UK: Blackwell Publishing Ltd chapter 6, pp. 141–171.

Kenworthy, Lane and Jonas Pontussen. 2005. "Rising Inequality and the Politics of Redistribution in Affluent Countries." *Perspectives on Politics* 3(3):449–471.

Limongi, Fernando, José Antonio Cheibub and Argelina Figueiredo. 2015. Participação Política no Brasil. In *Trajetórias da Desigualdade: Como o Brasil Mudou nos Últimos Cinquenta Anos*, ed. Marta Arretche. São Paulo: Editora Unesp/CEM chapter 1, pp. 23–50.

Litschig, Stephan and Kevin Morrison. 2013. "The Impact of Intergovernmental Transfers on Education Outcomes and Poverty Reduction." *American Economic Journal: Applied Economics* 5(4):206–240.

Meltzer, Allan H. and Scott F. Richard. 1981. "A Rational Theory of the Size of Government." *The Journal of Political Economy* 89(5):914–927.

Moene, Karl Ove and Immanuel Wallerstein. 2001. "Inequality, Social Insurance, and Redistribution." *American Politial Science Review* 95(4):859–874.

Monteiro, Joana and Claudio Ferraz. 2012. "Does oil make leaders unaccountable? Evidence from Brazil's offshore oil boom." `http://www.parisschoolofeconomics.eu/docs/ydepot/semin/texte1213/CLA2012DOE.pdf` (Accessed March 2017).

National Treasury (DFOFM). 2017. "Municipal expenditure." Brazilian Ministry of Finance.

Nickson, R. Andrew. 1995. *Local Government in Latin America*. Boulder: Lynne Rienner Publishers.

Nicolau, Jairo. 2012. *Eleições no Brasil: do Império aos Dias Atuais*. Rio de Janeiro: Zahar.

Oliven, Ruben, Marcelo Ridenti and Gildo Marçal Brandão, eds. 2008. *A Constituição de 1988 na Vida Brasileira*. São Paulo: Editora Hucitec.

Oster, Emily. 2017. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* pp. 1–18.

Perotti, Roberto. 1996. "Inequality, Redistribution and Growth: What the Data Say." *Journal of Economic Growth* 1(2):149–187.

Persson, Torsten and Guido Tabellini. 2003. *Economic Effects of Constitutions*. Cambridge, Mass.: Cambridge University Press.

Power, Timothy and Cesar Zucco Jr. 2009. "Estimating Ideology of Brazilian Legislative Parties, 1990-2005: A Research Communication." *Latin American Research Review* 44(1):218–246.

Power, Timothy and Cesar Zucco Jr. 2012. "Elite Preferences in a Consolidating Democracy: The Brazilian Legislative Surveys, 1990-2009." *Latin American Politics and Society* 54(4):1–27.

Rodríguez, Alfredo and Fabio Velásquez, eds. 1995. *Municípios y Servícios Públicos: Gobiernos Locales en Ciudades Intermédias de América Latina*. Santiago: Ediciones Sur.

Romer, Thomas. 1975. "Individual welfare, majority voting, and the properties of a linear income tax." *Journal of Public Economics* 4(2):163–185.

Saiegh, Sebastián. 2015. "Using Joint Scaling Methods to Study Ideology and Representation: Evidence from Latin America." *Political Analysis* 23(3):363–384.

Samuels, David. 2004. The Political Logic of Decentralization in Brazil. In *Decentralization and Democracy in Latin America*, ed. Alfred Montero and David Samuels. Indiana: University of Notre Dame Press pp. 67–93.

Samuels, David and Cesar Zucco Jr. 2014. "The Power of Partisanship in Brazil: Evidence from Survey Experiments." *American Journal of Political Science* 58(1):212–225.

Scheve, Ken and David Stasavage. 2006. "Religion and Preferences for Social Insurance." *Quarterly Journal of Political Science* 1(3):255–286.

Smolka, Martim and Fernanda Furtado. 1996. "Argumentos para a reabilitação do IPTU e do ITBI como instrumentos de intervenção urbana (progressista)." *Revista Espaço & Debates* 39(16):87–103.

Souza, Pedro and Marcelo Medeiros. 2015. "Top Income Shares and Inequality in Brazil, 1928-2012." *Journal of the Brazilian Sociological Society* 1(1):119–132.

TSE do Brasil. 2016. "Tribunal Superior Eleitoral do Brasil." `http://www.tse.jus.br/` (Accessed January 2017).

Biographical Statements

Florian M. Hollenbach is an assistant professor in the Department of Political Science at Texas A&M University, College Station, TX, 77843-4348.

Thiago Silva is a Ph.D. candidate in the Department of Political Science at Texas A&M University, College Station, TX, 77843-4348.

Supplementary Online Appendix: Fiscal Capacity and Inequality: Evidence from Brazilian Municipalities

Florian M. Hollenbach & Thiago Silva

July 17, 2018

A Appendix: Additional Models

Table A.1: Inequality and Fiscal Capacity in Brazilian Municipalities (Cross-Sectional Models for 2000 and 2010) – Non-imputed Data

	Dependent variable: IPTU Revenue (log)					
	Model 1 2000	Model 2 2000 (robust)	Model 3 2000 (cluster)	Model 4 2010	Model 5 2010 (robust)	Model 6 2010 (cluster)
Gini	-3.657***	-3.657***	-3.657***	-6.190***	-6.190***	-6.190***
	(0.579)	(0.619)	(0.922)	(0.493)	(0.624)	(1.396)
Population (log)	-0.994***	-0.994***	-0.994***	-0.080	-0.080	-0.080
	(0.087)	(0.105)	(0.349)	(0.075)	(0.081)	(0.185)
GDP (log)	2.230***	2.230***	2.230***	1.503***	1.503***	1.503***
	(0.074)	(0.091)	(0.226)	(0.059)	(0.068)	(0.190)
Left Party	-0.100	-0.100	-0.100	-0.033	-0.033	-0.033
	(0.087)	(0.086)	(0.122)	(0.058)	(0.059)	(0.059)
Rural Share	-2.823***	-2.823***	-2.823***	-2.734***	-2.734***	-2.734***
	(0.197)	(0.209)	(0.488)	(0.158)	(0.182)	(0.401)
Housing and Urbanization (log)	0.008	0.008	0.008	0.052***	0.052**	0.052*
	(0.017)	(0.018)	(0.027)	(0.015)	(0.023)	(0.027)
Transfers (log)	-0.071	-0.071	-0.071	-0.228*	-0.228*	-0.228
	(0.140)	(0.181)	(0.365)	(0.123)	(0.134)	(0.195)
Oil Revenue (log)	-0.033***	-0.033***	-0.033	-0.020***	-0.020***	-0.020
	(0.011)	(0.012)	(0.031)	(0.007)	(0.008)	(0.018)
Constant	-0.679	-0.679	-0.679	2.391**	2.391*	2.391
	(1.250)	(1.532)	(2.883)	(1.162)	(1.226)	(1.864)
N	4845	4845	4845	4269	4269	4269
R^2	0.507	0.507	0.507	0.641	0.641	0.641

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
Model 2 and Model 5 with robust standard errors.
Model 3 and Model 6 with standard errors clustered by state.
The negative and significant estimates for Gini in all models indicate that, as inequality increases, the state's ability to raise revenue from citizens decreases substantially.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.2: Inequality and Fiscal Capacity in Brazilian Municipalities (Cross-Sectional Models for 2000 and 2010) – Imputed data

	Dependent variable: IPTU Revenue (log)					
	Model 7 2000	Model 8 2000 (robust)	Model 9 2000 (cluster)	Model 10 2010	Model 11 2010 (robust)	Model 12 2010 (cluster)
Gini	-3.736***	-3.736***	-3.736***	-6.491***	-6.491***	-6.491***
	(0.584)	(0.615)	(0.894)	(0.519)	(0.639)	(1.388)
Population (log)	-1.117***	-1.117***	-1.117***	-0.366***	-0.366***	-0.366*
	(0.076)	(0.084)	(0.278)	(0.068)	(0.079)	(0.189)
GDP (log)	2.070***	2.070***	2.070***	1.327***	1.327***	1.327***
	(0.063)	(0.073)	(0.220)	(0.055)	(0.063)	(0.183)
Left Party	-0.097	-0.097	-0.097	-0.055	-0.055	-0.055
	(0.088)	(0.087)	(0.122)	(0.062)	(0.063)	(0.073)
Rural Share	-2.876***	-2.876***	-2.876***	-2.912***	-2.912***	-2.912***
	(0.197)	(0.208)	(0.475)	(0.168)	(0.188)	(0.386)
Housing and Urbanization (log)	0.011	0.011	0.011	0.061***	0.061**	0.061**
	(0.017)	(0.019)	(0.028)	(0.018)	(0.024)	(0.028)
Transfers (log)	0.345***	0.345***	0.345***	0.396***	0.396***	0.396***
	(0.073)	(0.080)	(0.086)	(0.083)	(0.102)	(0.099)
Oil Revenue (log)	-0.032***	-0.032***	-0.032	-0.023***	-0.023***	-0.023
	(0.011)	(0.012)	(0.029)	(0.007)	(0.008)	(0.018)
Constant	-4.193***	-4.193***	-4.193***	-3.383***	-3.383***	-3.383***
	(0.749)	(0.790)	(0.993)	(0.830)	(0.996)	(1.088)
N	5114	5114	5114	4580	4580	4580
R^2	.	.	.	.	.	.

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
Model 8 and Model 11 with robust standard errors.
Model 9 and Model 12 with standard errors clustered by state.
The results are consistent with the models using non-imputed data: The negative and significant estimates for Gini in all models indicate that as inequality increases, the state's ability to raise revenue from citizens decreases substantially.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.3: Results from Spatial Autoregressive Models

	Dependent variable: IPTU (log)			
	Model 1	Model 2	Model 3	Model 4
	2000 Binary	2000 Row-standardized	2010 Binary	2010 Row-standardized
Gini	-2.859^{***}	-1.437^{***}	-5.304^{***}	-3.446^{***}
	(0.563)	(0.551)	(0.488)	(0.474)
Population (log)	-0.805^{***}	-0.277^{***}	-0.077	0.255^{***}
	(0.083)	(0.084)	(0.073)	(0.071)
GPD (log)	1.844^{***}	1.535^{***}	1.358^{***}	1.094^{***}
	(0.074)	(0.074)	(0.058)	(0.058)
Left Party	-0.049	-0.116	-0.038	-0.070
	(0.084)	(0.082)	(0.057)	(0.055)
Rural Share	-2.770^{***}	-2.426^{***}	-2.776^{***}	-2.546^{***}
	(0.170)	(0.166)	(0.155)	(0.151)
Housing and Urbanization (log)	0.001	0.003	0.047^{***}	0.046^{***}
	(0.016)	(0.016)	(0.015)	(0.014)
Transfers (log)	-0.053	-0.103	-0.198^{*}	-0.251^{**}
	(0.136)	(0.133)	(0.120)	(0.116)
Oil Revenue (log)	-0.019^{*}	-0.036^{***}	-0.013^{**}	-0.028^{***}
	(0.011)	(0.011)	(0.006)	(0.006)
Intercept	-0.438	-3.961^{***}	2.281^{**}	-0.368
	(1.208)	(1.176)	(1.139)	(1.099)
N	4838	4838	4261	4261
Log-Likelihood	$-11{,}220.330$	$-11{,}124.920$	$-8{,}278.976$	$-8{,}151.819$
σ^2	6.027	5.693	2.849	2.642
Akaike Inf. Crit.	22,462.660	22,271.840	16,579.950	16,325.640
Wald Test (df = 1)	239.216^{***}	479.604^{***}	132.134^{***}	416.207^{***}
LR Test (df = 1)	238.898^{***}	429.717^{***}	131.314^{***}	385.629^{***}

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
This table shows the results from four spatial autoregressive models with neighbors based on contiguous boundaries between the municipalities, using 2000 and 2010 cross-sectional data. The results in Model 1 and Model 3 are based on a binary neighbor matrix, while the results in Model 2 and Model 4 are based on a row-standardized weights matrix. The results are in line with our findings: the coefficients for inequality (*Gini*) are still substantively meaningful, negative, and precisely estimated.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table A.4: Bivariate Cross-Sectional Models: Benchmark to Compare the Results

	Model 1 2000	Model 2 2000 (robust)	Model 3 2000 (cluster)	Model 4 2010	Model 5 2010 (robust)	Model 6 2010 (cluster)
Gini	-5.402***	-5.402***	-5.402**	-7.468***	-7.468***	-7.468***
	(0.745)	(0.735)	(2.417)	(0.594)	(0.623)	(2.377)
Constant	12.085***	12.085***	12.085***	14.607***	14.607***	14.607***
	(0.410)	(0.399)	(1.533)	(0.295)	(0.297)	(1.160)
N	5304	5304	5304	5211	5211	5211
R^2	0.010	0.010	0.010	0.029	0.029	0.029

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).

Model 2 and Model 5 with robust standard errors.

Model 3 and Model 6 with standard errors clustered by state.

To increase confidence in the results, we present OLS estimations without any controls, as a benchmark to compare the results. The results are consistent with our previous models including controls.

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

B Appendix: Additional Controls

Table B.1: Original Cross-Sectional Models Including the Control Variable "Turnout"

	Dependent variable: IPTU Revenue (log)			
	Model 1 2000 (robust)	Model 2 2000 (cluster)	Model 3 2010 (robust)	Model 4 2010 (cluster)
Gini	-3.202***	-3.202***	-6.019***	-6.019***
	(0.627)	(0.851)	(0.632)	(1.375)
Turnout	3.343***	3.343***	1.316**	1.316*
	(0.789)	(1.067)	(0.566)	(0.752)
Population (log)	-0.833***	-0.833**	-0.014	-0.014
	(0.109)	(0.339)	(0.084)	(0.187)
GDP (log)	2.162***	2.162***	1.497***	1.497***
	(0.093)	(0.223)	(0.068)	(0.186)
Left Party	-0.109	-0.109	-0.025	-0.025
	(0.086)	(0.118)	(0.059)	(0.060)
Rural Share	-2.807***	-2.807***	-2.689***	-2.689***
	(0.207)	(0.467)	(0.181)	(0.393)
Housing and Urbanization (log)	0.006	0.006	0.052**	0.052*
	(0.018)	(0.026)	(0.023)	(0.028)
Transfers (log)	-0.081	-0.081	-0.272**	-0.272
	(0.180)	(0.360)	(0.135)	(0.200)
Oil Revenue (log)	-0.035***	-0.035	-0.020***	-0.020
	(0.012)	(0.031)	(0.008)	(0.019)
Constant	-4.466**	-4.466	1.322	1.322
	(1.769)	(3.146)	(1.296)	(1.598)
N	4844	4844	4250	4250
R^2	0.509	0.509	0.642	0.642

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
$Turnout = \frac{\text{total number of voters in the municipal election}}{\text{total number of the electorate in the municipal election}}$
Model 1 and Model 3 with robust standard errors.
Model 2 and Model 4 with standard errors clustered by state.
The results from models including the independent variable *turnout* are consistent with our previous models: more unequal municipalities have a lower capacity to collect taxes.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.2: Electoral Competition (Cross-Sectional Model for 2010)

	Model 1	Model 2 (robust)	Model 3 (cluster)
Gini	-6.305***	-6.305***	-6.305***
	(0.509)	(0.645)	(1.410)
Electoral Competition	0.310	0.310	0.310
	(0.203)	(0.192)	(0.191)
Population (log)	-0.049	-0.049	-0.049
	(0.078)	(0.083)	(0.185)
GDP (log)	1.533***	1.533***	1.533***
	(0.061)	(0.070)	(0.195)
Left Party	-0.030	-0.030	-0.030
	(0.059)	(0.060)	(0.061)
Rural Share	-2.700***	-2.700***	-2.700***
	(0.163)	(0.188)	(0.392)
Housing and Urbanization (log)	0.049***	0.049**	0.049*
	(0.016)	(0.024)	(0.028)
Transfers (log)	-0.294**	-0.294**	-0.294
	(0.128)	(0.139)	(0.195)
Oil Revenue (log)	-0.017***	-0.017**	-0.017
	(0.007)	(0.008)	(0.019)
Constant	2.916**	2.916**	2.916
	(1.201)	(1.261)	(1.834)
N	4074	4074	4074
R^2	0.642	0.642	0.642

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).

Electoral Competition $= \frac{\text{elected candidate's vote share - runner up candidate's vote share}}{\text{total number of the electorate in the municipal election}}$

Model 2 with robust standard errors, and Model 3 with standard errors clustered by state.

The results for the models that include the independent variable *electoral competition* are consistent with our previous results: more unequal municipalities have a lower capacity to collect taxes.

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.3: Vulnerability to Poverty (%), Cross-Section and Panel Models

	Model 1 2000 (robust)	Model 2 2000 (cluster)	Model 3 2010 (robust)	Model 4 2010 (cluster)	Model 5 1991-2000-2010 (FE & cluster)
Gini	-2.982***	-2.982***	-5.302***	-5.302***	-3.142**
	(0.617)	(0.883)	(0.631)	(1.358)	(1.167)
Vulnerability to Poverty (%)	-0.021***	-0.021***	-0.015***	-0.015***	-0.014***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)
Population (log)	-0.751***	-0.751**	0.069	0.069	2.024***
	(0.106)	(0.337)	(0.079)	(0.173)	(0.464)
GDP (log)	1.950***	1.950***	1.304***	1.304***	-0.043
	(0.096)	(0.223)	(0.070)	(0.169)	(0.171)
Left Party	-0.103	-0.103	-0.006	-0.006	
	(0.086)	(0.119)	(0.058)	(0.058)	
Rural Share	-2.596***	-2.596***	-2.479***	-2.479***	-0.412
	(0.205)	(0.424)	(0.181)	(0.336)	(1.004)
Housing and Urbanization (log)	0.014	0.014	0.052**	0.052*	-0.005
	(0.018)	(0.026)	(0.023)	(0.027)	(0.015)
Transfers (log)	-0.029	-0.029	-0.162	-0.162	0.371**
	(0.178)	(0.352)	(0.131)	(0.187)	(0.161)
Oil Revenue (log)	-0.024**	-0.024	-0.017**	-0.017	
	(0.012)	(0.030)	(0.008)	(0.017)	
2000					3.878*
					(2.213)
2010					4.440*
					(2.349)
Constant	0.143	0.143	2.171*	2.171	-16.439***
	(1.508)	(2.723)	(1.195)	(1.790)	(3.223)
N	4845	4845	4269	4269	8138
R^2	0.518	0.518	0.650	0.650	0.878

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).

Vulnerability to Poverty (%) = The proportion of individuals with a per capita household income equals to or less than R$255.00 per month, in Brazilian reais as of August 2010, which is equivalent to half of the average minimum salary in Brazil as of that date. The sample of individuals is limited to those who live in permanent private households.

Model 1 and Model 3 cross-sectional models with robust standard errors. Model 2 and Model 4 cross-sectional models with standard errors clustered by state. Model 5 Panel model (1991-2000-2010) with Year and Municipal fixed-effects and standard errors clustered by state.

The results for models including the independent variable *vulnerability to poverty (%)* are consistent with previous results: more unequal municipalities have a lower capacity to collect taxes.

Standard errors in parentheses. Two-tailed test. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.4: Bolsa Família Cash Transfer Program, Cross-Sectional Models for 2010

	Model 1	Model 2 (robust)	Model 3 (cluster)	Model 4	Model 5 (robust)	Model 6 (cluster)
Gini	-3.134***	-3.134***	-3.134*	-2.521***	-2.521***	-2.521*
	(0.500)	(0.683)	(1.587)	(0.501)	(0.670)	(1.463)
Number of Families (log)	-1.277***	-1.277***	-1.277***			
	(0.067)	(0.069)	(0.200)			
Cash Benefits Amount (log)				-1.221***	-1.221***	-1.221***
				(0.058)	(0.057)	(0.159)
Population (log)	1.727***	1.727***	1.727***	1.732***	1.732***	1.732***
	(0.118)	(0.116)	(0.373)	(0.112)	(0.104)	(0.322)
GDP (log)	0.753***	0.753***	0.753***	0.686***	0.686***	0.686***
	(0.068)	(0.073)	(0.204)	(0.068)	(0.072)	(0.194)
Left Party	-0.011	-0.011	-0.011	-0.018	-0.018	-0.018
	(0.055)	(0.057)	(0.053)	(0.055)	(0.056)	(0.053)
Rural Share	-2.356***	-2.356***	-2.356***	-2.242***	-2.242***	-2.242***
	(0.153)	(0.179)	(0.230)	(0.152)	(0.178)	(0.230)
Housing and Urbanization (log)	0.045***	0.045**	0.045*	0.043***	0.043*	0.043*
	(0.015)	(0.023)	(0.026)	(0.015)	(0.022)	(0.025)
Transfers (log)	-0.072	-0.072	-0.072	-0.031	-0.031	-0.031
	(0.118)	(0.127)	(0.206)	(0.117)	(0.126)	(0.193)
Oil Revenue (log)	0.001	0.001	0.001	0.002	0.002	0.002
	(0.006)	(0.007)	(0.017)	(0.006)	(0.007)	(0.016)
Constant	-1.870	-1.870	-1.870	2.946***	2.946***	2.946
	(1.137)	(1.188)	(2.084)	(1.106)	(1.132)	(1.851)
N	4269	4269	4269	4269	4269	4269
R^2	0.669	0.669	0.669	0.675	0.675	0.675

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).

Model 2 and Model 5 with robust standard errors.

Model 3 and Model 6 with standard errors clustered by state.

The results when including the number of families that receives the Bolsa Família cash transfer (*number of families*) or the amount of cash benefits in Brazilian reais (cash benefits amount) are consistent with our previous results: more unequal municipalities have a lower capacity to collect taxes.

Standard errors in parentheses. Two-tailed test. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.5: GDP Growth

	Model 1 2000	Model 2 2000 (robust)	Model 3 2000 (cluster)	Mode 4 2010	Model 5 2010 (robust)	Model 6 2010 (cluster)
Gini	-2.286***	-2.286***	-2.286**	-5.798***	-5.798***	-5.798***
	(0.661)	(0.702)	(0.909)	(0.492)	(0.630)	(1.326)
GDP Growth	-0.143***	-0.143***	-0.143***	-0.240***	-0.240***	-0.240**
	(0.027)	(0.043)	(0.046)	(0.024)	(0.075)	(0.088)
Population (log)	-0.750***	-0.750***	-0.750**	-0.221***	-0.221***	-0.221
	(0.098)	(0.130)	(0.279)	(0.076)	(0.083)	(0.181)
GDP (log)	2.316***	2.316***	2.316***	1.629***	1.629***	1.629***
	(0.081)	(0.101)	(0.214)	(0.060)	(0.075)	(0.196)
Left Party	-0.085	-0.085	-0.085	-0.015	-0.015	-0.015
	(0.091)	(0.091)	(0.110)	(0.057)	(0.058)	(0.059)
Rural Share	-2.939***	-2.939***	-2.939***	-2.563***	-2.563***	-2.563***
	(0.229)	(0.241)	(0.463)	(0.158)	(0.188)	(0.377)
Housing and Urbanization (log)	0.002	0.002	0.002	0.057***	0.057**	0.057**
	(0.020)	(0.022)	(0.031)	(0.015)	(0.023)	(0.027)
Transfers (log)	-0.504***	-0.504**	-0.504	-0.219*	-0.219	-0.219
	(0.148)	(0.228)	(0.320)	(0.122)	(0.135)	(0.179)
Oil Revenue (log)	-0.036***	-0.036***	-0.036	-0.020***	-0.020**	-0.020
	(0.011)	(0.013)	(0.031)	(0.006)	(0.008)	(0.017)
Constant	2.293*	2.293	2.293	2.018*	2.018	2.018
	(1.294)	(1.880)	(2.649)	(1.155)	(1.228)	(1.648)
N	3695	3695	3695	4243	4243	4243
R^2	0.546	0.546	0.546	0.649	0.649	0.649

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
GDP Growth $= \frac{GDP - GDP_{t-1}}{GDP_{t-1}}$
Model 2 and Model 5 with robust standard errors.
Model 3 and Model 6 with standard errors clustered by state.
Results for models including the independent variable *GDP growth* are consistent with those reported previously: more unequal municipalities have a lower capacity to collect taxes.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.6: Adding ITBI and Total Tax as Independent Variables into the Original Cross-Sectional Model for 2010

	Model 1 IPTU	Model 2 2010 (robust)	Model 3 2010 (cluster)	Model 4 2010	Model 5 2010 (robust)	Model 6 2010 (cluster)
Gini	-6.190***	-6.190***	-6.190***	-5.154***	-5.154***	-5.154***
	(0.493)	(0.624)	(1.396)	(0.442)	(0.534)	(0.998)
ITBI (log)				0.288***	0.288***	0.288***
				(0.012)	(0.026)	(0.037)
Total Tax (log)				0.673***	0.673***	0.673***
				(0.042)	(0.059)	(0.127)
Population (log)	-0.080	-0.080	-0.080	0.051	0.051	0.051
	(0.075)	(0.081)	(0.185)	(0.066)	(0.067)	(0.136)
GDP (log)	1.503***	1.503***	1.503***	0.535***	0.535***	0.535***
	(0.059)	(0.068)	(0.190)	(0.061)	(0.074)	(0.153)
Left Party	-0.033	-0.033	-0.033	-0.040	-0.040	-0.040
	(0.058)	(0.059)	(0.059)	(0.051)	(0.052)	(0.049)
Rural Share	-2.734***	-2.734***	-2.734***	-1.701***	-1.701***	-1.701***
	(0.158)	(0.182)	(0.401)	(0.144)	(0.172)	(0.345)
Housing and Urbanization (log)	0.052***	0.052**	0.052*	0.006	0.006	0.006
	(0.015)	(0.023)	(0.027)	(0.014)	(0.021)	(0.018)
Transfers (log)	-0.228*	-0.228*	-0.228	-0.426***	-0.426***	-0.426**
	(0.123)	(0.134)	(0.195)	(0.112)	(0.125)	(0.167)
Oil Revenue (log)	-0.020***	-0.020***	-0.020	-0.015**	-0.015**	-0.015
	(0.007)	(0.008)	(0.018)	(0.006)	(0.007)	(0.016)
Constant	2.391**	2.391*	2.391	2.518**	2.518**	2.518
	(1.162)	(1.226)	(1.864)	(1.046)	(1.094)	(1.521)
N	4269	4269	4269	4265	4265	4265
R^2	0.641	0.641	0.641	0.715	0.715	0.715

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).

ITBI = Tax Revenue on Real Estate Transfers in Brazilian reais.

Total Tax = Total taxes revenue collected by the municipality.

Model 2 and Model 5 with robust standard errors.

Model 3 and Model 6 with standard errors clustered by state.

Results for models including the independent variable *ITBI* are consistent with previous models: more unequal municipalities have a lower capacity to collect taxes.

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table B.7: Industry Value Added as Percentage of GDP

	Model 1 2000 (robust)	Model 2 2000 (cluster)	Model 3 2010 (robust)	Model 4 2010 (cluster)	Model 5 1991-2000-2010 (FE & cluster)
Gini	-4.229***	-4.229***	-6.529***	-6.529***	-3.443**
	(0.629)	(0.979)	(0.627)	(1.422)	(1.261)
Industry (% GDP)	-2.908***	-2.908***	-1.960***	-1.960***	-1.778***
	(0.394)	(0.804)	(0.241)	(0.341)	(0.524)
Population (log)	-1.140***	-1.140***	-0.198**	-0.198	1.788***
	(0.103)	(0.359)	(0.080)	(0.178)	(0.452)
GDP (log)	2.475***	2.475***	1.668***	1.668***	0.066
	(0.099)	(0.264)	(0.074)	(0.200)	(0.175)
Left Party	-0.140	-0.140	-0.037	-0.037	
	(0.086)	(0.117)	(0.059)	(0.057)	
Rural Share	-2.947***	-2.947***	-2.750***	-2.750***	-0.818
	(0.210)	(0.461)	(0.180)	(0.397)	(0.858)
Housing and Urbanization (log)	0.016	0.016	0.054**	0.054*	-0.003
	(0.018)	(0.028)	(0.023)	(0.028)	(0.015)
Transfers (log)	-0.112	-0.112	-0.227*	-0.227	0.370**
	(0.180)	(0.367)	(0.133)	(0.189)	(0.159)
Oil Revenue (log)	-0.023*	-0.023	-0.017**	-0.017	
	(0.012)	(0.030)	(0.008)	(0.017)	
2000					3.884*
					(2.209)
2010					4.673*
					(2.376)
Constant	-0.566	-0.566	2.153*	2.153	-15.652***
	(1.517)	(2.898)	(1.210)	(1.752)	(2.909)
N	4845	4845	4269	4269	8138
R^2	0.513	0.513	0.646	0.646	0.879

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).

Industry (% GDP) = Industry value added, as % of GDP.

We could not find data for *municipal GDP* and *Industry (% GDP)* for the early 1990s. We thus have to rely on a GDP measurement from 1985 in the panel data for 1991.

Model 1 and Model 3 cross-sectional models with robust standard errors. Model 2 and Model 4 cross-sectional models with standard errors clustered by state. Model 5 Panel model (1991-2000-2010) with Year and Municipal fixed-effects and standard errors clustered by state.

Results for models including the independent variable *Industry (% GDP)* are consistent with our results reported in the manuscript: more unequal municipalities have a lower capacity to collect taxes. The results are consistent when dropping the 10 observations below 0 and 4 observations above 1 for *Industry (% GDP)*.

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

C Appendix: Additional Dependent Variables

Table C.1: IPTU as a Ratio DV, Cross-Sectional Models for 2010

	Model 1 $\frac{IPTU}{GDP}$	Model 2 $\frac{IPTU}{GDP}$ (robust)	Model 3 $\frac{IPTU}{GDP}$ (cluster)	Model 4 $\frac{IPTU}{Tax}$	Model 5 $\frac{IPTU}{Tax}$ (robust)	Model 6 $\frac{IPTU}{Tax}$ (cluster)
Gini	-5.156**	-5.156***	-5.156*	-0.432***	-0.432***	-0.432***
	(2.213)	(1.766)	(2.911)	(0.034)	(0.035)	(0.073)
Population (log)	0.418	0.418	0.418	0.017***	0.017***	0.017
	(0.344)	(0.299)	(0.399)	(0.005)	(0.006)	(0.012)
Left Party	-0.439*	-0.439*	-0.439	-0.002	-0.002	-0.002
	(0.266)	(0.254)	(0.365)	(0.004)	(0.004)	(0.004)
Rural Share	-9.056***	-9.056***	-9.056***	-0.178***	-0.178***	-0.178***
	(0.689)	(1.028)	(2.947)	(0.011)	(0.011)	(0.045)
Housing and Urbanization (log)	0.206***	0.206***	0.206*	0.002*	0.002*	0.002
	(0.070)	(0.048)	(0.120)	(0.001)	(0.001)	(0.002)
Transfers (log)	0.520	0.520	0.520	-0.083***	-0.083***	-0.083***
	(0.439)	(0.429)	(0.569)	(0.009)	(0.009)	(0.019)
Oil Revenue (log)	0.083***	0.083	0.083	-0.002***	-0.002***	-0.002
	(0.029)	(0.064)	(0.136)	(0.000)	(0.000)	(0.001)
GDP (log)				0.068***	0.068***	0.068***
				(0.004)	(0.004)	(0.014)
Constant	-6.608	-6.608	-6.608	0.888***	0.888***	0.888***
	(4.563)	(5.012)	(6.577)	(0.081)	(0.089)	(0.196)
N	4269	4269	4269	4267	4267	4267
R^2	0.113	0.113	0.113	0.344	0.344	0.344

Notes: Dependent variables: IPTU Revenue/GDP (Model 1, Model 2, and Model 3); IPTU Revenue/Total tax revenue (Model 4, Model 5, and Model 6)
Model 2 and Model 5 with robust standard errors.
Model 3 and Model 6 with standard errors clustered by state.
Results when using an alternative measures of tax capacity (IPTU as a ratio of GDP and as a ratio of total tax) are consistent with previous results: more unequal municipalities have a lower capacity to collect taxes.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.2: Alternative Local Tax: Original Cross-Sectional Model for 2010 and Model Using ITBI as Dependent Variable

	Model 1 IPTU	Model 2 IPTU (robust)	Model 3 IPTU (cluster)	Model 4 ITBI	Model 5 ITBI (robust)	Model 6 ITBI (cluster)
Gini	-6.190***	-6.190***	-6.190***	-4.378***	-4.378***	-4.378***
	(0.493)	(0.624)	(1.396)	(0.566)	(0.673)	(1.502)
Population (log)	-0.080	-0.080	-0.080	-0.341***	-0.341***	-0.341*
	(0.075)	(0.081)	(0.185)	(0.085)	(0.093)	(0.173)
GDP (log)	1.503***	1.503***	1.503***	1.865***	1.865***	1.865***
	(0.059)	(0.068)	(0.190)	(0.067)	(0.077)	(0.219)
Left Party	-0.033	-0.033	-0.033	0.062	0.062	0.062
	(0.058)	(0.059)	(0.059)	(0.066)	(0.064)	(0.079)
Rural Share	-2.734***	-2.734***	-2.734***	-1.782***	-1.782***	-1.782***
	(0.158)	(0.182)	(0.401)	(0.181)	(0.193)	(0.354)
Housing and Urbanization (log)	0.052***	0.052**	0.052*	0.062***	0.062***	0.062*
	(0.015)	(0.023)	(0.027)	(0.017)	(0.021)	(0.030)
Transfers (log)	-0.228*	-0.228*	-0.228	-0.685***	-0.685***	-0.685**
	(0.123)	(0.134)	(0.195)	(0.141)	(0.150)	(0.271)
Oil Revenue (log)	-0.020***	-0.020***	-0.020	-0.019**	-0.019**	-0.019
	(0.007)	(0.008)	(0.018)	(0.007)	(0.008)	(0.020)
Constant	2.391**	2.391*	2.391	6.917***	6.917***	6.917**
	(1.162)	(1.226)	(1.864)	(1.332)	(1.397)	(2.520)
N	4269	4269	4269	4267	4267	4267
R^2	0.641	0.641	0.641	0.524	0.524	0.524

Notes: Dependent variables: IPTU Revenue (logged) (Model 1, Model 2, and Model 3); ITBI Revenue (logged) (Model 4, Model 5, and Model 6)

ITBI = Tax Revenue on Real Estate Transfers in Brazilian reais.

Model 2 and Model 5 with robust standard errors.

Model 3 and Model 6 with standard errors clustered by state.

Results using an alternative local tax *(ITBI)* as our dependent variable are consistent with our previous models: more unequal municipalities have a lower capacity to collect tax.

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.3: Correlation Matrix

	IPTU	**ITBI**
IPTU	1.0000	
ITBI	0.9702 0.000	1.0000

Considering the high positive correlation between IPTU and ITBI (Pearson's r = 0.97) it is not surprising that the results using ITBI as the dependent variable are consistent with the main results from our original models.

Table C.4: IPTU Collected by Number of Buildings as DV, 2000

	DV: IPTU Collected (By Buildings)	
	Model 1	Model 2
	(robust)	(cluster)
Gini	-0.391***	-0.391***
	(0.063)	(0.076)
Population (log)	-0.163***	-0.163***
	(0.010)	(0.022)
GDP (log)	0.149***	0.149***
	(0.008)	(0.021)
Left Party	0.002	0.002
	(0.009)	(0.009)
Rural Share	0.124***	0.124**
	(0.021)	(0.057)
Housing and Urbanization (log)	0.003*	0.003
	(0.002)	(0.003)
Transfers (log)	0.001	0.001
	(0.015)	(0.013)
Oil Revenue (log)	-0.007***	-0.007***
	(0.001)	(0.002)
Constant	0.526***	0.526***
	(0.130)	(0.101)
N	4005	4005
R^2	0.175	0.175

Notes: Dependent variables: $\frac{\text{total number of buildings that paid IPTU}}{\text{total number of buildings that could be charged}}$

Model 1 with robust standard errors. Model 2 with standard errors clustered by state.

The results when using *IPTU Collected by buildings* as our dependent variable are consistent with those in our previous models: more unequal municipalities have a lower capacity to collect taxes.

Standard errors in parentheses. Two-tailed test. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.5: IPTU Collection Rate as DV

	Dependent variable: IPTU Collection Rate						
	Model 1 2000	Model 2 2000 (robust)	Model 3 2000 (cluster)	Model 4 2010	Model 5 2010 (robust)	Model 6 2010 (cluster)	Model 7 Panel (2000-2010) (FE & cluster)
Gini	-0.735**	-1.047**	-1.047*	-0.613**	-0.676**	-0.676*	-1.460**
	(0.315)	(0.428)	(0.556)	(0.250)	(0.317)	(0.392)	(0.544)
Population (log)		-0.092	-0.092		-0.088**	-0.088***	0.085
		(0.056)	(0.055)		(0.042)	(0.024)	(0.192)
GDP (log)		0.098	0.098*		0.162***	0.162***	-0.134***
		(0.061)	(0.048)		(0.046)	(0.044)	(0.032)
Left Party		-0.039	-0.039		-0.055*	-0.055**	-0.054***
		(0.035)	(0.029)		(0.028)	(0.025)	(0.016)
Rural Share		-0.237	-0.237		-0.349**	-0.349**	-0.975*
		(0.325)	(0.379)		(0.158)	(0.159)	(0.506)
Housing and Urbanization (log)		0.013	0.013		-0.003	-0.003	0.003
		(0.018)	(0.017)		(0.005)	(0.005)	(0.002)
Transfers (log)		0.003	0.003		-0.042	-0.042	0.043
		(0.083)	(0.086)		(0.079)	(0.069)	(0.053)
Oil Revenue (log)		-0.001	-0.001		-0.007***	-0.007***	0.010***
		(0.004)	(0.004)		(0.002)	(0.002)	(0.002)
2010							0.018
							(0.058)
Constant	0.910***	0.606	0.606	0.937***	0.633	0.633	1.421
	(0.176)	(0.532)	(0.652)	(0.130)	(0.609)	(0.556)	(1.968)
N	180	142	142	180	142	142	238
R^2	0.030	0.201	0.201	0.033	0.352	0.352	0.498

Notes: Dependent variables: IPTU Collection Rate as measured by Carvalho, Jr. (2017).
Model 2 and Model 5 with robust standard errors.
Model 3 and Model 6 with standard errors clustered by state.
Model 7 with year and municipality fixed-effects and standard errors clustered by state.
Results when using *IPTU Collection Rate* as our dependent variable are consistent with the results presented in the manuscript: more unequal municipalities have a lower capacity to collect taxes.
Standard errors in parentheses. Two-tailed test. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table C.6: Full Table: Municipal Applications to the Capacity-Building Program (PMAT)

	Dependent variable: PMAT Application			
	Model 1 OLS (robust)	Model 2 OLS (cluster)	Model 3 Logit (robust)	Model 4 Logit (cluster)
Gini	-0.242***	-0.242***	-2.786**	-2.786**
	(0.050)	(0.070)	(1.166)	(1.230)
IPTU Revenue (log)	0.057***	0.057***	0.449***	0.449***
	(0.007)	(0.010)	(0.084)	(0.126)
Population (log)	0.032***	0.032***	0.231**	0.231
	(0.006)	(0.009)	(0.114)	(0.196)
GDP (log)	0.019***	0.019**	0.489***	0.489***
	(0.005)	(0.008)	(0.099)	(0.138)
Rural Share	-0.035*	-0.035	-0.699*	-0.699
	(0.019)	(0.026)	(0.423)	(0.526)
Transfers (log)	-0.017***	-0.017**	-0.209***	-0.209
	(0.004)	(0.008)	(0.069)	(0.131)
Constant	-0.270***	-0.270***	-8.390***	-8.390***
	(0.057)	(0.067)	(1.042)	(1.560)
N	4047	4047	4047	4047
R^2	0.193	0.193		
Log-likelihood			-755.391	-755.391

Notes: Dependent variable: Binary variable PMAT (1 = municipality applied to PMAT, 0 = municipality didn't apply to PMAT).

Model 1 and Model 3 with robust standard errors.

Model 2 and Model 4 with standard errors clustered by state.

Results when *PMAT* as our dependent variable indicate that greater inequality is associated with a lower likelihood of application to PMAT.

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

D Appendix: Panel Models

Table D.1: Panel Fixed Effects Models (1991, 2000, and 2010)

	Dependent variable: IPTU Revenue (log)			
	Model 1 1991-2000-2010 (non-imputed)	Model 2 1991-2000-2010 (imputed)	Model 3 2000-2010 (non-imputed)	Model 4 2000-2010 (imputed)
Gini	-3.479***	-5.118***	0.805	0.084
	(1.220)	(1.113)	(0.622)	(0.665)
Population (log)	1.996***	1.455***	0.642	0.042
	(0.459)	(0.435)	(0.494)	(0.511)
GDP (log)	-0.033	0.026	0.654**	0.697***
	(0.171)	(0.168)	(0.238)	(0.207)
Rural Share	-0.468	-0.801	-2.809**	-2.662**
	(0.959)	(0.789)	(1.012)	(1.089)
Housing and Urbanization (log)	-0.006	-0.080***	0.013	0.033*
	(0.015)	(0.025)	(0.015)	(0.016)
Transfers (log)	0.363**	0.349***	0.188	0.293***
	(0.163)	(0.081)	(0.305)	(0.078)
2000	4.113*	5.404***		
	(2.257)	(1.023)		
2010	4.936*	6.271***	1.107***	0.902***
	(2.427)	(1.072)	(0.376)	(0.132)
Constant	-17.067***	-11.379***	-6.170	-2.545
	(3.167)	(3.704)	(5.916)	(4.685)
N	8138	9706	8599	9154
R^2	0.878		0.358	

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
Year and Municipal fixed-effects included in all models. Standard errors clustered by state.
The results of Model 1 and Model 2—including all the data—are consistent with the cross-sectional models, indicating that more unequal municipalities have a lower capacity to collect tax. The estimates for *Gini* in Model 3 and Model 4—2000 and 2010 data only—are not significant.
We dropped municipalities that were founded after 1985, resulting in a smaller number of observations in Model 1, i.e., those municipalities that did not exist for part of the time period used in the panel. The results do not change significantly if we do not drop these municipalities.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table D.2: Bivariate Fixed Effects Panel Models: Benchmark to Compare the Results

	IPTU (1991-2000-2010)	IPTU (1991-2000-2010)	IPTU (2000-2010)	IPTU (2000-2010)
Gini	-16.745**	-3.479***	-13.298***	0.805
	(6.696)	(1.220)	(1.083)	(0.622)
Population (log)		1.996***		0.642
		(0.459)		(0.494)
GDP (log)		-0.033		0.654**
		(0.171)		(0.238)
Rural Share		-0.468		-2.809**
		(0.959)		(1.012)
Transfers (log)		0.363**		0.188
		(0.163)		(0.305)
Housing and Urbanization (log)		-0.006		0.013
		(0.015)		(0.015)
2000		4.113*		
		(2.257)		
2010		4.936*		1.107***
		(2.427)		(0.376)
Constant	16.149***	-17.067***	16.921***	-6.170
	(3.518)	(3.167)	(0.565)	(5.916)
N	9378	8138	8655	8599
R^2	0.028	0.878	0.128	0.358

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).
Year and Municipal fixed-effects included in all models. Standard errors clustered by state.
The coefficient on inequality is larger in the bivariate models and remains significant. For the bivariate model, even the two-period panel model (2000-2010) results are in line with our argument.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table D.3: Original Panel Models (1991-2000-2010) with Municipality-specific Linear and Quadratic Time Trends

	DV: IPTU Revenue	
	Model 1	Model 2
	(Time Trend)	(Time Trend^2)
Gini	-3.317**	-3.479***
	(1.296)	(1.220)
Population (log)	1.829***	1.996***
	(0.427)	(0.459)
GDP (log)	-0.100	-0.033
	(0.202)	(0.171)
Rural Share	-0.426	-0.468
	(1.015)	(0.959)
Transfers (log)	0.685***	0.363**
	(0.062)	(0.163)
Housing and Urbanization (log)	0.018	-0.006
	(0.021)	(0.015)
Time Trend	0.425**	9.050
	(0.172)	(5.395)
Time Trend^2		-1.646
		(1.047)
Constant	-16.929***	-24.472***
	(3.177)	(6.540)
N	8138	8138
R^2	0.876	0.878

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged). All models with municipality fixed-effects and standard errors clustered by municipality.
This table shows the results for a panel model with both linear and quadratic time trends instead of year fixed effects. The results do not change substantially. Our independent variable of interest is still substantially large and significant.
Standard errors in parentheses. Two-tailed test.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E Appendix: Sub-sample Analysis

Between 1982 and 2007, the number of municipalities in Brazil increased by 41 percent. It could be problematic for our analysis if the split of municipal units was somehow associated with the level of inequality. In general, the increase in the number of municipalities has been attributed to a number of different factors.

Since income is often concentrated geographically, it is possible that a redrawing of municipalities could split the older municipality into two very different new municipalities. For example, one high inequality municipality could be split into two low inequality units. Or it could be split into one high inequality and one low inequality municipality. To rule out the possibility that our results are affected by these splits, we first show the densities of our main independent variable of interest, inequality, for subsamples of municipalities split up by on municipality age (until 2010).

Figure E.1 depicts the distribution of years since each municipality in our dataset was created. The trimodal distribution reveals the thee most often values in our data: 1. municipalities over 70 years old; 2. municipalities between 40 and 60 years old, and; 3. municipalities between 10 and 20 years old.

Figure E.2, in turn, shows the distribution of the GINI coefficient by municipality ages in decades of age. The non-relationship between age and inequality (captured by the relatively similar distributions in each graph) indicates that a possible split of municipalities due to high inequality are most likely not driving our results.

In addition, we run our original model on a sub-sample of those municipalities that were created prior to 1970. The results are presented in column 2 in Table E.1 and are consistent

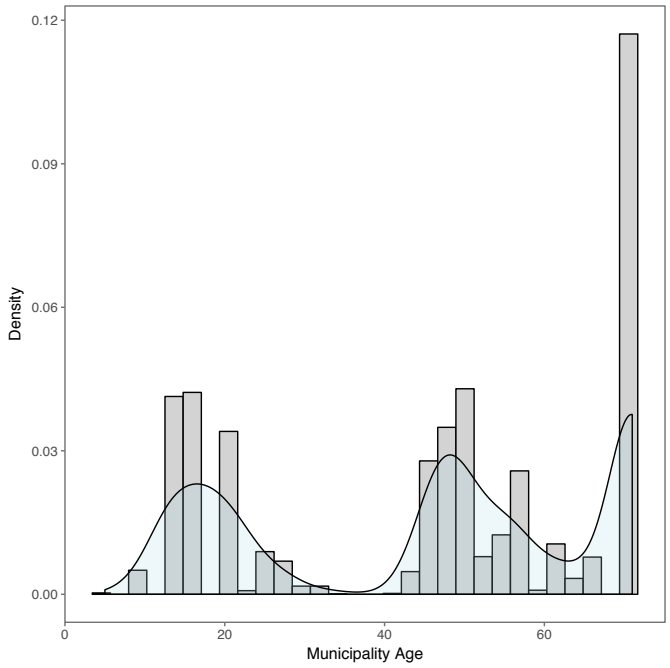


Figure E.1: Municipality Age in 2010: This plot shows the distribution of years since each municipality in the data set was created.

with those in the original sample (column 1). The results for the sub-sample analysis are consistent with the results using our original sample: a consistent negative effect of inequality on municipal IPTU revenue.

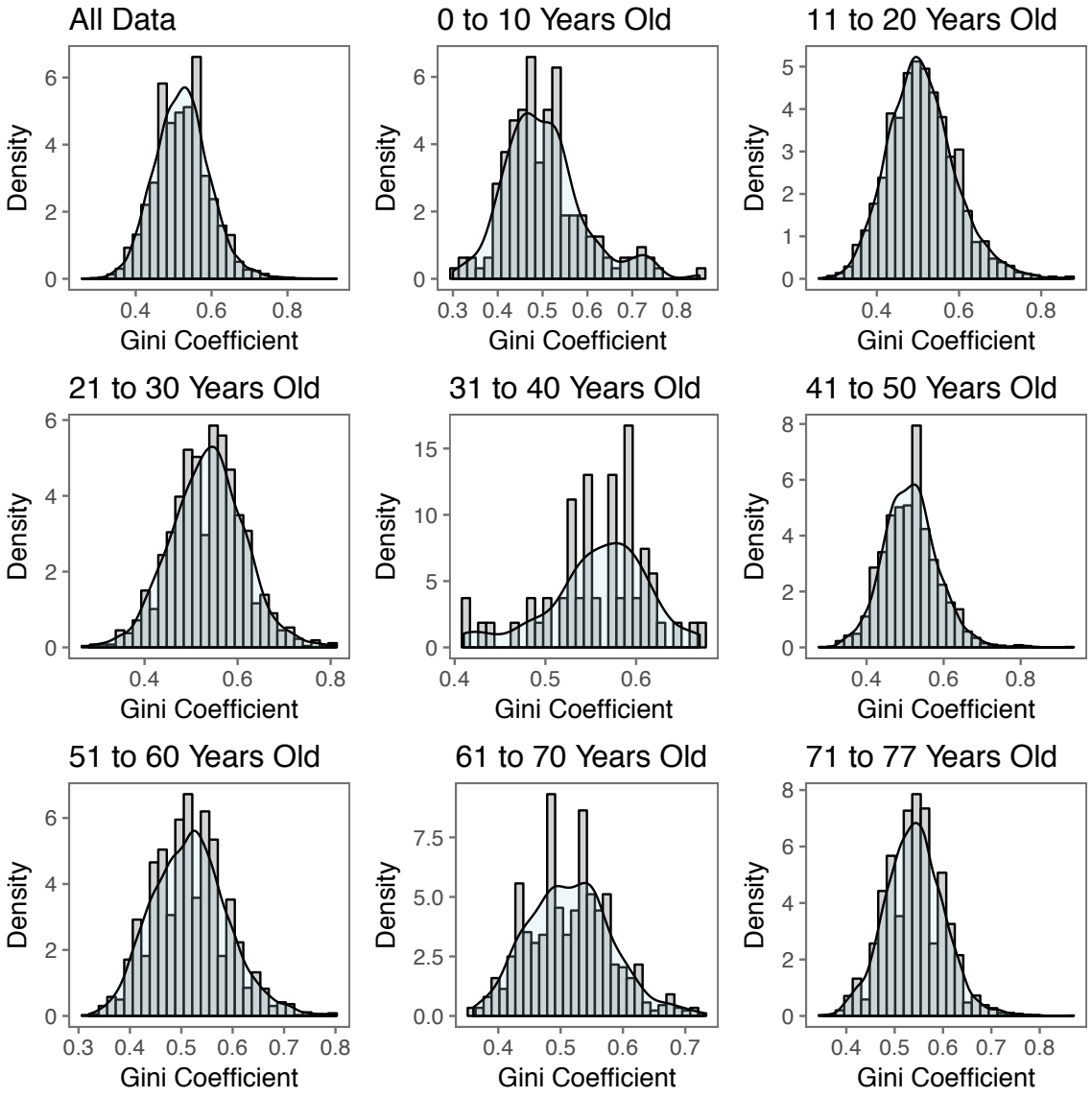


Figure E.2: GINI Coefficient by Municipality Age in 2010: This plot shows the distribution of the GINI coefficient by municipality ages (in decades of age). The non-relationship between age and inequality (captured by the relatively similar distributions in each graph) indicates that the split of municipalities due to high inequality is most likely not driving our results.

Table E.1: Sub-Sample Analysis Based on Municipality Age (until 2010)

	DV: IPTU Revenue (log)	
	Model 1 (Original 2010 Cross-Sectional Model)	Model 2 (Sub-Sample Model)
Gini	-6.190***	-4.880***
	(1.396)	(1.336)
Population (log)	-0.080	0.023
	(0.185)	(0.154)
GDP (log)	1.503***	1.570***
	(0.190)	(0.208)
Left Party	-0.033	-0.006
	(0.059)	(0.061)
Rural Share	-2.734***	-2.874***
	(0.401)	(0.390)
Housing and Urbanization (log)	0.052*	0.049
	(0.027)	(0.031)
Transfers (log)	-0.228	-0.437**
	(0.195)	(0.186)
Oil Revenue (log)	-0.020	-0.028
	(0.018)	(0.018)
Constant	2.391	3.690**
	(1.864)	(1.650)
N	4269	3337
R^2	0.641	0.677

Notes: Dependent variable: IPTU Revenue in Brazilian reais (logged).

Models with standard errors clustered by state.

The results for the sub-sample analysis are consistent with the results using our original sample: a consistent negative effect of inequality on municipal IPTU revenue.

Standard errors in parentheses. Two-tailed test.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

F Appendix: Zero Values in the Dependent Variable

Our measure of IPTU revenue collection, made available by the Institute of Applied Economic Research (IPEA), is in the current Brazilian currency called *real* (in plural, *reais*). The *real* was introduced on July 1, 1994. The data for 1991, therefore, was converted by IPEA from the former currency *cruzeiro* to *reais*, and also deflated to controlling for the high inflation in Brazil in 1991. In personal correspondence with the IPEA staff responsible for the data collection, they acknowledged that zero values in the revenue data should mean that no revenue was collected, but could not rule out the possibility that some of these zero values should actually be missing values. This is in addition to the missing values that do exist in the data.

We therefore decided to conduct an additional robustness check to our results. In an attempt to identify observations with actual zero revenue, we used the data on IPTU revenue in nominal values (in their original currencies), as originally released by the Brazilian Ministry of Finance through the National Treasury Secretariat. For the robustness check we created a corrected version of the IPTU revenue variable. We set values in the original IPTU data to NA (missing value) if the corresponding observation in the nominal IPTU data is NA.

This correction reduces the number of zero values in the data significantly—specifically, by about two-thirds for 1991. The original data have $1,527$ zeros in 1991, 458 in 2000, and 71 in 2010. The replacement of zeros with NA when the nominal data is missing, results in 443 zeros for 1991.

Given that only the data for 1991 is affected, we present the panel model with these

changes in Table F.1. We are calling this changed dependent variable *IPTU_corrected (log)*. The first model is the fixed effects panel model (1991-2000-2010) using the corrected dependent variable with non-imputed data, i.e., dropping the missing observations. The second column shows the results when the corrected dependent variable is also imputed, i.e., the original zero values that were set as missing observations are replaced with imputed values.

While the magnitude of the coefficient for *Gini* in the panel model is smaller than from our original panel model, the results from the analysis using our new dependent variable *IPTU_corrected (log)* are consistent with the results we found originally: a consistent negative effect of inequality on municipal IPTU revenue in both models (either using non-imputed or imputed data). In addition, the selection on unobservables test results in a δ value of 3.34 for the non-imputed panel model.

Table F.1: Replacing the observations in our original dependent variable (IPTU revenue) to NA (missing value) when the observation is NA in the original data using nominal values

	Dependent variable: IPTU_corrected (log)	
	Model 1	Model 2
	FE Panel Model	FE Panel Model
	(1991-2000-2010)	(1991-2000-2010)
	Non-Imputed Data	*Imputed Data*
Gini	-2.851***	-3.614***
	(0.994)	(0.945)
Population (log)	2.283***	2.096***
	(0.481)	(0.398)
GDP (log)	0.081	0.120
	(0.127)	(0.120)
Rural Share	-0.927	-1.033
	(0.765)	(0.673)
Housing and Urbanization (log)	0.007	-0.068***
	(0.014)	(0.022)
Transfers (log)	-0.581	-0.841***
	(0.484)	(0.279)
2000	13.796**	17.562***
	(5.472)	(3.037)
2010	15.881**	20.080***
	(6.057)	(3.344)
Constant	-16.369***	-13.076***
	(2.868)	(3.073)
N	8102	9221
R^2	0.879	.

Notes: Dependent variable: *IPTU_corrected (log)*.

Both models with standard errors clustered by state and fixed-effects. The results for the analysis using *IPTU_corrected (log)* as the dependent variable are consistent with the results using our original dependent variable: a negative effect of inequality on municipal IPTU revenue in both models.

Standard errors in parentheses. Two-tailed test. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Spatial Interdependence and Instrumental Variable Models*

Timm Betz † Scott J. Cook‡ Florian M. Hollenbach§

Conditionally accepted, *Political Science Research and Methods*

Abstract

Instrumental variable (IV) methods are widely used to address endogeneity concerns. Yet, a specific kind of endogeneity – spatial interdependence – is regularly ignored. We show that ignoring spatial interdependence in the outcome results in asymptotically biased estimates even when instruments are randomly assigned. The extent of this bias increases when the instrument is also spatially clustered, as is the case for most widely-used instruments: rainfall, natural disasters, economic shocks, and regionally- or globally-weighted averages. Because the biases due to spatial interdependence and predictor endogeneity can offset, addressing only one can increase the bias relative to ordinary least squares. We demonstrate the extent of these biases both analytically and via Monte Carlo simulation. Finally, we discuss a general estimation strategy – Spatial-2SLS – that accounts for both outcome interdependence and predictor endogeneity, thereby recovering consistent estimates of predictor effects.

Key Words: Instrumental Variables, Spatial Analysis, Spatial Modeling, Two-Stage Least Squares

*Thanks to Vincent Arel-Bundock, Patrick Brandt, Rob Franzese, Kosuke Imai, Stephen Jessee, Tom Pepinsky, Piero Stanig, Vera Troeger, Guy Whitten, and participants at the Annual Conference of the European Political Science Association in 2016, the Annual Conference of the Society for Political Methodology in 2016, and the Texas Methods Meeting in 2017 for their helpful comments. All remaining errors are ours alone. Authors are listed in alphabetical order, equal authorship is implied. Portions of this research were conducted with high performance research computing resources provided by Texas A&M University (`http://hprc.tamu.edu`).

†Assistant Professor of Political Science, Department of Political Science, Texas A&M University, College Station, TX 77843. Email: timm.betz@tamu.edu, URL: `www.people.tamu.edu/~timm.betz`

‡Assistant Professor of Political Science, Department of Political Science, Texas A&M University, College Station, TX 77843. Email: sjcook@tamu.edu, URL: `scottjcook.net`

§Assistant Professor of Political Science, Department of Political Science, Texas A&M University, College Station, TX 77843. Email: fhollenbach@tamu.edu, URL: `fhollenbach.org`

1 Introduction

As political scientists increasingly focus on the identification of causal effects, instrumental variable (IV) models are becoming commonplace (e.g., Sovey and Green, 2011). IV models hinge on the validity of the instrument. While researchers are usually aware of conditional independence and relevance as general requirements for valid instruments, we identify a specific threat that is frequently ignored: spatial interdependence in the outcome variable. Our review of IV models in leading political science journals reveals that authors rarely discuss and never empirically address spatial interdependence as a threat to inference (see Figure 1), even as theories of spatial interdependence and diffusion proliferate across political science (see, e.g., Siverson and Starr 1990; Starr 1991; Ward and O'Loughlin 2002; Ward and Gleditsch 2002; Simmons, Dobbin and Garrett 2006; Franzese and Hays 2007; Plümper and Neumayer 2010).[1]

This is is not a trivial oversight. We show that failing to model outcome interdependence produces estimates that are asymptotically biased, even when the instrument is randomly assigned. When, in addition, the instrument exhibits spatial dependence similar to that of the outcome, the bias in IV estimates increases and can even surpass that of ordinary least squares. This concern applies to many popular instruments, including geographic, meteorologic, and economic variables (see, e.g., Ramsay 2011; Hansford and Gomez 2010; Ahmed 2012), as well as any instrument measured at a higher level of aggregation than the outcome, such as regional or global economic, political, and institutional shocks (see, e.g., Stasavage 2005; Büthe and Milner 2008; Boix 2011; Ramsay 2011). Because these instruments are not randomly distributed across space, they risk increased bias even when they are otherwise plausibly exogenous.

Our results connect more general findings in the otherwise distinct literatures on spatial interdependence and instrumental variables. Ignored spatial interdependence constitutes an omitted variables problem (e.g., Franzese and Hays 2007). While IV models are commonly thought to be

[1]We analyzed each article on the basis of whether prior theories of spatial interdependence or diffusion had been established for and could reasonably apply to the outcome of interest. The articles using IV models that are not at risk of the issues we discuss here include pure time-series analyses, survey experiments, most field experiments, etc.

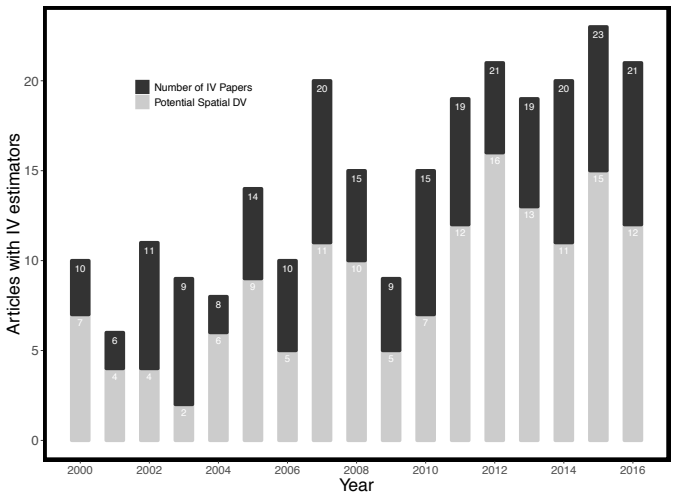


Figure 1: The plot shows the number of articles published in the APSR, AJPS, JOP, IO, BJPS, and World Politics between 2000 and 2016 that use IV models (light grey bars), and the number of those articles at risk of spatial interdependence in the dependent variable (dark grey bars).

immune to omitted variable bias, and indeed frequently used to overcome it (Wooldridge, 2002), this intuition does not always hold. Instead, as we show below, IV models can augment omitted variables bias from unmodeled spatial interdependence.

Because of its reciprocal relationship with the outcome, ignored spatial interdependence also ensures any instrument violates the exclusion restriction. As is well known, even mild violations of the exclusion restriction can produce substantial bias (Bartels, 1991; Bound, Jaeger and Baker, 1995). When these violations are caused by spatial interdependence, however, solutions are available to recover asymptotically unbiased estimates if one is willing to make assumptions about the nature of spatial relationships in the outcome variable. Recent work in the spatial econometrics literature has generalized spatial models to allow for endogenous predictors (e.g., Kelejian and Prucha 2004; Anselin and Lozano-Gracia 2008; Fingleton and Le Gallo 2008; Drukker, Egger and Prucha 2013; Liu and Lee 2013). These same methods – hereafter spatial-two stage least squares (S-2SLS) – are useful when addressing endogenous predictors even when researchers are otherwise uninterested in spatial dependence theoretically.[2] In short, with S-2SLS researchers instrument for

[2]To clarify, Franzese and Hays (2007) and others have previously used S-2SLS to indicate a spatial autoregressive (SAR) model estimated via 2SLS. Here, we use this term more broadly to

both the endogenous predictor and the spatial-lag of the outcome, thereby obtaining consistent estimates of the desired causal effect.

In addition to accounting for possible outcome interdependence, this approach has two attractive features. First, it nests the standard spatial-autoregressive (SAR) model and the standard IV model, allowing researchers to explicitly test restrictions rather than proceed by assumption.[3] Second, because it is an instrumental variables approach, it should be straightforward to understand and implement for those already pursuing IV strategies. Our simulations demonstrate that this approach consistently outperforms estimation strategies that neglect interdependence – even under conditions unfavorable to spatial models.

We therefore advocate that researchers consider S-2SLS as a general, conservative strategy when confronting endogenous predictors and existing theories suggest the possibility of interdependence in the outcome variable. In the conclusion, we discuss some of the implications for the use of IV models in applied research.

2 OLS and Multifarious Endogeneity

In order to better understand the problems that arise from neglecting spatial interdependence in IV estimation, it is useful to first clarify that unmodeled interdependence is itself an omitted variables problem. Consider a simple linear-additive model

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is an n-length vector of outcomes, $\mathbf{x}$ the predictor, and $\mathbf{e}$ the disturbance. The OLS

include instances where at least one of the non-spatial predictors is also endogenous.

[3]We focus on the spatial-autoregressive model (SAR) for two reasons. First, it is the most widely-used spatial model in political science. Second, it is the interdependence in the outcome – as in the SAR – that induces the simultaneity that is at the heart of the problem we discuss. Other models that contain a spatial lag of the outcome and additional features, such as autoregressive disturbances (the SAR-AR model), are extensions of the SAR model and could also be estimated. Drukker, Egger and Prucha (2013) discuss the estimation of a SAR-AR model with an endogenous predictor, which can be estimated using the same software routines we discuss below.

estimator of β is the sample covariance of $\mathbf{x}$ and $\mathbf{y}$ over the sample variance of $\mathbf{x}$,

$$\hat{\beta}_{ols} = \frac{\widehat{\text{cov}}(\mathbf{x}, \mathbf{y})}{\widehat{\text{var}}(\mathbf{x})}. \tag{2}$$

Substituting the right-hand side of equation (1) in for $\mathbf{y}$ yields the probability limit

$$\text{plim}_{n\to\infty} \hat{\beta}_{ols} = \beta + \underbrace{\frac{\text{cov}(\mathbf{x}, \mathbf{e})}{\text{var}(\mathbf{x})}}_{\text{endogeneity bias}}, \tag{3}$$

showing that $\hat{\beta}_{ols}$ is asymptotically unbiased if $\text{cov}(\mathbf{x}, \mathbf{e}) = 0$, that is, if $\mathbf{x}$ is exogenous.[4] This result should be familiar to readers. It is presented in any introductory econometrics textbook along with common sources of bias: confounding due to omitted variables, simultaneity or reverse causality, and measurement error in the variable of interest.

We are concerned with a special case of confounding: unmodeled interdependence between outcomes. Spatial, or cross-sectional, interdependence occurs when a unit's outcome affects the choices, actions, or decisions of other units (Kirby and Ward, 1987; Ward and O'Loughlin, 2002; Beck, Gleditsch and Beardsley, 2006; Franzese and Hays, 2007; Plümper and Neumayer, 2010). Theories of interdependence are "ubiquitous, and often quite central, throughout the substance of political science" (Franzese and Hays, 2007, p. 141): the contagion of conflict and crises, the spread of domestic institutions and ideologies, economic integration and resulting policy coordination, and participation in international agreements all provide examples. Ignoring this spatial interdependence induces cross-sectional correlation in the residuals and, more problematically, covariance between the predictors and the disturbances. As a consequence, coefficient estimates are both inefficient and biased; in the following, we focus on the latter concern.

To distinguish confounding due to spatial interdependence from other sources of endogeneity

[4]When we discuss bias, we refer to asymptotic bias. All IV estimators have small-sample bias.

of $\mathbf{x}$, we decompose the error term in equation (1) as

$$\mathbf{e} = \rho \mathbf{W}\mathbf{y} + \mathbf{u}, \tag{4}$$

where ρ is the effect of outcomes $\mathbf{y}$ in surrounding units j on unit i, weighted by $\mathbf{W}$, an n-by-n connectivity matrix which identifies the relationship between units i and j. As usual in spatial econometrics, we refer to $\mathbf{W}\mathbf{y}$ as the spatial lag, with $\mathbf{W}$ determining which other-unit outcomes y_j are likely to influence the choices, actions, behaviors of unit i.

Then, we can rewrite equation (3) as

$$\text{plim}_{n\to\infty}\, \hat{\beta}_{ols} - \beta = \underbrace{\left[\frac{\text{cov}(\mathbf{x}, \mathbf{u})}{\text{var}(\mathbf{x})}\right]}_{\substack{\text{Non-spatial} \\ \text{endogeneity bias}}} + \underbrace{\rho \left[\frac{\text{cov}(\mathbf{x}, \mathbf{W}\mathbf{y})}{\text{var}(\mathbf{x})}\right]}_{\substack{\text{Spatial} \\ \text{endogeneity bias}}}. \tag{5}$$

Equation (5) separately identifies spatial and non-spatial endogeneity as two potential sources of bias in the OLS estimator.[5] First, bias can result from more familiar, non-spatial sources of endogeneity of $\mathbf{x}$, that is, correlation between $\mathbf{x}$ and $\mathbf{u}$. This is represented by the first term in equation (5), which drops out if $\text{cov}(\mathbf{x}, \mathbf{u})$ is zero. Second, bias can arise from spatial interdependence in $\mathbf{y}$. As indicated by the second term on the right-hand side of equation (5), this bias drops out if $\rho = 0$; that is, when there is no spatial interdependence.[6] In what follows, we show that addressing the former while neglecting the latter fails to recover unbiased estimates of the effect. In many cases, it magnifies the bias relative to ordinary least squares.

3 Spatial Bias in IV Models

Following Sovey and Green (2011), we introduce IV estimation using familiar notation from structural equation models, assuming linear-additive relationships between the variables. Suppose a

[5] This derivation of the bias is only approximate, as $\mathbf{W}$ also increases in n.

[6] It is only when ρ is zero that this term drops out. $\text{cov}(\mathbf{x}, \mathbf{W}\mathbf{y})$ is non-zero, because $\mathbf{W}\mathbf{y}$ is a function of $\mathbf{x}$. While the most obvious solution to address the bias from interdependence may be including $\mathbf{W}\mathbf{y}$ as a variable, this would not be sufficient, because $\mathbf{W}\mathbf{y}$ itself is endogenous in the outcome equation; see, e.g., Franzese and Hays (2007).

suitable instrument $\mathbf{z}$ is available, resulting in the following system of equations:

$$\mathbf{y} = \beta\mathbf{x} + \mathbf{e}, \tag{6}$$

$$\mathbf{x} = \gamma\mathbf{z} + \mathbf{v}. \tag{7}$$

As before, suppose that the disturbance can be decomposed as $\mathbf{e} = \rho\mathbf{W}\mathbf{y} + \mathbf{u}$ and interdependence is ignored in the estimation. Then, non-spatial endogeneity arises if $\mathrm{cov}(\mathbf{u}, \mathbf{v}) \neq 0$ and therefore $\mathrm{cov}(\mathbf{x}, \mathbf{u}) \neq 0$. We assume in the following that the variable $\mathbf{z}$ satisfies the usual assumptions for a valid instrument – $\mathrm{cov}(\mathbf{z}, \mathbf{x}) \neq 0$ and $\mathrm{cov}(\mathbf{z}, \mathbf{u}) = 0$ – such that $\mathbf{z}$ is correlated with the endogenous predictor $\mathbf{x}$ but uncorrelated with the disturbance $\mathbf{u}$.

The IV estimator is obtained as two-stage least squares (2SLS), such that

$$\hat{\beta}_{2sls} = \frac{\mathrm{cov}(\mathbf{y}, \mathbf{z})}{\mathrm{cov}(\mathbf{x}, \mathbf{z})}. \tag{8}$$

Inserting the expression for $\mathbf{y}$ yields

$$\mathrm{plim}_{n\to\infty}\, \hat{\beta}_{2sls} - \beta = \frac{\rho \times \mathrm{cov}(\mathbf{W}\mathbf{y}, \mathbf{z})}{\mathrm{cov}(\mathbf{x}, \mathbf{z})} + \frac{\mathrm{cov}(\mathbf{u}, \mathbf{z})}{\mathrm{cov}(\mathbf{x}, \mathbf{z})}, \tag{9a}$$

$$= \underbrace{\rho \left[\frac{\mathrm{cov}(\mathbf{W}\mathbf{y}, \mathbf{z})}{\mathrm{cov}(\mathbf{x}, \mathbf{z})} \right]}_{\substack{\text{Spatial} \\ \text{endogeneity bias}}}, \tag{9b}$$

which shows that, by assumption, 2SLS does not suffer from the non-spatial endogeneity bias of OLS: because $\mathrm{cov}(\mathbf{u}, \mathbf{z}) = 0$ and $\mathrm{cov}(\mathbf{x}, \mathbf{z}) \neq 0$, the second term on the right-hand side of equation (9a) disappears. This result, of course, is well appreciated and motivates the use of 2SLS where $\mathbf{x}$ is suspected to be endogenous.

Less appreciated is that 2SLS is biased in the presence of (ignored and hence unmodeled) interdependence. In short, the instrument violates the exclusion restriction, because it is related to the outcome disturbances via the omitted interdependence term $\mathbf{W}\mathbf{y}$. To see why, note that after substituting and rearranging terms, equation (6) can multiplied through by $\mathbf{W}$ and written as

$$\mathbf{Wy} = \mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}[\beta\gamma\mathbf{z} + \beta\mathbf{v} + \mathbf{u}]. \tag{10}$$

That is, we can re-express the spatial lag, $\mathbf{Wy}$, in terms of the spatially weighted instrument $\mathbf{z}$ and stochastic terms $\mathbf{u}$ and $\mathbf{v}$. Substituting this expression into the definition of the spatial bias in 2SLS and rearranging, we obtain

$$\text{plim}_{n\to\infty}\,\hat{\beta}_{2sls} - \beta = \beta\left[\rho\frac{\text{cov}(\mathbf{Wz},\mathbf{z})}{\text{var}(\mathbf{z})}\right] + \beta\sum_{k=2}^{\infty}\left[\rho^k\frac{\text{cov}(\mathbf{W}^k\mathbf{z},\mathbf{z})}{\text{var}(\mathbf{z})}\right]. \tag{11}$$

2SLS is biased unless the terms on the right-hand side are zero. For clarity in the following exposition, we have split the bias into two terms – the first representing the first-order bias and the second representing higher-order terms. Both terms disappear if $\rho = 0$, such that no interdependence exists. If interdependence in the outcome does exist, such that $\rho \neq 0$, however, 2SLS is biased.

Notably, this bias persists even when $\mathbf{z}$ is randomly assigned and, therefore, independently distributed and otherwise exogenous. It is in this case that the two-term expression of the bias in equation (11) becomes useful. When $\mathbf{z}$ is independently distributed, the first term drops out, because independence in $\mathbf{z}$ implies that any specification of $\mathbf{W}$ yields $\text{cov}(\mathbf{Wz},\mathbf{z}) = 0$.[7] That is, the value of $\mathbf{z}$ on unit i is uncorrelated with the value of $\mathbf{z}$ on any other unit (and their weighted-sum $\sum_j w_{ij}z_j$). However, this is not true of the second term in equation (11). While $\mathbf{W}$ is a hollow matrix – all elements along the diagonal equal zero – higher-order multiples of $\mathbf{W}$ are not hollow matrices as ties between units are not uni-directional.[8] Because $\mathbf{W^k}$ has non-zero diagonal elements, it follows that $\mathbf{W^kz}$ is, for unit i, a function of z_i, and therefore correlated with $\mathbf{z}$,

[7] Recall that $\mathbf{W}$ is the connectivity matrix of the outcome – based on, e.g., contiguity, neighbors, or inverse distance – defining how y_i is related to all $y_{j\neq i}$. In connectivity matrices like $\mathbf{W}$ the diagonal elements are always zero, that is, you can not be a direct neighbor of yourself.

[8] If spatial ties were unidirectional – $\mathbf{W}$ is upper- or lower-triangular – the higher-order multiples would remain independent of z_i. However, *inter*dependence generally rules out unidirectional ties. The importance of reciprocal relationships between units – i.e., interdependence – for our results can also be seen in the contrast to temporal dependence. With temporal dependence, the current value of the outcome is a function of past values of the outcome, but past outcomes are not a function of the current value. Hence, a randomly assigned instrument poses no problems under temporal dependence.

regardless of the distribution of $\mathbf{z}$.

To gain more intuition for why this is the case, recall that $\mathbf{W}$ can be thought of as defining 'neighbors': non-zero entries indicate which units on the outcome variable are related to one another. Then, for each unit, the respective row of $\mathbf{W}$ defines a set of neighbors. Heuristically, powers of $\mathbf{W}$ then represent neighbors-of-neighbors. For example, the i^{th} row of $\mathbf{W}^2$ indicates i's neighbors' neighbors. This is important because, intuitively, a unit always is a neighbor of its own neighbors. Consequently, if $\mathbf{W}$ links unit i to j and unit j to unit i, then $\mathbf{W}^2$ (and higher powers of $\mathbf{W}$) links unit i back to itself. Therefore, even under independence of $\mathbf{z}$, some $\mathbf{W}^k\mathbf{z} \not\perp \mathbf{z}$ as long as $\mathbf{W}$ is non-triangular. Put simply, even if unit i is not related to any of the neighbors defined by $\mathbf{W}$, unit i is always related to itself through these higher powers of $\mathbf{W}$.

That is, for $\rho \neq 0$, any instrument that is randomly assigned is (only) first-order unbiased, providing a lower bound on the spatial bias. While the bias is relatively mild, spatial interdependence on the outcome variable renders IV models biased, even under conditions most favorable to IV models, such as experimental or quasi-experimental designs.

RESULT 1 *With unmodeled spatial interdependence in the outcome, 2SLS is asymptotically biased.*

However, the instruments often used in practice are not independently distributed, risking greater bias still. Specifically, the more the values z_i are similar to neighboring values $z_{j\neq i}$ (where neighboring values are defined by $\mathbf{W}$, the matrix defining relationships among units for the outcome), the greater the bias will be: the first term in equation (11) no longer drops out, and all of the terms in the expression increase in magnitude.

To understand this result, it helps to think of 2SLS broken down into two stages. The first stage is a regression of the endogenous predictor $\mathbf{x}$ on the instrument $\mathbf{z}$, which yields fitted values $\widehat{\mathbf{x}}$. The second stage is a regression of the outcome variable $\mathbf{y}$ on the fitted values $\widehat{\mathbf{x}}$. We make two observations. First, if $\mathbf{z}$ follows a spatial distribution, the projected values $\widehat{\mathbf{x}}$ inherit some of that spatial pattern. Second, in a regression with an (erroneously) omitted spatial lag, the bias in coefficient estimates is reinforced for variables that have a spatial distribution similar to that of the

outcome (see, e.g., Franzese and Hays 2007). It follows that the bias in 2SLS becomes most severe if the fitted values $\widehat{\mathbf{x}}$ have a spatial distribution similar to that of the outcome – which, in turn, is the case if the instrument has a spatial distribution similar to the outcome.

It is not crucial that the instrument and the outcome follow identical spatial patterns, merely that the instrument and the outcome have some similarity in their spatial patterns. That is, the bias in 2SLS increases if the $\mathbf{W}$ that characterizes the relationships in $\mathbf{y}$ also partially characterizes the relationships in $\mathbf{z}$. In practice, when considering the extent of the spatial bias in 2SLS, one can therefore remain largely agnostic about the nature of the spatial relationships on the instrument – in particular, it is not necessary to determine whether $\mathbf{z}$ and $\mathbf{y}$ are truly governed by identical $\mathbf{W}$s or even which specific $\mathbf{W}$ applies to the instrument (and our empirical approach, detailed in the next section, is consistent with this view). Our point is much simpler: if the outcome is spatially interdependent, then the bias in 2SLS will be more severe for instruments with spatial patterns similar to that of the outcome.

These concerns apply to a large set of common instruments. Researchers often draw on geographic, meteorologic, or economic variables, such as natural disasters (Ramsay, 2011), rainfall data (Hansford and Gomez, 2010), or commodity price shocks (Ahmed, 2012), where spatial dependence among units is likely – natural disasters, rainfall, and price shocks do not stop at territorial borders. The same problem arises for instruments that are measured at a higher level of aggregation than the endogenous predictor. If, for instance, the instrument is based on regional political or institutional shocks, such as waves of democratization (Stasavage, 2005) or membership in international institutions in neighboring countries (Büthe and Milner, 2008), the instrument induces spatial correlation in the projection $\widehat{\mathbf{x}}$ by construction: the value of the instrument is identical or nearly identical for each of the lower-order observations within the cluster. Since many outcome variables of interest in political science also cluster regionally – e.g., democratization, economic growth, or policy – regional-level instruments are likely to reinforce the bias in 2SLS.

To illustrate, consider the use of meteorological variables as instruments for democratization ($\mathbf{z}$) in models of economic development ($\mathbf{y}$). Contiguous states (a widely used $\mathbf{W}$) likely have

both similar levels of development ($\mathbf{y}$) and common weather patterns ($\mathbf{z}$), where the former implies $\rho > 0$ and the latter implies $\text{cov}(\mathbf{Wz}, \mathbf{z}) > 0$. It is under these conditions that the bias will be most severe; as can be seen in equation (11), the bias increases in the strength of the interdependence in the outcome (ρ) and the strength of the spatial dependence in the instrument ($\text{cov}(\mathbf{Wz}, \mathbf{z})$).

RESULT 2 *With unmodeled spatial interdependence in the outcome, the more similar are the spatial distributions of the instrument and the outcome, the greater is the bias in 2SLS.*

We add three additional observations. First, these biases are usually inflationary, which can be seen from equation (11). The bias terms are multiplied by powers of ρ, which is positive in most applications (Franzese and Hays, 2007). And, if $\mathbf{z}$ is governed by a similar pattern of spatial dependence as the outcome, the covariances between $\mathbf{W^k z}$ and $\mathbf{z}$ are non-negative. Consequently, the right-hand side of equation (11) should have the same sign as β and be proportional to β. Thus, in most applications the bias in 2SLS that arises from spatial interdependence exaggerates the true parameter value – where β is negative, 2SLS produces smaller coefficient estimates, and where β is positive, 2SLS produces larger coefficient estimates.

Second, the spatial bias induced from the instrument can exceed the spatial bias in ordinary least squares. Consider the relative spatial bias of OLS (the left-hand side) and 2SLS (the right-hand side):

$$\frac{\text{cov}(\mathbf{Wy}, \mathbf{x})}{\text{var}(\mathbf{x})} \lessgtr \frac{\text{cov}(\mathbf{Wy}, \mathbf{z})}{\text{cov}(\mathbf{x}, \mathbf{z})}. \tag{12}$$

To focus on the comparison of the spatial bias between 2SLS and OLS, suppose that no non-spatial endogeneity exists. Re-expressing both terms, condition (12) becomes

$$\sum_{k=1}^{\infty} \left[\rho^k \frac{\text{cov}(\mathbf{W}^k\mathbf{x}, \mathbf{x})}{\text{var}(\mathbf{x})} \right] \lessgtr \sum_{k=1}^{\infty} \left[\rho^k \frac{\text{cov}(\mathbf{W}^k\mathbf{z}, \mathbf{z})}{\text{var}(\mathbf{z})} \right]. \tag{13}$$

Simply put, differences in the spatial distribution of the instrument and the endogenous variable inform the relative degree of spatial bias. This is similar to Bartels's (1991) recognition that,

because $\mathbf{x}$ can be considered its own instrument, when using an invalid instrument $\mathbf{z}$ the gains relative to OLS are a function of the relative difference in how $\mathbf{z}$ and $\mathbf{x}$ covary with the disturbance of $\mathbf{y}$. Again thinking of the second stage in 2SLS as a regression of $\mathbf{y}$ on the projection $\widehat{\mathbf{x}}$ further clarifies the role of spatial dependence in the instrument: the bias of 2SLS relative to OLS increases as the spatial distribution of the instrumented predictor, $\widehat{\mathbf{x}}$, becomes more similar to the spatial distribution of the outcome than the original predictor, $\mathbf{x}$. Then, IV models augment the spatial bias, because $\widehat{\mathbf{x}}$ is more similar to the omitted spatial lag than $\mathbf{x}$ is. The reverse, of course, also holds: if the instrument is randomly assigned, then the similarity between the spatial pattern of the instrumented predictor, $\widehat{\mathbf{x}}$, and the outcome decreases, and the bias of 2SLS relative to OLS declines. Nonetheless, even in that case, as we emphasize in Result 1, 2SLS remains biased.

Finally, because spatial and non-spatial endogeneity biases may attenuate or reinforce each other, ignoring spatial interdependence in the outcome risks unpredictable and possibly greater overall bias than OLS. When the endogenous variable, $\mathbf{x}$, is spatially less clustered than the instrument, $\mathbf{z}$, the severity of the difference in the spatial biases may be sufficiently large to surmount the gains from addressing non-spatial endogeneity. And because the spatial and non-spatial bias may have different directions, resolving one of the biases may easily produce results further from the truth than resolving none. Perhaps most problematically, these offsetting effects mean that the OLS and 2SLS estimates will not even be sufficient to obtain bounds on the true parameter value.

4 Spatial Models with Additional Endogenous Predictors

What, then, can researchers concerned with endogeneity in a key predictor and spatial interdependence in the outcome do? The solution is actually quite simple: estimate a modified instrumental variables model. While early work in spatial econometrics assumed exogenous predictors, methods for estimating models with additional endogenous predictors have become increasingly common (Kelejian and Prucha, 2004; Anselin and Lozano-Gracia, 2008; Fingleton and Le Gallo, 2008; Drukker, Egger and Prucha, 2013; Liu and Lee, 2013). To date, however, these models have not received much attention in applied spatial work in political science, and even less so

in contexts where researchers are not theoretically interested in spatial relationships. In short, to redress the concerns above, researchers need to account for the spatial interdependence of the outcome. Yet, including a spatial lagged-outcome produces a system of simultaneously-determined, non-separable equations. That is, $\mathbf{Wy}$ is itself is an endogenous predictor, no different than a simultaneously-determined $\mathbf{x}$. Consequently, in spatial modeling, researchers exploit the same strategies generally used when confronting endogenous predictors (such as maximum likelihood, 2SLS, or GMM). As such, one can simply extend the familiar IV framework, applying it to account for outcome interdependence and predictor endogeneity. As this is simply a special case of multiple endogenous variables, a spatial-two stage least squares (S-2SLS) model can be estimated as in standard IV analysis: instrumenting for $\mathbf{Wy}$ and $\mathbf{x}$ simultaneously.

While other solutions are also available – e.g., purging the spatial dependence of the outcome equation via eigenvector filtering – we prefer S-2SLS for several reasons.[9] First, the mechanics of estimating this model are already familiar to researchers using IV estimation for an endogenous variable $\mathbf{x}$, because the estimator, 2SLS, is the same. Second, S-2SLS nests the non-spatial IV model a researcher would have otherwise estimated. Rather than restrict ρ – the spatial effect – to be zero by assumption, as in 2SLS, S-2SLS allows researchers to explicitly test this. As we demonstrate in simulations, this nesting helps ensure that – even if no spatial interdependence is present and $\rho = 0$ – the model recovers the same estimates as the original 2SLS, with only minimal efficiency loss due to the additional parameter. Finally, the S-2SLS model, as well as several extensions, can be estimated in both Stata (`spivreg`) and R (`sphet`).

The only practical hurdles to estimating a S-2SLS are in the specification stage: i) what are appropriate instruments for the spatial lag, and ii) what is the appropriate connectivity matrix $\mathbf{W}$ for the outcome variable. The first, instrument selection, is comparatively simple. While instruments for the endogenous predictor usually require finding additional exogenous variables,

[9]Limited and full information estimators allowing for both spatial and non-spatial endogeneity have been established, with Kelejian and Prucha (2004) the first to derive formal large sample results; see also Drukker, Egger and Prucha (2013) for a GMM estimator. Franzese, Hays and Cook (2016) discuss the complications of modeling spatial interdependence in discrete-choice models.

instruments for the spatial lag can typically be found from transformations to the existing data. Specifically, spatial lags of the exogenous predictors serve as instruments for the spatial lag of the outcome. To see the basic intuition for this, just multiply $\mathbf{W}$ by both sides of the simple linear-additive model – i.e., $\mathbf{y} = \beta\mathbf{x} + \mathbf{e} \Rightarrow \mathbf{Wy} = \beta\mathbf{Wx} + \mathbf{We}$. Just as $\mathbf{x}$ is related to $\mathbf{y}$, $\mathbf{Wx}$ is related to $\mathbf{Wy}$, the spatial lag.[10]

The second practical hurdle, the selection of $\mathbf{W}$, is already familiar to researchers with exposure to spatial models. For those less familiar, we briefly sketch out the basics. To undertake spatial econometric modeling, researchers must pre-specify how units are related to one another (i.e., the network). Geographic proximity (e.g., contiguity) is commonly used, though researchers should specify connections that are most theoretically appropriate for their data. These relational measures for 'space' are then supplied to the model as the elements in $\mathbf{W}$ – an n-by-n connectivity matrix which identifies the relationship between units i and j. S-2SLS clearly performs best when $\mathbf{W}$ reflects the true network, yet gains are still likely even when researchers do not have full information on the ties between units. First, in the worst-case (and unlikely) scenario that a researcher completely mischaracterizes $\mathbf{W}$, this would still do no worse in expectation than 2SLS – S-2SLS recovers a zero estimate of ρ due to misspecified $\mathbf{W}$, while 2SLS does so by assumption. Second, due the high correlation across different possible network structures, even a mis-specified $\mathbf{W}$ has power against the truth (LeSage and Pace, 2014). We revisit this concern in the simulated experiments in the next section.[11]

[10] A more complete derivation can be seen by noting that the reduced form of the spatial-lag model discussed in section 3,

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}[\mathbf{x}\beta_x + \mathbf{u}],$$

can be re-expressed using an infinite series and multiplied through by $\mathbf{W}$ to produce

$$\mathbf{Wy} = \mathbf{Wx}\beta_x + \rho\mathbf{W}^2\mathbf{x}\beta_x + \rho^2\mathbf{W}^3\mathbf{x}\beta_x + \ldots + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u},$$

indicating how spatial lags of $\mathbf{x}$ (and their higher-order powers) effectively instrument for $\mathbf{Wy}$.

[11] Note that the researcher need not specify the spatial distribution of the instrument. $\mathbf{W}$ specifies connections among units with respect to the outcome variable. As we highlighted in the previous section, the extent of the bias in 2SLS depends on the similarity in the spatial pattern between the instrument and the outcome. But it is not necessary to determine the specific spatial pattern of the instrument.

Once specified, estimation of the S-2SLS model proceeds without additional complications. Because S-2SLS is estimated via the 2SLS estimator, it inherits the asymptotic and small-sample properties of 2SLS (including consistency, but finite sample bias and the sensitivity to weak instruments).[12] Similarly, standard variance estimators – robust to heteroskedasticity or non-independence, for instance – are easily applicable. We demonstrate the gains that can be realized from S-2SLS in the following sections.

5 Simulation Experiments

To assess the performance of OLS, 2SLS, and S-2SLS, we undertake a series of Monte Carlo experiments with varying levels of spatial and non-spatial endogeneity. In particular, the data for our simulations are generated as follows:

$$\mathbf{y} = (\mathbf{I} - \rho_y \mathbf{W})^{-1}[\mathbf{x}\beta + \lambda_1 \mathbf{Q} + \mathbf{u}_1] \tag{14a}$$

$$\mathbf{x} = \gamma \mathbf{z} + \lambda_2 \mathbf{Q} + \mathbf{u}_2, \tag{14b}$$

$$\mathbf{z} = (\mathbf{I} - \rho_z \mathbf{W})^{-1}\mathbf{v}, \text{ where } \mathbf{v} \sim N(0, 1) \tag{14c}$$

where $\mathbf{y}$ is the outcome, $\mathbf{x}$ is the endogenous predictor, $\mathbf{Q}$ is a matrix of exogenous predictors, $\mathbf{W}$ is a row-standardized connectivity matrix, and $\mathbf{z}$ is the instrument.[13] Consistent with our discussion above, we only consider the consequences of spatial interdependence in $\mathbf{y}$ and $\mathbf{z}$, which are the key

[12] For an approach addressing weak instruments in IV models, see Betz (2013). Drukker, Egger and Prucha (2013) and Liu and Lee (2013) allow for both additional residual spatial error autocorrelation and/or heteroskedasticity. These extensions are GMM-plus-IV, implemented in Stata's spreg. While we do not discuss this at length here, the first step is the S-2SLS we present, which provides the initial, consistent estimates of the spatial interdependence in the outcome that can then be used in the second step estimation of the error autocorrelation, with successive iteration over both steps until convergence is obtained.

[13] Locations for the units are generated by twice taking n draws from a standard uniform, with the combined results producing xy-coordinate points. Connections between the units are then generated using a k-Nearest Neighbor algorithm with $k = 5$, returning a binary n-by-n matrix with each element in a row coded as 1 for the five closest units or 0 for all others (including zeros along the diagonal). The matrix is then row-standardized.

attributes for bias in 2SLS.[14]

The extent of spatial interdependence in the outcome and the instrument is given by parameters ρ_y and ρ_z, respectively, with larger values of ρ_y and ρ_z resulting in greater spatial interdependence in $\mathbf{y}$ and $\mathbf{z}$. We do not vary the specification of the $\mathbf{W}$ that governs the spatial pattern of $\mathbf{y}$ and $\mathbf{z}$, respectively. Non-spatial endogeneity is induced through draws of $(u_1, u_2)^T = N(0, \Sigma)$, where Σ is the covariance matrix of a bivariate normal random variable. We decompose Σ such that we can specify the correlation (δ) between u_1 and u_2 directly. We vary δ to induce different degrees of non-spatial endogeneity. If $\delta = 0$, $\mathbf{x}$ is exogenous and OLS (or standard spatial) models should be preferred. With non-zero δ and non-zero ρ_y, the assumptions of neither OLS nor 2SLS hold.

This setup allows us to consider various scenarios that correspond to our results above. ρ_y = ρ_z = 0 produces the standard IV model with an i.i.d. instrument, such that 2SLS should perform well. $\rho_y \neq 0$ but $\rho_z = 0$ implies interdependence in the outcome but an i.i.d. instrument. Following Result 1, we should still observe some bias in 2SLS in this scenario, whereas S-2SLS should perform better. As ρ_z increases, the bias in 2SLS should increase, both in absolute terms (Result 2) and relative to OLS, because the instrument becomes more similarly distributed to the outcome relative to the predictor. Finally, varying δ, the extent of non-spatial endogeneity, allows us to evaluate scenarios under which OLS – which produces spatial and non-spatial endogeneity – should perform worse than 2SLS – which produces only spatial endogeneity.

The remaining parameters $\{\beta, \gamma, \lambda_1, \lambda_2\}$ are the coefficients of the predictors of $\mathbf{x}$ and $\mathbf{y}$, respectively.[15] Our main focus is on the estimate of β, which we hold constant across experiments at 2. Table 1 shows the different parameter values which we use to create simulated data sets. There are 108 different combinations of the parameters shown in Table 1 (with the bolded values indi-

[14] As discussed above, the relative spatial pattern of $\mathbf{x}$ and $\mathbf{z}$ only matters for the performance of 2SLS relative to OLS. For the simulations, to illustrate Result 2, we only consider scenarios where 2SLS performs relatively poorly due to the spatial pattern in $\mathbf{z}$.

[15] In the first stage, we specify the intercept as 2. The two exogenous predictors have coefficients 3 and -2.5. For the second stage the intercept is -2 and the exogenous predictors are -3 and 2.5. For the plots presented in the manuscript we set the coefficient on the instrument to $\gamma = 1.5$ and the number of observations to $n = 200$. Observations for the predictors are drawn from standard normal distributions.

cating those used in the subsequent figures). For each combination we generate $1,000$ data sets, which results in a total of $108,000$ simulation runs. On each data set we estimate β using OLS, 2SLS, and our preferred method, S-2SLS.

Table 1: Parameter Values for Simulations

n	50	**200**	
ρ_y	**0**	**0.3**	**0.6**
ρ_z	**0**	**0.3**	**0.6**
γ		0.75	**1.5**
δ	**-0.5**	**0**	**0.5**

Note: Bold values used in Figures 2 & 3.

The results are presented in Figures 2 and 3, which report the median absolute error and coverage probabilities for the estimators, respectively.[16] The figures vary along three dimensions. First, δ – the non-spatial endogeneity – increases across the three rows from -0.5 in the top row, to 0 in the middle row, to 0.5 in the bottom row. Second, each column shows results for a different value of ρ_z – the spatial pattern of the instrument – ranging from 0 in the column on the left over 0.3 in the middle to 0.6 on the right. Finally, within each individual plot, ρ_y – the spatial interdependence in the outcome – increases from left to right across the x-axis.

[16]We prefer MAE as it limits the influence of potential outliers. In the Online Appendix we also present model performance in RMSE terms.

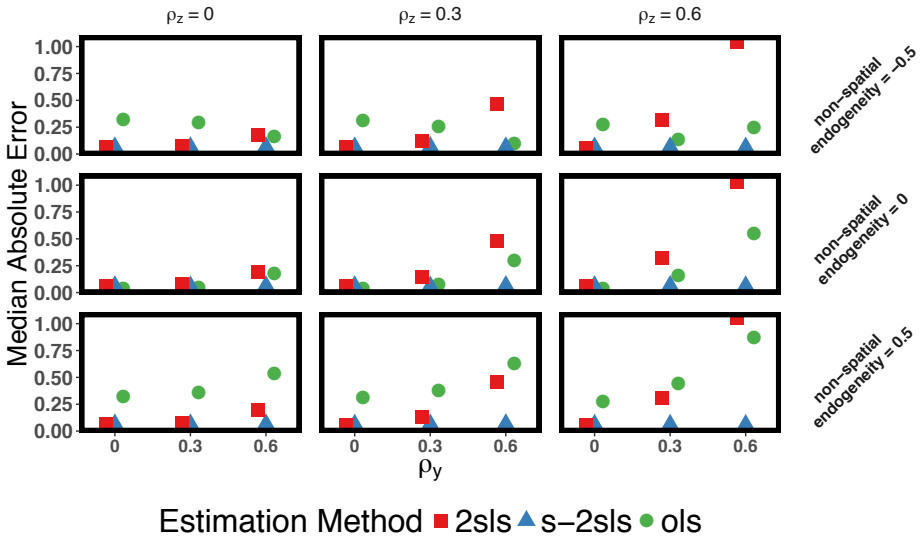


Figure 2: Median Absolute Error. Rows: δ, non-spatial endogeneity. Columns: ρ_z, spatial pattern of the instrument. Horizontal axis within each plot: ρ_y, spatial interdependence in the outcome. Vertical axis within each plot: MAE.

Several observations stand out from the plots. Turning to the median absolute error (MAE) in Figures 2 first, across all levels of non-spatial endogeneity (δ), the error of 2SLS grows as ρ_y increases, dramatically so as ρ_y *and* ρ_z increase together. This is consistent with our theoretical results: under interdependence in the outcome, the 2SLS model always returns biased estimates (Result 1), with the severity of these biases increasing in the similarity of the spatial pattern in the instrument and the outcome (Result 2). Importantly, when both the instrument and outcome are characterized by spatial dependence, a situation that in our view is not uncommon in the literature, the bias in 2SLS increases quickly. Conversely, the MAE of S-2SLS is stable, as its performance does not suffer under high interdependence in $\mathbf{y}$, $\mathbf{z}$, or both. In fact, S-2SLS weakly dominates 2SLS, besting it when spatial interdependence is present and matching it when there is not. Thus, when non-spatial endogeneity is present and IV models may be warranted, S-2SLS performs better than or as good as 2SLS. Across all scenarios considered in the simulations, 2SLS performs better in terms of MAE only when $\rho_y = 0$, and even then the maximum difference in median absolute error between 2SLS and S-2SLS is 0.03. While not surprising, this bolsters our claim that S-2SLS

is a useful conservative specification, robust under non-spatial and spatial endogeneity, because it nests both cases.

The OLS estimator performs poorly when either non-spatial or spatial endogeneity is present. However, and as discussed above, the bias can be larger for 2SLS than for OLS, even in the case of strong non-spatial endogeneity, where OLS should perform poorly. This occurs under higher levels of ρ_z and ρ_y – as we move from the left to the right in each box, and as we move from the left column to the right column – where the spatial and, in turn, total bias of 2SLS is greater due to the spatial interdependence of the instrument.

The top and middle rows of Figure 2 present two particularly interesting scenarios. In the top row, with negative non-spatial endogeneity and positive spatial interdependence, the relative performance of OLS improves, both in absolute terms and relative to 2SLS, as the spatial interdependence increases. The two biases are countervailing, combining to produce a result closer to the truth. Under these conditions, 2SLS produces relatively worse results, as it addresses one type (and therefore direction) of bias, while neglecting the other. As a result, 2SLS produces *more* biased estimates even while – in fact, due to – addressing one of the sources of that bias.

In the middle row, we have no non-spatial endogeneity bias, and relying on 2SLS is unnecessary. Usually, using 2SLS instead of OLS is not much of a concern, aside from a slight efficiency loss. This changes with interdependence. If the instrument is spatially more similar to the outcome than the predictor (as in the second and third column), 2SLS produces more total bias than OLS. In this case, 2SLS not only was unnecessary, but results in worse estimates than OLS. (Of course, this result hinges on the simulation setup, which consistent with our discussion allowed for a spatial pattern in $\mathbf{z}$ but not in $\mathbf{x}$ – if the reverse was the case, 2SLS would perform relatively better.)

These results are particularly problematic, as researchers relying on 2SLS over OLS estimates will be more confident about results that are further from the truth and dismissive of results that were closer to it. Frequently, a difference between 2SLS and OLS estimates is accepted as evidence of suspected non-spatial endogeneity (such as measurement error or reverse causality) that was successfully removed by 2SLS. While 2SLS removes non-spatial endogeneity, such argu-

ments ignore that 2SLS may come with biases of its own, and that these biases need not be less pronounced than the biases in OLS. Where outcomes are interdependent, there is no guarantee that 2SLS produces better estimates than OLS. S-2SLS, by contrast, does not confront this issue and consistently outperforms both OLS and 2SLS.

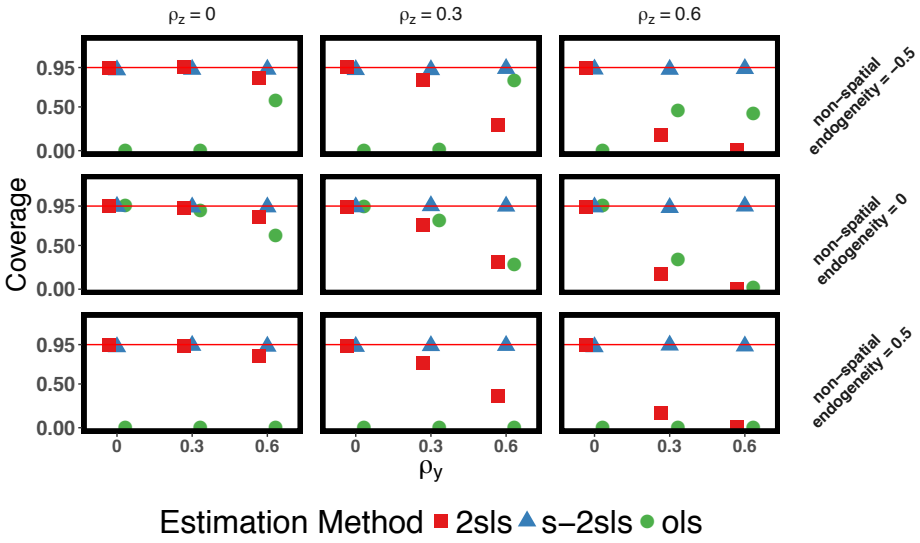


Figure 3: Coverage Probabilities. Rows: δ, non-spatial endogeneity. Columns: ρ_z, spatial pattern of the instrument. Horizontal axis within each plot: ρ_y, spatial interdependence in the outcome. Vertical axis within each plot: Coverage rate.

Figure 3 shows the coverage probabilities for each estimator. The coverage statistic measures the share of estimates for which the true parameter falls within the 95% confidence interval of the estimate. If perfectly calibrated, we would expect this to be true for 95% of cases. The results are generally consistent with our expectations. First, the coverage of OLS is generally poor under either spatial or non-spatial endogeneity. However, for the reason just discussed, when the spatial and non-spatial bias are oppositely signed (top row), the coverage of OLS improves with higher spatial interdependence. Second, with interdependence in the outcome, the 2SLS estimator under-covers, with the severity of this increasing ρ_z. Finally, S-2SLS has very good coverage throughout and is not affected by interdependence in $\mathbf{z}$ or $\mathbf{y}$. In fact, the coverage of S-2SLS is consistently

around 95%, ranging between 92% and 96%.

Robustness Checks

In the Online Appendix we discuss a series of additional experimental conditions. First, what if the researcher has incomplete information on the spatial network (i.e. the $\mathbf{W}$ matrix)? To evaluate this, we undertake additional experiments where we vary the level of misspecification – from no error to total error – of the spatial network used in estimation. As Figure A.1 and Figure A.2 in the Online Appendix shows, S-2SLS weakly dominates 2SLS. This demonstrates what we articulated earlier: because S-2SLS nests 2SLS, it only suffers minor efficiency losses when it is the incorrect model. Next, we explore the consequences of a weak instrument (i.e., $\gamma = 0.75$). As expected, IV methods perform worse, yet the overall order in performance between the different methods does not change. Finally, we evaluate how the performance of the estimators vary with changes to sample size. All the results presented in text hold.

6 Application

To illustrate how failing to account for spatial interdependence when using IV models can induce bias in published research, we replicate Ramsay's (2011) "Revisiting the Resource Curse: Natural Disasters, the Price of Oil, and Democracy."[17] A long-standing literature in political science has considered the effects of natural resource revenues on political order. Ramsay (2011) identifies reverse causality as one of the main threats to inference: changes in resource revenues may cause political change, but politics may also affect resource revenues.

The main independent variable of interest is a country's annual oil income per capita (specifically the price of crude oil times annual production divided by the population). The dependent variable is a country's level of democracy, measured as a normalized score of Polity IV. A valid instrument would have sufficient power to explain oil revenues; and fulfill the exclusion restriction such that it only affects changes in democracy via the path through oil revenues. In light of these

[17] In the online Appendix we provide an additional replication of Ashraf and Galor's article "Dynamics and Stagnation in the Malthusian Epoch" (2011).

requirements, Ramsay (2011) introduces out-of-region natural disasters as the instrumental variable (where regions are defined as Europe, Middle East, North Africa, sub-Saharan Africa, Asia, or the Americas). The rationale is that natural disasters, by reducing oil production in the affected countries, change world oil prices, and therefore oil revenues of individual countries; at the same time, natural disasters should have no direct effect on oil production in remote countries.

We highlight three concerns with these IV models. First, levels and changes of democracy and changes cluster in space (Gleditsch and Ward, 2006). Second, natural disasters, the instrument of choice, likely correlate in space. As Ramsay (2011) notes, the effects of disasters are likely to spill over, affecting neighboring states. Finally, Ramsay (2011) aggregates the variable to the regional level induces a spatial pattern by construction. By designing the instrument as "out of region disaster damage estimates" (Ramsay, 2011, 514), all countries within each of the five regions have the same value on the instrument, thus inducing spatial correlation in the instrument by design. As we discussed above, with a spatially interdependent outcome and instrument, we generally expect inflationary bias in the non-spatial IV estimator.

The results from our analysis are presented in Table 2.[18] First, we estimate a linear model (via OLS), with our results (see Model 1) reproducing those found in Ramsay (2011).[19] Next, we determine whether these findings hold once we account for spatial interdependence. Before undertaking spatial analysis we need to select the connectivity matrix, $\mathbf{W}$.[20] Here, we use a row-standardized geographic binary contiguity matrix, as these are widely used in the literature.[21] Given $\mathbf{W}$, we estimate a SAR model (Model 2), which returns a significant value for the spatial effect parameter ($\rho = 0.173$; p-value < 0.05). Oil income per capita is still negative and statistically significant.[22]

[18] As in Ramsay (2011) the predictors include log oil income per capita, GDP per capita, GDP growth, a lagged polity variable, and year fixed effects

[19] This is Model 4 in Ramsay (2011).

[20] A detailed discussion on $\mathbf{W}$ selection is beyond the scope of this paper, interested readers should consult Neumayer and Plümper (2016).

[21] We are unable to merge 4 observation to the shapefiles and therefore drop these from all models.

[22] Due to the non-linear nature of the SAR model, the average direct effect – that is, the average effect of a one-unit change in x_i on y_i – is not the coefficient estimate, but instead: $\mathrm{Tr}\{(\mathbf{I} - \rho\mathbf{W})^{-1}\beta_x\}/n$. For the variable of logged oil income per capita, this results in a value of -0.042, slightly smaller than the OLS effect estimate of -0.046.

This indicates that there does appear to be spatial interdependence in the model, so now we consider the consequence for this on IV estimator. In Model 3 of Table 2, we replicate the 2SLS model in Ramsay (2011).[23] Here the estimated coefficient of log oil income is -0.36 and statistically significant. That is, the 2SLS model presents an effect estimate that is almost *eight times larger* than the original OLS estimates. Finally, we estimate our preferred S-2SLS model, with the results given in Model 4 of Table 2.[24] The instrumented coefficient of logged oil income is now estimated to be -0.088. That is, while the effect is still significant and in the expected direction, its magnitude is much smaller than in the 2SLS model that ignores spatial interdependence.[25] Furthermore, we see substantial efficiency gains in the estimate – as indicated by the standard errors – once we account for spatial interdependence.

In sum, failing to account for spatial interdependence resulted in substantial inflationary bias in the estimates of interest. We do not overturn the central finding presented in Ramsay (2011) that oil revenue is negatively associated with the polity score, but the magnitude of the effect is reduced considerably and the purported gains from IV estimation are significantly reduced.

Conclusion

IV models are now a frequently used tool in political science research. IV methods are especially common in observational research, where endogeneity often threatens causal inference. However, observational data is also where concerns of spatial interdependence are the most salient and where instruments are unlikely to be randomly assigned. Consequently, IV methods are most widely used where the biases due to unmodelled outcome interdependence discussed above are the most likely. This problem may be especially pronounced in published research as researchers are disproportionately likely to publish IV research where gains over OLS are the most pronounced – as can and

[23] In Ramsay (2011) this is presented in column 4 of Table 3.

[24] $\mathbf{W}$ is the same as in Model 2.

[25] Table D.1 in the Appendix displays the results comparing the 2SLS model and the S-2SLS model for the robustness checks presented in Table 5 in Ramsay (2011). Again, the differences between 2SLS and S-2SLS are stark and reflect the same pattern as shown in Table 2.

Table 2: Replication of OLS and IV results Table 1 & 3 in Ramsay (2011)

	(1) OLS Replication	(2) SAR Spatial	(3) 2SLS Replication	(4) S-2SLS Spatial
Log oil income per capita	-0.0457*** (0.00575)	-0.0420*** (0.00354)	-0.358** (0.167)	-0.0878*** (0.0114)
Log GDP per capita	0.0653*** (0.00877)	0.0662*** (0.00533)	0.356** (0.155)	0.108*** (0.0115)
GDP growth	-0.00363*** (0.00106)	-0.00391*** (0.000921)	-0.0118** (0.00518)	-0.00499*** (0.00102)
Polity at entry	0.666*** (0.0280)	0.658*** (0.0165)	-0.00517 (0.373)	0.564*** (0.0288)
Constant	-0.171*** (0.0518)	-0.237*** (0.0499)	-0.962** (0.439)	-0.330*** (0.0591)
Spatial ρ_y		0.173***		0.119***
Observations	1263	1263	1263	1263
Year dummies	Yes	Yes	Yes	Yes

W matrix for spatial models based on contiguous neighbors.

Instrumental variable: out-of-region natural disasters.

Table shows coefficient estimates, standard errors in parentheses

$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

will occur if outcome interdependence is present and unaddressed.[26] We discuss a simple strategy researchers should employ to avoid these biases: S-2SLS. This estimation strategy offers few complications for researchers already pursuing IV methods, inherits the properties of 2SLS familiar to those using IV models, and ensures results are robust to spatial interdependence. Our simulations evidence that S-2SLS performs well across a variety of situations and presents a conservative and robust alternative.

Our discussion adds to growing concerns over spatially dependent instruments (Cooperman, 2017; Betz, Cook and Hollenbach, 2018). While we have identified challenges to credible inference using observational data, we emphasize that we do not discourage analyses using these data. Instead, our purposes in this paper are twofold. First, we highlighted the unique problems posed by spatial interdependence for instrumental variable models. In our reading of the literature, these problems have largely been ignored by applied researchers. Second, we want to encourage researchers to consider more carefully the potential drawbacks of instrumental variables. Frequently, instrumental variable estimates are conjectured to be superior to results from ordinary least squares. This assumption is often wrong. The estimates obtained from IV models can quickly, and under fairly general circumstances, be worse than ordinary least squares, even with instruments that are plausibly exogenous. Instrumental variables can, under specific circumstances, identify causal effects. But these circumstances are more limited than is often realized, which should give researchers some pause in advocating the use of instrumental variable models. Instrumental variables may cause more problems than they solve.

[26]This is a variant of the file drawer problem, with published findings suffering from substantial selection effects: biased IV estimates, driven by unmodeled spatial interdependence, are those which are most likely to be reported.

References

Ahmed, Faisal Z. 2012. "The Perils of Unearned Foreign Income: Aid, Remittances, and Government Corruption." *American Political Science Review* 106(1):146–165.

Anselin, Luc and Nancy Lozano-Gracia. 2008. "Errors in variables and spatial effects in hedonic house price models of ambient air quality." *Empirical Economics* 34(1):5–34.

Ashraf, Quamrul and Oded Galor. 2011. "Dynamics and Stagnation in the Malthusian Epoch." *American Economic Review* 101(5):2003—2041.

Bartels, Larry M. 1991. "Instrumental and "Quasi-Instrumental" Variables." *American Journal of Political Science* 35(3):777–800.

Beck, Nathaniel, Kristian Skrede Gleditsch and Kyle Beardsley. 2006. "Space is more than geography: Using spatial econometrics in the study of political economy." *International Studies Quarterly* 50(1):27–44.

Betz, Timm. 2013. "Robust Estimation with Nonrandom Measurement Error and Weak Instruments." *Political Analysis* 21(1):86–96.

Betz, Timm, Scott J. Cook and Florian Hollenbach. 2018. "On the Use and Abuse of Spatial Instruments." *Political Analysis* (forthcoming).

Boix, Carles. 2011. "Democracy, development, and the international system." *American Political Science Review* 105(04):809–828.

Bound, John, David A. Jaeger and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation between the Instruments and the Endogeneous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430):443–450.

Büthe, Tim and Helen V. Milner. 2008. "The Politics of Foreign Direct Investment into Developing Countries: Increasing FDI through International Trade Agreements?" *American Journal of Political Science* 52(4):741–762.

Cooperman, Alicia Dailey. 2017. "Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation." *Political Analysis* 25(3):277–288.

Drukker, David M, Peter Egger and Ingmar R Prucha. 2013. "On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors." *Econometric Reviews* 32(5-6):686–733.

Fingleton, Bernard and Julie Le Gallo. 2008. "Estimating spatial models with endogenous variables, a spatial lag and spatially dependent disturbances: Finite sample properties*." *Papers in Regional Science* 87(3):319–339.

Franzese, Robert J. Jr. and Jude C. Hays. 2007. "Models of Cross-Sectional Interdependence in Political Science Panel and Time-Series-Cross-Section Data." *Political Analysis* 15(2):140–164.

Franzese, Robert J., Jude C. Hays and Scott J. Cook. 2016. "Spatial- and Spatiotemporal-Autoregressive Probit Models of Interdependent Binary Outcomes." *Political Science Research and Methods* 4(1):151–173.

Gleditsch, Kristian Skrede and Michael D. Ward. 2006. "Diffusion and the International Context of Democratization." *International Organization* 60(4):911–933.

Hansford, Thomas G. and Brad T. Gomez. 2010. "Estimating the Electoral Effects of Voter Turnout." *American Political Science Review* 104(02):268–288.

Kelejian, Harry H and Ingmar R Prucha. 2004. "Estimation of simultaneous systems of spatially interrelated cross sectional equations." *Journal of Econometrics* 118(1):27–50.

Kirby, Andrew M. and Michael D. Ward. 1987. "The Spatial Analysis of Peace and War." *Comparative Political Studies* 20(3):293–313.

LeSage, James P and R Kelley Pace. 2014. "The biggest myth in spatial econometrics." *Econometrics* 2(4):217–249.

Liu, Xiaodong and Lung-Fei Lee. 2013. "Two-stage least squares estimation of spatial autoregressive models with endogenous regressors and many instruments." *Econometric Reviews* 32(5-6):734–753.

Malthus, Thomas R. 1798. *An Essay on the Principle of Population*. Vol. Reprint, ed. Geoffrey Gilbert,1999 Oxford, UK: Oxford University Press.

Neumayer, Eric and Thomas Plümper. 2016. "W." *Political Science Research and Methods* 4(1):175193.

Plümper, Thomas and Eric Neumayer. 2010. "Model Specification in the Analysis of Spatial Dependence." *European Journal of Political Research* 49(3):418–442.

Ramsay, Kristopher W. 2011. "Revisiting the Resource Curse: Natural Disasters, the Price of Oil, and Democracy." *International Organization* 65(3):507–530.

Simmons, Beth A, Frank Dobbin and Geoffrey Garrett. 2006. "Introduction: The international diffusion of liberalism." *International Organization* 60(4):781–810.

Siverson, Randolph M and Harvey Starr. 1990. "Opportunity, willingness, and the diffusion of war." *American Political Science Review* 84(1):47–67.

Sovey, Allison J. and Donald P. Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55(1):188–200.

Starr, Harvey. 1991. "Democratic dominoes: Diffusion approaches to the spread of democracy in the international system." *Journal of Conflict Resolution* 35(2):356–381.

Stasavage, David. 2005. "Democracy and education spending in Africa." *American Journal of Political Science* 49(2):343–358.

Ward, Michael D. and John O'Loughlin. 2002. "Spatial Processes and Political Methodology: Introduction to the Special Issue." *Political Analysis* 10(3):211–216.

Ward, Michael D. and Kristian Skrede Gleditsch. 2002. "Location, Location, Location: An MCMC Approach to Modeling the Spatial Context of War and Peace." *Political Analysis* 10(3):244–260.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

Supplementary Online Appendix: Random? As if – Spatial Interdependence and Instrumental Variables

A Additional Plots Simulation: Misspecified W

What if the researcher is unsure about the spatial network underlying the modeled processes? In the simulations above, we estimated the S-2SLS model based on the correct connectivity matrix. Effectively, this assumes the researcher has complete information on the spatial network, which is often an unrealistic assumption in applied research. Therefore, we perform the same set of simulation experiments as above but we also vary the level of misspecification of the spatial network in the estimation. To do so, we draw a second set of random spatial locations and its corresponding $\mathbf{W}$ matrix. We then create the $\mathbf{W}$ matrix for the model estimation based on binary draws from either the correct $\mathbf{W_c}$ or the false $\mathbf{W_f}$ matrix. The probability of each cell value being drawn from the false matrix is the misspecification parameter. We set this parameter to three different values: 0, 0.5 and 1. The results presented above assumed no misspecification, such that the probability of drawing from the false $\mathbf{W_f}$ matrix is 0.

As Figure A.1 shows, S-2SLS generally outperforms or is equivalent to 2SLS. In this scenario we consider positive correlation between the first and second stage, i.e. sufficient non-spatial endogeneity. In the worst case (bottom row in Figure A.1), when the $\mathbf{W}$ matrix is completely misspecified, S-2SLS parallels 2SLS in performance. As the median absolute error in 2SLS increases, so does the median absolute error for S-2SLS. Similarly, as Figure A.2 shows, as the coverage for 2SLS worsens, so does the coverage for S-2SLS. This demonstrates what we articulated earlier: because S-2SLS nests 2SLS, it only suffers minor efficiency losses when it is the incorrect model.[27]

Put differently, even if researchers have no knowledge of the spatial network in their data and

[27]For some of the simulations we were unable to estimate the spatial model. This only occurred when the spatial matrix was drawn from two different W matrices. We drop these observations before calculating the performance statistics. The relative performance of S-2SLS does not change if we also drop the corresponding results for OLS and 2SLS.

chose a spatial matrix at random, S-2SLS does not perform significantly worse than 2SLS. Conversely, where there is spatial interdependence and some knowledge of the connectivity matrix, the gains from S-2SLS are considerable – S-2SLS provides the more robust and conservative modeling strategy.

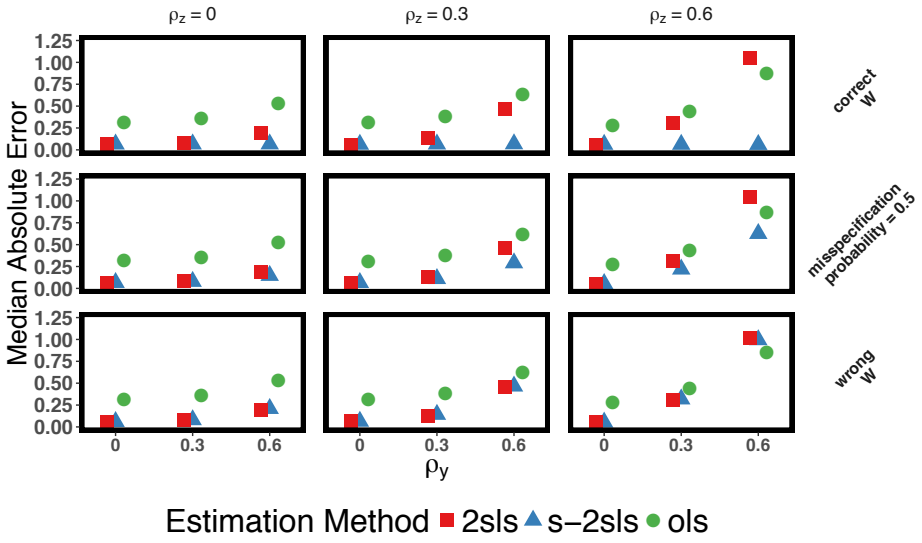


Figure A.1: Median Absolute Error over Misspecification of W ($\lambda = 1.5$ & $\delta = 0.5$ & $N = 200$)

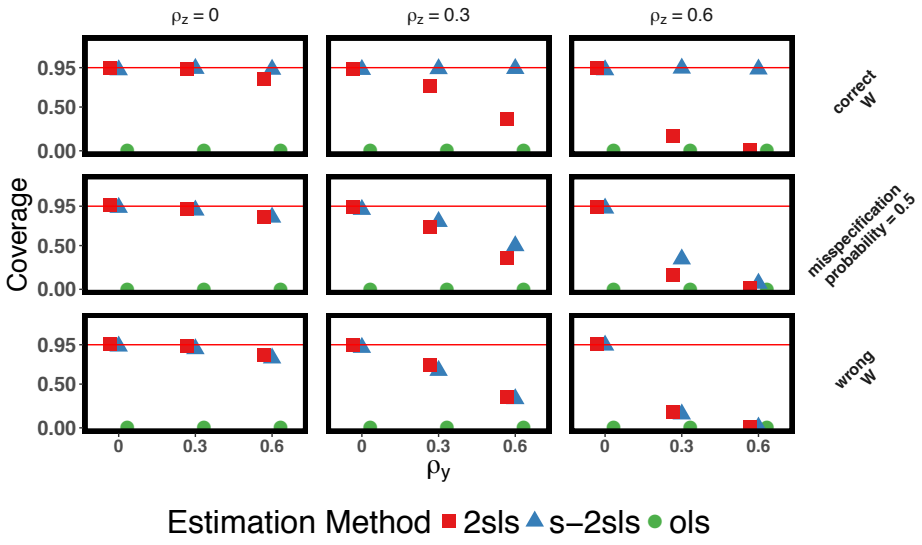


Figure A.2: Coverage over Misspecification of W ($\lambda = 1.5$ & $\delta = 0.5$ & $N = 200$)

B Additional Plots Simulation: MedAE & Coverage

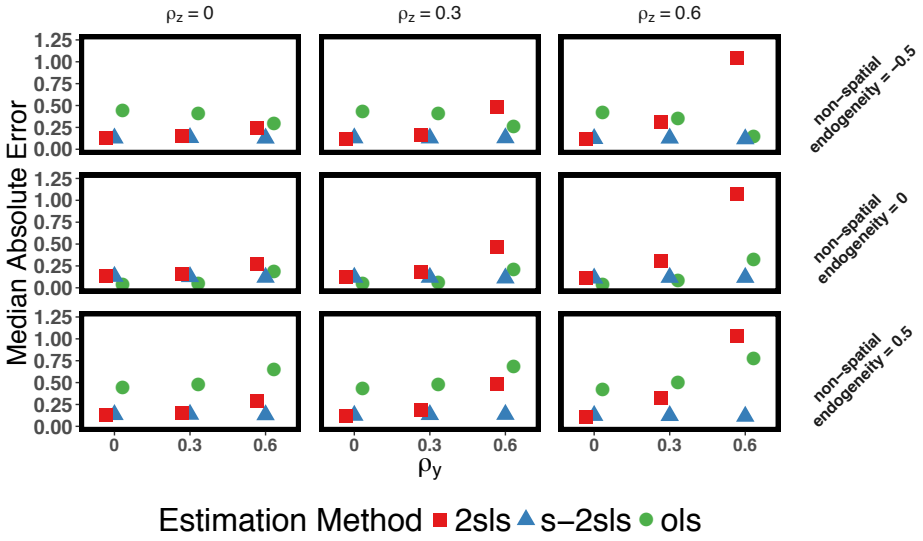


Figure B.3: Median Absolute Error over δ ($\lambda = 0.75$ & $N = 200$)

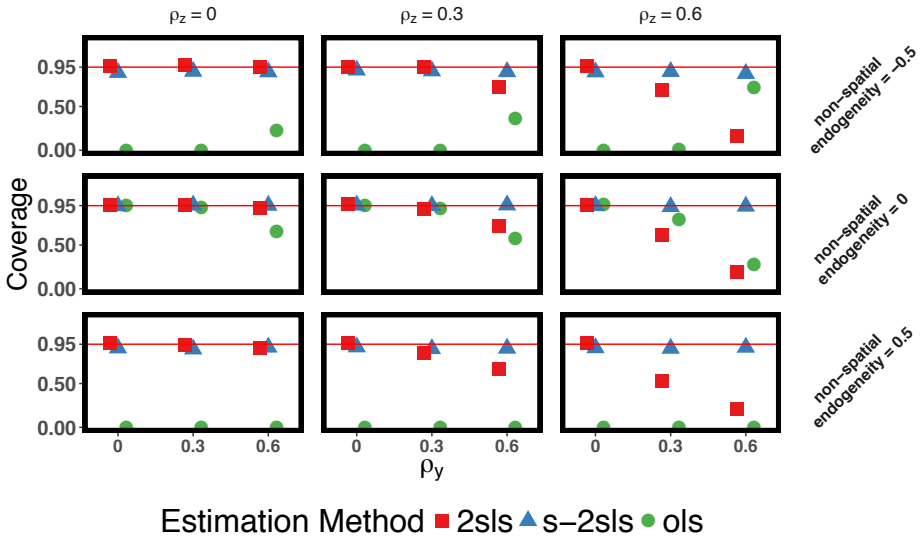


Figure B.4: Coverage over δ ($\lambda = 0.75$ & $N = 200$)

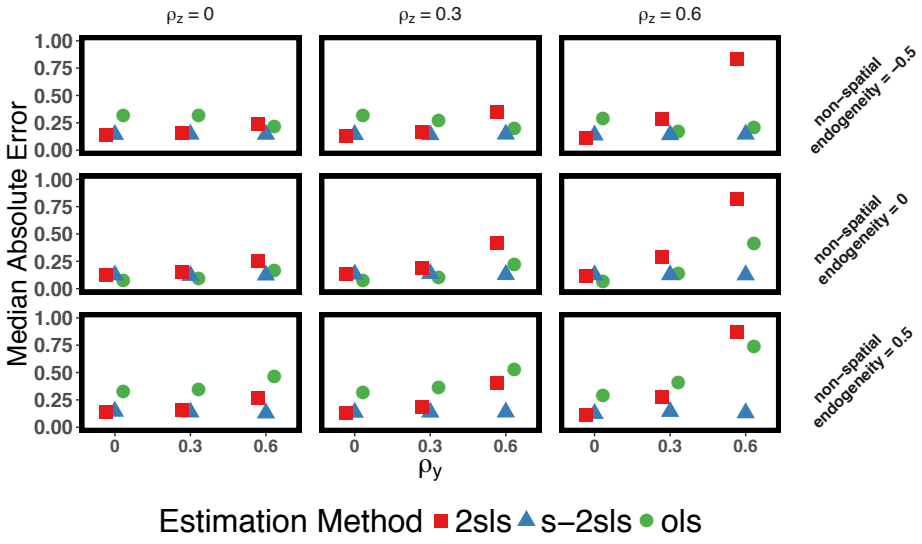


Figure B.5: Median Absolute Error over δ ($\lambda = 1.5$ & $N = 50$)

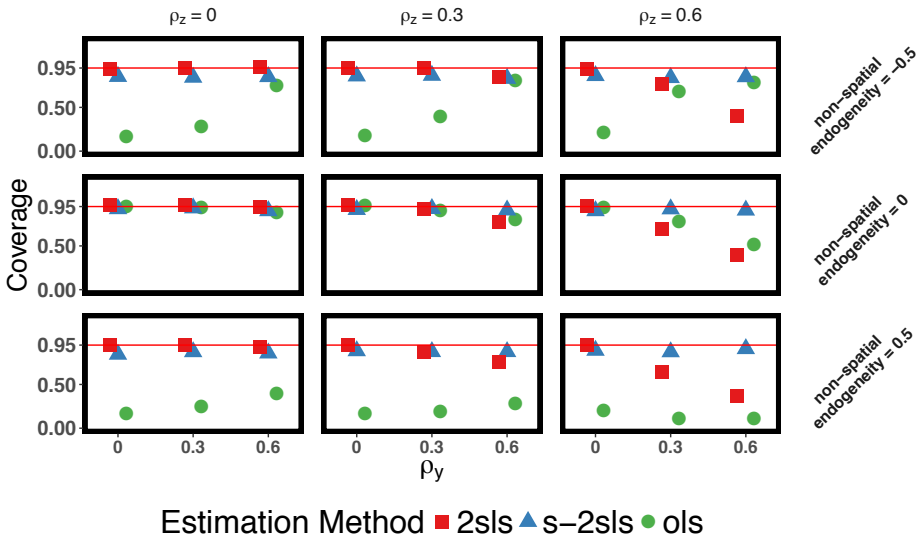


Figure B.6: Coverage over δ ($\lambda = 1.5$ & $N = 50$)

C Additional Plots Simulation: RMSE

Since the RMSE is very sensitive to outliers, we drop any simulation where the absolute error is greater than 10. The only estimation method for which this occurs is standard 2SLS, for which we drop 2217 simulated data sets. Thus, this adjustment actually improves the results for 2SLS.

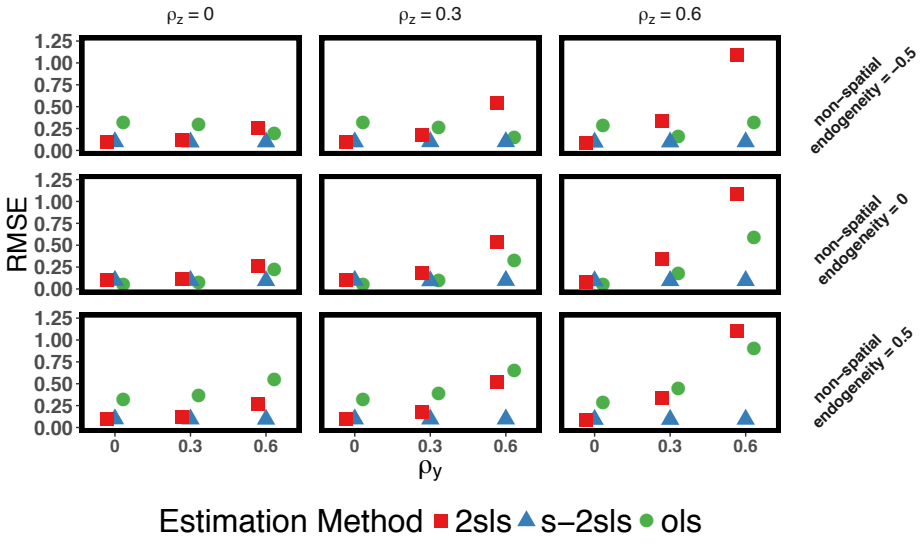


Figure C.7: RMSE over δ ($\lambda = 1.5$ & $N = 200$)

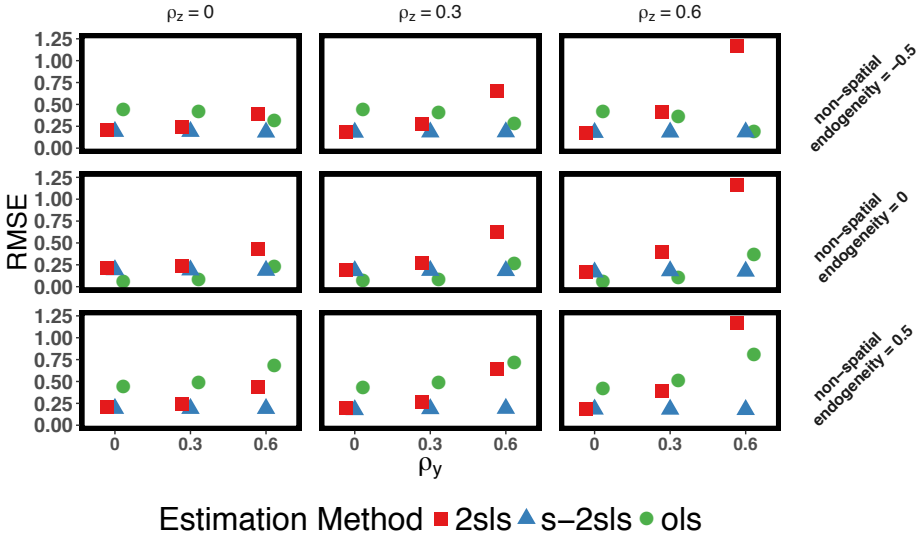


Figure C.8: RMSE over δ ($\lambda = 0.75$ & $N = 200$)

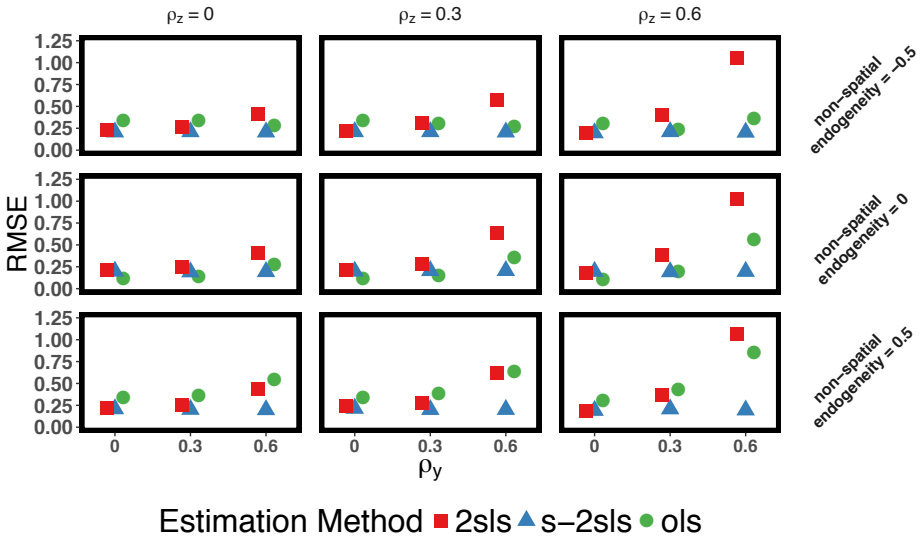


Figure C.9: RMSE over δ ($\lambda = 1.5$ & $N = 50$)

D Additional Tables for Applications

Table D.1: Replication of Robustness Checks, Table 5 Ramsay (2011)

Respective Column Table 5 in Ramsay (2011)	(1) 2SLS	(1) S-SLS	(2) 2SLS	(2) S-SLS	(3) 2SLS	(3) S-SLS	(4) 2SLS	(4) S-SLS	(5) 2SLS	(5) S-SLS	(6) 2SLS	(6) S-SLS	(7) 2SLS	(7) S-SLS
log oil income	-0.358**	-0.0867***	-0.261***	-0.0997***	-0.358**	-0.0725***	-0.358**	-0.0867***	-0.260***	-0.0897***	-0.357***	-0.177***	-0.453*	-0.0989***
per capita	(0.167)	(0.0113)	(0.0837)	(0.00970)	(0.167)	(0.00914)	(0.167)	(0.0113)	(0.0823)	(0.00883)	(0.0769)	(0.0161)	(0.252)	(0.0115)
log gdp	0.356**	0.106***	0.300***	0.129***	0.350**	0.0944***	0.356**	0.106***	0.294***	0.118***	0.371***	0.202***	0.492*	0.122***
per capita	(0.155)	(0.0113)	(0.0914)	(0.0114)	(0.151)	(0.00932)	(0.155)	(0.0113)	(0.0878)	(0.0104)	(0.0767)	(0.0172)	(0.265)	(0.0133)
gdp growth	-0.0118**	-0.00496***	-0.00851***	-0.00503***	-0.0118**	-0.00461***	-0.0118**	-0.00496***	-0.00851***	-0.00482***	-0.0112***	-0.00601***	-0.0127*	-0.00514***
	(0.00518)	(0.00102)	(0.00268)	(0.00103)	(0.00518)	(0.000975)	(0.00518)	(0.00102)	(0.00265)	(0.000998)	(0.00361)	(0.00155)	(0.00655)	(0.00103)
polity at entry	-0.00517	0.567***	0.226	0.548***	0.00684	0.595***	-0.00517	0.567***	0.239	0.567***			-0.0507	0.549***
	(0.373)	(0.0284)	(0.184)	(0.0251)	(0.363)	(0.0242)	(0.373)	(0.0284)	(0.177)	(0.0232)			(0.463)	(0.0271)
Latitude			-0.00622**	-0.00192***					-0.00630**	-0.00164***				
			(0.00247)	(0.000457)					(0.00249)	(0.000440)				
top5 oil producers					0.127	-0.0160			0.105*	-0.000276				
					(0.0948)	(0.0224)			(0.0627)	(0.0233)				
coldwar dummy							0	-0.327***						
							(.)	(0.0588)						
west dummy											-0.101	0.201***	-0.235	0.00790
											(0.165)	(0.0402)	(0.218)	(0.0238)
sub-Saharan Africa													0.462*	0.0949***
													(0.276)	(0.0233)
Constant	-0.210	-0.327***	-0.245*	-0.0670	-0.172	-0.0539	-0.210	0	-0.215	-0.0687	-0.318	-0.504***	-0.755	0
	(0.183)	(0.0512)	(0.149)	(0.0608)	(0.170)	(0.0500)	(0.183)	(.)	(0.140)	(0.0517)	(0.205)	(0.102)	(0.535)	(.)
Spatial ρ		0.119***		0.0989***		0.131***		0.119***		0.108***		0.113***		0.130***
		(0.0201)		(0.0214)		(0.0193)		(0.0201)		(0.0208)		(0.0346)		(0.0201)
Observations	1263	1263	1263	1263	1263	1263	1263	1263	1263	1263	1263	1263	1263	1263
Year Dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

E Additional Application: "Dynamics and Stagnation in the Malthusian Epoch"

In this section of the Appendix, we provide an additional application. In a 2011 article in the *American Economic Review*, Ashraf and Galor (2011) aim to test a central prediction of the famous Malthusian theory. Thomas R. Malthus (1798) argues that the main reason for stagnating incomes, prior to the industrial revolution, is that when incomes increase, population size rises as well. Since resources are limited, higher populations induce declining living standards. As a result, technological progress or the discovery of new resources only temporarily improves living standards, but does not produce sustained gains (Ashraf and Galor, 2011). As Ashraf and Galor (2011, p. 2004) outline, their article "exploits exogenous sources of cross-country variation in land productivity and technological levels to examine their hypothesized differential effects on population density versus income per capita during the time period 11500 CE." To test the Malthusian theory in pre-industrial societies, Ashraf and Galor (2011) investigate two predictions: 1) a country's improvements in productivity should lead to larger populations, but not higher living standards; and 2) countries with higher land productivity, or better technology, should have higher population densities, but again, should not be significantly richer.

In their empirical analysis, Ashraf and Galor (2011) use the timing of the onset of the neolithic revolution to proxy for technological change. Consistent with their expectations, the authors show that both the onset of the neolithic revolution and land productivity are positively (and significantly) associated with population density, but not with income per capita. In addition, Ashraf and Galor (2011) use instrumental variables to estimate the causal effect of technological progress on population density. They argue that "prehistoric biogeographical endownments," in particular the "availability of domesticable species of plants and animals," have had an important effect on technological progress and are otherwise exogenous (Ashraf and Galor, 2011, pp. 2029-2031). The use of the instrumental variable is primarily motivated by the authors to estimate the "causal impact of technology on population density" (Ashraf and Galor, 2011, p. 2031).

However, the authors ignore possible spatial interdependence in both the instrumental variables and the dependent variable. Both population density and natural wildlife are likely to be spatially clustered. In other words, it is likely that the animal and plant species found in one country are similar to those in adjacent regions. Likewise, in pre-historic times (i.e., 1000 CE), it is likely that some parts of the planet had higher population density than others, again reflecting positive spatial correlation. This does not mean that these variables are similarly clustered in space, but rather that by themselves they might exhibit (positive) spatial dependence. If correct, this would induce bias in their IV models for the reasons outlined above.

To test this, we first need to specify an appropriate weights matrix. Here, we create a binary-contiguity matrix, with neighbors defined as having adjacent borders.[28] As a preliminary test of spatial autocorrelation, we estimate Moran's I based on the residuals of the original OLS model – with logged population density in 1000 CE as the dependent variable (column 2, Table 9 in Ashraf and Galor (2011)). Based on a Moran's I value of 0.4 with an associated p-value smaller than 0.001, we are able to reject the null of independence of the residuals.

Table E.2 shows the results of replicating the models with population density in 1000 CE (Table 9 in Ashraf and Galor (2011)). Column 1 replicates the original OLS model on the restricted sample (column 2 in Table 9 in Ashraf and Galor (2011)). As a first step, column 2 in Table E.2 shows the results when we estimate a spatial autoregressive (SAR) model instead of the standard OLS model. As one can see, the main coefficients of interest (technological index) have the same levels of significance as in the original OLS results. The effect estimates, however, are quite different. For the linear-additive model, the direct effect is simply the reported coefficient estimate on the log technology index (3.21). The average direct effect of the SAR model – calculated as above – is 3.52 – larger than the coefficient estimate for SAR (due to expected feedback effects), but still substantially smaller than the OLS effect estimates of 4.198.

Columns 3 and 4 replicate the instrumental variable model for population density in 1000 CE as presented in Table 9 in Ashraf and Galor (2011). The differences in results between the original

[28] We have also replicated the results with a k(=5) nearest-neighbor matrix or a row-standardized contiguous neighbor matrix.

2SLS model and the spatial 2SLS model are stark. The coefficient on technological progress (log of technological index) in the original 2SLS model is 14.53, almost 3.5 times as large as the OLS coefficients. Ashraf and Galor (2011) argue that the difference in estimated coefficients is "a pattern that is consistent with measurement error in the transition-timing variable and the resultant attenuation bias afflicting OLS coefficient estimates" (Ashraf and Galor, 2011, p. 2031). Column 4, however, shows the results from the model estimated with S-2SLS. Here the coefficient for technological progress is much smaller compared to 2SLS, with the average direct effect – calculated as before – being 7.03. In fact, the average effect estimate of technological progress in the spatial 2SLS model is comparable to that in the original OLS estimates. Recall, that, as we show above, the non-spatial and spatial bias in OLS can be offsetting. This may be the case here. If the non-spatial measurement bias is attenuating and the spatial bias is upward, the OLS model ends up being less biased than the 2SLS model due to the countervailing forces of both biases on the coefficient estimate.

We note that the overall conclusion of Ashraf and Galor (2011) still stands.[29] The Malthusian theory for pre-industrial times is supported by these data. On the other hand, the causal effect of technological progress on population density is smaller than the standard 2SLS model indicates and is about the same size the original estimates in the OLS models in Ashraf and Galor (2011).

[29]In the Appendix we also replicate the results using population density in 1CE as the dependent variable (Table E.3), producing similar results.

Table E.2: Replication of Table 9 (1000 CE) in Ashraf and Galor (2011)

	(1) Original OLS	(2) SAR	(3) Original 2SLS	(4) S-2SLS
Log technology index in relevant period	4.198***	2.856***	14.53***	4.303***
	(1.164)	(0.953)	(4.437)	(1.328)
Log land productivity	0.498***	0.397***	0.572***	0.397***
	(0.139)	(0.0963)	(0.148)	(0.0987)
Log absolute latitude	−0.185	−0.093	−0.209	−0.086
	(0.151)	(0.106)	(0.209)	(0.108)
Mean distance to nearest coast or river	−0.363	−0.341	−1.155*	−0.462
	(0.426)	(0.360)	(0.640)	(0.368)
Percentage of land within 100 km of coast or river	0.442	0.472	0.153	0.431
	(0.422)	(0.341)	(0.606)	(0.344)
Constant	−1.820***	−1.286**	−5.507***	−1.796***
	(0.641)	(0.531)	(1.702)	(0.630)
Spatial ρ_y		0.151***		0.169***
		(0.0246)		(0.0334)
Observations	92	92	92	92
Continent dummies	Yes	Yes	Yes	Yes

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table E.3: Replication of Table 9 (1 CE) in Ashraf and Galor (2011)

	(1)	(2)	(3)	(4)
	Original OLS	SAR	Original 2SLS	S-2SLS
Log technology index in	3.947***	3.369***	10.80***	3.010***
relevant period	(0.983)	(0.760)	(2.857)	(0.978)
Log land productivity	0.350**	0.311***	0.464**	0.294***
	(0.172)	(0.106)	(0.182)	(0.105)
Log absolute latitude	0.0834	-0.0152	-0.0521	-0.0505
	(0.170)	(0.115)	(0.214)	(0.114)
Mean distance to nearest	-0.625	-0.300	-0.616	-0.175
coast or river	(0.434)	(0.394)	(0.834)	(0.388)
Percentage of land within	0.146	0.0986	-0.172	0.0867
100 km of coast or river	(0.424)	(0.357)	(0.642)	(0.351)
Constant	-2.719***	-1.749***	-4.770***	-1.334**
	(0.601)	(0.500)	(0.980)	(0.544)
Spatial ρ		0.182***		0.252***
		(0.0275)		(0.0358)
Observations	83	83	83	83
Continent dummies	Yes	Yes	Yes	Yes

Standard errors in parentheses

$^{*}\ p < 0.1$, $^{**}\ p < 0.05$, $^{***}\ p < 0.01$

Multiple Imputation Using Gaussian Copulas*

Florian M. Hollenbach†
Department of Political Science, Texas A&M University
and
Iavor Bojinov
Department of Statistics, Harvard University
and
Shahryar Minhas
Department of Political Science, Michigan State University
and
Nils W. Metternich
Department of Political Science, University College London
and
Michael D. Ward
Department of Political Science, Duke University
and
Alexander Volfovsky
Department of Statistical Science, Duke University

September 10, 2018

*Accepted for publication at *Sociological Methods & Research*. Florian M. Hollenbach is an Assistant Professor, Department of Political Science, Texas A&M University, College Station, TX 77843-4348 (email: fhollenbach@tamu.edu); Iavor Bojinov is a PhD Student, Department of Statistics, Harvard University, Cambridge, MA 02138 (email: bojinov@fas.harvard.edu); Shahryar Minhas is an Assistant Professor, Department of Political Science, Michigan State University, East Lansing, MI, 48824 (email: minhassh@msu.edu); Nils W. Metternich is a Senior Lecturer, Department of Political Science, University College London, London, UK WC1H 9QU (email: n.metternich@ucl.ac.uk); Michael D. Ward is a Professor, Department of Political Science, Duke University, Durham, NC 27708 (email: michael.d.ward@duke.edu); and Alexander Volfovsky is an Assistant Professor, Department of Statistical Sciences, Duke University, Durham, NC 27708 (email: av136@stat.duke.edu). This project was partially supported by the the Office of Naval Research (holding grants to the Lockheed Martin Corporation, Contract N00014-12- C-0066). Nils W. Metternich acknowledges support from the Economic and Social Research Council (ES/L011506/1). The work was completed while Alexander Volfovsky was supported by a NSF MSPRF under DMS-1402235. For helpful insights we thank Philippe Loustaunau, among the first of our colleagues to encourage this effort. Stephen Shellman was a strong critic who deserves our thanks too: his criticisms helped us to improve our approach. John Ahlquist, Matt Blackwell, Andreas Beger, Cassy Dorff, Gary King, and Jacob Montgomery provided helpful comments on previous versions of this paper.

†Corresponding author

Abstract

Missing observations are pervasive throughout empirical research, especially in the social sciences. Despite multiple approaches to dealing adequately with missing data, many scholars still fail to address this vital issue. In this paper, we present a simple-to-use method for generating multiple imputations using a Gaussian copula. The Gaussian copula for multiple imputation (Hoff, 2007) allows scholars to attain estimation results that have good coverage and small bias. The use of copulas to model the dependence among variables will enable researchers to construct valid joint distributions of the data, even without knowledge of the actual underlying marginal distributions. Multiple imputations are then generated by drawing observations from the resulting posterior joint distribution and replacing the missing values. Using simulated and observational data from published social science research, we compare imputation via Gaussian copulas with two other widely used imputation methods: `MICE` and `Amelia II`. Our results suggest that the Gaussian copula approach has a slightly smaller bias, higher coverage rates, and narrower confidence intervals compared to the other methods. This is especially true when the variables with missing data are not normally distributed. These results, combined with theoretical guarantees and ease-of-use suggest that the approach examined provides an attractive alternative for applied researchers undertaking multiple imputations.

Keywords: missing data, Bayesian statistics, categorical data

1 Introduction

Missing data problems are ubiquitous in observational data and common among social science applications. Statistical inference that does not adequately account for the missing data is widely known to lead to biased results, and inflated (or deflated) variance estimates (Rubin, 1976, King et al., 2001, White and Carlin, 2010, Molenberghs et al., 2014). Even though most statistical software platforms provides methods that adequately handle missing data (the most popular of these is multiple imputations (MI)), they are often ignored by applied researchers.[1]

In Figure E.1, we illustrate the number of articles published in five top sociology and political science journals since 1990 that contain "multiple imputations" in the body of the paper.[2] Our survey of the literature shows the rapid growth of the use of multiple imputations in the social sciences. Nevertheless, as missing data is a feature of almost any observational data set, the annual counts of articles mentioning multiple imputations per year still point to significant underutilization of this method in the social sciences. This may be due to a lack of understanding of the benefits (and assumptions) of common

[1] Principled approaches to missing data have existed for over three decades. First formalized by Rubin (1976), the number of readily available statistical softwares to deal with missing data has rapidly grown since the 1990s (e.g. King et al., 2001, Honaker and King, 2010, Van Buuren and Groothuis-Oudshoorn, 2011, Kropko et al., 2014). Further, see the special issue on the *State of Multiple Imputation Software* in the *Journal of Statistical Software* in 2011 (Yucel, 2011).

[2] The five journals we reviewed from sociology are *Annual Review of Sociology*, *American Sociological Review*, *American Journal of Sociology*, *Sociological Methodology*, and *Sociological Methods & Research*. In political science we examined the *American Political Science Review*, *American Journal of Political Science*, *Political Analysis*, *British Journal of Political Science*, and the *Journal of Politics*.

imputation methods.

[Figure 1 about here.]

This article has two aims. First, we introduce applied researchers in the social sciences to a specific copula method for imputation and discuss its advantages over other methods. The method discussed is easy to implement using the `sbgcop` package (Hoff, 2010) in $\mathcal{R}$ (R Development Core Team, 2004) [3] and has theoretical properties that make it attractive. Second, we conduct a systematic evaluation and comparison of the copula method to two commonly used imputation software packages (`MICE` Van Buuren and Groothuis-Oudshoorn (2011) and `AMELIA II` (Honaker et al., 2012)) in sociology and political science.

Copulas are often used for the estimation of dependency between variables and are particularly useful in the generation of imputations as they allow for the construction of valid joint distributions of the data, even if the researcher has little knowledge about the actual joint distribution of the variables. Given the joint distribution of the data, we can generate imputations by sampling from the conditional distribution of the missing data given the observed data.

We highlight a semi-parametric Gaussian copula approach to missing data imputation. The Gaussian copula is one particular way of constructing a joint distribution from which missing values can be easily drawn. The method was initially developed by Hoff (2007) to estimate empirical models on multivariate data.

In particular, the Gaussian copula defines the dependence among the distributions of

[3] For inexperienced users, our `gcImp` (`https://github.com/bojinov/gcImp`) package provides a simple interface for generating imputations using `sbgcop`.

a set of variables which may contain missing values. These variables can include normal, ordinal, and binary variables. Rather than using the distributions themselves, a rank likelihood approximation is used. As a result, the technique does not require the specification of marginal or conditional distributions. This is in stark contrast to other imputation methods using copulas that either require knowledge of the marginals or correlation structure (Käärik, 2006, Käärik and Käärik, 2009, Robbins et al., 2013) or target different copula parameters via pseudolikelihood methods (Di Lascio et al., 2015). The proposed approach allows applied researchers to undertake imputations of their data without relying on pre-specification or ad-hoc decisions.

The potential use of copulas for multiple imputation applications has not been thoroughly discussed within the social sciences. The copula methods we describe are easy to use and are more likely to provide a good representation of the joint distribution of the data than existing methods. Moreover, provided the Markov Chain Monte Carlo (MCMC) converges, the output from the copula model represents a valid posterior density. Simply put, this means that we have theoretical guarantees about the posterior distribution from which the imputations are generated that other methods can not provide. Based on an extensive simulation exercise, we show that the method presented here is generally at least as accurate as other commonly used methods—it is often better. It also provides better uncertainty estimates for the imputations. Lastly, as is shown in Bojinov et al. (2017), the copula method can also be used to test some of the underlying assumptions about the appropriateness of imputations for a given data set.

2 Common Approaches to Multiple Imputation

The standard techniques employed to deal with missing data require an assumption regarding the missing data pattern; these were first formalized in Rubin (1976).[4] To briefly summarize these terms, missing data are missing completely at random (**MCAR**) when the probability of the observed missing data pattern is unchanged regardless of what values both the observed and missing data take (Marini et al., 1980). The missing data are missing at random (**MAR**) when the probability of observing the missing data pattern is unchanged no matter what values the missing data take. Finally, the missing data are missing not at random (**MNAR**) when the probability of observing the missing data pattern changes for some values of the missing data.

These definitions are important both from a theoretical and a practical point of view. The most basic methods, such as listwise deletion, generally lead to biased regression coefficients if the missingness process is not **MCAR** (Graham, 2009). To achieve valid inference under the Bayesian and likelihood paradigms, while ignoring the missing data mechanism, we require the weaker **MAR** assumption.[5]

The most common appropriate approach to dealing with missing data is multiple imputation (MI), which refers to any method that replaces the set of missing values with various

[4] Little and Rubin (2002) provide a more up to date treatment and Mealli and Rubin (2015) an in-depth discussion on the different missing data mechanisms.

[5] A further assumption of parameter distinctness—the parameter governing the data and the parameter governing the missingness mechanism are *a priori* independent—is required to ensure that valid statistical inference whenever the data are **MAR** or **MCAR**. See Little and Rubin (2002) for more details on this assumption.

plausible values, thus obtaining m completed data sets (Rubin, 1996). Rubin (1987) initially suggested creating five imputations, but more recently authors recommended using closer to twenty imputations (Van Buuren, 2012).[6] The completed data sets are then separately analyzed using the standard full data techniques and the resulting quantities of interest from each data set are combined to obtain an overall, average estimate as well as its associated variance.

Before moving to introduce the copula method below, we briefly outline two important methods for generating multiple imputations here.

MI with EM This approach uses iterative expectation maximization (EM) to create complete data sets based on assuming a particular joint distribution. A widely used method for imputation in the social sciences is the `Amelia II` $\mathcal{R}$ package by Honaker and King (2010) In `Amelia II`, the joint distribution of the data is modeled as a multivariate normal distribution. `Amelia II` provides an implementation of the EM approach by the use of bootstrapping to derive solutions quickly. One of the disadvantages of imputation via EM is that for large data sets with significant amounts of missing data, it is computationally intensive. This is a trait of EM algorithms in general, as the rate of convergence is proportional to the amount of missing information in the model. Moreover, it is often unclear to what degree modeling the joint distribution of the data as a multivariate normal distribution is appropriate, especially since the data may include binomial and ordinal variables.

[6] This was based on examining large sample relative efficiency when using a finite number of proper imputations compared to an infinite number, from a Bayesian Gaussian model. In practice, non-normal data combined with non-Bayesian methods can lead to a decrease in the relative efficiency.

Conditional Approaches to Multiple Imputation An alternative method is to model each variable's imputation via its conditional distribution based on all other variables in the data. One such approach is developed in Multiple Imputation via Chained Equations (MICE) (Van Buuren, 2012), another was developed as the *MI* package in $\mathcal{R}$ (Goodrich et al., 2012). Imputations for fully conditional specification (FCS) methods, such as `MICE` or `MI`, are created based on an "appropriate generalized linear model for each variable's conditional distribution" (Kropko et al., 2014, 501). This is done for all variables and iterated until the model converges.

One of the main drawbacks of the FCS is that only under certain conditions do the individual conditional models define a valid joint distribution. This often leads to pathologies in the convergence of the algorithms (Li et al., 2012, Chen and Ip, 2015). For example, if $Y|X$ is specified to be an Exponential random variable with rate X and $X|Y$ is specified to be an Exponential random variable with rate Y, it is well known that no joint distribution exists and sequentially sampling from these two distributions generates draws that tend to infinity (Casella and George, 1992). More strikingly, Example 1 of Li et al. (2012) demonstrates that even when all the conditionals are normal, the order in which the variables are updated in MICE can determine whether the chain will converge to a stationary distribution.

One of the advantages of conditional model specification is that it allows each variable to be modeled based on its specific distribution, which is specified by the researcher. However, this also means the imputation model for each variable in the data has to be correctly specified, which can be "labor-intensive and challenging with even a moder-

ate number of variables" (Murray, 2013, 41). Moreover, coefficients estimates in the conditional models can suffer significantly when the number of missing observations is large, especially for categorical variables (Murray, 2013).

3 A copula approach to missing data imputation

One of the key issues with conditional approaches to imputation, such as `MICE`, is that they do not necessarily specify a valid joint distribution (such as the example in the previous section).[7] When a valid joint distribution does not exist, then there are no guarantees that the MI procedure is proper (as defined in Rubin, 2004). A natural approach to overcoming a possibly incompatible conditional specification is by specifying the joint distribution directly. For example, this is done in most EM approaches, such as `Amelia II`, by simply assuming a multivariate normal distribution. However, while an approximation, most social science data include binary and ordinal variables, and thus cannot have a multivariate normal joint distribution. As a result, this misspecification of the joint distribution is problematic. Moreover, specifying the correct joint distribution becomes increasingly complicated as the number of covariates in the model increase.

[7] Some theoretical results for `MICE` are available, but they do not allow too much misspecification in the conditional models. For example, Liu et al. (2013) showed that for valid semicompatible models (*i.e.*, models which are compatible when some of the parameters in the conditional distributions are set to zero, and the joint model obtained from the compatible conditionals contains the correct joint distribution) the combined imputation estimator is consistent. Further, Zhu and Raghunathan (2015) extend these results to more incompatible models at the expense of the type of missingness patterns allowed (restricting the theoretical results to missingness patterns where each individual is missing at most one variable).

It is therefore valuable to decouple the specification of the marginal distribution of each covariate from the function that describes the joint behavior of all covariates together. One of the main advantages of using copulas for imputation is that they allow us to do exactly that. Sklar's (1959) theorem guarantees that every joint distribution can be decomposed in this way:

Theorem 3.1 (Sklar's Theorem)**.** *Let F be a p-dimensional joint distribution function with marginals $F_1, \ldots, F_p$. Then there exists a copula C with uniform marginals such that*

$$F(x_1, \ldots, x_p) = C(F_1(x_1), \ldots, F_p(x_p))$$

Sklar's theorem guarantees that the function C is unique if the marginal distributions $F_1, \ldots, F_p$ are continuous. If they are discrete, then it is unique on the cross product of the ranges of the F_j.

Much work has been done studying the class of Gaussian copulas where the multivariate dependence is defined by C via the multivariate normal distribution with a correlation matrix R (Klaassen et al., 1997, Pitt et al., 2006, Chen et al., 2006, Hoff, 2007). That is, we define the Gaussian copula function as $C(\cdot|R) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)|R)$ for $u_1, \ldots, u_p \in (0,1)^p$ where Φ is the univariate normal CDF and $\Phi_p(\cdot|R)$ is the p-dimensional CDF with correlation matrix R. This means that the joint distribution of the p variables is given by $F(x_1, \ldots, x_p) = \Phi_p(\Phi^{-1}(F_1(x_1)), \ldots, \Phi^{-1}(F_p(x_p))|R)$. Simply put, the univariate CDFs $F_1, \ldots, F_p$ of the individual variables are bound together as a multivariate normal CDF where R determines the correlation between the individual variables on the normal scale.

As previously noted, the specification of marginal distributions is difficult in applied

settings and so of particular interest is the setting where the researcher does not need to specify the marginal distributions for $F_1, \ldots, F_p$. In fact, one big advantage to the method discussed here is that we consider a semiparametric approach that does not require parameterizing the p marginal distributions.

In this flexible setting, the estimation procedures described below provide consistent and likely asymptotically efficient estimates of the dependence parameters in the Gaussian copula, *i.e.*, R above (Murray et al., 2013; Hoff et al., 2014). These dependence parameters directly impact the imputation of the missing data, and thus these theoretical results are extremely appealing. The estimation approach we explore below was developed by Hoff (2007) by extending the ideas of the rank likelihood of Pettitt (1982) to the copula setting.

The rank likelihood (Pettitt, 1982) is a type of marginal likelihood that bases inference on the ranks of data rather than the full data. In a univariate setting it is defined as follows: consider $z_1, \ldots, z_n | \theta \sim p(z|\theta)$ be a sample from some distribution. Instead of observing the actual values $z_1, \ldots, z_n$, however, consider only observing the ordering of the data $x_1, \ldots, x_n$ (*i.e.* their rank). Then the rank likelihood is given by

$$L(\theta; x_1, \ldots, x_n) = \int_D p(z_1, \ldots, z_n|\theta) dz_1, \ldots, dz_n$$

where $D = \{z_{\alpha_1} < \cdots < z_{\alpha_n}\}$ and $\alpha_i = j$ if and only if z_j is the ith smallest of $z_1, \ldots, z_n$.

Hoff (2007) extends the rank likelihood to the multivariate setting by considering the semiparametric Gaussian copula. Let $z_1, \ldots, z_n | R \sim N(0, R)$, with $z_i = (z_{i1}, \ldots, z_{ip})$, and let $x_{ij} = F_j^{-1}(\Phi(z_{ij}))$. That is, latent data are drawn from a multivariate normal distribution with correlation R and are transformed to the observed scale via an inverse transformation as in the definition of the Gaussian copula above. One can consider the

observed data as the ranks of the unobserved latent Zs and define

$$D = \{Z \in R^{n \times p} : \max\{z_{kj} : x_{kj} < x_{ij}\} < z_{ij} < \min\{z_{kj} : x_{ij} < x_{kj}\}\}.$$

It is easy to see that all $Z \in D$ respect the order of the variables on the observed scale. Hoff (2007) shows that $P(Z \in D|R, F_1, \ldots, F_p) = P(Z \in D|R)$ which in turn allows for the decomposition

$$P(X|R, F_1, \ldots, F_p) = P(Z \in D|R)P(X|Z \in D, F_1, \ldots, F_p).$$

The aforementioned results guarantee that inference about R can proceed simply via $P(Z \in D|R)$. This leverages the ordering of the observed values $x_{1j}, \ldots, x_{nj}$ of each variable to make inference about the parameter R without estimating the CDFs $F_1, \ldots, F_p$.

This means that regardless of the marginal distributions of the individual variables, all we need is their ordering to facilitate the use of the Gaussian copula model to make inferences about the dependence between these variables, *i.e.*, the correlation matrix R. A Bayesian approach to estimating R specifies an inverse Wishart prior for a covariance matrix V such that R is its correlation matrix and a normal prior for the latent z_{ij}. Updates are performed via a Gibbs sampler since full conditional distributions can be derived by conditioning on the ranks of the data alone.[8]

To paraphrase and summarize the method in less technical terms. Assume we have two vectors Z_1 and Z_2 which come from a bivariate normal distribution with correlation R. We observe $X_i = F_i^{-1}(\Phi(Z_i))$ implying that X_i is distributed according to F_i. If the F_i are continuous and known, we can recreate the vectors Z_1 and Z_2 by using the

[8]Further details of the algorithm for estimation are available in Hoff (2007).

pseudo-inverse CDFs on the original data ($Z_1 = \Phi^{-1}(F_1(X_1))$ and $Z_2 = \Phi^{-1}(F_2(X_2))$). We could then generate a good estimate of R using the transformed vectors Z_1 and Z_2 and maximum likelihood estimation, for example $\Sigma_{i=1}^{N} \frac{Z_1^{(i)} Z_2^{(i)}}{N}$ would be a natural estimate for the correlation. However, when either vector is not continuous, the simple pseudo-inverse transformation does allow for correct estimation of the correlation. Now assume X_2 is a binary or ordinal variable as in many of our cases but that the marginal is not known. Instead of using a plug in value for Z_2 (say, by estimating the marginal F_2), we contend that the ranks of our latent continuous Z_2 are the same as those of the observed variable X_2. The estimation procedure then iterates the following two steps: Using the ranks of X_1 and X_2 and the current estimate of the multivariate correlation R, we can draw values of the latent variables (Z_1 and Z_2) that preserve the rank ordering of the observed data. The second step uses the sampled underlying latent variables to sample the correlation R. These steps are iterated until stationarity is reached. Relying on the ranks and latent scale allows us to not specify the marginal distributions of the individual variables and still arrive at a proper solution to estimating R.

When values of x_{ij} are missing at random, imputation can be performed first on the latent z_{ij} scale (since the latent variables are normal, sampling from the conditional distribution of the missing data given the observed data requires a multivariate normal draw) and are then transformed to the observed scale using the empirical cumulative density functions. As this is a Bayesian procedure we produce a posterior for the missing data. To make our approach comparable to the standard conditional approaches we only employ a few samples from this posterior and use those as multiply-imputed datasets. However, it

is natural to consider posterior predictive distributions of parameters of interest or other posterior summaries on a case-by-case basis. For example, the conditional independence graphs of Hoff (2007) succinctly summarize the relationships among many variables.

4 Comparing `Amelia II`, `sbgcop`, and `MICE`

In this section, we compare the working properties of the copula based imputation with those of `Amelia II` and `MICE` packages. We evaluate each method based on an extensive simulation study as well as an empirical example from the social sciences, discussed in the next section.

4.1 Evaluating Imputations

Multiple imputation procedures are specifically designed to yield valid statistical inference (meaning, asymptotically unbiased with correct standard errors and coverage) for population quantities of interest. Since correct estimation of the coefficients and standard errors is critical for obtaining valid statistical inference, any analysis of MI procedures must focus on studying its frequentist properties. Properties such as empirical coverage, average bias, and average interval length of the estimate of the scientific estimand over repeat samples will be of cardinal interest.

We therefore use the following approach to assess the validity of an MI procedure through simulation:

1. Define a full data quantity of interest, θ. In our setting, θ is a set of regression coefficients.

2. Generate a complete data set and apply a pre-specified missing data mechanism to remove some observations.

3. Use the MI procedure to create m completed data sets with the missing values replaced by imputed values.

4. Use each of the m data sets to obtain an estimate of θ as well as its associated variance and combine them using Rubin's combining rules (Rubin, 2004) to obtain $\hat{\theta}$ and a 95% confidence interval (CI).

5. Report the bias of $\hat{\theta}$, the CI interval length and whether or not the CI covered the true value (Van Buuren, 2012, Section 2.5.2).

We repeat Steps 2-5 S times to obtain the empirical coverage rate. By varying the full data model and the missing data mechanism, in Step 2, we can control the two paths that influence the effectiveness of the MI procedures.

4.2 Simulation Study

In regression settings, an outcome Y can depend on many explanatory variables $\mathbf{X} = X_1, \ldots, X_J$ some of which can be costly to measure. As such, it is common that while the outcome Y is measured for all variables, some entries of the design matrix $\mathbf{X}$ are missing. In this simulation, we exclusively focus on this situation and restrict the missingness to the explanatory variables. We will further assume that the missingness mechanism does not allow for the missingness to depend on the outcome Y.

In this situation complete case analysis (or listwise deletion) provides an unbiased estimate of the regression coefficients; however, the reduced sample size often leads to losses in efficiency, through higher standard errors. Another disadvantage of using complete case analysis whenever the number of explanatory variables J is of moderate size is that the probability of having enough complete cases to estimate the regression coefficients is low. In this setting using a MI procedure is paramount and leads to a significant reduction in the standard errors; however, this can induce a slight bias. White and Carlin (2010) show through an extensive simulation study that the increase in bias often time leads to a decreased empirical coverage rate for both **MAR** and **MNAR** data sets.

For our simulation study we set $J = 40$, $N = 1000$, and consider X_j that include both continuous and discrete variables to demonstrate the versatility of the copula approach without specifying any of the marginal distributions. This is precisely the scenario we described above; the probability of enough complete cases existing to estimate the regression coefficients is effectively 0.[9]

The distributions we consider for the elements of the design matrix are Gaussian, Bernoulli, Poisson and ordinal. To make imputation feasible we require the variables to be correlated. To generate correlated variables we first construct a matrix of correlated Gaussian random variables and then transform the variables to have the appropriate marginals. For example, to generate a pair of correlated Poisson random variables A and B with

[9]The reason is that with a high probability of missingness for each variable and a large enough number of variables, the probability of observing all variables for one particular case quickly becomes very small. Specifically, with probability of missingness p and k covariates, the probability of all observations being present for one case is $(1-p)^k$.

mean λ we construct $(Z_1, Z_2) \sim \mathcal{N}(0, \Sigma)$ where $\sigma_{11} = \sigma_{22} = 1$ and $\sigma_{12} = \sigma_{21} = \rho$ and set $A = F^{-1}_{\text{Pois},\lambda}(F_{\mathcal{N}}(Z_1))$ and $B = F^{-1}_{\text{Pois},\lambda}(F_{\mathcal{N}}(Z_2))$. The data generating process thus leads to the following marginal distributions for the entries in $\mathbf{X}$: for $j = 1, \ldots, 10$

$$X_j \sim \mathcal{N}(0, \sigma_j^2) \qquad X_{j+10} \sim \text{Bern}(p_j)$$

$$X_{j+20} \sim \text{Pois}(\lambda_j) \qquad X_{j+30} \sim \text{ordinal}(0, 1)$$

$$\mathbf{X} = (X_1, \ldots, X_{40}) \qquad Y \sim \mathcal{N}\left(\sum_{i=1}^{40} X_i, 1\right),$$

where $\sigma_j = 1 + (j-1)/9$, $\lambda_j = 0.2 + 2(j-1)/90$ and $p_j = 2 + 3(j-1)/9$. Both the amount of missingness (MC) and correlation (ρ) between the different variables is varied according to the specified values given in Table E.1.

[Table 1 about here.]

We consider two missing data mechanisms for $\mathbf{X}$, one that produces **MAR** data sets and another one that generates **MNAR** data sets, see Appendices A and B for details. The MI procedures we considered are only valid under the MAR assumption; however, it is useful to check how each method performs when this assumption is violated - as is often the case in practice.

4.3 Results

We performed 1,000 simulations under each of the possible combinations of the correlation and missingness coefficient, as detailed in Table E.1, under both **MAR** and **MNAR** missing data mechanisms. For `MICE`, we specified the correct marginal distributions (for

example ordered logit model for the ordinal variables). For `Amelia II`, we used the appropriate variable transformation in accordance with the package help files. For the copula approach, we did not need to specify any distributions/transformations. Using each of the three procedures, we created 20 completed data sets that were used to estimate the regression coefficients and a corresponding 95% confidence interval.[10]. None of the simulations had enough complete cases to estimate the regression coefficients using listwise deletion.

The most significant source of variation in the simulation was due to the different classes of variables, followed by the correlation and the missingness coefficient. There is only a small difference in the results obtained from the **MAR** and **MNAR** data; therefore, our discussion will focus on the former, with the figures for the latter included in Appendix C. Figures E.3 and E.2 illustrate how the bias, coverage and interval length, vary across the interaction of the different variable classes, the correlation, and the missingness coefficient, respectively. Overall the copula method achieved an empirical coverage rate of 93% which was much higher than that of `MICE`, 87%, and `Amelia II`, 83%. Less adversarial regimes were previously studied in White and Carlin (2010), by reducing the number of covariates in our simulation we can recover similar coverage rates for the MI procedures as are reported there. Both the copula and `MICE` methods had an absolute average bias of 0.17. `Amelia II` performed worse and had a bias of 0.25. On average, all three methods had approximately the same interval length.

The copula imputations were obtained using $10,000$ iterations from Hoff's (2010) package whose convergence was checked on a subset of simulations. The lag-10 autocorrelation

[10]Throughout the simulation, the `Amelia II` software crashed numerous times, as detailed in Table D.1 in Appendix D. Due to this the results for `Amelia II` are only on a subset of the $36,000$ simulations

for the thinned chains was less than 0.18 in absolute value for each of the elements of the latent correlation matrix, and the effective sample size was always above 200 (97.6% of the entries were above 500). Since the copula method is sampling from the posterior distribution which requires the MCMC algorithm to converge to the stationary distribution, its computation time depends on the rate of convergence as well as the desired number of imputations. Running multiple MCMC chains in parallel to generate independent imputations can reduce the computation time. This approach is slightly slower than `Amelia II` but is substantially faster than the standard `MICE` algorithm where all $J-1$ variables are used to impute the j^{th} variable. Fortunately, the copula algorithm scales well as the sample size and the number of explanatory variables increases. The copula method had the lowest bias, highest coverage rate and often the longest interval length. It is noteworthy that even though the semi-parametric estimation procedure did not require specification of the marginals, any data transformations, or tuning, it still outperformed the other two procedures.

Since the `MICE` procedure is iterative, users need to check that the model parameters fully explore the parameter space. Unlike the Bayesian copula method, there are no explicit convergence criteria that can be tracked. We performed a visual check that revealed no abnormalities and also ran each `MICE` chain for 20 iterations as recommended in Van Buuren and Groothuis-Oudshoorn (2011). The `MICE` method performed almost as well as the copula method but had slightly lower coverage rate, meaning the estimated standard errors were too small. `MICE` also had the smallest average bias for the normal and Poisson variables. Again, however, these results are contingent on specifying the correct conditional

distribution which can often be challenging.

`Amelia II` had the lowest coverage and highest bias both on average and in most scenarios that we considered. It had the smallest average interval length of 1.23, which shows that it was systematically underestimating the variance: leading to the low coverage rates.

Figure E.2 shows that the average bias and the interval length increases as a function of the proportion of missing values. This leads to a decrease in the empirical coverage as the bias increases at a faster rate than the interval length. One notable exception was the correct coverage of the copula approach for the regression parameters of the ordinal and binomial variables, both `Amelia II` and `MICE` undercovered. Given that these types of variables are frequently encountered in social science applications, these results especially suggest that using a copula approach can lead to better statistical conclusions. Moreover, the overall simulation results indicate that when a normal distribution does not well approximate the data, then the copula approach will consistently outperform both `Amelia II` and `MICE`.

Somewhat surprisingly, there seems to be less variation in the bias and the interval length as a function of the correlation, as is shown in Figure E.3. Except for the normally distributed variables, the bias decreases as the correlation increases due to the reduction in the relative loss of information from the missing data.

[Figure 2 about here.]

[Figure 3 about here.]

Breaking the **MAR** assumption did not lead to drastically worse results. We observe

a decrease of about 3% in the coverage of all three methods and a slight decrease in the average bias. This shows that the methods are somewhat robust to violations of MAR assumption when it is not too severe. Figures C.1 and C.2 in the Appendix C show the results of the simulations when the MAR assumption is violated.

5 Application Study

In this section, we provide a comparison of the three imputation methods using an application from political science. The empirical example shows how copula methods can be used to generate imputations in a large data set with a variety of variable types.

5.1 Inequality and Democratic Support

As we have elaborated above, imputation methods are still underused, especially in the social sciences. There is, however, some visible progress. One example where scholars have taken advantage of one of the imputation methods currently available is "Economic Inequality and Democratic Support" by Krieckhaus et al. (2014) published in the *Journal of Politics*. Krieckhaus et al. (2014) explore whether the support for democracy within countries is affected by the level of inequality. The authors combine country level variables (such as inequality) with individual level survey data from 40 democracies around the world. For multiple countries several survey waves are included, resulting in 57 country-years and a total of 77,642 observations (Krieckhaus et al., 2014, 144). For this replication exercise we replicate *Model* 1 in *Table 1* in Krieckhaus et al. (2014). The dependent variable is a "13-point additive index (ranging from 0 to 12) of democratic support", which the authors treat as a continuous variable (Krieckhaus et al., 2014, 144). The main independent variables of interest are *Inequality* at the country level, and an ordinal *Income* scale at the individual level (ranging from 1 to 10). Additionally, the authors control for *Age*, *Gender*, *Institutional Confidence*, *Interest in Politics*, *Interpersonal Trust*, *Education*, *Prior Regime Evaluation*, and *Leftist Ideology* all drawn from the *World Values Survey* (World Values Survey, 2012).

As in the original article, all individual level variables are demeaned "using group-mean centering" after the imputation (Krieckhaus et al., 2014, 145). The data are analyzed using a random-coefficients model.

[Table 2 about here.]

Most importantly for this study, the original data suffers from a relatively high number of missing observations. Table E.2 shows the share of missing observations for variables included in the replication exercise. We can see that many of the variables have a large share of missing observations. If instead of multiple imputations, the authors used in listwise deletion then the number of observation in the regression model would have been approximately halved. Instead, Krieckhaus et al. (2014) use `Amelia II` to multiple impute five data sets which they analyze. Estimates are then combined using Rubin's rule.

This is an excellent setting for our comparison of multiple imputation techniques. The number of missing observations is quite large, and the data set includes different types of variables, continuous, binary, as well as ordinal. We created 20 multiple imputed data sets using each of the imputation techniques: `Amelia II`, `MICE`, and `sbgcop`. We then re-estimate *Model 1* in *Table 1* in Krieckhaus et al. (2014, 147) and combine the estimation results for each method's multiple imputed data sets via Rubin's rule.

For `Amelia II` we specify the type of each variable and then generate 20 imputed data sets using the full original data. Similarly, we declare each variable's type for `MICE` and estimate the default model for each. We use all variables except the one to be imputed as independent variables in the chained equations. Again, we create 20 multiple imputed data sets and set the maximum number of iterations to 20.

Lastly, we use our preferred method, imputation via the semi-parametric Gaussian copula, to generate 20 imputed data sets. We run the MCMC chain for 2100 iterations and randomly draw 20 data sets from the posterior. Note that, again, we do not have to declare any of the variable types or make any other specification or transformation of the data.

Figure E.4 shows the coefficient estimates and 95% intervals for the replicated model based on each of the imputation techniques, as well as when list-wise deletion is used. First, the results are quite similar for the *Inequality*, *Income*, and *Age* variables. Even for the models based on listwise deletion. For the two main variables of interest, inequality, and income, the results based on different imputation techniques are virtually the same.

On the other hand, there are several significant differences for the other variables included in the model. First, the effect of gender is essentially zero according to the models estimated on the copula imputed data. Based on the data imputed using `MICE` or `Amelia II`, females have higher ratings of democracy satisfaction (though the confidence intervals just cover zero). According to the non-imputed data, the effect of gender is quite strong.

Based on the data imputed with the copula method, the estimated association of *Institutional Confidence* with *Democracy Satisfaction* is significantly stronger compared to the models based on listwise deletion or other imputation methods. Similarly, the estimated effect of *Leftist Ideology* is also substantially larger according to the copula imputed data. On the other hand, the association of *Education* levels with *Democracy Satisfaction* is significantly smaller. Based on the copula, the relationships of *Interest in Politics*, and *Prior Regime Evaluation* with the dependent variable of *Democracy Satisfaction* are all modeled

to be weaker, compared to the other methods (and the non-imputed data), though the confidence intervals overlap.

[Figure 4 about here.]

It is interesting to note, that, except for one variable (*Interpersonal Trust*), whenever the estimated coefficient for the copula imputed data differs from the coefficients based on the other imputation methods, it is in the opposite direction of the difference to the list-wise deletion coefficient. This is especially easy to see for the *Gender* and *Leftist Ideology* variables, where the effect is strongest (weakest) according to the model estimated on the list-wise deleted data and weakest (strongest) for the copula based models.

Based on the simulation results, especially with respect to binary and ordinal variables, and the theoretical properties we are confident in the accuracy of the copula imputation method. These results suggest then that *Gender* is not associated with people's satisfaction with democracy, whereas *Institutional Confidence* and *Left* ideology both have much stronger effects.

6 Conclusion

What practical lessons can we learn about how to deal with missing data? In this article, we re-emphasize the importance of dealing with missing data and present a copula based approach, developed by Hoff (2007), that is elegant and requires little pre-specification of the data. With the rank based approach introduced by Hoff (2007), the Gaussian copula can be used to impute binary, ordinal, and continuous variables. We discuss the theoretical properties of the copula method and its theoretical attractiveness compared to other commonly employed techniques. In particular, the Gaussian copula introduced here enables researchers to make imputation via draws from a valid posterior of the joint distribution without specifying the distributions of the individual variables. Moreover, we present evidence from simulations that it performs better than either `Amelia II` or `MICE`, especially when it comes to non-normally distributed data.

While the three imputation methods perform relatively similarly, throughout the simulation, the Copula method does have the lowest average bias (tied with `MICE`) and the highest coverage rate (93%). More so, `MICE` requires specification of the conditional distributions whereas the copula method does not. Recent theoretical results for `MICE` suggest that good performance heavily relies on being approximately correct in the choice of conditionals (Li et al., 2012). On the other hand, theoretical guarantees for good behavior of copula methods are available. In particular, information bounds for rank-based estimators are the same as the information bounds for estimators based on the full (scale and rank) data (Hoff et al., 2014). Under **MAR** and **MCAR** we inherit all the properties of the full data, and by introducing structure to the imputation, we are likely to have good behavior

even under **MNAR**.

One aspect that we have not addressed herein is the validity and sensitivity to the unassessable assumptions made when analyzing data with missing values (Molenberghs et al., 2014), *i.e.* the type of missingness mechanism. Bojinov et al. (2017) show that the Gaussian copula approach can be used to assess the validity of the missing always at random assumption (a slightly stronger assumption that implies **MAR**). Their results suggest that by using a Gaussian copula for generating imputations, the analyst can also easily diagnose the assumptions they made and quickly identify variables which are likely to break these assumptions. This adds another benefit to using the method discussed in this paper.

Consideration must also be given to the computational cost of any procedure. As indicated by Graham (2009) the disadvantages of EM approaches are especially large when imputing databases with many variables or applications of "big data". `MICE` can be computationally less expensive but suffers when the number of variables increases as the correct choice for each of the conditionals becomes increasingly unlikely. The semiparametric copula approach described here relies on MCMC, its speed does not depend on the fraction of missing data and scales nicely in the dimension of the dataset. This makes it possible to impute even large database in a relatively timely manner and no pre-specification of the data. Moreover, using the copula model to multiply impute missing values provides some of the advantages (such as a proper posterior distribution of the data) but is less burdensome on scholars than imputing values in a fully Bayesian approach (Erler et al., 2016).

Finally, the copula approach is quite flexible and can be employed at different stages of

the analysis process. First, it can be used to generate a single estimate of the missing data or the mean of a large number of draws, which is exactly what might be needed in some situations. Second, per the recommendation of Rubin, it can be used to construct multiple databases. As with `Amelia II`, the copula imputations can be analyzed separately and the results combined using either `mitools` or `Zelig` (Imai et al., 2008) in $\mathcal{R}$. Thus, the copula approach to missing data can be explicitly integrated into the modeling and analysis of observational data in a simplistic, organic fashion.

SUPPLEMENTARY MATERIAL

Code will be provided on the author's Dataverse.

R-packages for Imputation: 3 R-packages used to impute the missing data: Amelia II, MICE, sbgcop

R-code for simulation in Section 4: R-code to replicate simulation study in section 4.

R-code for Application in Section 5: R-code to replicate application in section 5.

A Missing at Random

We now describe a missing data mechanism that always produces **MAR** data. Our goal is to make the simulations as realistic as possible; therefore some variables will be fully observed, and others will have different amounts of missing values.

1. Given a fully observed data set $\mathbf{X}$ randomly select four variables, one from each of the four classes, that will be fully observed; without loss of generality relabel them X_1, X_{11}, X_{21} and X_{31}.

2. Randomly select four variables from the remaining thirty six, one from each of the four classes, that will have a 5-6% missingness; without loss of generality relabel them X_2, X_{12}, X_{22} and X_{32}. The probability that the i^{th} observation for each variable is missing is based on a logistic regression on the fully observed variables, X_1, X_{11}, X_{21} and X_{31}, adjusted so that the mean number of missing variables is between 5-6%. The missingness indicators are then sampled from independent Bernoulli random variables with the appropriate probabilities. Let $\mathbf{X}^{(1)} = (X_1, X_2, X_{11}, X_{12}, X_{21}, X_{22}, X_{31}, X_{32})$ and $\mathbf{X}^{(1)}_{\text{cc}}$ be the complete cases after removing the any rows that have missing values.

3. The probability of the i^{th} observation missing for the remaining thirty two variables is proportional to a logistic regression on the fully observed $\mathbf{X}^{(1)}_{\text{cc}}$. The probabilities are then adjusted so that the mean number of missing variables is equal to the Missingness Coefficient (MC) (see Table 1 for the range of values that we considered). The missingness indicators are sampled from independent Bernoulli random variables with the appropriate probabilities. If the i^{th} row of $\mathbf{X}^{(1)}$ has been removed in $\mathbf{X}^{(1)}_{cc}$

then that row is always observed for the thirty-two variables.

The proportion of missing values is slightly lower than the MC as four variables are fully observed, and four others only have 5-6% of their values missing.

B Missing not at Random

We now describe a missing data mechanism that produces **MNAR** data with extremely high probability.

1. Given a fully observed data set $\mathbf{X}$ randomly select four variables, one from each of the four classes, that will be fully observed; without loss of generality relabel them X_1, X_{11}, X_{21} and X_{31}.

2. Randomly select four variables from the remaining thirty six, one from each of the four classes, that will have a small amount of missingness; without loss of generality relabel them X_2, X_{12}, X_{22} and X_{32}. The probability that the i^{th} observation is missing is given by,

$$P(R_2 = 1|\mathbf{X}) = 1_{X_2>0} p_{MC},$$

$$P(R_{12} = 1|\mathbf{X}) = 1_{X_{12}=0} p_{MC},$$

$$P(R_{22} = 1|\mathbf{X}) = 1_{X_{22}>3} p_{MC},$$

$$P(R_{32} = 1|\mathbf{X}) = 1_{X_{32}=3} p_{MC},$$

where the value of p_{MC} is given by the MC in Table 1.

3. For the remaining thirty two variables the probability of the i^{th} observation missing is based on a logistic regression on $\mathbf{X}^{(1)}$ adjusted so that the mean number of missing variables is equal to the MC (see Table 1). The missingness indicators are again sampled from independent Bernoulli random variables with the appropriate probabilities. In contrast to the MAAR mechanism if the i^{th} row of $\mathbf{X}^{(1)}$ has missing values then other variables in that row can still be missing.

C Plots of MNAR Simulation Results

[Figure 5 about here.]

[Figure 6 about here.]

D Number of Simulations for which `Amelia II` crashed

[Table 3 about here.]

E Example `sbgcop` Application

In this section, we discuss how to use the 'sbgcop' package for multiple imputation in the context of conducting inferential analysis on data with missingness. Specifically, we show how to conduct regression analysis in the presence of missing data using an example dataset. First we simulate a dataset in which we introduce missingness.

```
# simulate data
set.seed(6886)
n <- 100
x1 <- rnorm(n) ; x2 <- rnorm(n) ; x3 <- rnorm(n)
y <- 1 + 2*x1 -1*x2 + 1*x3 + rnorm(n)

## organize into matrix
raw <- cbind(y, x1, x2, x3)

## simulate missingness
naMat <- matrix(rbinom(n*4,1,.7),
    nrow=nrow(raw),ncol=ncol(raw))
naMat[naMat==0] <- NA

## remove observations
data <- raw * naMat

## summarize missingness
missStats <- apply(data, 2, function(x){sum(is.na(x))/nrow(data)})
missStats <- matrix(missStats,
    ncol=1,
    dimnames=list(colnames(data),'Prop. Missing')
    )
```

Using this simulated dataset, our goal is to show how to conduct inference on the effect

of x_1, x_2, and x_3 on y after imputing the missing values with the `sbgcop` package in R. `sbgcop` is available on CRAN and can be installed and loaded into your R session just as any other package.

```
install.packages('sbgcop')
library(sbgcop)
```

The key function in this package is `sbgcop.mcmc` and there are four arguments that should always be set (for a full list of arguments run `?sbgcop.mcmc`):

- Y: a matrix with missing values to be imputed
- `nsamp`: number of iterations of the Markov chain
- `odens`: number of iterations between saved samples
- `seed`: an integer for the random seed

The Y argument specifies the dataset to be imputed. The object passed to the argument must be in **matrix** format. Additionally, users should only include variables that can provide information to the imputation algorithm. For example, this can include lags and leads of a variable in the case of time-series-cross-sectional data. Identification variables, such as actor names, abbreviations, or years, should not be included in the **matrix**.

The imputation procedure in `sbgcop.mcmc` is a Bayesian estimation scheme, so users must pass the number of iterations for which they want the Markov chain to be run to the `nsamp` argument. If `nsamp` is set to 100, then the Markov chain will run for 100 iterations and 100 imputed datasets will be produced. The `odens` argument specifies how often an

iteration from the Markov chain should be saved. Thus, if `nsamp` is set to 100 and `odens` is set to 4, 25 imputed datasets will be returned by `sbgcop.mcmc`. Last, since this is a Bayesian model and we will be sampling from distributions to arrive at parameter values, one should always pass an integer to the `seed` argument. This way when users rerun `sbgcop.mcmc` they will arrive at the same results.

To impute missingness in our example dataset, we pass our `data` object to the `sbgcop.mcmc` function. We run the Markov chain for 2000 iterations and save every 10th iteration. We store the output from `sbgcop.mcmc` to `sbgcopOutput`.

```
26  sbgcopOutput <- sbgcop.mcmc(Y=data, nsamp=2000, odens=10, seed=6886)
```

This is quite simple to do as the output from `sbgcop.mcmc` is simply a list. The first element in this list is `C.psamp`, which contains posterior samples of the correlation matrix. The `C.psamp` is structured as an array of size p x p x `nsamp`/`odens`. Where p indicates the number of variables included in the imputation process. In our case, the `data` object includes 4 variables and we ran the Markov chain for 2000 iterations saving every tenth. Thus giving us dimensions of: 4 x 4 x 200.

Each value in this array is providing us with the estimated association between a pair of parameters at every saved iteration of the Markov chain. We show an example below using the 100th and 200th saved iterations.

```
sbgcopOutput$C.psamp[,,c(100,200)]

## , , 100
##
##             y          x1          x2          x3
## y   1.0000000  0.78961179 -0.43494151  0.36593885
## x1  0.7896118  1.00000000 -0.08686933  0.05172101
## x2 -0.4349415 -0.08686933  1.00000000 -0.14619182
## x3  0.3659389  0.05172101 -0.14619182  1.00000000
##
## , , 200
##
##             y         x1          x2         x3
## y   1.0000000 0.68269537 -0.46139236  0.4138161
## x1  0.6826954 1.00000000  0.08754115  0.1495993
## x2 -0.4613924 0.08754115  1.00000000 -0.1278238
## x3  0.4138161 0.14959933 -0.12782384  1.0000000
```

To generate a trace plot of this data we need to restructure our dataframe into a long format. We can do so using the `reshape2` package:

```
  library(reshape2)
  sbgcopCorr = reshape2::melt(sbgcopOutput$'C.psamp')

  # remove cases where variable is the same in both columns
  sbgcopCorr = sbgcopCorr[sbgcopCorr$Var1 != sbgcopCorr$Var2,]

  # construct an indicator for pairs of variables
  sbgcopCorr$v12 = paste(sbgcopCorr$Var1, sbgcopCorr$Var2, sep='-')

  #
  print(head(sbgcopCorr))

##   Var1 Var2 Var3       value   v12
## 2   x1    y    1  0.62439270  x1-y
## 3   x2    y    1 -0.43347850  x2-y
## 4   x3    y    1  0.28013565  x3-y
## 5    y   x1    1  0.62439270  y-x1
## 7   x2   x1    1  0.03581958 x2-x1
## 8   x3   x1    1  0.15626246 x3-x1
```

Using the `reshape2` package we have reformatted the array into a dataframe, in which the first two columns designate the variables for which a correlation is being estimated, the third an indicator of the saved iteration, the fourth the correlation, and the fifth an indicator designating the variables being compared.

Next, we use `ggplot2` to construct a simple trace plot shown in Figure E.7.

```
63  library(ggplot2)
64
65  ggplot(sbgcopCorr, aes(x=Var3, y=value, color=v12)) +
66      geom_line() +
67      ylab('Correlation') + xlab('Iteration') +
68      facet_wrap(~v12) +
69      theme(legend.position='none')
```

[Figure 7 about here.]

Based on these trace plots we can see that the Markov chain tends to converge rather quickly in this example. The `coda` package provides an excellent set of diagnostics to test convergence in more depth.

After conducting the imputation and evaluating convergence, our goal is now to use the imputed datasets to conduct inferential analysis. For the purpose of this example, we estimate the effect of x_1, x_2, and x_3 on y. By using `sbgcop` as above we have generated 200 copies of our original dataset in which posterior samples of the original missing values have been included. Each of these copies are saved in the output from `sbgcop.mcmc`, which has dimensions of 100 x 4 x 200.

The first two dimensions of this object correspond to the original dimensions of our `data` object, and the third corresponds to the number of saved iterations from the Markov

chain.

Having generated a set of imputed datasets, our next step is to use a regression model to estimate the effect of our independent variables on y. We cannot just use one of the imputed datasets – as this would not take into account the uncertainty in our imputations. Instead we run several regression on as many of the imputed datasets generated by `sbgcop.mcmc` that we think are appropriate. For the sake of this example, we utilize all 200 imputed datasets, but typically randomly sampling around 20 imputed datasets should be be sufficient.

Each time we run the regression model, we will save the coefficient and standard errors for the independent variables and organize the results into a matrix as shown below.

```
  coefEstimates <- NULL
  serrorEstimates <- NULL
  for( copy in 1:dim(sbgcopOutput$'Y.impute')[3]){
          # extract copy from sbgcopOutput
          copyDf <- data.frame(sbgcopOutput$'Y.impute'[,,copy])
          names(copyDf) <- colnames(sbgcopOutput$Y.pmean)
          # run model
          model <- lm(y~x1+x2+x3,data=copyDf)
          # extract coefficients
          beta <- coef(model)
          coefEstimates <- rbind(coefEstimates, beta)
          # extract standard errors
          serror <- sqrt(diag(vcov(model)))
          serrorEstimates <- rbind(serrorEstimates, serror)
  }

  print(head(coefEstimates))

##      (Intercept)       x1         x2        x3
## beta   0.6576411 1.449662 -1.1290934 0.4569379
## beta   0.7436243 1.661250 -1.0542155 0.6866980
## beta   0.8299671 1.613892 -1.1363969 0.7454211
## beta   0.8073597 1.513452 -0.7512275 0.6331863
## beta   0.8112010 1.583065 -0.9608251 0.6529509
```

```
## beta   0.7882072 1.509635 -0.5152139 0.8897130
```

The last step is to combine each of the estimates using using Rubin's rule. Many existing packages have implemented functions to aid in this last step, one could use the `pool` function from `mice` or the `mi.meld` function from `Amelia II` as below.

```
  paramEstimates <- Amelia::mi.meld(q=coefEstimates, se=serrorEstimates)
  print(paramEstimates)

## $q.mi
##      (Intercept)      x1         x2        x3
## [1,]    0.892732 1.70032 -0.9023761 0.7235922

## $se.mi
##      (Intercept)        x1        x2        x3
## [1,]   0.1680402 0.1965969 0.2213771 0.1588638
```

The resulting parameter estimates take into account the uncertainty introduced through the imputation process, and we can interpret them just as we would interpret the results from a typical regression.

Below we show the full set of steps required to conduct a regression analysis in the context of missing data using `sbgcop`.

```
library(sbgcop)
sbgcopOutput <- sbgcop.mcmc(Y=data, nsamp=2000, odens=10, seed=6886)

## restructure posterior samples of correlation matrix
library(reshape2)
sbgcopCorr = reshape2::melt(sbgcopOutput$'C.psamp')
sbgcopCorr = sbgcopCorr[sbgcopCorr$Var1 != sbgcopCorr$Var2,]
sbgcopCorr$v12 = paste(sbgcopCorr$Var1, sbgcopCorr$Var2, sep='-')

## trace plot of C.psamp
library(ggplot2)
```

```
ggplot(sbgcopCorr, aes(x=Var3, y=value, color=v12)) +
        geom_line() +
        ylab('Correlation') + xlab('Iteration') +
        facet_wrap(~v12) +
        theme(legend.position='none')

## conduct regression analysis
coefEstimates <- NULL
serrorEstimates <- NULL
for( copy in 1:dim(sbgcopOutput$'Y.impute')[3]){
                copyDf <- data.frame(sbgcopOutput$'Y.impute'[,,copy])
        names(copyDf) <- colnames(sbgcopOutput$Y.pmean)
                model <- lm(y~x1+x2+x3,data=copyDf)
                beta <- coef(model)
                coefEstimates <- rbind(coefEstimates, beta)
                serror <- sqrt(diag(vcov(model)))
                serrorEstimates <- rbind(serrorEstimates, serror) }

## combine estimates using Rubin's rules
paramEstimates <- Amelia::mi.meld(q=coefEstimates, se=serrorEstimates)
```

References

Bojinov, I., N. Pillai, and D. Rubin (2017). Diagnosing missing always at random in multivariate data. *arXiv preprint arXiv:1710.06891*.

Casella, G. and E. I. George (1992). Explaining the Gibbs Sampler. *The American Statistician 46*(3), 167–174.

Chen, S.-H. and E. H. Ip (2015). Behaviour of the Gibbs sampler when conditional distributions are potentially incompatible. *Journal of statistical computation and simulation 85*(16), 3266–3275.

Chen, X., Y. Fan, and V. Tsyrennikov (2006). Efficient Estimation of Semiparametric Multivariate Copula Models. *Journal of the American Statistical Association 101*(475), 1228–1240.

Di Lascio, F., S. Giannerini, and A. Reale (2015). Exploring copulas for the imputation of complex dependent data. *Statistical Methods & Application 24*(1), 159–174.

Erler, N. S., D. Rizopoulos, J. v. Rosmalen, V. W. Jaddoe, O. H. Franco, and E. M. Lesaffre (2016). Dealing with Missing Covariates in Epidemiologic Studies: A Comparison Between Multiple Imputation and a Full Bayesian Approach. *Statistics in Medicine*.

Goodrich, B., J. Kropko, A. Gelman, and J. Hill (2012). mi: Iterative Multiple Imputation from Conditional Distributions. R package.

Graham, J. W. (2009). Missing Data Analysis: Making it Work in the Real World. *Annual Review of Psychology 60*(1), 549–576.

Hoff, P. (2010). *sbgcop: Semiparametric Bayesian Gaussian Copula Estimation and Imputation*. R package version 0.975. https://CRAN.R-project.org/package=sbgcop.

Hoff, P. D. (2007). Extending the Rank Likelihood for Semiparametric Copula Estimation. *Annals of Applied Statistics 1*(1), 265–283.

Hoff, P. D., X. Niu, and J. A. Wellner (2014). Information Bounds for Gaussian Copulas. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability 20*(2), 604.

Honaker, J. and G. King (2010, April). What to do About Missing Values in Time-Series Cross-Section Data. *American Journal of Political Science 54*(2), 561–581.

Honaker, J., G. King, and M. Blackwell (2012). AMELIA II: A Program for Missing Data – Documentation.

Imai, K., G. King, and O. Lau (2008). Toward A Common Framework for Statistical Analysis and Development. *Journal of Computational and Graphical Statistics 17*(4), 892–913.

Käärik, E. (2006). Imputation algorithm using copulas. *Metodoloski zvezki 3*(1), 109.

Käärik, E. and M. Käärik (2009). Modeling dropouts by conditional distribution, a copula-based approach. *Journal of Statistical Planning and Inference 139*, 3830–3835.

King, G., J. Honaker, A. Joseph, and K. Scheve (2001, March). Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Politial Science Review 95*(1), 49–69.

Klaassen, C. A., J. A. Wellner, et al. (1997). Efficient Estimation in the Bivariate Normal Copula Model: Normal Margins are Least Favourable. *Bernoulli 3*(1), 55–77.

Krieckhaus, J., B. Son, N. Bellinger, and J. Wells (2014). Economic Inequality and Democratic Support. *The Journal of Politics 76*(1), 139–151.

Kropko, J., B. Goodrich, A. Gelman, and J. Hill (2014). Multiple Imputation for Continuous and Categorical Data: Comparing Joint Multivariate Normal and Conditional Approaches. *Political Analysis 22*(4), 497–519.

Li, F., Y. Yu, and D. B. Rubin (2012). Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines. *Duke University Department of Statistical Science Discussion Paper 1124*.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (second ed.). New York: Wiley.

Liu, J., A. Gelman, J. Hill, Y.-S. Su, and J. Kropko (2013). On the Stationary Distribution of Iterative Imputations. *Biometrika 101*(1), 155–173.

Marini, M. M., A. R. Olsen, and D. B. Rubin (1980). Maximum-likelihood estimation in panel studies with missing data. *Sociological methodology 11*, 314–357.

Mealli, F. and D. B. Rubin (2015). Clarifying Missing at Random and Related Definitions, and Implications when Coupled with Exchangeability. *Biometrika 102*(4), 995–1000.

Molenberghs, G., G. Fitzmaurice, M. G. Kenward, A. Tsiatis, and G. Verbeke (2014). *Handbook of Missing Data Methodology*. Boca Raton, FL: Chapman and Hall/CRC.

Murray, J. S. (2013). Some Recent Advances in Non- and Semiparametric Bayesian Modeling with Copulas, Mixtures, and Latent Variables. Dissertation. Department of Statistical Science Duke University. http://dukespace.lib.duke.edu/dspace/handle/10161/8253.

Murray, J. S., D. B. Dunson, L. Carin, and J. E. Lucas (2013). Bayesian Gaussian Copula Factor Models for Mixed Data. *Journal of the American Statistical Association 108*(502), 656–665.

Pettitt, A. (1982). Inference for the Linear Model using a Likelihood Based on Ranks. *Journal of the Royal Statistical Society. Series B (Methodological) 44*(2), 234–243.

Pitt, M., D. Chan, and R. Kohn (2006). Efficient Bayesian Inference for Gaussian Copula Regression Models. *Biometrika 93*(3), 537–554.

R Development Core Team (2004). *R: A language and environment for statistical computing*. Vienna, Austria.

Robbins, M. W., S. K. Ghosh, and J. D. Habiger (2013). Imputation in High-Dimensional Economic Data as Applied to the Agricultural Resource Management Survey. *Journal of the American Statistical Association 108*(501), 81–95.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika 63*(3), 581–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American statistical Association 91*(434), 473–489.

Rubin, D. B. (2004). *Multiple imputation for Nonresponse in Surveys*, Volume 81. John Wiley & Sons.

Sklar, A. (1959). Fonctions de Répartition à N Dimensions et Leur Marges. *Publications de l'Institut Statistique de l'Université Paris 8*, 229–231.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC Press.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software 45*(3), 1–67.

Van Buuren, S. and K. Groothuis-Oudshoorn (2011). MICE: Multivariate imputation by chained equations in R. *Journal of statistical software 45*(3), 1–67.

White, I. R. and J. B. Carlin (2010). Bias and Efficiency of Multiple Imputation Compared with Complete-Case Analysis for Missing Covariate Values. *Statistics in Medicine 29*(28), 2920–2931.

World Values Survey (2012). 1981-2008 Integrated Questionnaire.

Yucel, R. M. (2011). State of the Multiple Imputation Software. *Journal of Statistical Software 45*(1), 1 – 7.

Zhu, J. and T. E. Raghunathan (2015). Convergence Properties of a Sequential Regression Multiple Imputation Algorithm. *Journal of the American Statistical Association 110*(511), 1112–1124.

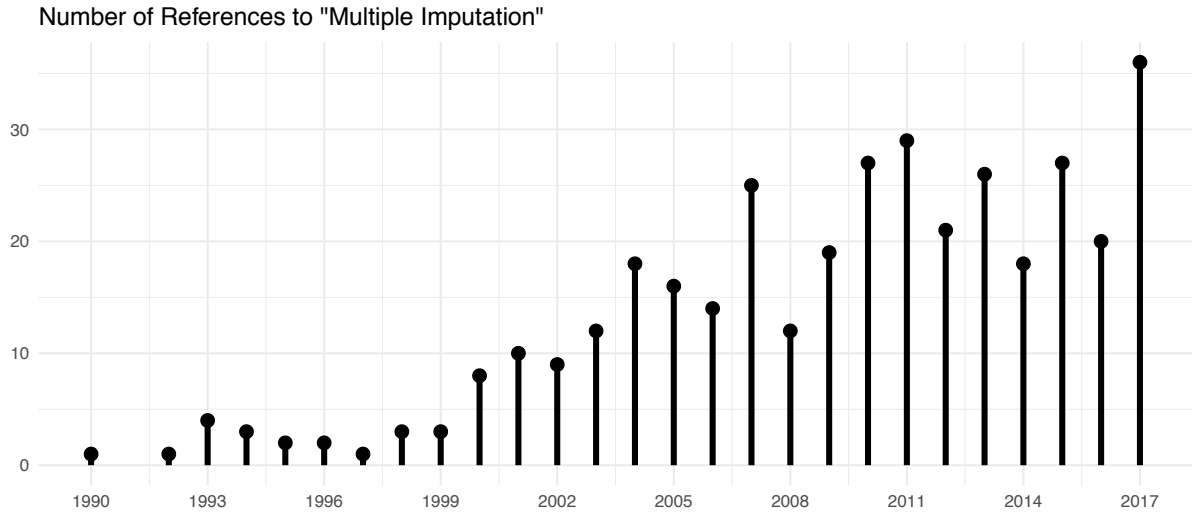


Figure E.1: Number of references to "multiple imputation" in articles from five top sociology and political science journals since 1990.

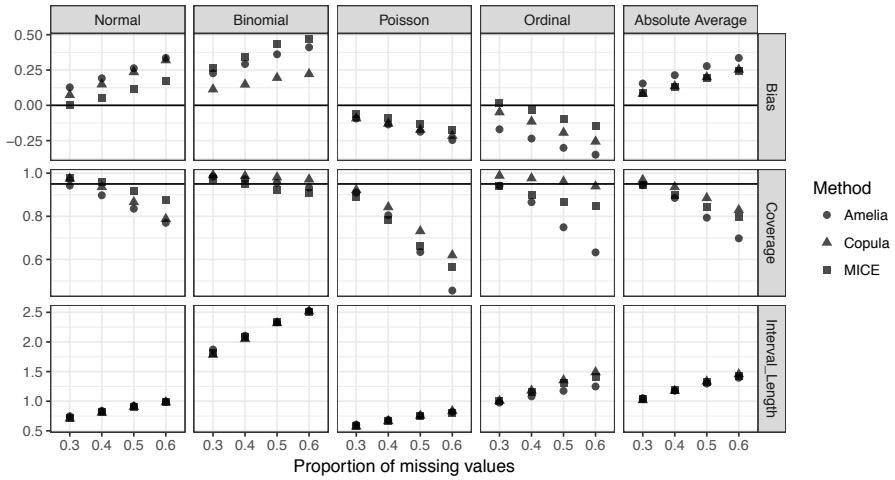


Figure E.2: Simulation study results for the **MAR** data as a function of the missingness coefficient, averaging over the correlation. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interested (the bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

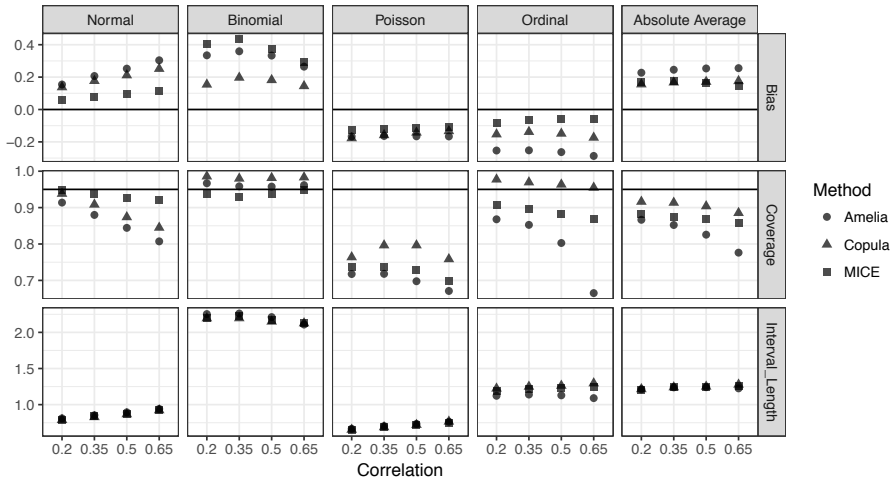


Figure E.3: Simulation study results for the **MAR** data as a function of the correlation, averaging over the missingness coefficient. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interested (the bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

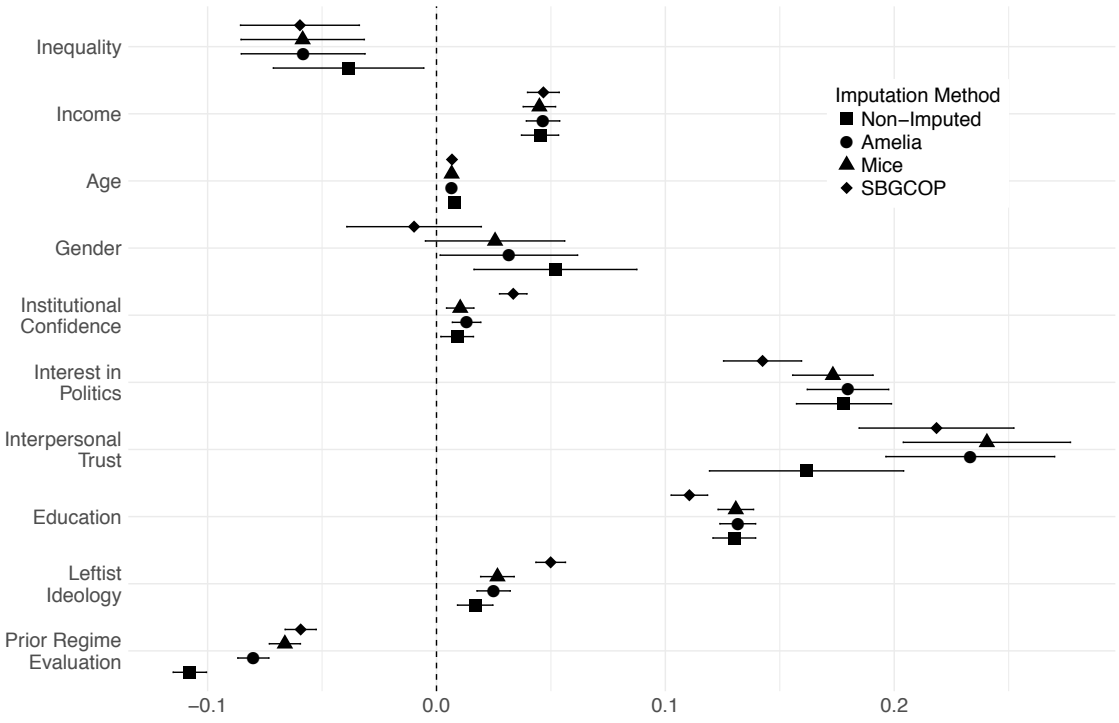


Figure E.4: Coefficient estimates and confidence intervals for *Model 1* in *Table 1* in Kriecken-haus et al. (2014) based on three imputation techniques and list-wise deletion

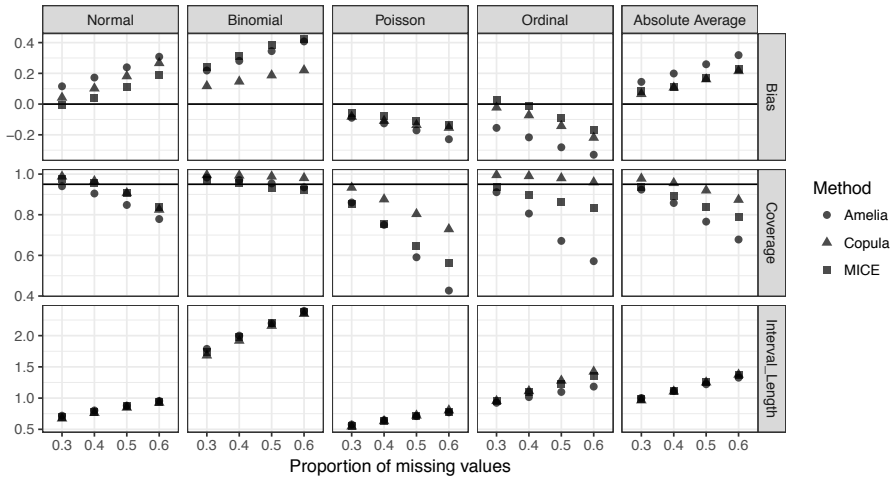


Figure E.5: Simulation study results for the **MNAR** data as a function of the missingness coefficient, averaging over the correlation. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interested (the bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

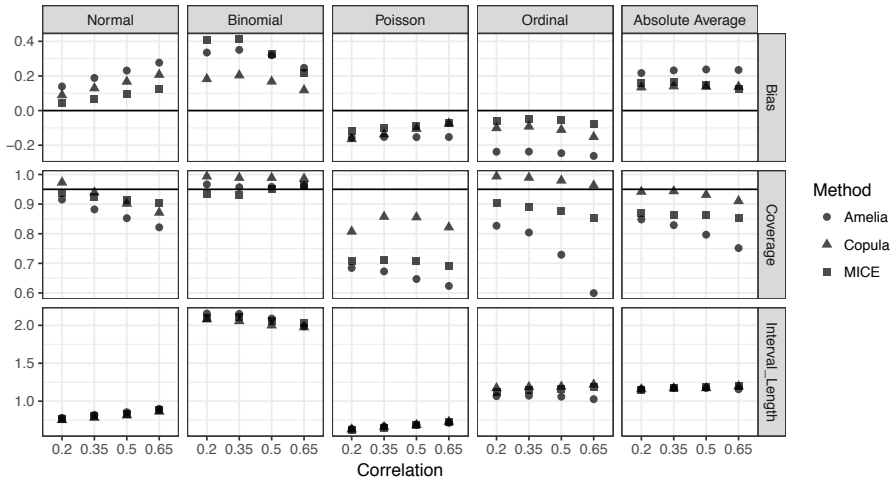


Figure E.6: Simulation study results for the **MNAR** data as a function of the correlation, averaging over the missingness coefficient. The plot is split by the different variable types (normal, binomial, Poisson and ordinal) and the three outcomes of interested (the bias, coverage and interval length). The rightmost panel shows the result averaging over the different variable types.

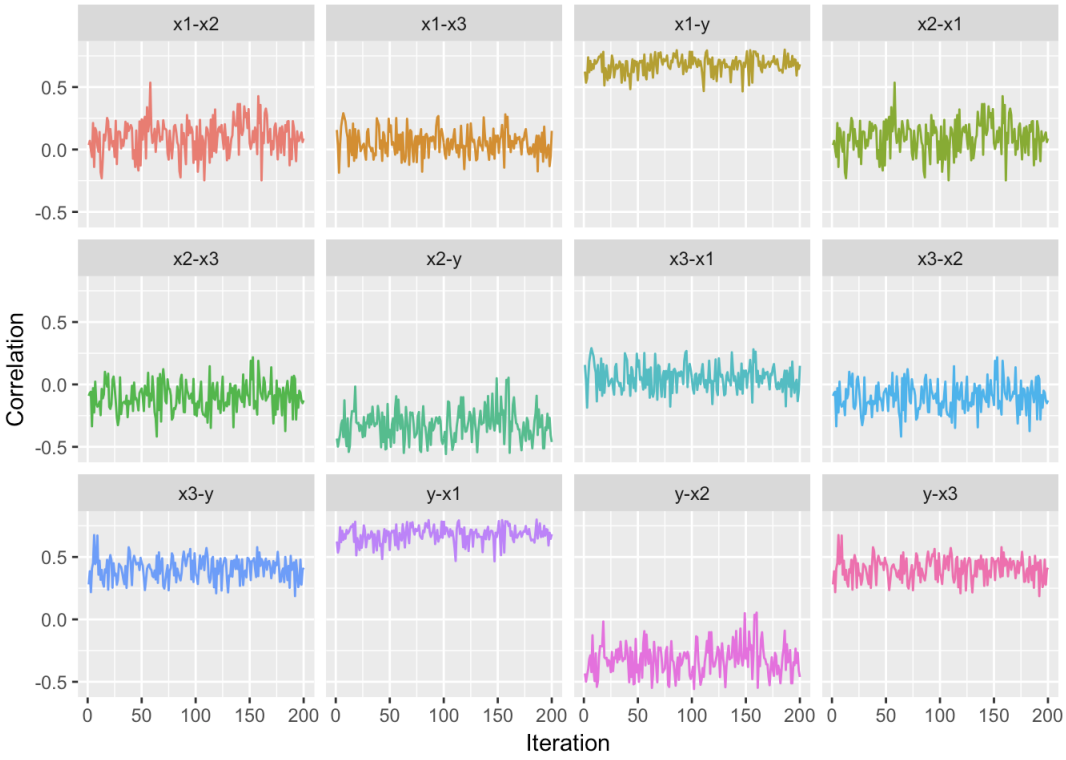


Figure E.7: Trace plot of correlation between variables.

Correlation (ρ)	Missingness Coefficient (MC)
0.2	0.3
0.35	0.4
0.5	0.5
0.65	0.6

Table E.1: Simulation Study configurations.

Table E.2: Share of Missingness in Variables of Interest

Democracy Support	Inequality	Income	Age
19.9	1.8	12.9	0.2
Gender	Institutional Confidence	Interest in Politics	Interpersonal Trust
0.1	11.7	2.5	3.7
Education	Leftist Ideology	Prior Regime Evaluation	
3.9	18.5	21.3	

		Correlation			
		0.2	0.35	0.5	0.65
Share of Missingness	0.3	2	0	0	7
	0.4	93	16	8	0
	0.5	285	138	37	13
	0.6	485	305	159	72

Table E.3: The number of `Amelia II` crashes out of the 1000 simulations under each of the possible scenarios.

B.J.Pol.S. **46**, 31–43 *Copyright © Cambridge University Press, 2014*
doi:10.1017/S0007123414000180
First published online 9 June 2014

Cabinet Formation and Portfolio Distribution in European Multiparty Systems

JOSH CUTLER, SCOTT DE MARCHI, MAX GALLOP, FLORIAN M. HOLLENBACH, MICHAEL LAVER AND MATTHIAS ORLOWSKI*

Government formation in multiparty systems is of self-evident substantive importance, and the subject of an enormous theoretical literature. Empirical evaluations of models of government formation tend to separate government formation *per se* from the distribution of key government pay-offs, such as cabinet portfolios, between members of the resulting government. Models of government formation are necessarily specified *ex ante*, absent any knowledge of the government that forms. Models of the distribution of cabinet portfolios are typically, though not necessarily, specified *ex post*, taking into account knowledge of the identity of some government 'formateur' or even of the composition of the eventual cabinet. This disjunction lies at the heart of a notorious contradiction between predictions of the distribution of cabinet portfolios made by canonical models of legislative bargaining and the robust empirical regularity of proportional portfolio allocations – Gamson's Law. This article resolves this contradiction by specifying and estimating a *joint* model of cabinet formation and portfolio distribution that, for example, predicts *ex ante* which parties will receive zero portfolios rather than taking this as given *ex post*. It concludes that canonical models of legislative bargaining do increase the ability to predict government membership, but that portfolio distribution between government members conforms robustly to a proportionality norm because portfolio distribution follows the much more difficult process of policy bargaining in the typical government formation process.

Government formation in multiparty systems is of self-evident substantive importance (especially to political scientists), and it is dealt with in a large and heterodox theoretical literature. Key outputs of the government formation process – for example the partisan composition of the cabinet, the distribution of portfolios among cabinet members, bargaining delays and cabinet durations – are easily and unambiguously observable. This enables a more clear-cut confrontation between theory and data than is typically the case in political science. It has also thrown into sharp relief a puzzle that Warwick and Druckman call the portfolio allocation paradox: a notorious contradiction between (1) predictions of the distribution of cabinet portfolios made by canonical alternating offers models of legislative bargaining and (2) the robust empirical regularity of proportional portfolio allocations that has been characterized as Gamson's Law (GL).[1]

Empirical models of government *formation* are necessarily specified *ex ante*, absent any knowledge of the government that forms.[2] Empirical models of the distribution of cabinet portfolios are typically specified *ex post*, given knowledge of the identity of some putative

* Cutler, De Marchi, Gallop and Hollenbach are at the Department of Political Science, Duke University. Laver is at the Department of Politics, New York University. Orlowski is at the Berlin Graduate School of Social Sciences, Humboldt-University Berlin (emails: josh.cutler@duke.edu, max.gallop@duke.edu, demarchi@duke.edu, florian.hollenbach@duke.edu, michael.laver@nyu.edu, mace84@gmail.com). Data replication sets and online appendices are available at http://dx.doi.org/doi: 10.1017/S0007123414000180.

[1] Warwick and Druckman 2006.

[2] See, for example, Martin and Stevenson 2001, 2010.

government *formateur* or of the full partisan composition of the cabinet.[3] We argue below that this disjunction lies at the heart of the portfolio allocation paradox, and set out to resolve the paradox by specifying and estimating a joint *ex ante* model of cabinet formation and portfolio distribution. An important feature of such a joint model is that it treats the list of parties *outside* the cabinet as informative, predicting *ex ante* which parties will receive zero portfolios rather than taking this as given. Our joint model allows us to conclude that canonical models of legislative bargaining do add to our ability to predict government membership, but that portfolio distribution between government members conforms robustly to a proportionality norm, which supports recent *ex ante* theoretical models of portfolio distribution.

RESOLVING THE PORTFOLIO ALLOCATION PARADOX

The Paradox

The essential features of the portfolio allocation paradox have been extensively discussed by Warwick and Druckman, among many others.[4] We confine ourselves here to its bare bones. A long tradition of empirical research linking legislative seat shares to distributions of cabinet positions originates in work published over fifty years ago by the sociologist William Gamson.[5] This established a strong and non-trivial empirical regularity: GL. Government parties tend to receive cabinet portfolios in strict proportion to the legislative seats they contribute to the government's aggregate seat total. Gamson did not provide a formal bargaining model that yielded GL as a prediction. Indeed, if party leaders are motivated to maximize Gamsonian (proportional to seats) pay-offs, the government coalition should command the smallest legislative seat total that is also a legislative majority. This is not true empirically, a finding that was first published over forty years ago.[6] Nonetheless, GL itself has proved extraordinarily robust to replication, even when different cabinet portfolios are assigned very different empirical weights.[7] On the other side of the paradox we find canonical alternating offers models of legislative bargaining, with an intellectual pedigree traceable to Rubinstein, adapted to legislative bargaining by Baron and Ferejohn.[8] This approach was applied explicitly to portfolio distribution by Ansolabehere et al., who derive two propositions about legislative bargaining over government formation in multiparty systems:[9]

(1) 'Elementary microeconomic theory teaches that in competitive situations perfect substitutes have the same price … We show that the noncooperative bargaining model of David P. Baron and John A. Ferejohn (1989) leads naturally to the result that expected payoffs are proportional to voting weights'.[10]

[3] See, for example, Ansolabehere et al. 2005; Falcó-Gimeno and Indridason 2013; Snyder, Ting, and Ansolabehere 2005; Warwick and Druckman 2001, 2006.

[4] Bassi 2013; Carroll and Cox 2007a; Falcó-Gimeno and Indridason 2013; Laver, Marchi, and Mutlu 2011; Warwick and Druckman 2006.

[5] Ansolabehere et al. 2005; Browne and Franklin 1973; Browne and Frendreis 1980; Fréchette, Kagel, and Morelli 2005b; Gamson 1961; Laver, Marchi, and Mutlu 2011; Schofield and Laver 1985; Snyder, Ting, and Ansolabehere 2005; Warwick and Druckman 2006.

[6] Fréchette, Kagel, and Morelli 2005a, 2005b; Taylor and Laver 1973.

[7] Ansolabehere et al. 2005; Browne and Franklin 1973; Browne and Frendreis 1980; Fréchette, Kagel, and Morelli 2005b; Gamson 1961; Laver, Marchi, and Mutlu 2011; Schofield and Laver 1985; Snyder, Ting, and Ansolabehere 2005; Warwick and Druckman 2006.

[8] Baron and Ferejohn 1989; Rubinstein 1982.

[9] Ansolabehere et al. 2005; Snyder, Ting, and Ansolabehere 2005.

[10] Snyder, Ting, and Ansolabehere 2005, 982.

(2) 'One key prediction of the model … is that the party that is recognized to form a coalition – the *formateur* – will receive a share of the cabinet posts that is much larger than its share of the voting weight'.[11]

On the face of things these two propositions seem contradictory: 'payoffs are proportional to voting weights' but 'the *formateur* gets a payoff much larger than its share of voting weight'. The apparent contradiction arises because the first proposition is stated *ex ante*, absent knowledge of the identity of the *formateur*. The second proposition is stated *ex post*, with knowledge of the identity of the *formateur*. The propositions are therefore not contradictory because they apply in different settings, but this highlights a critical distinction between *ex ante* and *ex post* models of cabinet portfolio distribution.

Over and above the special position claimed for the *formateur*, Ansolabehere et al. also diverge from typical work on GL by arguing that cabinet pay-offs should be associated with theoretical voting weights, specifically minimum integer weights (MIWs), rather than the 'raw' legislative seat shares of each party. Parties' MIWs are derived from their raw seat shares and the winning threshold for passing votes in the legislature. The list of raw seat shares is used to calculate the list of winning coalitions; the raw seat shares are replaced by the list of smallest integers, one per party, that generates the same set of winning coalitions. These integers are the parties' MIWs.[12] Ansolabehere et al. therefore make two simultaneous moves away from traditional research on portfolio distribution: they propose a *formateur* advantage and they propose using MIWs rather than raw seat shares.

Notwithstanding the empirical findings published by Ansolabehere et al., analyses by subsequent authors show little if any empirical support for either of the two theoretical propositions stated above. In relation to the first –the use of MIWs rather than raw seat shares – Warwick and Druckman's results, replicated by Laver et al., are 'clear and strong: cabinet portfolios, in both number and value, are allocated in very close proportion to the seat contributions of cabinet parties, and the bargaining strengths of these parties distort this allocation principle only very slightly (or very occasionally)'.[13] In relation to the second proposition, on the *ex post formateur* effect, Warwick and Druckman find 'the power that ought to come with *formateur* status appears to yield little in terms of portfolios'.[14] Laver et al. also show that estimating an *ex post formateur* effect is compounded by the methodological problem that *formateur* party status is coded endogenously in relevant datasets as the party of the eventual prime minister.[15] Thus Golder reports that 'it is not uncommon for the *formateur* to fail to form a coalition on the first or even the second attempt. As an example, it took seven different coalition proposals more than 106 days for a government to form after the 1979 Belgian legislative elections'.[16] None of these failed attempts by *formateurs* is analyzed by Snyder, et al. or Warwick and Druckman when they estimate the *formateur* advantage, and only one of 250 *formateur* parties in the dataset they use was *not* the party of the eventual prime minister.[17] The prime ministerial portfolio thus appears on both sides of the relevant regressions, as both one of the key independent variables

[11] Snyder, Ting, and Ansolabehere 2005, 992.

[12] For example, in a five-party 100-seat legislature with a simple majority-winning threshold and a raw seat vector of (43, 35, 8, 8, 6), the MIW vector is (3, 1, 1, 1, 0). The smallest party has zero MIW because it is never pivotal; adding its seats never turns a losing coalition into a winning one. The largest party can form a winning coalition with any of the three middle parties; all three of the middle parties must combine to exclude the largest.

[13] Laver, Marchi, and Mutlu 2011; Warwick and Druckman 2006, 659.

[14] Warwick and Druckman 2006, 659.

[15] Laver, Marchi, and Mutlu 2011.

[16] Golder 2010, 8.

[17] Snyder, Ting, and Ansolabehere 2005; Warwick and Druckman 2006.

(*formateur* status) and as part of the dependent variable (portfolio share). Correcting for this, Laver, et al. find that the empirical *formateur* effect disappears.[18]

The paradox is therefore both simple and striking. As Carroll and Cox say, 'all modern bargaining models predict that Gamson's Law should *not* hold'.[19] But as Bassi says, 'the most important empirical law in government-formation studies is that coalition partners share cabinet portfolios in proportion to their relative seat shares, which contradicts the predictions of the entire theoretical literature'.[20]

Resolving the Paradox

Two theoretical questions underlie this paradox, as we have seen. The first is the claim that there is some *formateur* in a privileged bargaining position. The second is the claim that parties' portfolio pay-offs respond to their theoretical voting weights (MIWs) rather than their raw seat shares. We argue above against predicting portfolio pay-offs *ex post* (after Nature has told us the identity of the randomly chosen *formateur* but before the government has formed), given the problem of identifying the list of exogenously selected *formateurs* in a way that is not endogenous to the bargaining outcome, which is compounded by the lack of any measurable effect once this problem is addressed. We therefore address the paradox by predicting portfolio allocation *ex ante*, absent any knowledge of the *formateur*, and developing a joint *ex ante* model of government formation and portfolio distribution.

This leads us to the second question, of whether raw seat shares or theoretical voting weights better predict parties' observed portfolio shares. Absent knowledge of the *formateur*, the canonical bargaining models referred to above predict that:

- pay-offs will be proportional to theoretical voting weights;
- only minimal-winning coalitions will form (minority cabinets face a majority opposition with both the incentive and the ability to capture all portfolios, while surplus coalitions contain members who are not needed yet are assigned some portfolios);
- parties with zero MIW will therefore not be part of any government coalition.

Unlike the canonical bargaining models, the GL prediction that portfolio distribution will reflect raw seat shares lacks deep theoretical underpinnings, though several scholars have recently addressed this. Anna Bassi assumes that *formateurs* are not picked exogenously, but emerge *endogenously* from negotiations between party leaders. Her model predicts no *formateur* advantage (in equilibrium, party leaders bargain this away) and Gamsonian pay-offs.[21] The latter arise because Bassi assumes that any proposed pay-off distribution must satisfy the full set of party legislators, not just a single party leader.[22] Carroll and Cox model bargaining over government formation that begins *before* an election. Their argument is that committing to Gamsonian pay-offs in pre-election deals provides incentives for all parties to the deal to expend maximum electoral effort on increasing their legislative seat total, thus increasing the probability that both they and their coalition partners will take office. Falco-Gimenó and Indridason argue that Gamsonian portfolio distributions act as natural 'focal points' in a difficult and complex bargaining environment involving complicated policy negotiations and many other matters, thereby taking one piece of complexity off the bargaining table.[23]

[18] Laver, Marchi, and Mutlu 2011.
[19] Carroll and Cox 2007a, 301.
[20] Bassi 2013, 778.
[21] Bassi 2013.
[22] Bassi 2013, 784.
[23] Falcó-Gimeno and Indridason 2013.

Given the theoretical arguments on both sides of the debate, it is helpful to resolve empirically the question of whether parties' portfolio pay-offs tend to respond to raw seat shares or their theoretical voting weights. Building on the argument of Falco-Gimenó and Indridason, it is also helpful to investigate whether Gamsonian proportional pay-offs become more likely as the bargaining environment becomes more complex. We address both of these questions below.

A Joint a Priori *Model of Government Formation and Portfolio Distribution*

Until now, all empirical work has predicted portfolio pay-offs *given a particular coalition of parties that has already formed*.[24] In other words, existing empirical work has modeled portfolio distribution *ex post*, conditional on government formation and regardless of whether any model predicted the government that formed. From a theoretical perspective, this is despite the fact that bargaining models predict portfolio distribution as an integral part of a model of government formation, so that (in)ability to predict government membership is highly relevant empirically. In practical terms, *ex post* modeling of pay-off distributions results in the deletion from the analysis of every party that does not receive a cabinet portfolio. This happens even when the theoretical model under investigation predicts that, on average, many of these parties should receive a positive bargaining pay-off and should (with some positive probability) be members of the government. Selecting on the dependent variable in this way ignores the possibility that there may be complexities in the bargaining process that systematically select different types of parties, which is contrary to the claim that parties should receive portfolio pay-off solely in proportion to their bargaining strength.

We therefore eschew *ex post* prediction of portfolio distributions and specify an *ex ante* statistical approach that includes all parties involved in government formation, whether or not they join the cabinet and receive a pay-off, and evaluates joint predictions of government membership and pay-off allocation between government members. This is analogous to an approach developed recently by Chiba et al., who develop a statistical model to evaluate joint predictions of government formation and duration.[25]

STATISTICAL MODEL

Our unit of analysis is a political party in a government formation situation. Our dependent variable is the party's observed share of cabinet portfolios, which ranges in theory from 0 to 1, and has a point mass at zero in the empirical data. A party with zero cabinet portfolios is by all conventional definitions considered not to be a member of the government; a party with non-zero portfolios is considered to be a government member; a party with all the portfolios constitutes a single-party cabinet. Thus our dependent variable is far from normally distributed. We present the results of ordinary least squares (OLS) regressions in the supplementary materials. But to facilitate comparison with previously published work, our core method involves maximum-likelihood models based on a mixed continuous–discrete distribution. The dependent variable is a proportion and therefore has a limited range. This suggests the choice of a beta distribution, which is flexible enough to suit the problem.[26] As a continuous function, however, it is not appropriate for modeling large numbers of values that are at a single point, and the dependent variable has a very large number of zeros, which arise whenever a given

[24] Ansolabehere et al. 2005; Carroll and Cox 2007b; Falcó-Gimeno and Indridason 2013; Laver, Marchi, and Mutlu 2011; Snyder, Ting, and Ansolabehere 2005; Warwick and Druckman 2006.

[25] Chiba, Martin, and Stevenson 2014.

[26] See Brehm and Gates 1993.

party is not included in the government. This implies using a mixed distribution. We adopt this approach, following work by Ospina and Ferrari,[27] and model these data as a mixture between a beta distribution and a degenerate distribution in 0. We thus use a zero-inflated beta model to directly model both the likelihood that a party will enter a coalition and its pay-off in cabinet seats if it does enter. Hence, the likelihood function we maximize is:

$$L(y; v, \mu, \sigma) = \prod_{i=1}^{n} v_i^{1-I(y_i)}(i-v_i)^{I(y_i)} \prod_{i;y_i\in(0,1)} \frac{\Gamma(\sigma_i)}{\Gamma(\mu_i\sigma_i)\Gamma((1-\mu_i)\sigma_i)} y_i^{\mu_i\sigma_i-1}(1-y_i)^{(1-\mu_i)\sigma_i-1},$$

where:

$$v_i = \log\left(\frac{\beta_{v0}+\beta_{v1}\cdot miw\ share_i+\beta_{v2}\cdot seat\ share_i}{1-(\beta_{v0}+\beta_{v1}\cdot miw\ share_i+\beta_{v2}\cdot seat\ share_i)}\right),$$

$$\mu_i = \log\left(\frac{\beta_{\mu0}+\beta_{\mu1}\cdot miw\ share_i+\beta_{\mu2}\cdot seat\ share_i}{1-(\beta_{\mu0}+\beta_{\mu1}\cdot miw\ share_i+\beta_{\mu2}\cdot seat\ share_i)}\right),$$

$$\sigma_i = \log(\beta_{\sigma0}+\beta_{\sigma1}\cdot miw\ share_i+\beta_{\sigma2}\cdot seat\ share_i).$$

Link functions were chosen to constrain parameters to the unit interval (in the case of ν_i and μ_i) and to be strictly positive in the case of the precision parameter σ_i. The parameter ν_i has a direct and relevant substantive interpretation: it gives the likelihood of a zero observation, hence the *inverse* likelihood of cabinet membership. The parameter μ_i is the mean for the beta distribution and shows the strength of the relationship between the independent variables and the party portfolio shares; σ_i is the precision of the beta distribution. We thereby achieve our core objective of deriving distinct joint estimates of the probability of cabinet membership, v_i, and the distribution of cabinet portfolios, μ_i.

Weighting Cases

Since our unit of analysis is a political party in a government formation situation, we might ask whether all observations in the data should, as in all previous empirical analyses, have equal weight. Different countries have different numbers of parties, and some have many more governments than others. Italy, for example, had more governments than most other European countries in the post-war era and also tends to have more parties than most other countries. The dataset we describe below covers sixteen countries, but Italy accounts for about 22 per cent of the observations. Italy, Belgium, Denmark and Finland together account for about 55 per cent of all observations. Empirical analyses that ignore this may be biased, and we do not want our 'cross-national' analysis of portfolio distribution to be mostly about these four countries. In order to control for the different weights of parties in the estimation due to the differences in party system sizes and the frequency of elections, we calculated a probability for each party to be selected into a subsample of our data. This selection probability is inversely related to the number of parties in parliament. It is calculated as:

$$\pi_i = \frac{1}{n_k}\cdot\frac{1}{K},$$

where n_k is the number of parties in a given cabinet k and K is the total number of cabinets in the truncated dataset. Based on the selection probabilities, 80 per cent of observations

[27] Ospina and Ferrari 2012.

(1,134 parties) in the truncated dataset are randomly selected in a subsample.[28] By providing results based on multiple, randomly drawn samples, we mitigate the possibility that one nation exercises undue influence on the results.[29] In many respects, our sampling strategy functions like the bootstrap. By creating empirical distributions from the data and generating multiple samples (with replacement), we hedge against any single nation (or subset of cabinets) having undue influence on our results.[30]

Estimation Approach

As indicated in the previous section, our estimation method involves creating multiple training sets based on the principle of the empirical bootstrap. We use these training sets to test three different models. The first includes all cabinets and is the main test of whether theoretical voting weights or raw seat shares best predicts outcomes, including both government membership and pay-off distribution. Addressing Falco-Gimenó and Indridason's argument that Gamsonian pay-offs are a response by party leaders to bargaining complexity, we test two further models using these training data, separating bargaining settings according to their complexity. We categorized 'simple' and 'complex' settings according to the difficulty of calculating the full set of $2^n - 1$ coalitions in an n-party system (and therefore of exploring all coalition possibilities) using a natural inflection point in the time taken by our computational algorithm to calculate MIWs. This led us to specify simple settings as those with fewer than eight pivotal parties and complex settings as those with eight or more pivotal parties.[31] For the full population of cases, as well as those reflecting simple and complex settings, we draw 1,000 subsamples from the dataset, split each subsample into training and test sets, and fit the model. Instead of evaluating our model performance using the training samples, we report results based only on the 20 per cent out-of-sample observations in the test set.[32] This is a further hedge against the possibility that results could be overfitted to non-systematic elements of the original sample.

DATA

Most of our data derive from the Parliament and Government Composition Database (Parlgov), constructed by Döring, Manow and collaborators.[33] This includes election results and government formation data for all EU members as well as many Organisation for Economic Co-operation and Development countries from 1945 onwards. Our dataset includes data from 1945 until the most recent available Parlgov data on cabinet seats for sixteen parliamentary

[28] We use a threshold of 80 per cent for the training set (and, accordingly, 20 per cent for the test set) to ensure that the test samples are large enough to make reasonable inferences. Splitting the sample in this way is accepted practice in the machine learning literature, though there are more complex ways to accomplish the same goals (see, for example, Bishop and Nasrabadi 2006).

[29] To provide comparability with previously published results that do not weight cases in this way, we provide unweighted (flat pi), bootstrapped OLS versions of all key estimates in the online supplementary materials (online Appendix 3, Tables A3.1 and A3.2).

[30] We also tried a more aggressive strategy that rebalanced the relative proportions of the various nations in the training sets to make them more even. This modified approach did not produce substantively different results but is available on request.

[31] A pivotal party P is one with non-zero MIW. For robustness, we examined models setting the 'complexity' threshold at seven or nine pivotal parties, and the results were similar.

[32] Given the vanishingly small probability that all members of any cabinet would be selected in all samples, this sampling strategy also addresses the methodological issue that the set of party portfolio proportions within any given cabinet generates compositional data.

[33] http://www.parlgov.org.

democracies in Western Europe.[34] While we rely on Parlgov for data on elections, seat share and party of the prime minister, this source does not include data on the distribution of cabinet seats between parties, which we collected from a reliable online database.[35]

Theoretical Voting Weights

A key independent variable in our analysis concerns parties' MIWs, which Ansolabehere et al. use to implement their claim that theoretical voting weights (rather than raw seat shares) inform government formation and subsequent portfolio allocation. There is a focused technical literature on MIWs, much of it outside mainstream political science.[36] Montero has shown, for example, that *ex post* predicted pay-offs under the canonical alternating offers bargaining protocol are proportional to agents' MIWs.[37] The vector of MIWs can be surprisingly difficult to calculate, especially in systems with more than a few parties. Computationally, as Strauss shows, this is because the problem of coalition enumeration is NP-hard, which of course generates problems for real politicians as well as for political scientists.[38] While Snyder et al. address this by programming a calculator for computing MIWs,[39] their 'perfect substitutes have the same price' argument highlights a theoretical distinction drawn by Freixas and Kurz between simple MIWs and MIWs preserving types. At issue is whether mutually substitutable parties should be given equal voting weights; Snyder et al. argue they should. If so, we should use MIWs preserving types, which are constrained so that parties that are perfect substitutes have the same voting weights.[40] Following Freixas and Kurz, therefore, we replace MIWs with MIWs preserving types.[41] Solving for these is an application of linear programming.[42] We programmed and verified our own algorithm and, assuming a simple majority-winning quota, calculated MIWs preserving types from party seat totals reported in Parlgov.

Single Party Majorities and Minority Governments

When a single political party wins a legislative majority, models of government formation invariably predict that it will form a single-party government and award itself all cabinet portfolios. This is almost always true empirically, and is not an open theoretical question. While we might think of plausible models that treat single-party majority governments as coalitions of party factions, this is neither the focus of the literature with which we are dealing nor our focus in this article. We therefore exclude from our analysis, as uninformative for our purposes, settings in which a single party won a legislative majority.

[34] These countries are: Austria, Australia, Belgium, Denmark, Finland, Germany, Great Britain, Iceland, Ireland, Italy, Luxembourg, Netherlands, Norway, Portugal, Spain and Sweden.

[35] The data on the number of cabinet seats were collected from http://www.kolumbus.fi/taglarsson/dokumentit/governm2.htm. Summary statistics are available in online Appendix 2.

[36] Freixas and Kurz 2011; Freixas and Molinero 2009a, 2009b; Montero 2002, 2006.

[37] Montero 2006.

[38] Strauss 2003.

[39] Ansolabehere et al. 2005; Snyder, Ting, and Ansolabehere 2005.

[40] Freixas and Kurz (2011) note that this problem is especially acute with 'non-homogenous' voting games and can occur with as few as eight parties. Non-homogenous voting games are those for which all MWCs do not have the same aggregate MIW. For example in the legislature (8: 4, 3, 3, 2, 2), expressed in MIWs, the coalitions (4, 2, 2), (4, 3, 2) and (4, 3, 3) are all MWCs, but each has a different aggregate MIW. Theoretical complications arise because a party with MIW 2 in this example can substitute in an MWC for a party with weight 3. Laver, Marchi, and Mutlu (2011) find that non-homogenous voting games arise in about one-third of real legislatures in the Snyder, Ting, and Ansolabehere replication dataset, so this is a non-trivial issue.

[41] For stylistic reasons, we refer to these weights as MIWs in what follows. The Freixas and Kurz definition should, in our opinion, be universally adopted.

[42] See online Appendix 1.

Minority cabinets pose a more difficult challenge for a joint model of government formation and portfolio distribution. Predicting the *formation* of minority governments remains a wide-open theoretical question for government-formation models. But models that focus on *portfolio distribution* predict no minority governments, as we have noted, since a minority government faces a majority opposition that is both willing and able to consume all portfolios. Including minority governments in our analysis would thus include a set of cases in which we know that the models of portfolio distribution we are evaluating make failed predictions, and it is not at all clear what portfolio distributions these models predict in cabinets they deem out of equilibrium. Excluding minority cabinets, however, amounts to selecting on the dependent variable for a model that jointly predicts government formation and portfolio distribution, and including an indicator variable for minority governments on the right-hand side of the relevant regressions violates our objective of *ex ante* prediction. Since our goal of *ex ante* prediction is paramount and we do not wish to select on the independent variable, we opt for the lesser of two evils and include all cases of minority coalitions in our analysis.[43]

RESULTS

Estimation results are reported in Tables 1–3 below; these exclude observations in which a single party forms the government.[44] The first parameter of interest is ν, which shows the relationship between the independent variables and a party being out of office (having no portfolio). The second parameter is μ, which shows the relationship between the independent variables and the strictly positive proportion of portfolios allocated to a party. Taking all cases together (Table 1), both the MIW and raw legislative seat share of a party are correlated with its inclusion in government.

The ν coefficients demonstrate a negative relationship between exclusion from office for both MIWs and raw seat shares; higher MIWs are more strongly connected with cabinet membership than higher raw seat shares. However, and consistent with Warwick and Druckman's findings, the μ coefficients in Table 1 show that raw seat shares are far better predictors than MIWs of the share of cabinet portfolios a party receives, conditional on it being in the government at all.

Our analysis therefore yields an unambiguous substantive finding about legislative bargaining. Independently of raw legislative seat shares, parties are more likely to get into the cabinet if they have a higher theoretical voting weight, measured using MIWs. In this sense, theoretical bargaining models add value to predictions about cabinet *membership*. Once a party is in the government, however, its *pay-off*, measured as its share of cabinet portfolios, is predicted by its raw legislative seat share and not at all by its theoretical bargaining weight. In this precise sense, the empirical robustness of GL has again been vindicated. While raw seat shares robustly predict portfolio distribution, they do not do so because they are associated with theoretical voting weights.

Comparing Tables 2 and 3, we see that bargaining complexity has a strong effect. In the more complex settings described in Table 3, the ν coefficients show that MIWs no longer even predict

[43] Results for models that exclude minority governments are available on request. In broad terms, and unsurprisingly, excluding minority governments improves overall model fit, though only by a small margin. The effects for the variables of interest – MIWs and raw seat shares – remain the same.

[44] As noted, we exclude cases in which one party receives a simple majority of the seats in parliament. A much smaller subset of exclusions ($n = 75$) concerns observations in which a single minority party receives all portfolios, since these do not conform to the assumptions of a zero-inflated beta model. We do, however, include such cases in the (substantively similar) results estimated using OLS regressions and reported in online Appendix 3 (Tables A3.1 and 3.2).

TABLE 1 *Distribution of 1,000 Parameter Estimates and Root Mean Squared Error: Out-of-Sample Predictions for all Cabinets*

		Mean	Std. Dev.
ν (Cabinet Membership)	Intercept	1.95	0.04
	MIW	−5.75	0.45
	Raw weights	−2.94	0.35
μ (Portfolio Share)	Intercept	−2.14	0.02
	MIW	1.32	0.14
	Raw weights	5.43	0.13
σ	Intercept	3.33	0.06
	MIW	0.15	0.50
	Raw weights	−2.51	0.43
	RMSE	0.16	0.003

TABLE 2 *Distribution of 1,000 Parameter Estimates and Root Mean Squared Error Out-of-Sample Predictions for Cabinets with at Most Eight Parties with Non-zero MIWs*

		Mean	Std. Dev.
ν	Intercept	2.31	0.08
	MIW	−8.14	0.58
	Raw weights	−1.42	0.45
μ	Intercept	−1.83	0.06
	MIW	0.73	0.32
	Raw weights	4.91	0.19
σ	Intercept	2.70	0.18
	MIW	4.01	1.14
	Raw weights	−4.41	0.71
	RMSE	0.18	0.004

cabinet membership, which is much better predicted by raw legislative seat shares. Table 2 shows the reverse for less complex settings, bearing in mind that these are the typical settings for post-war Western Europe. *Membership* of the government is much better predicted by MIWs, controlling for raw legislative seat shares. In both types of settings, however, the μ coefficients show that it is raw legislative seat shares, and not MIWs, that predict portfolio pay-offs.

To facilitate comparison with previously published results, we present estimates derived from OLS regressions in online Appendix 3, with the caveat that these models are mis-specified for reasons we discuss above. The OLS estimates comport with those derived from zero-inflated beta models presented here. Theoretical voting weights predict government membership, especially in the lower-complexity settings with eight or fewer pivotal parties. But raw legislative seats, rather than MIWs, predict distributions of cabinet portfolios within a given coalition.

TABLE 3 *Distribution of 1,000 Parameter Estimates and Root Mean Squared Error Out-of-Sample Predictions for Cabinets with More Than Eight Parties with Non-zero MIWs*

		Mean	Std. Dev.
v	Intercept	1.89	0.05
	MIW	–4.19	0.80
	Raw weights	–5.13	0.63
μ	Intercept	–2.25	0.02
	MIW	–0.02	0.21
	Raw weights	7.17	0.20
σ	Intercept	3.46	0.07
	MIW	–1.80	0.80
	Raw weights	0.35	0.69
	RMSE	0.116	0.004

CONCLUSIONS

Our zero-inflated beta model jointly estimates membership of the government and, conditional on this, the distribution of cabinet portfolios between government members. The results are unambiguous. First, controlling for raw legislative seat shares, the theoretical voting weights used by non-cooperative models of legislative bargaining predict *membership* of the government emerging from bargaining between parties, especially in settings with fewer than eight pivotal parties (which is typical in post-war Western Europe). Indeed, Table 1 shows that theoretical voting weights predict government membership much better than raw seat shares in this common setting. In a nutshell, controlling for raw seat shares, *parties with higher theoretical voting weights are better able to force their way into government coalitions.* This is a theoretically significant empirical finding. Secondly, whatever the setting – and conditional on the government that forms – *raw legislative seat shares, not theoretical voting weights, predict the distribution of cabinet portfolios between government parties.* The latter finding is, in effect, a further replication of the robust GL results using a more sophisticated and appropriate statistical apparatus. Thirdly, Tables 2 and 3 show that, while MIWs predict cabinet membership in simple bargaining settings, they are much less effective than raw seat shares at predicting cabinet membership in more complex settings.

We draw the following conclusions from these findings. First, our findings imply that the scientific community should accept and exploit the robust empirical regularity that raw seat shares (and not theoretical voting weights) predict the distribution of cabinet portfolios in European governments. The theoretical task is not to make this empirical result go away, but to find a rigorous theoretical model of government formation that explains it. Secondly, and newly emerging from our joint modeling of government membership and portfolio distribution, the good news for legislative bargaining models is that theoretical voting weights – more precisely, MIWs preserving types – help to explain government membership.

We see no necessary contradiction between these findings. One way to reconcile them is to infer that the distribution of cabinet portfolios is neither the only, nor even the most important, matter that concerns party leaders who are bargaining over the formation of coalition cabinets.

Golder's analysis of bargaining delays over government formation in Western Europe shows that, in post-electoral settings, these delays range from 86 days on average in the Netherlands though sixty-one days in Belgium to less than three days in Norway.[45] One recent Belgian government took more than a year to form. Informal press reports of what is happening during these often lengthy bargaining periods suggest that the party leaders are *not* spending their time bargaining over the distribution of cabinet portfolios, but rather are trying to negotiate a joint policy program for government that reconciles conflicting party manifestos. This is detailed and time-consuming work, and it appears from these reports that party leaders typically turn to the distribution of cabinet portfolios only after a joint policy program has been agreed. For us to go further would require setting out a model of bargaining over government formation that takes into account both portfolio distribution and the need for a joint policy program. This is clearly not our task in this article, though we are exploring it in ongoing work. Our point here is that our twin empirical findings can be reconciled if portfolio distribution is not the only pay-off that concerns party leaders when they bargain over government formation. If it were, then we would expect parties' theoretical voting weights to predict portfolio distribution, but they do not. Rather, while these voting weights have a significant bearing on which parties get into government, the distribution of portfolios between the parties in government very systematically conforms to a norm of proportionality with raw legislative seat shares.

REFERENCES

Ansolabehere, Stephen, James M. Snyder, Aaron B. Strauss, and Michael M. Ting. 2005. Voting Weights and Formateur Advantages in the Formation of Coalition Governents. *American Journal of Political Science* 49 (3):550–63.

Baron, David P., and John A. Ferejohn. 1989. Bargaining in Legislatures. *American Political Science Review* 83 (4):1181–1206.

Bassi, Anna. 2013. A Model of Endogenous Government Formation. *American Journal of Political Science* 57 (4):777–93.

Bishop, Christopher M., and Nasser M. Nasrabadi. 2006. *Pattern Recognition and Machine Learning, Volume 1*. New York: Springer.

Brehm, John, and Scott Gates. 1993. Donut Shops and Speed Traps: Evaluating Models of Supervision on Police Behavior. *American Journal of Political Science* 37 (2):555–81.

Browne, Eric, and John Frendreis. 1980. Allocating Coalition Payoffs by Conventional Norm: An Assessment of the Evidence from Cabinet Coalition Situations. *American Journal of Political Science* 24:753–68.

Browne, Eric, and Mark Franklin. 1973. Aspects of Coalition Payoffs in European Parliamentary Democracies. *American Political Science Review* 67:453–69.

Carroll, Royce, and Gary W. Cox. 2007a. The Logic of Gamson's Law: Pre-Election Coalitions and Portfolio Allocations. *American Journal of Political Science* 51 (2):300–13.

——. 2007b. The Logic of Gamson's Law: Pre-Election Coalitions and Portfolio Allocations. *American Journal of Political Science* 51 (2):300–13.

Chiba, Daina, Lanny W. Martin, and Randolph T. Stevenson. 2014. *A Copula Approach to the Problem of Selection Bias in Models of Government Survival*. Colchester: University of Essex.

Falcó-Gimeno, Albert, and Indridi H. Indridason. 2013. Uncertainty, Complexity, and Gamson's Law: Comparing Coalition Formation in Western Europe. *West European Politics* 36 (1):221–47.

Fréchette, Guillaume, John H. Kagel, and Massimo Morelli. 2005a. Behavioral Identification in Coalitional Bargaining: An Experimental Analysis of Demand Bargaining and Alternating Offers. *Econometrica* 73 (6):1893–937.

[45] Golder 2010, 8.

——. 2005b. Gamson's Law Versus Non-Cooperative Bargaining Theory. *Games and Economic Behavior* 51:365–90.

Freixas, J., and S. Kurz. 2011. On Minimal Integer Representations of Weighted Games. In *Proceedings of CoRR*.

Freixas, Josep, and Xavier Molinero. 2009a. On the Existence of a Minimum Integer Representation for Weighted Voting Systems. *Annals of Operations Research* 166 (1):243–60.

——. 2009b. Weighted Games Without a Unique Minimal Representation in Integers. *Optimization Methods and Software* 25 (2):203–15.

Gamson, William A. 1961. A Theory of Coalition Formation. *American Sociological Review* 26:373–82.

Golder, Sona N. 2010. Bargaining Delays in the Government Formation Process. *Comparative Political Studies* 43 (1):3–32.

Laver, Michael, Scott de Marchi, and Hande Mutlu. 2011. Negotiation in Legislatures Over Government Formation. *Public Choice* 147 (3):285–304.

Martin, Lanny W., and Randolph Stevenson. 2010. The Conditional Impact of Incumbency on Government Formation. *American Political Science Review* 104:503–18.

Martin, Lanny W., and Randolph T. Stevenson. 2001. Government Formation in Parliamentary Democracies. *American Journal of Political Science* 45 (1):33–50.

Montero, Maria. 2002. Non-Cooperative Bargaining in Apex Games and the Kernel. *Games and Economic Behavior* 41 (2):309–21.

——. 2006. Noncooperative Foundations of the Nucleolus in Majority Games. *Games and Economic Behavior* 54:380–97.

Ospina, Raydonal, and Silvia L. P. Ferrari. 2012. A General Class of Zero-or-One Inflated Beta Regression Models. *Computational Statistics & Data Analysis* 56 (6):1609–23.

Rubinstein, Ariel. 1982. Perfect Equilibrium in a Bargaining Model. *Econometrica* 50 (1):97–109.

Schofield, Norman, and Michael Laver. 1985. Bargaining Theory and Portfolio Payoffs in European Coalition Governments. *British Journal of Political Science* 15:51–72.

Snyder, James M., Michael Ting, and Stephen Ansolabehere. 2005. Legislative Bargaining Under Weighted Voting. *American Economic Review* 95 (4):981–1004.

Strauss, Aaron. 2003. *Applying Integer Programming Techniques to Find Minimum Integer Weights of Voting Games*. Cambridge, MA: Massachusetts Institute of Technology.

Taylor, Michael, and Michael Laver. 1973. Government Coalitions in Western Europe. *European Journal of Political Research* 1:205–48.

Warwick, Paul V., and James N. Druckman. 2001. Portfolio Salience and the Proportionality of Payoffs in Coalition Governments. *British Journal of Political Science* 31:627–49.

——. 2006. The Portfolio Allocation Paradox: An Investigation into the Nature of a very Strong but Puzzling Relationship. *European Journal of Political Research* 45:635–65.

doi:10.1017/S0003055413000075

Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa

JAN H. PIERSKALLA *German Institute of Global and Area Studies (GIGA)*
FLORIAN M. HOLLENBACH *Duke University*

The spread of cell phone technology across Africa has transforming effects on the economic and political sphere of the continent. In this paper, we investigate the impact of cell phone technology on violent collective action. We contend that the availability of cell phones as a communication technology allows political groups to overcome collective action problems more easily and improve in-group cooperation, and coordination. Utilizing novel, spatially disaggregated data on cell phone coverage and the location of organized violent events in Africa, we are able to show that the availability of cell phone coverage significantly and substantially increases the probability of violent conflict. Our findings hold across numerous different model specifications and robustness checks, including cross-sectional models, instrumental variable techniques, and panel data methods.

‘

The mobile industry changed Africa. I must admit we were not smart enough to foresee that. What we saw was a real need for telecommunication in Africa, and that need had not been fulfilled. For me that was a business project.” Mo Ibrahim, as quoted by Livingston (2011, 10).

This quote from Mo Ibrahim, a Sudanese-born cell phone magnate, exemplifies the increasing influence new media technologies have in Africa. During the recent events of the Arab Spring, cell phones and other new media technologies have worked as catalysts for political collective action (Aday *et al.* 2012; Breuer, Landman, and Farquhar 2012). While many commentators describe the effect of modern communication technologies on political action, social scientific research is only slowly catching up (but see, for example, Aday *et al.* 2012; Breuer, Landman, and Farquhar 2012; Shirky 2008). In this article we ask whether modern communication technology has affected political collective action in Africa. Specifically we ask if the rapid spread of cell phone technology has increased *organized* and *violent* forms of collective action.

Jan H. Pierskalla is a Postdoctoral Fellow, German Institute of Global and Area Studies (GIGA), FSP 2, Neuer Jungfernstieg 21, 20354 Hamburg, Germany (jan.pierskalla@giga-hamburg.de).

Florian M. Hollenbach is a Ph.D. candidate, Department of Political Science, Duke University, Perkins Library 326, Box 90204, Durham NC 27708, USA (florian.hollenbach@duke.edu).

Authors’ names are listed in reverse alphabetical order; equal authorship is implied. This project was in part funded by the Program for the Study of Democracy, Institutions and Political Economy (DIPE) at Duke University. We thank Andreas Forø Tollefsen and colleagues for sharing the PRIO-GRID dataset with us. The availability of these data made life much easier on us. We are very grateful for comments and criticisms from four anonymous reviewers, the editors of the APSR, Cassy Dorff, Vincent Gawronski, Evan Lieberman, Nils W. Metternich, Brittany N. Perry, Audrey Sacks, Michael D. Ward, Erik Wibbels, and William Wittels. Their comments and suggestions helped to substantially improve our research and this paper. A previous version of this paper was presented at MPSA 2012 in Chicago. All remaining errors are our own. Data and replication files are available on the authors’ dataverse subject to dissemination restrictions by the GSMA.

We focus on the connection between communication technology and violent, organized forms of collective action for several reasons. While scholars in economics and other fields have been concerned with the beneficial effects of cell phones for various development outcomes (Abraham 2007; Aker 2010; Aker, Ksoll, and Lybbert 2012; Aker and Mbiti 2010), the implications of increased cell phone communication are much less clear when it comes to politics.[1] The existing discussion in political science on new media and collective action is rather qualitative and lacks a specific focus on cell phones and their relationship to political violence.[2] Much of this literature stresses the possible positive effects of new media and technology for democracy, transparency, and accountability. While the quick and cheap spread of communication technology can improve political accountability through various mechanisms, private communication technology (and cell phones specifically) may also facilitate organized violence.

The vast literature on civil conflict onset and duration has explored structural determinants such as economic development, growth shocks, natural resources, elections, ethnic diversity, and political exclusion (see, for example, Cederman, Weidmann, and Gleditsch 2011; Collier and Hoeffler 2007; Collier *et al.* 2003; Fearon, Kasar, and Laitin 2007; Fearon and Laitin 2003; Metternich 2011; Ross 2006; Sambanis 2002; Weidmann 2009; Wucherpfennig *et al.* 2012). A smaller, but growing, body of research has investigated important factors at the individual and group level (Blattman 2009; Weinstein 2007; Wood 2003). On the other hand, little explicit attention has been given to the role of technology in facilitating violence. While some recent studies analyze the potential effects of mass media, like television and radio broadcasting (Warren 2013; Yanagizawa-Drott

[1] However, recent research has tested the possible impact of mobile phones on voter information and participation, as well as its possible impact on fighting electoral fraud and corruption (Aker, Collier, and Vincente 2011; Bailard 2009).

[2] See, for example, Earl and Kimport (2011) and Diamond and Plattner (2012).

2012), hardly any empirical research deals with individual-to-individual communication.[3]

We argue that private, mobile long-distance communication addresses crucial free-rider and coordination problems endemic to insurgent activity. Similar to other organizational technologies (Weinstein 2007), cell phones facilitate in-group organization and the implementation of insurgent activity against the greater power of the state. Given the motivation and opportunity for political violence through structural context conditions, cell phone coverage, *ceteris paribus*, should then increase the likelihood of violent collective action.

To test this proposition we use highly spatially disaggregated data on cell phone coverage and violent conflict in Africa. Today, Africa is the largest growing cell phone market in the world, with yearly growth rates of around 20% and an estimated 732 million subscribers in 2012 (The Economist 2012). What makes Africa special in this context is that cell phones not only provide a new way for communication, but in many areas are the only way for interpersonal, direct communication over distance. Many areas that are now covered by cell phone networks were never connected to land lines. At the same time, Africa is host to a large number of active or simmering civil conflicts (The World Bank 2011), often in areas with newly expanded access to cell phone technology. This directly poses the question: how does the introduction of easy interpersonal communication affect the incidence of organized violence on the continent?

We match proprietary data from the GSM Association (GSMA), an interest group of cell phone providers, on the spatial extent of GSM2 network coverage on the African continent and conflict events from the UCDP Georeferenced Event Dataset (Melander and Sundberg 2011; Sundberg, Lindgren, and Padskocimaite 2011) to a lattice of 55 km × 55 km grid cells in Africa (PRIO-GRID), created by the Peace Research Institute Oslo (PRIO) (Tollefsen, Strand, and Buhaug 2012).

We then implement three complementary estimation strategies to assess the potential effect of cell phone coverage on violent collective action. First, we exploit spatial variation in conflict and cell phone coverage by estimating a series of statistical models and adjusting for important covariates using cross-sectional data. Second, to safeguard against reverse causality and to improve the identification of a causal effect, we rely on an instrumental variable strategy. Prior research on the spread of cell phone technology in Africa has established the importance of regulatory quality and competitive private markets (Buys, Dasgupta, and Thomas 2009), which we use as an instrument for the extent of cell phone coverage. Third, we expand our analysis to a three-year panel of grid cells to exploit variation in cell phone coverage over time, controlling for any grid-level time-invariant factors.

We are able to document a clear positive and statistically significant effect of cell phone coverage on violent collective action across all three approaches. In other words, modern means of private long-distance communication not only have economic benefits, but also facilitate overcoming collective action and coordination problems in the political realm. Under specific structural context conditions this translates to more organized violence.

This finding carries meaningful implications for research on civil conflict and collective action more generally. Our research indicates the importance of technological shifts for organized violence and calls for further research on the role of modern communication technology for both enabling and curbing violence. Echoing the findings of research on civil society (Berman 1997), we find that improvements in the ability to organize collective action do not automatically produce purely beneficial effects for overall society. Rather they empower political agents and groups more generally, which can raise the human costs of political struggles.

CELL PHONES AND COLLECTIVE ACTION

Given the breathtaking spread of cell phone technology worldwide and the particularly fast expansion on the African continent, citizens across many regimes have vastly improved means for private, direct, and immediate long-distance communication. The availability of cell phone technology and networks to citizens in some of the poorest regions in the world has been lauded as an important transformative force for economic development. In particular, the decrease in communication costs associated with the rising availability of cell phones has been linked to a boost in labor and consumer market efficiency (Abraham 2007; Aker 2010; Aker and Mbiti 2010). This research emphasizes the diminishing effect of cell phone technology on information asymmetries between market participants. For example, in the case of Indian fishers, cell phone technology allowed for the monitoring of prices in nearby markets without the need to personally attend the market, while also giving fishers access to sell goods to markets at further distances (Abraham 2007). Similar developments have been noted for agricultural markets in Africa (Aker 2010; Aker and Mbiti 2010). In addition, African entrepreneurs are developing ways in which cell phones can be used to increase market efficiencies and deliver services to customers.[4] The increasing availability of cell phone coverage has gone hand in hand with the use of mobile money and mobile banking (Donner and Tellez 2008). In fact, the development to make payments and transfers via mobile

[3] To our knowledge, the only research that explicitly engages this question is a working paper by Shapiro and Weidmann (2012) on insurgent activity and cell phone towers in Iraq. The authors document a decrease in insurgent violence in areas with improved access to cell phone communication, which is attributed to the reduced cost of communicating information to counterinsurgency agents. We discuss their work in more detail further below.

[4] A particular success story is a startup company named Esoko Ltd. from Ghana, which is active in 15 African countries and provides a mobile internet platform to share, collect, and analyze data regarding prices of agricultural goods (Mutua 2011).

money instead of cash or credit cards has proliferated widely in Africa and lowers transaction costs for many market participants and citizens.[5]

Interestingly, in Africa, this digital revolution is largely driven by private entrepreneurs, which have built up an extensive wireless infrastructure in a matter of years, often independent of governments or government-funded infrastructure.[6] Private cell phone providers have increased coverage at a vastly faster rate than landline providers. Today many areas that had never been connected to landline communication networks are covered by cell phone networks (Africa Partnership Program 2008; The World Bank 2010). One of the advantages of cell phone networks is that the expansion is much costly in terms of infrastructure investments and thus a more decentralized expansion is possible.

Existing economics research provides evidence of lower transaction costs through the provision of cell phones. While much of this work has emphasized the positive effects of this new technology on economic outcomes, research on the direct effects of cell phones in the political sphere is not quite as common. However, Aker, Collier, and Vincente (2011) show that in the case of Mozambique, cell phones can be used for voter education and can increase political participation in elections, as well as demands for accountability. The major takeaway is that cell phone usage, the availability of hotlines to voters, and text messaging can have positive effects on the political information available to voters as well as their political participation. In a similar vein Bailard (2009), using country level analysis as well as provincial data for Namibia, finds that the use of cell phones by citizens can decrease corruption. She argues that cell phones change the information environment, as they decentralize and increase the spread of information. In addition, the proliferation of cell phones increases the probability of detection of corrupt officials and thus alters "the cost-benefit calculus of corrupt behavior by strengthening oversight and punishment mechanisms" (Bailard 2009, 337). Evidence further suggests that, through text messaging services, cell phones have been used to inform citizens of government wrongdoings, monitor elections, or report violence in many African states (Diamond 2012, 11).

More generally, observers of current events have linked cell phone technology to collective action, in particular peaceful protest, producing a new "protest culture" (Lapper 2010). In the context of authoritarian regimes, examples of cell phones aiding the organization of protests around the world are abundant, ranging from China in 1999, where Falun Gong was able to stage a large protest in a secure government complex, to Manila, Philippines in 2001 (Philippine Daily Inquirer 2001), or Kiev, Ukraine during the Orange Revolution (Diamond 2012, 12).[7]

Yet, cell phone technology does not only affect collective action in authoritarian governments. Protesters in Madrid, Spain in 2004 were able to organize quickly using text messaging (Shirky 2008, 180). The increased organization capabilities of protesters have been noted by the police in the riots in London in the summer of 2011, as well as protests over G20 summits (Bradshaw 2009; Sherwood 2011).

The link between political behavior and cell phone usage is also borne out in survey data. The 2008 wave of the Afrobarometer public opinion survey includes a question on the usage of mobile phone technology and protest behavior (Mattes *et al.* 2010). A simple regression of the protest item on cell phone usage, controlling for a number of socioeconomic factors, reveals a positive and highly statistically significant effect, i.e., cell phone users are more likely to participate in protests.[8]

While these observations and emerging scholarship highlight the positive effects of cell phone technology for peaceful forms of collective action, we argue that cell phones have another important effect: improved communication through cell phones can facilitate organization and coordination of groups for the purpose of *violent* collective action.

In a recent working paper, Shapiro and Weidmann (2012) pose a similar question about the spread of cell phone coverage and political violence in Iraq. The authors start from a theoretically ambiguous point. On the one hand, they emphasize that the availability of cell phones could lead to increased violence as it strengthens the position of insurgents against the coalition forces. On the other hand, cell phones could allow for better insurgent surveillance by U.S. and Iraqi forces, as well as lower the cost of whistle blowing on terrorists for the local population. Using district level data and a difference-in-difference design, the authors find that the expansion of the cell phone network in Iraq is associated with decreases in successful violent attacks by insurgent forces. Shapiro and Weidmann (2012) contend that this is due to the extensive use of cell phone surveillance by U.S. and Iraqi anti-insurgent forces as well as successful whistle-blower programs. Similarly, in the African context, Livingston (2011) argues, that while cell phones might empower rebel groups and produce more violence, there also exists the potential for a reduction in violence through improved monitoring for international peacekeeping or governmental forces, although such efforts have been rare so far.

While improved monitoring and well-organized counterinsurgency activities can leverage cell phone coverage to increase the capacity of the state to uphold the monopoly of violence, it is unclear how easily this

[5] The spread of this technology is exemplified by recent investments by Visa (Alliy 2011; Quandzie 2011). Further positive examples can be found in an OECD report on information technology and infrastructure in Africa (Africa Partnership Program 2008).

[6] However, one should note that governments are always involved to some degree, even if it is only through granting permits and regulating the creation of cell phone networks.

[7] More examples of protest mobilization via information technology in general and cell phones in particular can be found in Diamond (2012).

[8] A detailed analysis of a simple cross-tab and the regression model can be found in the online Appendix (http://www.journals.cambridge.org/psr2013007).

can be achieved in the African context. Furthermore, there exist strong theoretical considerations that suggest the marginal benefits of improved communication technology are substantial for insurgent groups.

Organizing violence is fraught with challenges. Successful insurgent activity requires solving various collective action and coordination problems (Kalyvas and Kocher 2007; Wood 2003), such as the free-riding problem (Olson 1965). This is particularly true when it comes to the organization of political violence, where participation is risky and benefits are often unclear (Shadmehr and Bernhardt 2011). Free riding within groups arises because members of insurgent groups have to endure the high costs of engaging in violence, but the potential payoffs for toppling the government will accrue to the wider population. Hence, rebel leaders have to ensure that group members actively contribute continuously throughout the conflict. Collective action problems also arise in the support network of rebel groups. Effective insurgencies rely strongly on the tacit support of the local population (Kalyvas 2006). Here, insurgents have to convince supporters to offer material support or valuable information from local residents, who themselves have an incentive to free ride.

In addition, insurgent groups suffer from strong coordination problems. Even if rebel groups can convince members to actively fight and the local population offers tacit support, military action needs to be carefully coordinated to be successful. Warfare against state forces with superior military technology, firepower, and training relies on careful plotting of attacks, appropriate timing, coordination of group movements in target areas, and managing the retreat to safe havens. While organizing protests is often about getting the right people together at the right time and place, insurgent violence requires the coordinated interplay of independent groups across distant geographic locations and time.

Recent work on mass media and violence has shown preliminary evidence on how radio and television can facilitate or block civil violence (Warren 2013; Yanagizawa-Drott 2012). Warren (2013) shows a reduction in militarized challenges to the state, if mass media access is widespread across its territory. The "soft power" of mass media enables the government to dissuade insurgent collective action through dissemination of progovernment propaganda. Observers of propaganda radio in Africa have highlighted the potential dangers for ethnic strife and violence (Livingston 2011). Going beyond qualitative accounts, Yanagizawa-Drott (2012) uses data on radio access in Rwandan villages to document the effects of "hate radio" on killings between Hutu and Tutsi during the genocide. Here, the use of mass media by one conflict faction shifted public perception and facilitated violent collective action. Both arguments emphasize the role of mass media in creating shared beliefs about the enemy and the convergence of privately held information. This can facilitate or hinder collective action and coordination. The formal literature on information and coordination problems in collective action against the government has also emphasized the importance of public (potentially government controlled) and private signals (Edmond 2012; Shadmehr and Bernhardt 2011).

We contend that, in contrast to mass media, access to individual communication technology like cell phones can undermine the effects of government propaganda and, more importantly, play an integral part in overcoming other specific collective action and coordination problems inherent in insurgent violence. Through improved communication and monitoring, cell phone technology aids overcoming internal collective action problems, allows the distribution of information to tacit supporters in the wider population, and, on an operational level, allows for real-time coordination of insurgent activity.

Several organizational technologies can be used to improve cooperation among group members when dealing with the free-riding dilemma. Selective incentives and external punishment can be used effectively by rebel leaders to elicit support from rank-and-file insurgents and civilian supporters. At the same time, free-riding behavior can also be curbed through repeated interaction, increased communication, and improvement in the monitoring of group member's actions. The cheap availability of cell phones naturally improves and increases the communication between group members and allows for the tightening of group networks. The interaction between group members becomes more likely as the provision of cell phones makes long distance communication easier, especially in the context of rural insurgencies in which factions operate apart from each other for longer periods of time. The reduction of transaction costs resulting from the access to cell phone technology is especially valuable in many infrastructure-poor African regions, where this development makes personal long-distance communication possible for the first time. It is important to note that this does not require each individual to own a cell phone device, as cell phones can be shared collectively between group members or villagers.

Enlarging the communication network of rebel groups as well as increasing the rate of communication by group members should raise in-group trust between individual participants. The possibility for fast and easy communication boosts the propensity and rate of information sharing within groups, creating a shared awareness among group members. As Shirky (2008, 51) writes, collective action is critically dependent on group cohesion. The expansion of within-group communication is likely to foster shared beliefs and awareness of groups, thus providing one channel of easing collective action. The higher rate of communication between individual group members also makes the transmission of messages and instructions from group leaders through the decentralized network more likely and efficient. Furthermore, the increase in two-way communication vastly raises opportunities for monitoring each other's behavior. Rebel leaders can exert better control over their rank and file and their wider support network, thus limiting free-riding behavior.

On a more general level, the spread of personal communication technology to the general population aids

the flow of information and the coordination of beliefs not only within the particular groups, but also in the population. In instances when public or corporate private news sources are unavailable or pro-regime, the increased possibility of cell phone communication can aid the distribution of news. Anecdotal evidence suggests that cell phone communication can be useful as a substitute to traditional media, where the press is suppressed.[9] Indeed, tipping-point models of protest and popular support (Kuran 1991; Lohmann 1994) suggest that if citizens are able to communicate their privately held beliefs about the regime, without fear of reprisal, public support for the regime can quickly transform into widespread opposition. The spread of cell phones makes the transmission of news to citizens throughout the country more likely. The support for insurgent activity can increase in the general population when news about government wrongdoings are communicated through citizen communication. For example, when news about indiscriminate killings by the government are more likely to travel through the population via cell phones, the general population may adjust the calculus of participation in nonviolent protests or even insurgent groups (Kalyvas and Kocher 2007). Reportedly, cell phones have been used effectively by Syrian rebels to spread information on government atrocities and rebel victories, greatly aiding insurgency efforts (Peterson 2012). The ability to spread information about government violence against civilians or other forms of repression through private communication networks should thus improve the position of the insurgents within the population.

Apart from affecting a group's ability to address collective action problems, the distribution of cell phones aids the coordination of actions, especially during asymmetric insurgent warfare. On a basic level, it allows insurgent commanders to better plan and implement operations. As noted above, successful insurgent warfare against the state requires high levels of coordination. The availability of cell phones can aid violent groups in the planning and execution of operations. Reportedly, Charles Taylor successfully utilized mobile phone technology to coordinate and control his rebel commanders in Liberia's civil conflict (Reno 2011, 4). Similarly, while Shapiro and Weidmann (2012) find a negative effect of cell phone availability on violence in Iraq, other research suggests that insurgents were aware and made use of the advantages of cell phones. One simple indication is that cell phone towers, in contrast to other infrastructure, were spared from insurgent attacks (Brand 2007). In addition, the use of cell phones to communicate enemy movements, scouting, and other intelligence has been emphasized (Cordesman 2005; Leahy 2005). Stroher (2007) highlights the use of cell phones by Iraqi insurgents as an organizational tool, for the spread of information, as well as to provide propaganda to group members and the population.

This also indicates that the gain of cell phone technology by rebels can possibly close the technological gap between government troops and the rebel movement. Prior to the availability of cell phone communication to private citizens, it is likely that the government had a significant advantage when it comes to in-group communication and group coordination. This likely affects combat strategies as well as, indirectly, the probability of winning for each side. The availability of cell phones may thus decrease or close the size of this gap. Common conflict models assume technology as an important factor in determining the probability of winning of the fighting parties (Blattman and Miguel 2010; Grossman 1991). Increasing the probability of winning by insurgents or rebels in turn should make the onset of conflict more likely.

While modern communication technology can play an important role for peaceful collective action in the form of protests, the marginal benefit of coordination is likely to be larger for organized violence. Protest in dense urban environments already enjoys several advantages for information sharing, monitoring, and coordination. Urban environments often offer other tools and opportunities to spread information and long-distance communication is less important in cities. Rural insurgents, on the other hand, can derive large benefits from private, mobile long-distance communication outside of major population settlements.

Empirical Implications

Given the logic laid out above we believe that, overall, the ability to communicate, monitor, coordinate, and spread information through private cell phone networks should improve the ability of rebel groups to organize political violence. Hence we contend that *local cell phone coverage will increase the probability of an occurrence of political violence*. We will test this proposition in the empirical section.

DATA

Testing the above specified argument requires a sample of cases in which violent collective action can conceivably be influenced by cell phone technology, as well as spatially disaggregated data on conflict and cell phone coverage. Given these requirements, focusing on the African continent offers several advantages over other world regions. The African continent has been, and still is, a major hotspot for organized violent conflict (The World Bank 2011), yet also exhibits strong temporal and spatial variation thereof. At the same time, cell phone technology has proliferated at a rapid pace across the continent in the last 15 to 20 years (Buys, Dasgupta, and Thomas 2009), including to regions with characteristics that make them more prone to hosting violent events (e.g., aggrieved populations,

[9] For example, text messaging and cell phone communication is often used to relay newsworthy events and government repression to media sources outside the country when traditional reporting is impossible. As journalists are unable to work from within the country, let alone attend protest or other violent events, the communication of news is left to actors themselves or bystanders via text messaging (see, for example, Fowler 2007).

poverty, difficult terrain, etc.). Often cell phones are the first long-distance communication device available in those areas. This confluence of factors creates an ideal environment to assess the impact of modern communication technology on facilitating violent collective action. Most other world regions lack such a high level of variance in conflict and access to cell phone technology. In addition, high-quality georeferenced data on conflict events is scarce for most regions of the world. Fortunately, recent years have seen an increase in the number of available conflict datasets that provide this type of information, in particular, for Africa.

For our primary analysis, we rely on the recently updated conflict data provided in the UCDP Georeferenced Event Dataset (UCDP GED) (Melander and Sundberg 2011; Sundberg, Lindgren, and Padskocimaite 2011). The UCDP GED includes yearly event data on *organized* violence in Africa from 1989 up to 2010. Violent events are included in the data if the conflict with which the event is associated has totaled 25 or more deaths and the event itself led to at least one death.[10] We use data on organized forms of violent collective action, instead of data on protests, for two reasons: First, our theoretical argument is geared specifically to the effects of cell phone communication technology on organized and violent forms of collective action. Second, quality and coverage of georeferenced data on organized violent collective action in Africa is higher than for other, more spontaneous and nonviolent forms of collective action.

Each event in the conflict dataset is specified to a location through longitude and latitude coordinates and by a date. We can use these data to map violent events across Africa for a number of years.[11]

Importantly, since the event data are based on news reports, one might expect the danger of measurement bias, as cell phone coverage may affect the probability of reporting of events. This is a valid concern, but we believe it is mitigated through several factors. For one, the UCDP coding team relies on a large number of print, radio, and television news reports from regional newswires, major and local newspapers, secondary sources, and expert knowledge, attempting to cover events even in remote locations without access to cell phone coverage. Furthermore, the focus on events with at least one death increases the likelihood of better event coverage in comparison to more low intensity events like peaceful protests or strikes.[12] A quality comparison of the UCDP-GED and ACLED data by Eck (2012) concludes that the UCDP data have higher quality and report often dramatically more events in rural or remote areas compared to ACLED. In addition, we also control for a number of other factors that would account for measurement bias in the event count in our empirical models, such as distance to the capital, local GDP per capita, or population size. Conditional on these factors, it is unlikely that cell phone coverage will be associated with any further over-reporting of events.[13]

Data on cell phone coverage are provided through *Collins Coverage* by *Harper Collins Publishers*. The data are made available by cell phone companies via the *GSMA* or *Collins Bartholomew*.[14] The availability and extent of coverage is represented via spatial polygons. We received data on GSM 2G network coverage for the first quarter of 2007, 2008, and 2009.[15] Our data only indicate the availability of cell phone services, not network traffic and usage by citizens. Information on usage is simply not available and without further information on the number of subscribers might also be misleading with regard to the role of cell phone communication for collective action in the wider population. As noted above, the argument does not require the ownership of cell phones by each individual, as phones can be shared within groups and villages. More importantly, we believe assessing the effect of coverage is more relevant from a policy perspective. While individual use of cell phones is hard to measure

[10] Violent events are defined by UCDP as the following: "The incidence of the use of armed force by an organised actor against another organized actor, or against civilians, resulting in at least 1 direct death in either the best, low or high estimate categories at a specific location and for a specific temporal duration" (Sundberg, Lindgren, and Padskocimaite 2011, 5). The UCDP data combine information on state-based armed conflict, nonstate conflict, and one-sided violence. We believe our theoretical argument applies to some degree to all forms of violence, but in future research we hope to differentiate.

[11] A number of robustness tests were performed by using the ACLED (Raleigh *et al.* 2010) conflictual event data as the dependent variable, as well as excluding those events with low precision on the conflict location. The results are presented in additional tables in the online Appendix.

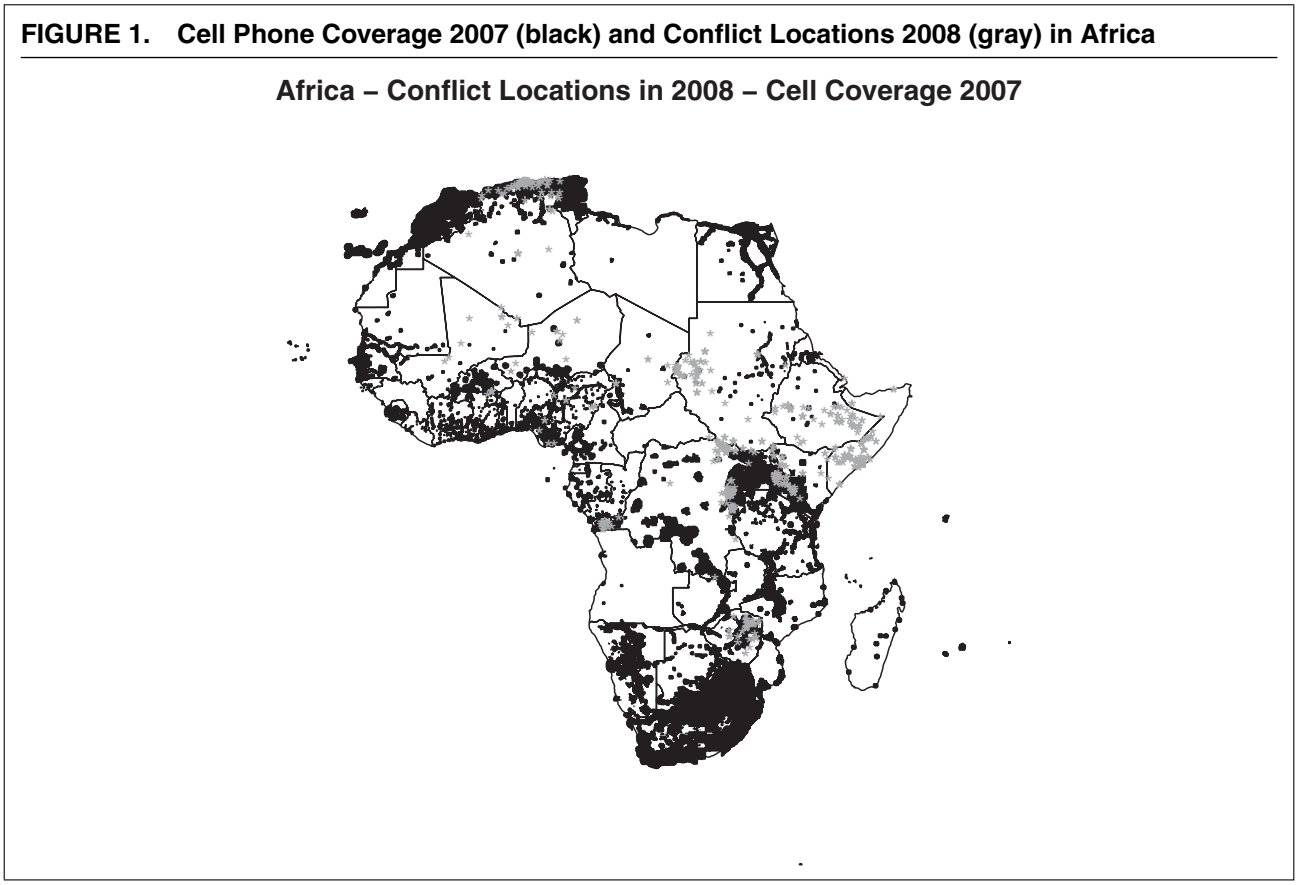
[12] Research in sociology and political science has evaluated the effect of "newsworthiness" on the likelihood of an event being reported in various news sources. The intensity or violence of an event is one of the important factors that often increases the chances of inclusion (Earl *et al.* 2004). Hence, by relying on UCDP-GED data, we maximize the chances that even events outside of areas with cell coverage are reported.

[13] In addition, in our main empirical models we collapse the event counts to a simple binary dummy variable, which reduces the scope of potential measurement bias: the exact reporting of event counts might be influenced by the availability of modern communication technology, but an information on the mere presence of any violent events in a grid-cell year is much less likely affected by underreporting. We also tested for an interaction between capital distance and cell phone coverage. We find that the effect of cell phone coverage is weaker in areas far from the capital, which is the opposite of what one would expect if a positive association were solely driven by measurement bias. Last, we also visually compared maps of violence in Sierra Leone's civil war (1991–2002) based on the UCDP-GED data with a map based on household-level survey data collected by the World Bank. The map is provided by Sacks and Larizza (2012). The visual comparison reveals that the UCDP-GED data very clearly track patterns of self-reported violence in the 153 chieftains in Sierra Leone (see online Appendix). Importantly, UCDP GED constructed these event data based on news reports without local cell phone coverage (1991–2002), i.e. UCDP GED is able to report violent events of sufficient quality irrespective of modern communication technology.

[14] GSMA website: http://www.gsma.com/home/; Collins Bartholomew website: http://www.bartholomewmaps.com/.

[15] In addition we have data on the 3G network coverage. 3G coverage though is much smaller and concentrated in a few countries, e.g., South Africa. Since areas with 3G coverage are a strict subset of 2G coverage, i.e., any area with 3G coverage also has 2G coverage, but not vice versa, 3G networks are unlikely to have any appreciable effect on collective action above and beyond 2G technology.

FIGURE 1. Cell Phone Coverage 2007 (black) and Conflict Locations 2008 (gray) in Africa



and control, coverage is the first and most important step in extending access of cell phone technology to the wider population.

Figure 1 shows the distribution of conflictual events in Africa in 2008, as well as areas with available GSM 2G coverage in 2007. As one can easily see, cell phone coverage is most widely spread in South Africa, Namibia, Kenya, as well as in northern Africa (specifically Morocco, Tunisia, and Egypt). However, coverage has expanded massively in the past years and has become more and more available in other areas of the continent. While coverage is more likely in coastal areas, the map clearly shows that it has been expanded further into the continent and away from population centers. Areas with a clear overlap in cell coverage and conflict events are in Algeria, the DRC, Kenya, Nigeria, Uganda, and Zimbabwe.

To analyze the relationship between the local availability of cell phone coverage and the occurrence of violent events we follow Buhaug and Rød (2006) in relying on spatially disaggregated grid cells as our units of analysis. Our grid is partitioned into 0.5×0.5 decimal degree resolution cells, i.e., each grid cell is approximately 55 km $\times$ 55 km large. Using such high-resolution spatial units of analysis allows us to avoid problems of data aggregation common in cross-national studies of violence. The grid was created by Tollefsen, Strand, and Buhaug (2012) at the Peace Research Institute Oslo (PRIO). The PRIO-Grid dataset provides grid cells and data for the whole world on a yearly basis from 1946 to 2008. Given our particular interest, we only use the data concerning Africa.

Using the grid provided in the PRIO-Grid dataset we create our dependent variables based on conflict locations in the UCDP GED dataset (Melander and Sundberg 2011; Sundberg, Lindgren, and Padskocimaite 2011). First, we generate a conflict indicator, a binary variable that takes the value of 1 in cases where one or more conflictual events were registered by UCDP GED in the given grid cell in 2008, and 0 otherwise. Our dataset consists of 10,674 cells, of which 3.3% experience violent conflict in 2008, thus conflict is quite rare. Second, for additional robustness checks we create a conflict count variable that counts the number of conflictual events in 2008 according to the UCDP GED data for each grid cell. Despite the increase in variation between grid cells, we use the count measure only as a secondary variable, because multiple counts within each grid-cell year are likely to be realizations of the same conflict process.

Our main independent variable of interest is generated in a similar manner. For each grid cell an indicator for cell phone coverage is created that takes the value of 1 if cell phone coverage existed in 2007 and 0 otherwise.[16] The distributions of cell phone coverage in 2007 and conflict locations in 2008 are presented in

[16] While it would certainly be preferable to use the percentage of area covered by cell phone networks as our main independent variable, we are confident that given the size of the individual grid cells

Figure 1. In 2007, cell phone coverage was available in 37% of grid cells; in 2008 this increased to 38%.

To identify a potential causal effect of cell phone coverage on violent conflict events we rely on three complementary strategies: First, we use a series of standard models on the cross-sectional data for 2008 and control for a number of potential confounding factors to approximate the potential causal effect of cell phone coverage. Second, we take the same cross-sectional data and implement an instrumental variable strategy that leverages exogenous variation in our main independent variable. Third, we use conflict data and lagged cell phone coverage for 2008, 2009, and 2010 to construct a short panel for African grid cells and implement a set of panel data approaches that exploit over-time variation.

Confounding Variables

A large literature on civil conflict has identified a collection of theoretically motivated factors that contribute to organized violence. It will be important to understand the effects of cell phone coverage in a context that provides motive and opportunity for violent collective action (Collier *et al.* 2003). The existing literature has emphasized structural factors that affect the motivation of parties potentially seeking violent conflict with the state, such as poverty, inequality, ethnic fractionalization, or ethnic exclusion. At the same time, other factors, for example mountainous terrain, forests, or natural resources, can impact the ability of groups to rebel and have also been identified as drivers of violence. For our first set of cross-sectional models it will be particularly important to control for other variables that contribute to conflict. It is reasonable that those variables which drive conflict are also likely to correlate with the availability of cell phone coverage and might thus induce omitted variable bias in our findings. The majority of control variables in our models are also provided in the PRIO-Grid dataset, but originally come from other sources. Time varying independent variables were lagged by one year (2007) to control for the possibility of reverse causality.

Our models include a measure of the distance to the capital as well as distance to the border for each grid cell, as certain conflicts are more likely to occur close to the capital or close to other countries (Buhaug and Rød 2006).[17] Similarly, conflict is more likely to occur in regions with larger populations (Fearon and Laitin 2003). Hence, an estimate of population size for each grid cell is included. These variables are particularly important since cell phone providers are most likely to build infrastructure around the capital and population centers. The data on capital and border distance are provided through the PRIO Grid, as are the population data, which originally stem from CIESIN (2005).[18] We also include a variable measuring prior conflict levels for each grid cell, based on UCDP conflict events in each grid cell from 1989 to 2000.

In addition, we include controls for the percent of mountainous terrain, as this may be advantageous to guerrilla warfare and thus may make fighting more likely. It may also affect the likelihood of coverage availability.[19] This variable was originally collected by the UN Environment Programme (UNEP-WCMC World Conservation Monitoring Centre 2002), but is available in the PRIO GRID. In addition, we control for the percent of area in a grid cell that is equipped for irrigation.[20] This variable is again provided in the PRIO-Grid dataset, but was originally collected by Siebert *et al.* (2007).

Violent conflict is often thought to be more likely in poorer regions, where the substitution costs for engaging in violence are particularly low and grievances with the current government are high (Blattman and Miguel 2010; Collier and Hoeffler 2004; Fearon and Laitin 2003). Cell phone coverage, on the other hand, is more likely in richer areas of the continent. Thus controlling for income is highly warranted. Economic data are provided in the PRIO-Grid dataset as well, and originally stem from the G-Econ dataset by Nordhaus (2006). We use per capita GDP for 2000 calculated for each grid cell.[21]

For further robustness checks we control for potential ethnic grievances by including a variable on the exclusion of ethnicities. To do so we match data on the identity of ethnic groups in each grid cell with data on the political exclusion of ethnic groups in a given country, recording how many local ethnic groups are politically excluded. The spatial data on settlement patterns of ethnic groups originally stems from Weidmann, Rød, and Cederman (2010) and were merged with data on political exclusion by Cederman, Wimmer, and Min (2010).

Furthermore, we include data on the location of natural resources. This may be warranted as cell phone companies are likely to extend coverage to areas with important economic activity. In addition, as rebel groups try to capture natural resources, fighting in these

(55 km × 55 km) using an indicator variable should not affect our results substantively.

[17] In addition this helps to control for under-reporting of events in remote areas.

[18] As an alternative to using simple population counts, we also consider a log transformation. One issue for the transformation is the presence of grid cells with zero population. To address this issue (even if insufficiently), we add 1 to each population count to allow for the log transformation. Using log transformed population counts instead of the original counts has no implication for the effect of population on conflict, but does weaken our main findings for cell phone coverage somewhat. Importantly though, for our most conservative models and the instrumental variable estimation, all main findings are unaffected.

[19] As an alternative measure we tested the share of forested land in each grid cell, which has no effect on our main results.

[20] Unfortunately this measure is only available for the year 2000, however it should be highly correlated with later data.

[21] Originally the GDP data are calculated for 1×1 decimal degree grid cells, thus each grid cell in the G-Econ dataset contains four grid cells of the PRIO-Grid dataset. We also consider a log transformation for the GDP variable, with no effect on the findings for GDP per capita or the cell phone coverage variable.

TABLE 1. Binary DV Models

	Logit, Robust SE	Logit, Robust SE	Re-Logit, Robust SE	Mixed Effects Logit	Mixed Effects Logit	Fixed Effects OLS, Robust SE
(Intercept)	−3.814***	−4.020***	−4.020***	−4.020***	−3.340***	−0.014†
	(−20.178)	(−21.449)	(−21.422)	(−21.652)	(−16.490)	(−1.649)
Pre-2000 Conflict	0.020†	0.019†	0.019†	0.019***	0.021***	0.002**
	(1.861)	(1.850)	(1.834)	(5.680)	(6.192)	(3.040)
Border Distance	0.000	0.000	0.000	0.000	−0.000	−0.000**
	(0.450)	(0.884)	(0.922)	(0.941)	(−0.416)	(−2.701)
Capital Distance	0.000	0.000*	0.000*	0.000*	0.000	−0.000
	(1.629)	(2.264)	(2.270)	(2.327)	(1.604)	(−0.014)
Population	0.000*	0.000**	0.000**	0.000***	0.000***	0.000*
	(2.482)	(2.733)	(2.611)	(4.510)	(4.776)	(2.545)
Pct Mountainous	1.641***	1.578***	1.578***	1.578***	1.698***	0.056***
	(8.518)	(8.410)	(8.413)	(8.391)	(8.793)	(5.305)
Pct Irrigation	−0.027†	−0.031†	−0.031†	−0.031†	−0.046*	−0.001***
	(−1.663)	(−1.851)	(−1.651)	(−1.834)	(−2.456)	(−3.558)
GDP pc	−0.000***	−0.000***	−0.000***	−0.000***	−0.000***	−0.000
	(−3.589)	(−3.915)	(−3.881)	(−5.590)	(−3.924)	(−0.404)
Cell Phone Coverage		0.390**	0.390**	0.390**	1.112***	0.027***
		(2.798)	(2.798)	(2.836)	(7.319)	(5.824)
Mean Cell Coverage					−2.806***	
					(−8.505)	
Country Fixed Effects	No	No	No	No	No	Yes
AIC	2269.560	2263.781	2263.781	2222.052	2147.475	−7590.326
BIC	2326.699	2328.063	2328.063	2293.476	2226.041	−7211.780
Deviance	2253.560	2245.781	2245.781	2202.052	2125.475	240.027
Log-likelihood	−1126.780	−1122.891	−1122.891	−1101.026	−1062.737	3848.163
N	9343	9343	9343	9343	9343	9343

†$p = 0.1$. *$p = 0.05$. **$p = 0.01$. ***$p = 0.001$.

areas is also more likely. We therefore include indicators for the location of diamond mines (Gilmore *et al.* 2005) as well as known gas and oil deposits (Lujala, Rød, and Thieme 2007). Summary statistics for all variables are included in the online Appendix.

CROSS-SECTIONAL ANALYSIS

Using the data described in the prior section, we estimate a series of cross-sectional statistical models to evaluate our hypothesis. The main measure of conflict we use is the simple binary conflict indicator for each grid cell. Naturally, we utilize the generalized linear model framework to formulate our probability models. The dependent variable y_i for each grid cell i in 2008 is binary,

$$y_i = \begin{cases} 1 & \text{conflict,} \\ 0 & \text{otherwise,} \end{cases}$$

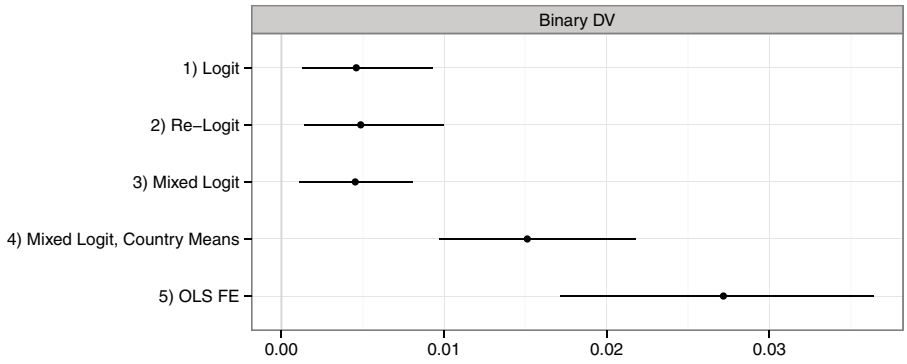
and is modeled as a binomial process. We link the response variable to observed covariates via a standard link function (i.e., logit) to the linear predictor η: $P(Y = y|X) = \mu = g(\eta)$. The linear predictor in turn is a function of control variables and our cell phone coverage indicator:

$$\eta_i = \mathbf{x}_i'\beta + c_i'\gamma,$$

where $\mathbf{x}_i$ is a vector of control variables and the intercept and c_i is the indicator of cell phone coverage. The parameter γ measures the impact of improved communication technology on violent collective action.

We consider five alternative estimation approaches. Each of these models has certain advantages and disadvantages. They address distinct issues present in our data and differ in the severity of assumptions needed to attribute causal effects to the estimated cell phone coverage parameter. Table 1 shows parameter estimates and z statistics for all models. The first column shows a baseline specification of only control variables, estimated with a standard logit model and robust standard errors to address issues of heteroskedasticity. The second column simply adds our cell phone coverage indicator as a covariate. The third column presents the results of a rare-events logistic regression to account for rare events bias (Tomz, King, and Zeng 2003). The fourth logit model includes country-level

FIGURE 2. First Differences, With and Without Cell Phone Coverage, Binary Dependent Variable, Baseline $P(Y = 1|X) \approx 0.01$



random effects to vary baseline levels of conflict across countries.[22] The standard random intercept model assumes zero correlation between the random effects and other covariates. To further control for potential omitted variable bias, we present in column 5 the results for a mixed effects logit model that also includes the country-level means of the cell phone coverage indicator. Including the country-level mean of the variable of interest allows for a correlation between the country random effect and the mean level of cell phone coverage (Gelman and Hill 2008, 506), removing the effects of country-level unobservables that affect cell phone coverage (Bell and Jones 2012). Last, we also include the estimates of a linear probability model estimated via OLS that allows the inclusion of country-level fixed effects to control for any unobserved time-invariant country characteristics that might bias our findings.[23]

The baseline specification in column 1 shows that a number of our control variables perform as expected. Prior levels of conflict have a statistically significant and positive effect on experiencing a conflict event in 2008. Similarly, population counts in the grid cell and mountainous terrain also increase the probability of conflict. On the other hand, in line with theoretical expectations and prior empirical findings the percentage of land with irrigation technology and GDP per capita reduce conflict (Buhaug and Rød 2006; Lujala, Buhaug, and Gates 2009; Buhaug *et al.* 2011; Fearon and Laitin 2003).

Across all models which include our measure of cell phone coverage, the cell phone coverage indicator is estimated to increase the probability of conflict and is precisely estimated—statistically significant below the 1% or even the 0.1% level. Even when controlling for the country level of cell phone coverage and only exploiting within country variation, as in the mixed effects logit model, or including country fixed effects, we always find a clear positive effect. Given that we control for a sizable number of confounding variables, as well as unobserved country-level factors, the results in Table 1 offer a good first approximation of the effect of cell phone coverage on political violence. In addition, including the cell phone coverage variable improves model fit statistics. A likelihood ratio test between a model including cell phone coverage and the nested model results in a significant test statistic at the 1% level in favor of the model including our variable of interest. We implement further analyses of the model fit, amongst others using separation plots (Greenhill, Ward, and Sacks 2011); these are displayed in the online Appendix.

Substantive Effects

Before implementing further robustness checks, we evaluate the substantive effects of cell phone coverage. To evaluate the impact of access to cell phone technology we simulate first differences of predicted probabilities for the cell phone coverage indicator, setting all control variables at their respective means (King, Tomz, and Wittenberg 2000). Figure 2 plots the mean effects and 95% confidence intervals for each model. The baseline probability of conflict in a grid cell with all variables at their means, but with no cell phone coverage, is approximately 1%. A grid cell with the same configuration of control variables, but with access to cell phone coverage is expected to see an increase of roughly 0.5 percentage points. The estimated effect is even larger (one to three percentage points) in the models where we control for the mean level of coverage or include country fixed effects. Thus, holding everything constant and extending cell phone coverage to a grid cell is estimated to increase the probability of a conflict event occurring by 50% for the standard logit model and up to nearly 300% for the fixed effects model.

These results imply that cell phone coverage facilitates violent collective action. Whereas the probability of conflict is still very low, the marginal effect of cell phone provision holding all other variables constant is quite large. Compared to the baseline probability with no cell phone coverage, areas with cell phones are much more likely to experience violent events. This

[22] We estimate the mixed effects model using the *lmer( )* function in *R*.

[23] We present standard robust standard errors but the results are very similar with standard errors clustered at the country level.

indicates that in areas with structural conditions that favor violence, cell phone coverage enables groups to overcome their collective action and coordination problems more easily, which translates to more organized conflict events.

Spatial Dependence

A common problem in the analysis of conflict, especially when using highly spatially disaggregated data, is spatial dependence between units of observation. The plot of conflict events in Figure 1 clearly shows a spatial clustering that suggests issues of nonindependence, i.e., a conflict event in one grid cell might increase the probability of a conflict event in a neighboring cell. Recently, the analysis of spatial dependence in comparative politics and international relations has gained increased attention (Beck, Gleditsch, and Beardsley 2006; Franzese and Hays 2008; Hays and Franzese 2007; Neumayer and Plümper 2012). Neglecting to account for spatial dependence in the data-generating process can lead to biased and inconsistent parameter estimates (LeSage and Pace 2009). A popular approach to modeling spatial dependence in the standard linear framework relies on the inclusion of a spatial lag. Usually a spatial lag represents the weighted average of the dependent variable in "neighboring" units. The neighborhood structure is defined through a spatial weights matrix and can be based on adjacency, nearest-neighbor, distance or other geographic or social connectivity concepts. The use of spatial lags for binary dependent variables or other distributions in the GLM setting has been employed in the analysis of conflict (Ward and Gleditsch 2002; Weidmann and Ward 2010) but presents formidable computational challenges (LeSage and Pace 2009). The correct estimation of parameters in the presence of the simultaneity between the dependent variable and the spatial lag becomes especially harrowing for large datasets.

Given that our African lattice has over 10,000 cells, computational hurdles become prohibitively high. Short of a correctly specified spatial lag model, many researchers rely on a simpler approach to avoid the computational issues of nonlinear spatial lag models. Any direct simultaneity can be avoided if the spatial lag is also temporally lagged. It is then only a standard covariate and can be included as such in the GLM specification. While not ideal, this approach is feasible and does capture some of the spatial dependence in the data. We calculate for our binary dependent variable a spatial conflict lag based on a six-nearest-neighbor spatial weights matrix in 2007. The online Appendix presents a table with the models from Table 1, including spatial lags. Across all models we find a statistically significant and positive effect of the spatial lag, suggesting that conflict in neighboring grids increases the likelihood of violence and underscores the presence of spatial dependence in the data. The cell phone coverage indicator is uniformly estimated to be positive, but just misses significance at the 10% level in the logit models. If we include country random effects and a control for the country mean of cell coverage to allow for correlated random effects or simply include country-level fixed effects, the effect is found to be significant below the 0.1% level.

Count Models

As an alternative to the indicator of conflict, we also employ the number of conflictual events in each grid cell as a dependent variable. The online Appendix shows parameter estimates and associated z statistics for a simple Poisson regression with robust standard errors, a negative binomial model to allow for overdispersion in the counts, and the same models with spatial lags. As with the binary dependent variable, the cell phone coverage indicator is estimated to have a positive effect on conflict counts and is statistically significant below the 0.1% level in the Poisson model and below the 10% level in the negative binomial model.

Alternative Measures, Natural Resources, and Ethnicity

One problem with using the UCDP conflict events as our dependent variable is potential measurement error in the conflict location. The UCDP data pinpoint latitude and longitude for each conflict event, but the accuracy of the location varies. For some events the exact location is identifiable, whereas for others, only the administrative unit or region is available. In the second case, UCDP uses the unit centroid as the location identifier. Fortunately, for each event the UCDP data record the quality of geographic information on a seven-point scale. We thus create a dependent binary and count variable that only considers events with fairly exact geographic information, to make sure our results are not biased by events where the exact location or spatial extent was unclear.[24]

In addition, we also consider the ACLED (Raleigh et al. 2010) conflict event data available for Africa from 1997 to 2010, which differs slightly in the definition of conflict events from UCDP. Importantly, ACLED also covers violent protests with death counts below the 25 person threshold. We repeat all analyses using binary and count dependent variables based on the alternative (*precise*) UCDP events and the ACLED events. Across both alternative measures, we obtain the same results in terms of substantive and statistical significance. For some models statistical significance even increases (all results are available in the online Appendix).

Our last alternative measure for the dependent variable takes advantage of new geo-coded data on social unrest in Africa (Salehyan *et al.* 2012). The "Social Conflict in Africa Database" (SCAD) codes news on a multitude of social conflict events, covering protests, riots, strikes, intercommunal conflict, and government violence against civilians, on the African continent from

[24] Given the 55-km × 55-km grid size, we use all events that were coded 1–3 on the geographic location quality variable, i.e., observations with exact know locations, or where the limited area around an exact location or the district/municipality is known.

1990 to 2011. Using this measure allows us to capture low intensity collective action that did not necessarily end in a large number of deaths. SCAD also provides the geographic coordinates of events. Identical to the previous analyses, we create an indicator and a count variable based on SCAD events for each grid cell. We use those observations that are geolocated with sufficient quality and are not already included in the ACLED database.[25] We run the same set of binary dependent variable models, count, and spatial lag models as before. While the results for our control variables change, reflecting the difference in processes between organized rebel violence and social unrest, for all models we still find a clear positive and highly statistically significant effect of cell phone coverage on the incidence of social conflict events (all results are available in the online Appendix).

Apart from considering alternative measures for the dependent variable, we also address two additional concerns of omitted variable bias. Recent studies on civil conflict and ethnicity have established a link between the political exclusion of ethnic minorities and the propensity for group conflict (Cederman and Girardin 2007; Cederman, Weidmann, and Gleditsch 2011; Cederman, Wimmer, and Min 2010). If politically excluded ethnic groups are more prone to engage in violent collective action and at the same time locations in which these groups are dominant are provided with less cell phone access, omitting information on ethnicity from our analysis might bias our estimates. To control for this possibility we utilize information on politically relevant local ethnic groups in each grid cell provided by the PRIO-GRID data, based on the Geo-referencing of Ethnic Groups (GREG) project (Weidmann, Rød, and Cederman 2010). Given the group identifier in each grid cell we join information on the status of political inclusion or exclusion at the national level in the Ethnic Power Relations Dataset for each group (Wucherpfennig *et al.* 2011). We repeat the analysis for the binary and count models including a variable that measures the share of local ethnic groups that are politically excluded. For most models, we find that political exclusion of ethnic groups increases the probability of conflict, but has no effect on the direction or statistical significance of the cell phone coverage indicator (results are available in the online Appendix).

Similarly, as noted above, local natural resources might provide motives for local rebel groups to engage in extraction to secure access to economic rents (Collier and Hoeffler 2004; Lujala, Gleditsch, and Gilmore 2005). In addition, regions with lucrative petroleum or diamond deposits might receive better cell phone coverage as mining and oil companies can influence the construction of cell phone towers. To correct for potential omitted variable bias, we use information on the geographic location of diamond mines (Gilmore *et al.* 2005) and oil and gas deposits (Lujala, Rød, and Thieme 2007) and include an indicator variable for grid cells that cover a known resource deposit. As before, we re-estimate all models and find some evidence that petroleum increases the probability of conflict and diamond mines surprisingly reduce the incidence of violence. Though neither variable has any effect on the role of cell phone coverage, which stays consistently positive and statistically significant (results available in the online Appendix).

[25] Specifically, we exclude events for which the geographic location was "nationwide" or "unknown."

Matching

Alternative to the estimation of parametric models, we also explore the effect of cell phone coverage using matching methods. In particular, we rely on "Coarsened Exact Matching" (CEM) (Iacus, King, and Porro 2012). CEM bins observations into coarsened strata and matches based on the new groupings. This matching approach reduces imbalance in the sample based on all properties of the covariate distributions, not just differences of means or similar univariate statistics (Iacus, King, and Porro 2011). Details on the matching procedure are reported in the online Appendix. Overall, our matching-based estimates are very similar in magnitude to our original estimates and confirm the main finding.

INSTRUMENTAL VARIABLES

One important concern is the potential endogeneity between conflict events and cell phone coverage in each grid cell. It is plausible that conflict destroys cell towers and reduces coverage, suggesting a potentially negative relationship in the data. Although we lag cell phone coverage by one year in our models to address causal ordering and the estimated positive effect suggests we might actually underestimate the true relationship, we aim to address fundamental endogeneity concerns with an explicit identification strategy relying on an instrumental variable.

Prior work on the spread of cell towers in Africa has identified a number of geographic characteristics that predict cell phone tower locations (Buys, Dasgupta, and Thomas 2009). However, in our case, for most of these variables the exclusion restriction of the instrumental variable estimator is likely to be violated since remoteness, difficult terrain, or population density have been found to predict conflict as well. Yet, in addition, Buys, Dasgupta, and Thomas (2009) identify the regulatory environment of African countries as an important predictor of cell phone coverage. A series of studies has found that healthy private competition in the cell phone market leads to better coverage and provision of cell phone services, in comparison to single-provider state-run systems (for references see Buys, Dasgupta, and Thomas (2009, 1495)). A World Bank study on telecommunications policy reform in 24 African economies finds important policy changes pertaining to privatization, increased competition, the formalization of regulations, and the creation of regulatory agencies, all contribute to improved coverage (The World Bank 2010). Buys, Dasgupta, and Thomas (2009) find the World Bank's Country Policy and

TABLE 2. Instrumental Variable Models

	2SLS, Robust SE	2SLS, Robust SE
(Intercept)	−0.141***	−0.100***
	(−6.55)	(−5.42)
Spatial Lag		0.522***
		(9.91)
Pre-2000 Conflict	0.003***	0.001
	(3.98)	(2.28)
Border Distance	.0001***	.0001***
	(4.31)	(3.63)
Capital Distance	.0001***	.0001***
	(7.93)	(6.45)
Population	−0.000†	0.000
	(−1.88)	(−0.79)
Pct Mountainous	0.065***	0.033**
	(4.69)	(2.71)
Pct Irrigation	−0.004***	−0.022**
	(−4.31)	(−2.73)
GDP pc	0.000	0.000
	(0.57)	(1.14)
Cell Phone Coverage	0.289***	0.183***
	(6.33)	(4.62)
Kleibergen-Paap rk LM Statistic	108.698	110.235
Kleibergen-Paap rk Wald F Statistic	125.442	126.637
N	6598	6598

†$p = 0.1$. *$p = 0.05$. **$p = 0.01$. ***$p = 0.001$.

Institutional Assessment (CPIA) measure of regulatory quality is an important variable that affects cell phone coverage, even for highly spatially disaggregated data. We argue that the CPIA regulatory coverage measure is a good instrument for our purposes, since it is a robust predictor of cell phone coverage, i.e., avoids "weak instrument" criticisms, and has a strong claim to justifying the exclusion restriction. While regulatory quality affects private market competitiveness and the extent of cell phone coverage, we believe it is unlikely that an alternative, unmeasured causal link to violence exists.[26]

A potential issue could be that poorer African countries were forced to introduce regulatory reform in light of budgetary pressures and demands by outside actors. However, we find no aggregate link between the level of development and a country's regulatory score, as measured by the World Bank data. Furthermore, the estimated models include a control for GDP per capita levels and hence any effects regulatory quality might exert on violent collective action through poverty is accounted for.

We use the average CPIA regulatory quality score from 2005 to 2007 as our instrument for cell phone coverage. Initially we estimate a simple linear probability model via two-stage least squares (2SLS) and robust standard errors, which generally does well in identifying marginal effects, even with binary dependent variables (Angrist and Pischke 2009, 197–205). We estimate models with and without the spatial conflict lag. For both models in the first stage regression of cell phone coverage on regulatory quality, our instrument, is highly statistically significant. The Kleibergen-Paap rk LM statistic and the Wald F statistic are very high and we are able to reject the null hypothesis of under- and weak identification. The results of the second stage are reported in Table 2, showing coefficient estimates and *z* statistics for the binary dependent variable models. Controls largely perform as expected, but more importantly the instrumented cell phone coverage variable is positive and highly statistically significant for both models. The exogenous variation in cell phone coverage induced by the regulatory quality increases the probability of local conflict events.[27]

[26] More precisely, we assume that the instrument is independent of potential outcomes, that regulatory quality does not affect conflict other than through cell phone coverage, that the instrument is a good predictor of cell phone coverage, and, last, monotonicity in the first stage, which then identifies the local average treatment effect (LATE) (Angrist and Imbens 1994). The LATE here is the effect of cell phone coverage on violence in grid cells that received coverage due to regulatory effects, but not through other sources of cell phone service provision.

[27] In addition, we obtain statistically positive results if conflict counts are the dependent variable. These results provide an important additional layer of confidence in our results. We also implement bivariate probit models with our regulatory quality score as a predictor in the equation for cell phone coverage. Using robust as well as clustered standard errors, we again find a clear positive and statistically significant effect of cell phone coverage on the probability of conflict (all results are available in the online Appendix).

TABLE 3. Panel Data

	(1) Binary DV, OLS, Clustered SE	(2) Count DV, OLS, Clustered SE
Cell Phone Coverage	0.0116*	0.0502**
	(0.00547)	(0.0158)
Cell & Year Effects	Yes	Yes
Observations	32022	32022
Adjusted R^2	0.004	0.001
F	21.33	5.732

*$p < 0.05$. **$p < 0.01$.
Clustered standard errors in parentheses.

PANEL DATA

Our last approach to estimate the effect of cell coverage on violent collective action exploits variation over time. Data on cell phone coverage is available for 2007, 2008, and 2009. We utilize these data to construct a short, three-year panel of our grid cells in 2008, 2009, and 2010. This offers us an additional opportunity to not only use geographic variation in coverage, but also changes over time. While in 2007 cell phone coverage existed in 36.9% of all grid cells, that value increased to 37.8% in 2008 and 42.5% in 2009. The expansion of cell phone coverage allows us to compare grid cells before and after the expansion. In a first step, we estimate OLS models for the binary and count dependent variable with country and year fixed effects and our standard set of controls, clustering standard errors at the country level. Again, we find a highly statistically significant and positive effect of cell phone coverage on violent conflict events (detailed results are presented in the online Appendix). More importantly, the panel structure allows us to now include grid-cell fixed effects. Here, we control for all observed and unobserved factors for each grid cell in the three-year period from 2008 to 2010. Since our standard control variables at the grid level are constant over time, we only include the cell phone coverage indicator as a predictor, apart from the grid cell and year effects. Table 3 shows the estimated coefficient for the cell phone coverage indicator for the binary and count measure.

For both models, the cell phone coverage variable is estimated to be positive and is statistically significant below the 5% and 1% level, respectively. The coefficient size suggests an increase of one percentage point for the linear probability model and 0.05 events with the count dependent variable, very much in line with our prior estimates of the effect size. The results are substantively unchanged if we log-transform the count to make the distribution appear more normally distributed or alternatively estimate a Poisson fixed effects model. Again, this confirms our previous findings that the expansion of cell phone coverage in Africa facilitates violent conflict events.

Overall, our quantitative models demonstrate a clear positive association between cell phone coverage and the occurrence of violent organized collective action. This effect persists when controlling for a series of standard explanations of violence, as well as unobserved, time-invariant factors at the country and even grid level. Plainly, our results suggest that local cell phone coverage facilitates violent collective action on the African continent.

CONCLUSION

Whereas prior research has emphasized the positive consequences of expanding cell phone coverage across the African continent, this article is concerned with possible negative externalities. In general, increasing cell phone coverage in developing countries has been associated with higher levels of market efficiency, especially across labor markets and private goods markets. Cell phones decrease information asymmetries between market participants and facilitate economic exchange. However, few works have been concerned with the effect of new communication technologies in the political sphere. In particular, to our knowledge only Shapiro and Weidmann (2012) have examined how cell phone technology affects the propensity for political violence. Shapiro and Weidmann (2012) find that in the case of Iraq, the location of cell phone towers is negatively associated with violence.

In contrast, in this article we argue and provide evidence to show that cell phone technology can increase the ability of rebel groups to overcome collective action problems. In particular, cell phones lead to a boost in the capacity of rebels to communicate and monitor in-group behavior, thus increasing in-group cooperation. Furthermore, cell phones allow for coordination of insurgent activity across geographically distant locations.

We test the empirical relationship between cell phone coverage and the location of violent conflict across the African continent. To do so we utilize a grid of 55 km $\times$ 55 km cells across Africa. Using data on GSM2 coverage provided by the GSMA and georeferenced data on conflictual events by UCDP (Melander and Sundberg 2011; Sundberg, Lindgren, and Padskocimaite 2011), we create measures for each grid cell indicating whether cell phone coverage was available in 2007, as well as an indicator and count of conflictual events for 2008. In addition we include numerous other covariates to avoid potential omitted variable bias.

Across a wide range of empirical models, including various control variables and robustness checks, we find that cell phone coverage has a significant and substantive effect on the probability of conflict occurrence. When cell phone coverage is present, the likelihood of conflict occurrence is substantially higher than otherwise. We consistently find a relationship between cell phone coverage and violent conflict across standard logit models, models including controls for spatial correlation, random or fixed effects, as well as count models. In addition to traditional robustness checks, we

furthermore include instrumental variable regressions to test for the possibility of endogeneity and panel data models.

The results in this article stand in contrast to the findings presented by Shapiro and Weidmann (2012) regarding the relationship between cell phones and violence in Iraq. However, we believe it is reasonable that the effects of cell phones are different across these cases. The context of political violence in African countries is much different from that in Iraq. The military capacity of the anti-insurgent forces is likely higher in the case of the U.S. military and government forces in Iraq. While government forces in Iraq have the ability to monitor cell phone activity of insurgents, this is much less likely for many African governments, especially with the more prominent role of private enterprises in spreading technology. In addition, explicit whistle blower programs have so far only been used rarely in Africa (Livingston 2011). Similarly, the technological and strategic capacity of anti-insurgency forces in Iraq is likely to be much higher than that of many African forces. Thus the expansion of cell phone coverage may be less advantageous to Iraqi insurgents, whereas in the right context, rebels can make great use of it. At a minimum our findings suggest that we need further research investigating the specific conditions under which modern technology plays a role in insurgent and counterinsurgency activities.

Numerous exciting avenues for future research exist. First, a better theoretical understanding on how communication technology can affect collective action is warranted. The underlying mechanism for our findings needs to be unpacked further. Distinguishing between collective action and coordination problems might be particularly important. Our results only imply an association at the aggregate level of the spatial unit and do not reveal the exact causal mechanism in operation or the role of individual-level behavior. Naturally, future research will have to engage these questions in more detail and bring different data to bear. We suspect that the use of communication technology varies across contexts, rebel and insurgent groups, as well as counterinsurgency strategies. Exploring potential interactions with country or group-level variables will further illuminate the effects of communication technology on violence. Prior research on internal rebel group organization and the use of violence has focused on the role of internal norms and discipline (Weinstein 2007). Similar to recruitment strategies and the use of violence against civilians, the adoption of technology and its effects on rebel group behavior appear as promising topics of research to complement our aggregate-level findings.

Second, cell phone coverage should similarly have an effect on other forms of collective action, such as nonviolent protests. We do present some auxiliary evidence on the link between cell phone coverage and protest behavior using aggregate data (SCAD), but more research is warranted. The marginal benefits of modern communication technology likely varies across violent and nonviolent activities, which could lead to important substitution effects.

We do not believe that the spread of cell phone technology has an overall negative effect on the African continent. The increase in violence induced by better communication might represent a short-term technological shock, while the positive effects of better communication networks on growth and political behavior may mitigate root causes of conflict in the long run.

If the economics literature is correct in assuming that cell phone technology increases the productivity of farmers or service-oriented industries, then the spread of cell phones throughout Africa increases the returns to productive economic activity in the long term. This implies that the opportunity costs to violence (i.e., lost wages) increase, reducing the incentive to fight. Several formal models have identified this potential link between violence and economic activity (Chassang and Padro-i-Miguel 2009; Dal Bó and Dal Bó 2011; Grossman 1991; Grossman and Kim 1995). Some empirical work has shown a link between increased returns to labor-intensive production and lower violence in Colombia (Dube and Vargas forthcoming), while another study on the link between unemployment and insurgent activity in Iraq and the Philippines finds the opposite effect (Berman, Felter, and Shapiro 2009). However, the effect of cell phones on incomes is likely to be a long-term process. If cell phone coverage increases economic activity and economic growth in the long run, it may indirectly lower political violence in the long term. However, we find that given contextual factors which make conflict likely, in the short run, cell phones increase the propensity for violent events.

Finally, the effect of communication technology on other aspects of the political arena is still quite unclear and has not been studied widely. More research is needed on whether the availability of widespread communication between citizens decreases the likelihood of electoral fraud or government repression, as, for example, found by Aker, Collier, and Vincente (2011) and Bailard (2009). Can the possibility of private communication serve as a substitute for free and fair media and what are the effects across different political regimes? The increasing availability of spatially disaggregated data in combination with these data on cell phone coverage in Africa should allow us to answer a number of these questions in future projects.

REFERENCES

Abraham, Reuben. 2007. "Mobile Phones and Economic Development: Evidence from the Fishing Industry in India." *Information Technologies and International Development* 4 (1): 5–17.

Aday, Sean, Farrell Henry, Lynch Marc, Sides John, and Deen Freelon. 2012. *Blogs and Bullets II: New Media and Conflict after the Arab Spring*. Tech. rept. United States Institute of Peace.

Africa Partnership Program. 2008. "ICT in Africa: Boosting Economic Growth and Poverty Reduction." Working Paper. 10th Meeting of the Africa Partnership Forum. http://www.oecd.org/dataoecd/46/51/40314752.pdf.

Aker, Jenny C. 2010. "Information from Markets Near and Far: Mobile Phones and Agricultural Markets in Niger." *American Economic Journal: Applied Economics* 2: 46–59.

Aker, Jenny C., Paul Collier, and Pedro C. Vincente. 2011. "Is Information Power? Using Cell Phones during an Election in

Mozambique." Working Paper. http://www.pedrovicente.org/cell.pdf (Accessed April 22, 2012).
Aker, Jenny C., Christopher Ksoll, and Travis J. Lybbert. 2012. "Can Mobile Phones Improve Learning? Evidence from a Field Experiment in Niger." *American Economic Journal: Applied Economics* 4 (4): 94–120.
Aker, Jenny C., and Isaac M. Mbiti. 2010. "Mobile Phones and Economic Development in Africa." *Journal of Economic Perspectives* 24 (3): 207–32.
Alliy, Mbwana. 2011. *Visa Gets Serious: Let the Africa Mobile Payment Wars Begin*. Afrinnovator: Putting Africa on the Map. http://afrinnovator.com/blog/2011/11/20/visa-gets-serious-let-the-africa-mobile-payments-wars-begin/ (Accessed March 22, 2011).
Angrist, Joshua D., and Guido Imbens. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.
Angrist, Joshua D., and Joern-Steffen Pischke. 2009. *Mostly Harmless Econometrics*. Princton, NJ: Princeton University Press.
Bailard, Catie Snow. 2009. "Mobile Phone Diffusion and Corruption in Africa." *Political Communication* 26 (3): 333–53.
Beck, Neal, Kristian Skrede Gleditsch, and Kyle Beardsley. 2006. "Space Is More than Geography: Using Spatial Econometrics in the Study of Political Economy." *International Studies Quarterly* 50 (1): 27–44.
Bell, Andrew, and Kelvyn Jones. 2012 (May). "Explaining Fixed Effects: Random Effects Modelling of Time-Series Cross-Sectional and Panel Data." Working Paper.
Berman, Sheri. 1997. "Civil Society and the Collapse of the Weimar Republic." *World Politics* 49 (3): 401–29.
Berman, Eli, Joseph Felter, and Jacob N. Shapiro. 2009. "Do Working Men Rebel? Insurgency and Unemployment in Iraq and the Philippines." NBER Working Paper.
Blattman, Christopher. 2009. "From Violence to Voting: War and Political Participation in Uganda." *American Political Science Review* 103 (2): 231–47.
Blattman, Christopher, and Edward Miguel. 2010. "Civil War." *Journal of Economic Literature* 48 (1): 3–57.
Bradshaw, Tim. 2009. "Twitter Used by Protest Groups to Galvanise Forces." *Financial Times*. http://search.proquest.com/docview/250171039?accountid=10598 (Accessed September 15, 2012).
Brand, Jon. 2007. "Iraqi Insurgents Target Water and Electricity, But Spare the Cell Phone." *PBS Online NewsHour*. http://www.pbs.org/newshour/extra/features/jan-june07/infrastructure_1- 29.html (Accessed October 13, 2012).
Breuer, Anita, Todd Landman, and Dorothea Farquhar. 2012. "Social Media and Protest Mobilization: Evidence from the Tunisian Revolution." Paper prepared for the 4th European Communication Conference for the Euopean Communication Research and Education Conference (ECREA) 2012.
Buhaug, Halvard, Kristian Skrede Gleditsch, Helge Holtermann, Gudrun Østby, and Andreas Forø. 2011. "It's the Local Economy, Stupid! Geographic Wealth Dispersion and Conflict Outbreak Location." *Journal of Conflict Resolution* 55 (5): 814–40.
Buhaug, Halvard, and Jan Ketil Rød. 2006. "Local Determinants of African Civil Wars, 1970–2001." *Political Geography* 25 (3): 315–35.
Buys, Piet, Susmita Dasgupta, and Timothy Thomas. 2009. "Determinants of a Digital Divide in Sub-Saharan Africa: A Spatial Econometric Analysis of Cell Phone Coverage." *World Development* 37 (9): 1494–505.
Cederman, Lars-Erik, and Luc Girardin. 2007. "Beyond Fractionalization: Mapping Ethnicity onto Nationalist Insurgencies." *American Political Science Review* 101 (1): 173–85.
Cederman, Lars-Erik, Nils B. Weidmann, and Kristian Skrede Gleditsch. 2011. "Horizontal Inequalities and Ethnonationalist Civil War: A Global Comparison." *American Political Science Review* 105 (3): 478–95.
Cederman, Lars-Erik, Andreas Wimmer, and Brian Min. 2010. "Why Do Ethnic Groups Rebel? New Data and Analysis." *World Politics* 62 (1): 87–119.
Chassang, Sylvain, and Gerard Padro-i-Miguel. 2009. "Economic Shocks and Civil War." *Quarterly Journal of Political Science* 4 (3): 211–28.
CIESIN. 2005. "Center for International Earth Science Information Network (CIESIN), Columbia University; and Centro Internacional de Agricultura Tropical (CIAT)." *Gridded Population of the World Version 3 (GPWv3): Population Density Grids*. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. http://sedac.ciesin.columbia.edu/gpw (Accessed June 20, 2011).
Collier, Paul, V. L. Elliott, Harvard Hegre, Anke Hoeffler, Marta Reynal-Querol, and Nicolas Sambanis. 2003. *Breaking the Conflict Trap. Civil War and Development Policy*. Washington, DC and New York, NY: World Bank and Oxford University Press.
Collier, Paul, and Anke Hoeffler. 2004. "Greed and Grievance in Civil War." *Oxford Economic Papers* 56 (4): 563–95.
Collier, Paul, and Anke Hoeffler. 2007. "Civil War." *Handbook of Defense Economics*, Vol. 2, eds. Sandler Todd, and Keith Hartley. New York, NY: Elsevier B.V., 712–738.
Cordesman, Anthony H. 2005. "Iraq's Evolving Insurgency." Working Paper. Center for Strategic and International Studies.
Dal Bó, Ernesto, and Pedro Dal Bó. 2011. "Workers, Warriors and Criminals: Social Conflict in General Equilibrium." *Journal of the European Economic Association* 9 (4): 646–77.
Diamond, Larry. 2012. "Liberation Technology." *Liberation Technology: Social Media and the Struggle for Democracy*, eds. Diamond Larry, and Marc F. Plattner. Baltimore, MD: The Johns Hopkins University Press, 3–17.
Diamond, Larry, and Marc F. Plattner, eds. 2012. *Liberation Technology: Social Media and the Struggle for Democracy*. Baltimore, MD: Johns Hopkins University Press.
Donner, Jonathan, and Camilo Andres Tellez. 2008. "Mobile Banking and Economic Development: Linking Adoption, Impact, and Use." *Asian Journal of Communication* 18 (4): 332.
Dube, Oeindrila, and Juan Vargas. "Commodity Price Shocks and Civil Conflict: Evidence from Colombia." *Review of Economic Studies*. Forthcoming.
Earl, Jennifer, and Katrina Kimport. 2011. *Digitally Enabled Social Change: Activism in the Internet Age*. Cambridge, MA: MIT Press.
Earl, Jennifer, Andrew Martin, John D. McCarthy, and Sarah A. Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30: 65–80.
Eck, Kristine. 2012. "In Data We Trust? A Comparison of UCDP GED and ACLED Conflict Events Datasets." *Cooperation and Conflict* 47 (1): 124–41.
Edmond, Chris. 2012. "Information Maniplation, Coordination, and Regime Change." Working Paper.
Fearon, James D., Kimuli Kasara, and David D. Laitin. 2007. "Ethnicity Minority Rule and Civil War Onset." *American Political Science Review* 101 (1): 187–93.
Fearon, James D., and David D. Laitin. 2003. "Ethnicity, Insurgency, and Civil War." *American Political Science Review* 97 (1): 75–88.
Foster, Vivien, and Cecilia Briceño Garmendia, eds. 2010. "Information and Communication Technologies: A Boost for Growth." *Africa's Infrastructure: A Time for Transformation*. Washington, DC: The International Bank for Reconstruction and Development/The World Bank, 165–180.
Fowler, Geoffrey A. 2007. " 'Citizen Journalists' Evade Blackout on Myanmar News." *Wall Street Journal*. http://online.wsj.com/article/SB119090803430841433.html?mod=hps_us_page one (Accessed May 29, 2012).
Franzese, Robert J., and Jude C. Hays. 2008. "Interdependence in Comparative Politics: Substance, Theory, Empirics, Substance." *Comparative Political Studies* 41 (4/5): 742–80.
Gelman, Andrew, and Jennifer Hill. 2008. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.
Gilmore, Elisabeth, Nils Petter Gleditsch, Päivi Lujala, and Jan Ketil Rød. 2005. "Conflict Diamonds: A New Dataset." *Conflict Management and Peace Science* 22 (3): 257–92.
Greenhill, Brian, Michael D. Ward, and Audrey Sacks. 2011. "The Separation Plot: A New Visual Method for Evaluating the Fit of Binary Models." *American Journal of Political Science* 55 (4): 991–1002.
Grossman, Herschel I. 1991. "A General Equilirium Model of Insurrections." *American Economic Review* 81 (4): 912–21.
Grossman, Herschel I., and Minseong Kim. 1995. "Swords or Plowshares? A Theory of the Security of Claims to Property." *Journal of Political Economy* 103 (6): 1275–88.

Hays, Jude, and Robert Franzese. 2007. "Spatial-Econometric Models of Cross-Sectional Interdependence in Political Science Panel and TSCS Data." *Political Analysis* 15 (2): 140–64.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2011. "Multivariate Matching Methods that are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106 (493): 345–361.

Iacus, Stefano M., Gary King, and Giuseppe Porro. 2012. "Causal Inference Without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20 (1): 1–24.

Kalyvas, Stathis N. 2006. *The Logic of Violence in Civil Wars*. Cambridge, MA: Cambridge University Press.

Kalyvas, Stathis N., and Matthew Adam Kocher. 2007. "How 'Free' is Free Riding in Civil Wars? Violence, Insurgency, and the Collective Action Problem." *World Politics* 59 (2): 177–216.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 347–61.

Kuran, Timur. 1991. "Now Out of Never: The Element of Surprise in the East European Revolution of 1989." *World Politics* 44 (1): 7–48.

Lapper, Richard. 2010 (January 29). *Youthful Protesters Help Shape New Kind of Politics*. Financial Times. http://search.proquest.com/docview/250264463?accountid=10598 (Accessed May 29, 2012).

Leahy, Kevin C. 2005. "The Impact of Technology on the Command, Control, and Organizational Structure of Insurgent Groups." Masters thesis. U.S. Army Command and General Staff College.

LeSage, James, and R. Kelley Pace. 2009. *Introduction to Spatial Econometrics*. Boca Raton, FL: CRC Press.

Livingston, Steven. 2011. "Africa's Evolving Infosystems: A Pathway to Security and Stability." Africa Center for Strategic Studies Research Paper No. 2.

Lohmann, Susanne. 1994. "Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-1991." *World Politics* 47: 42–101.

Lujala, Päivi, Halvard Buhaug, and Scott Gates. 2009. "Geography, Rebel Capability, and the Duration of Civil Conflict." *Journal of Conflict Resolution* 53 (4): 544–69.

Lujala, Päivi, Nils Petter Gleditsch, and Elisabeth Gilmore. 2005. "A Diamond Curse? Civil War and a Lootable Resource." *Journal of Conflict Resolution* 49 (4): 538–62.

Lujala, Päivi, Jan Ketil Rød, and Nadia Thieme. 2007. "Fighting Over Oil: Introducing a New Dataset." *Conflict Management and Peace Science* 24 (3): 239–56.

Mattes, Robert, Michael Bratton, Yul Derek Davids, and Cherrel Africa. 2010. *Afrobarometer: Round IV 2008*. Dataset.

Melander, Erik, and Ralph Sundberg. 2011. "Climate Change, Environmental Stress, and Violent Conflict—Test Introducing the UCDP Georeferenced Event Dataset." Paper presented at the International Studies Association, March 16–19, Montreal, Canada.

Metternich, Nils W. 2011. "Expecting Elections: Interventions, Ethnic Support, and the Duration of Civil Wars." *Journal of Conflict Resolution* 55 (6): 909–37.

Mutua, Will. 2011. "Startup Watch: Esoko Ltd: Powering an Agric Revolution." *Afrinnovator: Putting Africa on the Map*. http://afrinnovator.com/blog/2011/08/13/startup-watch-esoko-ltd-powerin g-an-agric-revolution/ (Accessed April 22, 2012).

Neumayer, Eric, and Thomas Plümper. 2012. "Conditional Spatial Policy Dependence: Theory and Model Specification." *Comparative Political Studies* 45 (7): 819–49.

Nordhaus, William D. 2006. "Geography and Macroeconomics: New Data and New Findings." *Proceedings of the National Academy of Sciences of the USA* 103 (10): 3510–17.

Olson, Mancur. 1965. *The Logic of Collective Action. Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Peterson, Scott. 2012. "Syria's iPhone Insurgency Makes for Smarter Rebellion." *The Christian Science Monitor*. http://www.csmonitor.com/World/Middle-East/2012/0801/Syria-s-iPhone-ins urgency-makes-for-smarter-rebellion (Accessed November 13, 2012).

Philippine Daily Inquirer. 2001. "Text Messages Bring Political Protests to Cyberspace." *Philippine Daily Inquirer* (January) http://news.google.com/newspapers?id=PX42AAAAIBAJ&sjid=hCUMAAAAIBAJ&dq=political 2012).

Quandzie, Ekow. 2011. "Africa Leads in Mobile Money Deployment as Users Hit Over 40 Million." *Ghana Business News*. http://www.ghanabusinessnews.com/2011/10/30/africa-leads-in-mobile-money-deployment-as-users-hits-over-40-million/ (Accessed April 22, 2012).

Raleigh, Clionadh, Andrew Linke, Havard Hegre, and Joakim Karlsen. 2010. "Introducing ACLED-Armed Conflict Location and Event Data." *Journal of Peace Research* 47 (5): 1–10.

Reno, William. 2011. *Warfare in Independent Africa*. Cambridge, UK: Cambridge University Press.

Ross, Michael L. 2006. "A Closer Look At Oil, Diamonds, and Civil War." *Annual Review of Political Science* 9: 265–300.

Sacks, Audrey, and Marco Larizza. 2012 (February). "Why Quality Matters: A Multi-Level Analysis of Decentralization, Local Government Performance and Trustworthy Government in Post-Conflict Sierra Leone." Working Paper.

Salehyan, Idean, Cullen S. Hendrix, Christina Case, Christopher Linebarger, Emily Stull, and Jennifer Williams. 2012. "The Social Conflict in Africa Database: New Data and Applications." *International Interactions* 38 (4): 503–11. www.scaddata.org (Accessed June 25, 2012).

Sambanis, Nicholas. 2002. "A Review of Recent Advances And Future Directions in the Quantitative Literature on Civil War." *Defence and Peace Economics* 13 (3): 215–43.

Shadmehr, Mehdi, and Dan Bernhardt. 2011. "Collective Action with Uncertain Payoffs: Coordination, Public Signals, and Punishment Dilemmas." *American Political Science Review* 105 (4): 829–51.

Shapiro, Jacob N., and Nils B. Weidmann. 2012. "Is the Phone Mightier than the Sword? Cell Phones and Insurgent Violence in Iraq." Working Paper. www.princeton.edu/~jns/papers/SW_2011_Cell_Phones_Insurgency_06SEP12.pdf (Accessed September 20, 2012).

Sherwood, Bob. 2011 (August 10). "Police Face Mobile Network Spreading its Angry Message." *Financial Times*. http://search.proquest.com/docview/882342765?accountid=10598 (Accessed May 29, 2012).

Shirky, Clay. 2008. *Here Comes Everybody: The Power of Organizing Without Organizations*. London, UK: Allen Lane, Penguin Group.

Siebert, Stefan, Petra Döll, Jippe Hoogeveen, J-M Frenken, Karen Frenken, and Sebastian Feick. 2007. "Development and Validation of the Global Map of Irrigation Areas." *Hydrology and Earth System Sciences* 9 (5): 535–47.

Stroher, Tiffany. 2007. "Cell Phone Use by Insurgents in Iraq." Working Paper. Urban Warfare Analysis Center.

Sundberg, Ralph, Mathilda Lindgren, and Ausra Padskocimaite. 2011. *UCDP GED Codebook version 1.0-2011*. Codebook. Department of Peace and Conflict Research, Uppsala University. http://ucdp.uu.se/ged/data/ucdp_ged_v.1.0-codebook.pdf (Accessed January 20, 2012).

The Economist 2012. "Business this Week." *The Economist*. http://www.economist.com/node/21538210 (Accessed March 24, 2012).

The World Bank 2011. *The World Development Report 2011: Conflict, Security, and Development* Washington, DC: The World Bank Group.

Tollefsen, Andrea Forø, Havard Strand, and Halvard Buhaug. 2012. "PRIO-GRID: A Unified Spatial Data Structure." *Journal of Peace Research* 49 (2): 363–74.

Tomz, Mike, Gary King, and Langche Zeng. 2003. "ReLogit: Rare Events Logistic Regression." *Journal of Statistical Software* 8 (2): 246–47.

UNEP-WCMC World Conservation Monitoring Centre. 2002. *Mountain Watch 2002*.http://www.unep-wcmc.org/mountains/mountain_watch/pdfs/WholeReport.pdf (Accessed September 15, 2012).

Ward, Michael D., and Kristian Skrede Gleditsch. 2002. "Location, Location, Location: An MCMC Approach to Modeling the Spatial Context of War." *Political Analysis* 10 (3): 244–60.

Warren, T. Camber. 2013. "Not by the Sword Alone: Soft Power, Mass Media, and the Production of State Sovereignty." *International Organization*. Forthcoming.

Weidmann, Nils B. 2009. "Geography as Motivation and Opportunity: Group Concentration and Ethnic Conflict." *Journal of Conflict Resolution* 53 (4): 526–43.

Weidmann, Nils B., Jan Ketil Rød, and Lars-Erik Cederman. 2010. "Representing Ethnic Groups in Space: A New Dataset." *Journal of Peace Research* 47 (4): 491–99.
Weidmann, Nils B., and Michael D. Ward. 2010. "Predicting Conflict in Space and Time." *Journal of Conflict Resolution* 54 (6): 883–901.
Weinstein, Jeremy M. 2007. *Inside Rebellion: The Politics of Insurgent Violence*. New York: Cambridge University Press.
Wood, Elisabeth J. 2003. *Insurgent Collective Action and Civil War in El Salvador*. New York: Cambridge University Press.
Wucherpfennig, Julian, Nils W. Metternich, Lars-Erik Cederman, and Kristian Skrede Gleditsch. 2012. "Ethnicity, the State, and the Duration of Civil War." *World Politics* 64 (1): 79–115.
Wucherpfennig, Julian, Nils B. Weidmann, Luc Girardin, Lars-Erik Cederman, and Andreas Wimmer. 2011. "Politically Relevant Ethnic Groups Across Space and Time: Introducing the GeoEPR Dataset." *Conflict Management and Peace Science* 10 (10): 1–15.
Yanagizawa-Drott, David. 2012. "Propaganda and Conflict: Theory and Evidence From the Rwandan Genocide." Working Paper.

Advance Access publication March 22, 2012

Political Analysis (2012) 20:271–291
doi:10.1093/pan/mps002

Improving Predictions Using Ensemble Bayesian Model Averaging

Jacob M. Montgomery
Department of Political Science, Washington University in St Louis, Campus Box 1063, One Brookings Drive, St Louis, MO 63130-4899

Florian M. Hollenbach and Michael D. Ward
Department of Political Science, Duke University, Perkins Hall 326, Box 90204, Durham, NC 27707-4330
e-mail: michael.d.ward@duke.edu (corresponding author)

Edited by R. Michael Alvarez

We present ensemble Bayesian model averaging (EBMA) and illustrate its ability to aid scholars in the social sciences to make more accurate forecasts of future events. In essence, EBMA improves prediction by pooling information from multiple forecast models to generate ensemble predictions similar to a weighted average of component forecasts. The weight assigned to each forecast is calibrated via its performance in some validation period. The aim is not to choose some ''best'' model, but rather to incorporate the insights and knowledge implicit in various forecasting efforts via statistical postprocessing. After presenting the method, we show that EBMA increases the accuracy of out-of-sample forecasts relative to component models in three applied examples: predicting the occurrence of insurgencies around the Pacific Rim, forecasting vote shares in U.S. presidential elections, and predicting the votes of U.S. Supreme Court Justices.

1 Introduction

Testing systematic predictions about future events against observed outcomes is generally seen as the most stringent validity check of statistical and theoretical models. Yet, political scientists rarely make predictions about the future. Empirical models are seldom applied to out-of-sample data and are even more rarely used to make predictions about future outcomes. Instead, researchers typically focus on developing and validating theories that explain past events.

In part, this lack of emphasis on forecasting results from the fact that it is so difficult to make accurate predictions about complex social phenomena. However, research in political science could gain immensely in its policy relevance if predictions were more common and more accurate. Improved forecasting of important political events would make research more germane to policymakers and the general public who may be less interested in explaining the past than anticipating and altering the future. From a scientific standpoint, greater attention to forecasting would facilitate stringent validation of theoretical and statistical models since truly causal models should perform better in out-of-sample forecasting.

In this article, we extend a promising statistical method—ensemble Bayesian model averaging (EBMA)—and introduce software that will aid researchers across disciplines in making more accurate forecasts. In essence, EBMA makes more accurate predictions possible by pooling information from multiple forecast models to generate ensemble predictions similar to a weighted average of component forecasts. The weight assigned to each forecast is calibrated via its performance in some prior period. These component models can be diverse. They need not share covariates, functional forms, or error structures. Indeed, the components may not even be statistical models but may be predictions generated by agent-based models or subject-matter experts.

Authors' note: For generously sharing their data and models with us, we thank Alan Abramowitz, James Campbell, Robert Erikson, Ray Fair, Douglas Hibbs, Michael Lewis-Beck, Andrew D. Martin, Kevin Quinn, Stephen Shellman, Charles Tien, and Christopher Wlezien. We especially want to thank Adrian Raftery and Brendan Nyhan for their encouragement and feedback as this project evolved. The editor and the reviewers of *Political Analysis* provided especially salient suggestions that substantially improved our research.

In the rest of this article, we briefly review existing political science research aimed at forecasting and then present the mathematical details of the EBMA method. We then illustrate the benefits of EBMA by applying it to predict insurgency events on the Pacific Rim, U.S. presidential elections, and voting on the U.S. Supreme Court.

2 Dynamic Forecasting in Political Science

Although forecasting is a rare exercise in political science, there are an increasing number of exceptions. In most cases, ''forecasts'' are conceptualized as an exercise in which the predicted values of a dependent variable are calculated based on a specific statistical model and then compared with observed values (e.g., Hildebrand, Liang, and Rosenthal 1976). In many instances, this reduces to an analysis of residuals. In others, the focus is on randomly selecting subsets of the data to be excluded during model development for cross-validation. However, there is also a more limited tradition of making true forecasts about events that have not yet occurred. (See Brandt, Freeman, and Schrodt 2011a for a recent and thorough survey of forecasts in political science and economics with a focus on strategies to perform more meticulous comparisons of their accuracy.)

An early proponent of using statistical models to make predictions in the realm of international relations (IR) was Stephen Andriole (Andriole and Young 1977). In 1978, a volume edited by Nazli Choucri and Thomas Robinson provided an overview of the then current work in forecasting in IR. Much of this work was done in the context of policy-oriented research for the U.S. government during the Vietnam War. Subsequently, there were a variety of efforts to create or evaluate forecasts of international conflict, including Freeman and Job (1979), Singer and Wallace (1979), and Vincent (1980). In addition, a few efforts began to generate forecasts of domestic conflict (e.g., Gurr and Lichbach 1986). Recent years, however, have witnessed increasing interest in prediction across a wide array of contexts in IR.[1] The 2011 special issue of *Conflict Management and Peace Science* on prediction in the field of IR exemplifies this growing emphasis on forecasting (cf. Bueno de Mesquita 2011; Brandt, Freeman, and Schrodt 2011b; Schneider, Gleditsch, and Carey 2011). Ward, Greenhill, and Bakke (2010) and Greenhill, Ward, and Sacks (2011) provide additional discussion of forecasting in IR.

Outside of IR, forecasting in political science has largely taken place in the context of election research. In the 1960s, de Sola Pool, Abelson, and Popkin (1964) published a volume describing their work on the 1960 and 1964 presidential elections. They reported their efforts to use a computer simulation to predict election outcomes, which was initially undertaken in the context of providing campaign management advice for the 1960 campaign of John F. Kennedy. Rosenstone (1983) published perhaps the most influential early work on elections forecasting, which surveyed the then state-of-the-art and included examples going back to 1932.

In the 1990s, political scientists renewed their interest in predicting presidential elections (Campbell and Wink 1990; Campbell 1992). This work was anticipated by the efforts of several economists, most notably the forecast established by Fair (1978). As we discuss below, predicting U.S. presidential and congressional elections has since developed into a regular exercise. Moreover, researchers have begun to forecast election outcomes in France (e.g., Jerome, Jerome, and Lewis-Beck 1999) and the United Kingdom (e.g., Whiteley 2005).[2]

Although efforts to predict future outcomes remain uncommon, research that combines multiple forecasts is nearly nonexistent. To our knowledge, the only non-IR examples are the PollyVote project (cf. Graefe et al. 2010), which combines multiple predictions using simple averages of forecasts to predict U.S. presidential elections, and Lock and Gelman (2010), who use Bayesian methodology to combine information from historical state-level election returns, current polling data, and forecasting models to generate election forecasts.

Yet, methods for combining forecasts, and ensemble models in particular, have been shown to substantially reduce prediction error in two important ways. First, across subject domains, ensemble predictions are

[1] An incomplete list of recent work would include Krause (1997), Davies and Gurr (1998), Pevehouse and Goldstein (1999), Schrodt and Gerner (2000), King and Zeng (2001), O'Brien (2002), Bueno de Mesquita (2002), Fearon and Laitin (2003), de Marchi, Gelpi, and Grynaviski (2004), Enders and Sandler (2005), Leblang and Satyanath (2006), Ward, Siverson, and Cao (2007), Brandt, Colaresi, and Freeman (2008), Bennett and Stam (2009), and Gleditsch and Ward (2010). A summary of classified efforts is reported in Feder (2002). An overview of some of the historical efforts, along with a description of current thinking about forecasting and decision support, is given by O'Brien (2010).

[2] Lewis-Beck (2005) provides a more in-depth discussion of election forecasting in a comparative context.

usually more accurate than any individual component model. Second, they are significantly less likely to be dramatically incorrect (Bates and Granger 1969; Armstrong 2001; Raftery et al. 2005).[3] Combining forecasts not only reduces reliance on single data sources and methodologies (which lowers the likelihood of dramatic errors) but also allows for the incorporation of more information than any one model is likely to include in isolation.

The idea of ensemble learning itself has a long history in the machine learning community. The most thorough treatment is found in Hastie, Tibshirani, and Friedman (2009). A wide range of statistical approaches including bagging, random forests, as well as boosting and penalized methods may be properly considered ensemble approaches. They are different from EBMA, however, which comes from another branch on the ensemble family tree—Bayesian statistics. Bayesian methods themselves can generally be viewed as ensemble methods since they produce a large number of candidate ''models'' that are averaged to create a posterior distribution of parameters (Hastie, Tibshirani, and Friedman 2009, 605).

These advances in the statistical literature parallel additional research in formal theory, which shows that groups of agents using diverse decision rules or composed of agents with different viewpoints on a problem can produce superior outcomes in difficult decision environments (Page, Sander, and Schneider-Mizell 2007; Page 2008, 2011). That is, social systems, organizations, and institutions that are better able to combine insights and knowledge from diverse actors are more functional, successful, and adaptive in complex environments.

This last strain of thought is related to research that suggests the use of prediction markets as a method of aggregating a large number of individual predictions about particular events. For example, Berg, Nelson, and Rietz (2008) discuss prediction markets and demonstrate that they can be more accurate than polls when forecasting elections. One important prediction market in political science is the Iowa Electronic Market, in which individuals buy futures on politicians which are paid after election results are revealed.

3 Ensemble Bayesian Model Averaging

Predictive models remain underutilized, yet an increasing number of scholars have developed forecasting models for specific research domains. As the number of forecasting efforts proliferates, however, there is a growing benefit from developing methods to pool across models and methodologies to generate more accurate forecasts. Very often, specific predictive models prove to be correct only for certain subsets of observations. Moreover, specific models tend to be more sensitive to unusual events or particular data issues than ensemble methods.

To aid the newfound emphasis on prediction in political science, we are advancing recent statistical research aimed at integrating multiple predictions into a single improved forecast. In particular, we are adapting an ensemble method first developed for application to the most mature prediction models in existence—weather forecasting models. To generate predictive distributions of outcomes (e.g., temperature), weather researchers apply ensemble methods to forecasts generated from multiple models (Raftery et al. 2005). Thus, state-of-the-art ensemble forecasts aggregate multiple runs of (often multiple) weather prediction models into a single unified forecast.

The particular ensemble method we are extending for application to political outcomes is ensemble Bayesian model averaging (EBMA). First proposed by Raftery et al. (2005), EBMA pools across various forecasts while meaningfully incorporating *a priori* uncertainty about the ''best'' model. It assumes that no particular model or forecasting method can fully encapsulate the true data-generating process. Rather, various research teams or statistical techniques will reflect different facets of reality. EBMA collects all the insights from multiple forecasting efforts in a coherent manner. The aim is not to choose some best model, but rather to incorporate the insights and knowledge implicit in various forecasting efforts via statistical postprocessing. In recent years, variants of the EBMA method have been applied to subjects as diverse as inflation (Wright 2009; Koop and Korobilis 2009; Gneiting and Thorarinsdottir 2010), stock prices (Billio et al. 2011), economic growth and policymaking (Brock, Durlauf, and West 2007; Billio et al. 2010), exchange rates (Wright 2008), industrial production (Feldkircher 2012), ice formation (Berrocal et al. 2010), visibility (Chmielecki and Raftery 2010), water catchment streamflow (Huisman et al. 2009), climatology (Min and Hense 2006;

[3]The case for using predictions heuristically can also be found in early work by Dawid (1982, 1984).

Min, Simonis, and Hense 2007; Smith et al. 2009), and hydrology (Zhang, Srinivasan, and Bosch 2009). Indeed, research is underway to extend the method to handle missing data (Fraley, Raftery, and Gneiting 2010; McCandless, Haupt, and Young 2011) as well as calibrate model weights on nonlikelihood criteria (e.g., Vrugt et al. 2006).

3.1 *Overview of Method*

EBMA is designed for application in the context of a subject domain with ongoing forecasting efforts. That is, it assumes the existence of multiple teams or individuals making regular predictions about a common set of outcomes. For example, there may be multiple analysts or teams making predictions about the likelihood of violent conflict in specific regions of the world, quarterly economic growth for the United States, or the votes of members on bills before Congress. As we show in our examples below, these predictions might originate from the insights and intuitions of individual subject experts, traditional statistical models, nonlinear classification trees, neural networks, agent-based models, or anything in between.

EBMA is a method for taking the predictions made by multiple teams and combining them—based on their past performance and uniqueness—to create a new ensemble forecasting model. This ensemble model can then make predictions about unobserved outcomes in the future and usually outperforms its components. Roughly speaking, it generates forecasts by creating weighted averages of component predictions or component predictive probability distribution functions (PDFs). The weight assigned to each component forecast, denoted w_k below, reflects two aspects of the components' past forecasts. First, *ceteris paribus*, the EBMA model will give greater weight to forecasts that were more accurate in the past. Second, *ceteris paribus*, it will assign a greater weight to models that made unique (but correct) predictions. That is, component forecasts that are too highly correlated may jointly have a large weight but will individually be penalized.

There are two important aspects of EBMA that distinguish it from the alternative model selection or averaging methods referenced above. First, EBMA is more flexible in not requiring any information about the actual covariates that go into the component models. A second, and related, point is that EBMA does not require researchers to develop metrics to penalize component forecasts for the number of parameters included, the number of covariates, or their complexity more generally. In the case of subject-expert opinions, there may not even be any covariates or statistical models involved, a point we return to in the Supreme Court example below. Another nonstatistical component model that researchers might include is prices on prediction markets. In other instances, predictions may come from models whose "complexity" is not easily defined or enumerated (e.g., agent-based models). Of course, overly complex "garbage can" models will generally perform poorly when making predictions over any outcome and therefore receive a lower weight in the ensemble forecast. Yet, this lower weight is a function of the model's predictive performance rather than a prespecified preference for parsimony. The upshot is that EBMA forecasts will implicitly penalize overfitting of component models since the weight assigned to component models is based on predictive performance. However, it can do so without explicitly penalizing components for complexity.[4]

3.2 *Mathematical Foundations*

EBMA itself is an extension of the Bayesian model averaging (BMA) methodology (cf. Madigan and Raftery 1994; Draper 1995; Raftery 1995; Hoeting et al. 1999; Clyde 2003; Raftery and Zheng 2003; Clyde and George 2004) that has received considerable attention in the field of statistics. BMA was first introduced to political science by Bartels (1997) and has been applied in a number of contexts (e.g., Bartels

[4]To fully capture the ability of EBMA to reduce overfitting when using statistical models, it is necessary to divide the data into three periods (Hastie, Tibshirani, and Friedman 2009). The first period, the training period, is used to fit the parameters for each component model. The second period, the validation period, is used to calculate model weights and other parameters for the EBMA model using out-of-sample predictions generated from the component models. We then generate ensemble predictions for the third period, the test period, using the EBMA model parameters calculated in period two. This approach is explicit in the insurgency forecasting example below and implicit in the Supreme Court example below since the subject experts and classification algorithm were "trained" on data not included in the study. This three-stage method is adjusted in the election forecasting example as the component models are already sparse (somewhat ameliorating concerns about overfitting), and there are far fewer observations. In this example, component models are trained over the period beginning in 1916 and the EBMA parameters are calculated only for the period beginning in 1952. However, there is significant overlap in the training and validation samples.

and Zaller 2001; Gill 2004; Imai and King 2004; Geer and Lau 2006). Montgomery and Nyhan (2010) provide a more in-depth discussion of BMA and its applications in political science.

Assume we have some quantity of interest to forecast, $\mathbf{y}^{t^*}$, in some future period $t^* \in T^*$. Further assume that we have extant forecasts for events $\mathbf{y}^t$ for some past period $t \in T$ that were generated from K forecasting models or teams, $M_1, M_2, \ldots, M_K$. Each model, M_k, is assumed to come from the prior probability distribution $M_k \sim \pi(M_k)$, and the PDF for $\mathbf{y}^t$ is $p(\mathbf{y}^t|M_k)$. The outcome of interest is distributed $p\left(\mathbf{y}^{t^*}|M_k\right)$. Applying Bayes's rule, we get that

$$p\left(M_k|\mathbf{y}^t\right) = \frac{p(\mathbf{y}^t|M_k)\pi(M_k)}{\sum_{k=1}^{K} p(\mathbf{y}^t|M_k)\pi\left(M_k\right)} \tag{1}$$

and the marginal predictive PDF is

$$p\left(\mathbf{y}^{t^*}\right) = \sum_{k=1}^{K} p\left(\mathbf{y}^{t^*}|M_k\right)p(M_k|\mathbf{y}^t). \tag{2}$$

The BMA PDF (equation (2)) can be viewed as the weighted average of the component PDFs where the weights are determined by each model's performance within the already observed period T. Likewise, we can simply make a deterministic estimate using the weighted predictions of the components, denoted $E\left(\mathbf{y}^{t^*}\right) = \sum_{k=1}^{K} E\left(\mathbf{y}^{t^*}|M_k\right)p(M_k|\mathbf{y}^t)$.

3.2.1 EBMA for dynamic settings

In generating predictions of future events, the task is to first build a set of statistical models for some set of observations S in the past time periods T', which we refer to as the training period. Using these models, we then generate predictions $\mathbf{f}_k^{s|t}$ for some period T, which has already occurred but which was not included in the training sample. We refer to this as the validation period, and we will use these data to calibrate the EBMA model.[5] Finally, using the same K models, we assume that there are true forecasts ($\mathbf{f}_k^{s|t^*}$) for observations $s \in S$ in future time periods $t^* \in T^*$.[6] We either (a) treat these raw predictions as a component model in the steps below or (b) statistically postprocess the predictions for out-of-sample bias reduction and treat these recalibrated predictions as a component model.

As a running example, let us assume that we have K forecasting efforts for modeling insurgencies in a set of countries S ongoing throughout the training (T') validation (T) and test (T^*) periods. We will associate each component forecast with a component PDF, $g_k(\mathbf{y}|\mathbf{f}_k^{s|t,t^*})$, which may be the original prediction from the forecast model or the bias-corrected forecast.

The EBMA PDF is then a finite mixture of the K component PDFs, denoted

$$p(\mathbf{y}|\mathbf{f}_1^{s|t}, \ldots, \mathbf{f}_K^{s|t}) = \sum_{k=1}^{K} w_k g_k(\mathbf{y}|\mathbf{f}_k^{s|t}). \tag{3}$$

The w_k's $\in [0, 1]$ are model probabilities and $\sum_{k=1}^{K} w_k = 1$. Roughly speaking, they are associated with each component model's predictive performance in the validation period controlling for the degree to which they offer unique insight (i.e., a model's predictions are distinct from those of other component models). We provide additional discussion about the model weights in the election forecasting example below.

[5] In the case of subject experts, the training period is implicitly the period over which experts have gained their experience. Forecasts will only be necessary for the validation period.

[6] Sloughter et al. (2007) make predictions for only one future time period and use only a subset of past time periods (they recommend 30) in their validation period. Thus, predictions are made sequentially with the entire EBMA procedure being recalculated for each future event as observations are moved from the test period into the validation period T. Another alternative is to simply divide all the data into discrete training, validation, and test periods for the entire procedure. We use both approaches in our examples below.

Details for parameter estimation are provided in the Appendix. The ensemble PDF for an insurgency in the test period t^* in country s is then

$$p\left(y|f_1^{s|t^*},\ldots,f_K^{s|t^*}\right) = \sum_{k=1}^{K} w_k g_k\left(y|f_k^{s|t^*}\right). \tag{4}$$

3.2.2 EBMA for normally distributed outcomes

To gain a fuller understanding of the EBMA method, it is easiest to imagine an effort to predict some normally distributed outcome. When forecasting outcomes that are distributed normally, Raftery et al. (2005) propose approximating the conditional PDF as a normal distribution centered at a linear transformation of the individual forecast, $g_k(\mathbf{y}|\mathbf{f}_k^{s|t}) = N(a_{k0}+a_{k1}\mathbf{f}_k^{s|t}, \sigma^2)$. Prior applications have found that this adjustment of the component models' forecasts reduces overfitting and improves the performance of the final ensemble forecasting model (Raftery et al. 2005).[7] Using equations (3) and (4) above, the EBMA PDF is then

$$p(\mathbf{y}|\mathbf{f}_1^{s|t},\ldots,\mathbf{f}_K^{s|t}) = \sum_{k=1}^{K} w_k N\left(ak_0+ak_1\mathbf{f}_k^{s|t}, \sigma^2\right), \tag{5}$$

and the predictive distribution for some observation y is

$$p\left(y|f_1^{s|t^*},\ldots,f_K^{s|t^*}\right) = \sum_{k=1}^{K} w_k N\left(ak_0+ak_1 f_k^{s|t^*}, \sigma^2\right). \tag{6}$$

Thus, the predictive PDF is a mixture of K normal distributions each of whose mean is determined by the component prediction ($f_k^{s|t^*}$) and whose "height" (i.e, the total area under the curve for component k) is determined by the model weight $w_{k.}$

4 Empirical Applications

In this section, we provide empirical applications of EBMA to predict insurgency in the Pacific Rim, presidential election outcomes in the United States, and votes of Justices of the United States Supreme Court. These three examples demonstrate the usefulness of the method for diverse domains of political science research, different types of outcomes of interest (i.e., dichotomous and continuous), and different forms of component models (i.e., statistical models versus expert predictions).

4.1 *Application to Insurgency Forecasting*

Our first example applies the EBMA method to data collected for the Integrated Crisis Early Warning Systems (ICEWS) project sponsored by the Defense Advanced Research Projects Agency (DARPA). The task of the ICEWS project is to train models on data (focusing on five outcomes of interest) for 29 countries for every month from 1997 through the present and to then make accurate predictions about expected crisis events.[8] For purposes of demonstration, we focus on only one of these outcomes—violent insurgency.

The bulk of the data for the ICEWS project is gleaned from natural language processing of a continuously updated harvest of news stories (primarily taken from Lexus/Nexus and Factiva archives). These are digested with a version of the TABARI processor for events developed by Philip Schrodt and colleagues in the context of the Event Data Project (see http://eventdata.psu.edu/ for more details). These data are augmented with a variety of covariates, including country-level attributes (coded on a monthly or

[7] Our adjustments to the basic EBMA method for application to dichotomous outcomes, as well as details of parameter estimation, are shown in the Appendix.

[8] The twenty-nine countries are Australia, Bangladesh, Bhutan, Cambodia, China, Comoros, Fiji, India, Indonesia, Japan, Laos, Madagascar, Malaysia, Mauritius, Mongolia, Myanmar, Nepal, New Zealand, North Korea, Papua New Guinea, Philippines, Russia, Singapore, Solomon Islands, South Korea, Sri Lanka, Taiwan, Thailand, and Vietnam. This set is not a random sample but rather constitutes the countries of population greater than 500,000 that are in the area of responsibility of the U.S. Pacific Command.

yearly basis) from the Polity and World Bank data sets, information about election cycles (if any), events in neighboring countries, and the length of shared borders with neighboring countries.

4.1.1 Component models

We apply EBMA to make predictions for the occurrence of insurgency in these 29 countries for each month in the year 2010 (the last year in the data set). As a first step in the process, we must choose some set of observations that we wish to treat as a training period for the component statistical models and a second set of data to treat as a validation set with which to calibrate the EBMA model.

Unfortunately, there is no clear guidance available on how to choose the relative sizes of the training set, the validation set, or the test set. Hastie, Tibshirani, and Friedman (2009, chap. 7) discuss this in some detail, wherein they note (p. 195): ''It is difficult to give a general rule on how to choose the number of observations in each of the . . . parts, as this depends on the signal-to-noise ratio in the data and the training sample size.'' The basic point is that there is no general rule. Moreover, the appropriate size of the training and test set depends on the prevalence of the signal in the training data. In the case of predictive studies like ours, the only general rule is: it depends.

In this case, we estimated three exemplar statistical models using data for the *training period* ranging from January 1999[9] to December 2007 and fit an EBMA model using the component model predictions for the *validation period* ranging from January 2008 to December 2009. We then make forecasts for the test period ranging from January 2010 to December 2010 using both the component and EBMA models. To provide variation in the complexity (as well as accuracy) of the components, we included the following models.

- Strategic Analysis Enterprise (SAE): This is one model developed as part of the ICEWS project and was designed by Strategic Analysis Enterprises. It is specified as a simple generalized linear logistic model including 27 different independent variables.[10] All the variables are taken from the ICEWS event-stream data.
- Generalized Linear Model (GLM): For the purposes of demonstrating the properties of the EBMA method, we estimated a crude logistic model that includes only *population size*, *GDP growth* (both lagged 3 months), the number of *minority groups at risk* in the country, and a measure of *anocracy* supplied in the *Polity IV* data set (Marshall, Jaggers, and Gurr 2009).
- Linear Mixed Effects Regression (LMER): This is a generalized linear mixed effects model using a logistic link function and including a random effects term for lagged *GDP per capita* and the lagged number of *conflictual events involving the United States* in the country of interest.[11] The list of additional covariates includes the number of *conflictual events involving the military* within the country of interest (lagged three months), the number of *days elapsed since the last election*,[12] the number of *new insurgencies* that began in the previous 2 years, and a *spatial lag* that reflects recent occurrences of domestic crises in the countries' geographic neighbors.[13]

4.1.2 Results

Table 1 shows the EBMA model parameters as well as fit statistics associated with the individual component models and the EBMA predictions for the validation time period (2008–09). The first column shows the weights that the EBMA model assigned to each component. As can be seen, the GLM model is effectively excluded, whereas the LMER model carries the greatest weight ($w_k = 0.85$) followed by the SAE model ($w_k = 0.14$). The constant term associated with each component corresponds to the term a_{k0} in equation (8), whereas the predictor corresponds to a_{k1}. The other columns in Table 1 are fit statistics. AUC

[9]Because some of the models include lagged data, this is the first year for which all the component models produce fitted values or predictions.
[10]See strategicanalysisenterprises.com for more details.
[11]It is worth noting that the mixed effects model is a kind of ensemble mixture in that it averages the so-called within model with the between model.
[12]This is calculated as the number of days between the middle of the current month and the last federal election regardless of the legitimacy of the election.
[13]Geographical proximity is measured in terms of the length of the shared border between the two countries.

Table 1 Validation period results (2008–09)

	Weight	*Constant*	*Predictor*	*AUC*	*PRE*	*Brier*	*% Correct*
LMER	0.85	−1.89	2.58	0.97	−0.58	0.08	87.07
SAE	0.14	−1.25	3.11	0.92	−0.21	0.07	90.09
GLM	0.00	−1.76	1.42	0.66	0.00	0.08	91.81
EBMA				0.96	0.65	0.04	97.13
$n = 696$							

Note. The table shows estimated model weights, parameters, and fit statistics for the EBMA deterministic forecast and all component forecasts of insurgency in 29 countries of the U.S. Pacific Command. EBMA outperforms any single model on most measures.

is the area under the receiver-operating characteristic (ROC) curve. The advantage of using ROC curves is that it evaluates forecasts in a way that is less dependent on an arbitrary cutoff point. A value of 1 would mean that all observations were predicted correctly at all possible cutoff points (King and Zeng 2001).

We compare the models using three additional metrics. The proportional reduction in error (PRE) is the percentage increase of correctly predicted observations relative to some predefined base model. In this case, the base model is predicting ''no insurgencies'' for all observations. Insurgencies are relatively rare events. Thus, predicting a zero for all observations leads to a 91.8% correct prediction rate. The Brier score is the average squared deviation of the predicted probability from the true event (0 or 1). Thus, a lower score corresponds to higher forecast accuracy (Brier 1950). Finally, we calculate the percentage of observations that each model would predict correctly using a 0.5 threshold on the predicted probability scale.

Note that the EBMA model does at least as well (and usually better) than all the component models on almost all of our model fit statistics. The EBMA model has the highest PRE and % correct and the lowest Brier score. Although the LMER model has a slightly higher AUC, that model's overall performance suggests that it may be overfit.

Figure 1 shows separation plots for the EBMA model and the individual components (Greenhill, Ward, and Sacks 2011). In each plot, the observations are ordered from left to right by increasing predicted probabilities of insurgency (as predicted by the particular model). The black line corresponds to the predicted probability produced by the relevant model for each observation, and actual occurrences of insurgencies are colored red. Figure 1 shows visually that the GLM model performs very poorly, whereas the LMER model performs very well but tends to assign high probabilities to a large number of observations where we observe no insurgencies (i.e., it overpredicts insurgencies). More importantly, the overall best performance is associated with the EBMA forecast. The separation plots show that the EBMA model produces few false negatives and significantly fewer false positives than any of the component models.

However, the more interesting evaluation of the EBMA method is its test-period predictive power. Table 2 shows fit statistics for the individual components as well as the EBMA forecasts for observations in the 12 months following the validation period. The EBMA model outperforms the component models on all metrics. In particular, the EBMA model has the highest PRE at 0.43. Since it is possible to predict 89.9% of these observations correctly by forecasting no insurgency, a 43% reduction of error relative to the baseline model is quite substantial.

Importantly, EBMA clearly outperforms all component models in regard to the Brier score. Research regarding scoring rules for forecasts has shown that the Brier score is one of the best statistically proper scoring rules for evaluating predictions of binary dependent variables (Gneiting and Raftery 2007). Thus, to generally compare and rank the different models, one should use the Brier score (Gneiting and Raftery 2007). As can be seen in Tables 1 and 2, the EBMA model has the lowest Brier score in both the validation and test-period forecasts.[14]

[14]One alternative approach to generating ensemble forecasts would be to use the simple average of each component forecast. However, this causes difficulties because the researcher must use their own judgment to decide which alternative models are sufficiently accurate and diverse for inclusion. EBMA offers a more statistically motivated and straightforward method for achieving the same end. In any case, these simple averages do not perform well against the EBMA forecast. In the current example, a simple unweighted average results in AUC = 0.885, PRE = 0.123, Brier = 0.052, and % Correct = 92.8 for the test period. This is not surprising given that simple averaging weights an inaccurate model the same as an accurate one. EBMA, on the other hand, is able to detect the superiority of components and calibrates weights accordingly. Likewise, simple averages cannot identify pairs or groups of highly correlated forecasts and will tend to give these groupings too much weight.

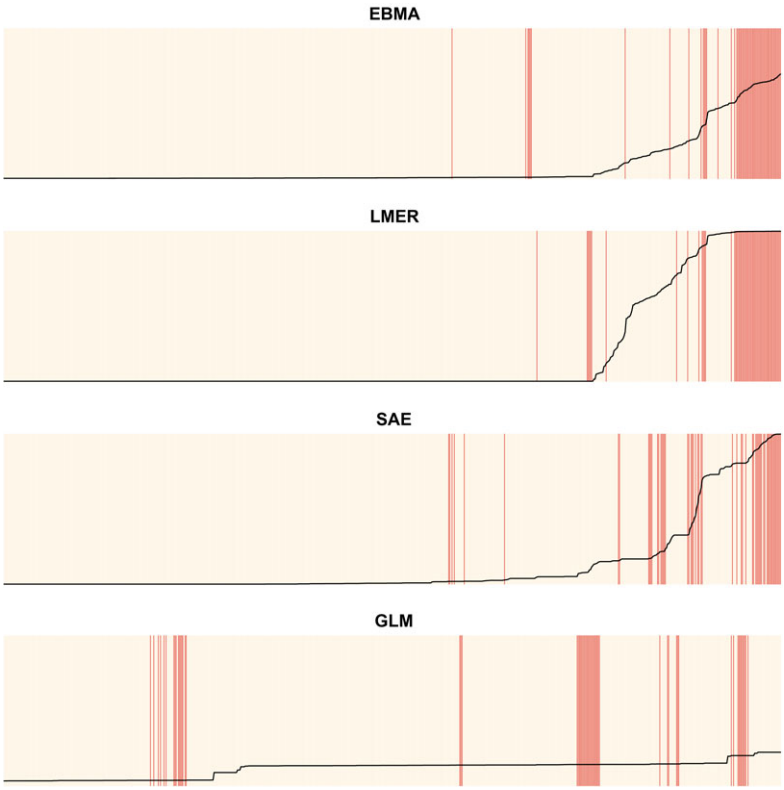


Fig. 1 Separation plots for validation-period predictions of the ICEWS data ($n = 696$). For each model, observations are shown from left to right in order of increasing predicted probability of insurgency (shown as the black line). Observations where insurgency actually occurred are shown in red. EBMA outperforms all component models in assigning high predicted probabilities to more observed insurgencies and to fewer non-insurgencies.

Figure 2 shows the separation plots for the components as well as the EBMA forecasts for 2010. The EBMA model again performs better than any of the individual components with very high predicted probabilities for the majority of actual events without overpredicting too many events. Taking both the fit statistics and the visual evidence together, we can conclude that the EBMA model leads to a substantial improvement in test-period forecasts relative to its components. This is true even in data sets with rare events and even when the individual components are already performing well.

4.2 *Application to U.S. Presidential Election Forecasts*

For the past several U.S. election cycles, a number of research teams have developed forecasting models and published their predictions in advance of Election Day. For example, before the 2008 election, a symposium of forecasts was published in *PS: Political Science and Politics* with forecasts of presidential and congressional vote shares developed by Campbell (2008), Norpoth (2008), Lewis-Beck and Tien (2008), Abramowitz (2008), Erikson and Wlezien (2008), Holbrook (2008), and Lockerbie (2008). Responses to the forecasts were published in a subsequent issue. Earlier, in 1999, an entire issue of the *International Journal of Forecasting* was dedicated to the task of predicting presidential elections (Brown and Chappell 1999). Predicting presidential elections has also drawn the attention of economists seeking to understand the relationship between economic fundamentals and political outcomes. Two prominent examples include work by Fair (2010) and Hibbs (2000).

Table 2 Test-period results (2010)

	AUC	*PRE*	*Brier*	*% Correct*
LMER	0.97	0.11	0.08	91.09
SAE	0.96	0.20	0.06	91.95
GLM	0.72	0.00	0.09	89.94
EBMA	0.97	0.43	0.04	94.25
$n = 348$				

Note. The table shows fit statistics for the EBMA deterministic forecast and all component model forecasts of insurgency in 29 countries of the Pacific Rim for the test period. EBMA equals or outperforms any single model on all measures.

4.2.1 Component models

In the rest of this subsection, we replicate several of these models and demonstrate the usefulness of the EBMA methodology for improving the prediction of single important events.[15] We include forecasting models from six of the most widely cited presidential forecasting teams. Note that we do not, in every case, replicate the particular model which the authors identify as their definitive forecast.

- Campbell: Campbell's "Trial-Heat and Economy Model" (Campbell 2008).
- Abramowitz: The "Time-for-Change Model" created by Abramowitz (2008).
- Hibbs: Hibbs's "Bread and Peace Model" (Hibbs 2000).
- Fair: Fair's presidential vote-share model.[16]
- Lewis-Beck/Tien: Lewis-Beck and Tien's "Jobs Model Forecast" (Lewis-Beck and Tien 2008).
- EWT2C2: Column 2 from Table 2 in Erikson and Wlezien (2008).

With the exception of the Hibbs forecast, the models are simple linear regressions. The dependent variable is the share of the two-party vote received by the incumbent-party candidate.[17]

4.2.2 Results

Rather than selecting a single partition of the data into training, validation, and test periods (as in the insurgency analysis), we generate sequential predictions. For each year from 1976 to 2008, we use all available prior data to fit the component models.[18] We then fit the EBMA model using the components' performances for election years beginning with 1952 (the year when all models begin generating predictions). For example, to generate predictions for the 1988 election, we used the performance of each component for the 1952–84 period to estimate model weights.[19]

[15]It is important to note that we attempted to replicate each of the models for the 2008 election as closely as possible given the model descriptions in the articles and the data provided by the authors. We then proceeded to use the same model specifications as used in the 2008 articles to forecast all elections previous to 2008. Thus, prior to 2008, the individual model results are not exact replications of the author's given prediction for that election year, and results may vary from what was presented by the authors as the forecast for a given election. This may be due to changes in the model specification over time and data updates. Thus, we neither attempted nor succeeded in replicating the exact forecasts for all election years for all components.

[16]The model here replicates equation (1) in Fair (2010).

[17]The data to replicate the models by Abramowitz (2008), Campbell (2008), Erikson and Wlezien (2008), and Lewis-Beck and Tien (2008) were provided in personal correspondence with the respective authors. The remaining data were downloaded from the Web sites of Ray C. Fair (http://fairmodel.econ.yale.edu/vote2012/tbl1.txt) and Douglas Hibbs (http://www.douglas-hibbs.com/).

[18]For example, the Fair model uses data for election results beginning in 1916 while the Abramowitz model begins with data from the 1952 election.

[19]See footnote 4 for additional discussion of the implications of the overlapping training and validation samples. Results in this section were computed using modifications of the "ensembleBMA" package (Fraley, Raftery, and Gneiting 2010, 2011). Because of the paucity of data, we did not apply any bias correction to these forecasts. Thus, the predictor and constant, denoted a_{0k} and a_{1k} above, are constrained to zero and one, respectively.

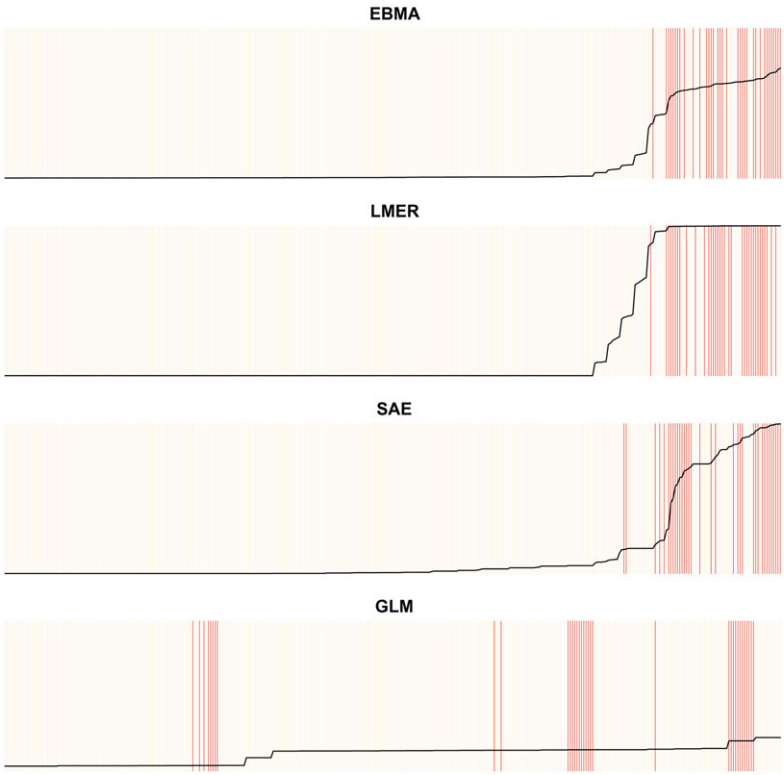


Fig. 2 Separation plots for the test-period predictions of the ICEWS data ($n = 348$). For each model, observations are shown from left to right in order of increasing predicted probability (shown as the black line). Observations where insurgency actually occurred are shown in red. EBMA outperforms all component models in assigning high predicted probabilities to more observed insurgencies and to fewer non-insurgencies.

Table 3 provides exemplar results for the 2004 and 2008 elections. Table 3 shows the weights assigned to each model as well as the validation period root mean squared error (RMSE) and mean absolute error (MAE) for the components and the EBMA forecasts.[20] It also shows the prediction errors, calculated as the difference between predicted and actual values in each year for each component model and the EBMA forecast.

The example results in Table 3 illustrate three important points. First, the EBMA model does better than any individual component on validation-sample measures of model fit such as RMSE and MAE.[20] Second, these results demonstrate that EBMA is not guaranteed to generate the most accurate prediction for any single observation. In each year, at least one component model comes closer to predicting the actual outcome. However, the EBMA forecasts will very rarely provide egregiously wrong predictions (e.g., as found in the Campbell model in 2008 and the Fair model in 2004) since it borrows strength from multiple components. Moreover, as we show below, in the aggregate, the EBMA model tends to provide the best forecast over time.

Third, Table 3 shows that there is not a clean relationship between validation-sample model performance and model weights. For instance, the weight for the Abramowitz model in 2008 is 0.001 even though it has the lowest RMSE and MAE of any component. The diminished relationship between

[20] As we noted above, these models are fit sequentially, so the validation periods change. For example, the validation period for the forecast of the 2004 election is 1952–2000. The validation period RMSE is therefore calculated for those observations. For the 2008 election, the validation period is 1952–2004.

Table 3 Test-period prediction errors, model weights, and validation-period fit statistics for component and EBMA forecasts of the 2004 and 2008 elections

	2004 Election				*2008 Election*			
	Weights	*RMSE*	*MAE*	*Prediction errors*	*Weights*	*RMSE*	*MAE*	*Prediction errors*
Campbell	0.40	1.71	1.33	0.53	0.36	1.65	1.28	6.33
Abramowitz	0.00	1.50	1.18	2.20	0.06	1.53	1.26	−2.37
Hibbs	0.12	1.95	1.38	1.54	0.25	1.92	1.38	−1.39
Fair	0.48	2.07	1.47	4.82	0.00	2.22	1.80	−2.02
Lewis-Beck/Tien	0.00	1.67	1.42	−0.41	0.17	1.61	1.33	−2.64
EWT2C2	0.00	2.67	2.06	4.76	0.17	2.81	2.18	−0.14
EBMA		1.29	1.02	2.08		1.30	1.01	−0.53

Note. The models are trained using all prior data, and the EBMA model is validated on the observations beginning in 1952. The EBMA model does better than all components on validation sample fit statistics. In addition, although it does not necessarily make the most accurate prediction for any given year, it is less likely to make dramatic forecasting errors for the test period.

validation-sample performance and weight is a result of high correlations between forecasts.[21] For instance, fitted values for the Abramowitz model are correlated at 0.94 with the Campbell model and at 0.96 with the Lewis-Beck/Tien model. Thus, conditioned on knowing these forecasts, the Abramowitz component provides limited additional information.

In general, as the number of models included as EBMA components increases, the risk of including highly correlated forecasts will also rise. Researchers should be aware of the fact that adding additional forecasts as components will not necessarily improve the performance of EBMA. EBMA performance will instead be improved by the inclusion of increasing numbers of diverse and accurate forecasts (see also Graefe et al. 2010). Including a large number of extremely correlated forecasts may actually reduce the benefits of ensemble forecasting. However, we note that in practice this is unlikely to be a significant concern as there are few domains in political science for which there are large numbers of ongoing forecasting efforts.

With the 2004 and 2008 examples in mind, we now turn to the relative test-sample performance of the EBMA and component forecasts across the entire 1976–2008 period. Table 4 shows the test-sample RMSE and MAE statistics as well as the percentage of observations that fall within the 67% and 90% predictive intervals for each model. For our purposes here, the main result in Table 4 is that the EBMA model again outperforms all components. The first two columns show this to be true in terms of prediction error (RMSE and MAE).[22]

In addition, the coverage statistics demonstrate better calibration of EBMA forecasts relative to its component models. For instance, the observed outcome falls within the 67% predictive interval for the Abramowitz model only three out of nine times, whereas it covers the observed values eight out of nine times for the Lewis-Beck/Tien model. Meanwhile, the EBMA 90% and 67% predictive intervals are nearly perfectly calibrated.

[21]The correlation matrix between fitted values of the models for the 1952–2004 period is

	C	A	H	F	L	E
Campbell	1.00					
Abramowitz	0.94	1.00				
Hibbs	0.91	0.93	1.00			
Fair	0.87	0.89	0.89	1.00		
Lewis-Beck/Tien	0.93	0.96	0.91	0.88	1.00	
EWT2C2	0.85	0.90	0.87	0.91	0.86	1.00

[22]Brandt, Freeman, and Schrodt (2011a) survey a variety of metrics in addition to those we employ here. These include measures of average prediction errors, measures using medians and geometric averages, measures that compare the complete difference in probability distributions, and sequential rank-based methods. Although there are many candidate metrics, at least for the alternative metrics we have calculated so far, the substantive conclusions we reach do not change for our examples and are not presented due to space constraints. However, as suggested by a helpful reviewer, we note that there are reasons to doubt that RMSE or MAE will necessarily provide a ranking of component models based on accuracy. A more complete approach to evaluating the accuracy of the component models is to examine the results displayed in Fig. 3 below.

Table 4 Fit statistics and observed coverage probabilities for sequentially generated test-sample predictions of presidential elections from 1976 to 2008

			Coverage	
	RMSE	*MAE*	*67%*	*90%*
Campbell	2.74	1.99	0.67	0.78
Abramowitz	2.27	2.05	0.33	0.78
Hibbs	2.81	2.24	0.22	0.56
Fair	4.01	3.20	0.44	0.78
Lewis-Beck	2.27	1.82	0.89	1.00
EWT2C2	2.88	2.16	0.78	1.00
EBMA	1.72	1.47	0.67	0.89

Note. EBMA outperforms its component models on all metrics.

In a well-calibrated forecasting model, out-of-sample outcomes should fall within predictive intervals at a rate corresponding to their size. For instance, the goal is for two-thirds of all out-of-sample observations to fall within their respective 67% predictive intervals. Poorly calibrated models will tend to produce predictive intervals that are either too narrow, generating inaccurate predictions, or too large, generating predictions that are accurate but too vague to be useful. The better calibration of the EBMA model can be seen visually in Fig. 3. The plot shows the point predictions and the 67% and 90% predictive intervals for each model in each year. The vertical dashed lines show the actual observed outcomes. Note that two of the most accurate forecasts, the Lewis-Beck/Tien and Erikson/Wlezien models, make very imprecise predictions. Thus, although they have very good coverage, it is at least partly because their estimates are so inexact. The Campbell, Abramowitz, and Hibbs models provide more reasonable predictive intervals but are less accurate than EBMA. Meanwhile, the Fair model falls somewhere between these two groupings.

Finally, it is worth noting an example—very noticeable in these data—of the kinds of problems that may arise when relying on a single model for making predictions. From 1952 to 2004, the Campbell model was consistently one of the strongest performers. Indeed, it made the most accurate forecast of the 2004 election. However, one of the crucial variables in this model comes from polling data measured in early September. As a result of the particularly late timing of the Republican Convention in 2008, it was the only model to forecast a victory for John McCain. By relying on a wider array of data sources and methodologies, EBMA reduces the likelihood of such large misses without completely eliminating the general insights captured by individual models that may on occasion be wide of the mark.

4.3 *Application to the Supreme Court Forecasting Project*

Our final application of EBMA is a reanalysis of data from the Supreme Court Forecasting Project (Ruger et al. 2004; Martin et al. 2004).[23] This example is especially interesting and important as it shows a particular strength of EBMA that was not utilized in the previous two examples. That is, not only is EBMA able to combine the forecasts from multiple statistical models, in addition, it can also combine statistical predictions with forecasts generated by classification trees, subject experts, and other sources. As is shown below, the EBMA model is able to combine the strength of a statistical forecasting model with the particular strength of subject-expert predictions and improves on the accuracy of both. Furthermore the Supreme Court Forecasting Project offers a clean example for our purpose. The weights for the EBMA model are calibrated on the performance of the components on actual predictions of Supreme Court Justice votes. That is, even for the validation period, the predictions were made before the court decisions were issued. Thus, we can use the performance of the component models on actual predictions to calculate the weights

[23] Additional details about the project, replication files, as well as a complete listing of cases and expert forecasts are available at http://wusct.wustl.edu/index.php.

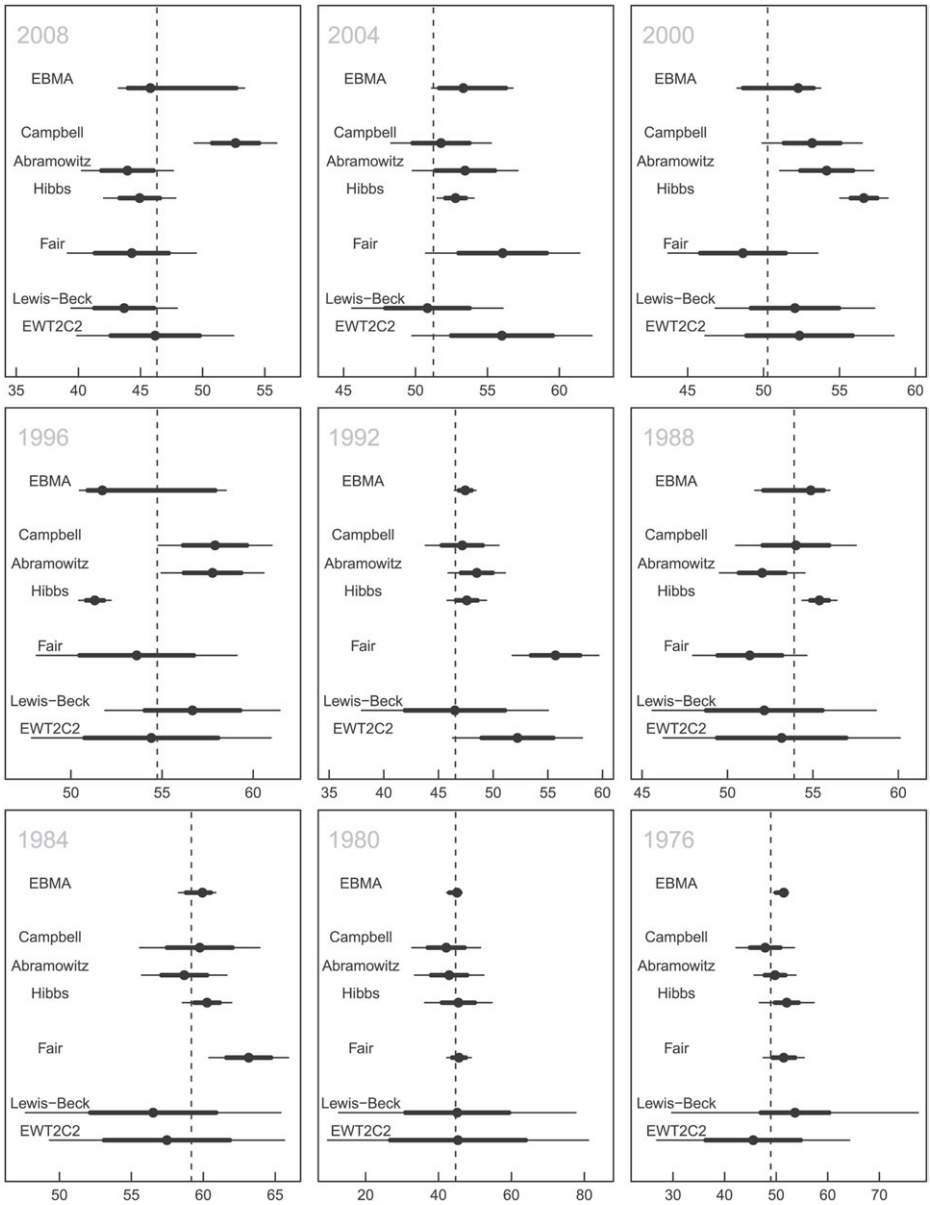


Fig. 3 The predicted and actual percentage of the two-party vote going to the incumbent party in U.S. presidential elections from six component models and the EBMA forecast. For each year, the plots show the point predictions (circles), 67% predictive intervals (thick horizontal lines), and 90% predictive intervals (thin horizontal lines). The vertical dashed line is the observed outcome. The EBMA model is better calibrated than its components.

for the EBMA model. We then compare the EBMA forecasts with the component model predictions on a separate test sample.

A large literature in political science is concerned with constitutional courts and the Supreme Court in particular, with one of the most prominent strands of the literature on courts trying to analyze and explain justices' voting behavior. In general, however, the theories and models attempt to explain behavior ex post (e.g., Hausegger and Baum 1999; Segal and Cover 1989; Richards and Kritzer 2002; Klein and Hume 2003; Songer, Segal, and Cameron et al. 1994).

In contrast to most of this literature, a research team consisting of Andrew Martin, Kevin Quinn, Theodore Ruger, and Pauline Kim (henceforward MQRK) set out to develop methods to predict Supreme Court decisions in the future. Throughout 2002–03, MQRK generated two sets of forecasts for every pending case. First, for each case, MQRK collected data on six different case characteristics, such as

the ''circuit of origin of the case'' or ''ideological direction of the lower courts ruling,'' which were then used as explanatory variables (Martin et al. 2004, 762). The authors then used classification trees to generate a binary forecast for the expected vote of each justice on each case (voting to affirm the lower court opinion is coded as a 1).

As a second method to forecast Supreme Court decisions, MQRK recruited a team of 83 legal experts. These experts would then predict the decision for each justice and the court as a whole for particular cases in their specialty area. The list of experts included academics, appellate attorneys, former Supreme Court clerks, and law school deans. MQRK attempted to recruit three expert forecasts for each case, although this was not possible for all cases.

The classification trees made predictions for all 67 cases included in the MQRK analysis. We include these binary model predictions as one component forecast. However, the individual legal experts made predictions on only a handful of cases. Owing to the paucity of the data for each expert, we pooled them together and treat all the expert opinions as part of a single forecasting effort. We coded the expert forecast to be the mean expert prediction. This implies that the expert forecast predicts a vote to affirm if a majority of experts polled for that case predict an affirming vote. We fit an EBMA model using all cases with docket numbers dating from 2001 ($n = 395$) and made EBMA forecasts for the remaining 214 cases with 2002 docket numbers.[24]

Table 5 shows the component weights for the two forecasts and the test-period fit statistics for the MQRK classification trees, subject experts, and EBMA forecasts. The EBMA model weights the subject experts about twice as much as the statistical model. Once again, the results show that the EBMA procedure outperforms all components (even when there are only two). In terms of AUC, Brier scores, and percent correct predictions, the EBMA forecast outperforms both the statistical model and the combined subject experts. In addition, EBMA scores substantially better on the PRE metric.[25]

There is a long-standing debate in many circles of the relative strengths and weaknesses of statistical models and subject experts for making predictions (e.g., Ascher 1978). Models that use quantifiable measurements and widely available (if sometimes crude) data to make predictions can make egregious errors in particular cases. Some cases may be decided by forces invisible to the statistical model but obvious to experts familiar with the case. Subject experts, on the other hand, can become too focused on minutiae and miss larger (if more subtle) trends in the data easily recognized by more advanced methodologies. The EBMA technique offers a theoretically motivated way to combine the strengths of both methods, while smoothing over their relative weaknesses, to make more accurate predictions.

5 Discussion

As currently implemented, EBMA already offers a method for aiding the accurate prediction of future events. However, we envision several paths forward for future research in this area. First, we are planning to extend EBMA into a fully Bayesian framework. Markov chain Monte Carlo estimation of EBMA

Table 5 Test-period results for U.S. Supreme Court example

	Weight	*AUC*	*PRE*	*Brier*	*% Correct*
MQRK model	0.32	0.66	−0.02	0.29	70.56
Subject experts	0.68	0.62	0.15	0.23	75.23
EBMA forecast		0.70	0.21	0.18	77.10
$n = 214$					

Note. The table shows fit statistics for the EBMA deterministic forecast and component forecasts of U.S. Supreme Court votes on cases in the 2002-2003 session with 2002 docket numbers. EBMA outperforms its component models on all metrics.

[24] As noted above, these cases were heard in the 2002–03 period. We note that the dates on the docket number do not necessarily reflect the order in which they were argued before the court and the order in which cases were argued did not correspond to when the decisions were handed down. Thus, there is no obvious way to partition the data into validation and test periods. However, in general, the docket numbers roughly correspond to the age of the case. Although partitioning the data in this manner is slightly arbitrary, it serves the limited purpose of demonstrating the method.

[25] The baseline model here is the prediction that all votes will be to reverse the lower court. This baseline model is correct for roughly 70% of the votes in the test period.

models promises to more efficiently handle a wider variety of outcome distributions and will provide additional information regarding our uncertainty about model weights and within-model variances (Vrugt, Diks, and Clark 2008).

Second, EBMA estimates model weights based exclusively on the point predictions of component forecasts. Even for continuous data (e.g., the presidential vote forecasts), the current procedure assumes that the within-forecast variance (σ^2) is constant across models. In other words, model weights do not reflect the uncertainty associated with each model's predictions. Applying both Bayesian and bootstrap methods, we intend to incorporate the entire predictive PDFs of component forecasts so that model weights reflect not only components' accuracy but also their precision. Poorly calibrated models should be penalized and receive less posterior weight.

However, EBMA as it is currently implemented shows considerable promise for aiding systematic social inquiry. For many important and interesting events, it is almost impossible for social scientists to find the ''true'' data-generating process. Socially determined events are inherently difficult to predict because of nonlinearities and the complexity of human behavior. This may be one of the main reasons political scientists so rarely make systematic predictions about the future. Yet, we believe it should be one ultimate goal of the discipline to make sensible and reliable forecasts. Doing so would make the discipline more relevant to policymakers and provide more avenues for rigorous testing of theoretical models and hypothesized empirical regularities.[26]

EBMA uses the accuracy of in-sample predictions of individual models to calibrate a combined weighted-average forecast and to make more accurate predictions. Moreover, it does so in a transparent and theoretically motivated manner that allows us to see which component models are most important in informing the broader EBMA model. Thus, EBMA can enhance the accuracy of forecasts in political science, while also allowing the continued development of multiple theoretical and empirical approaches to the study of important topics. In addition, we have adjusted EBMA to work for dichotomous dependent variables. The EBMA model, therefore, can now be used in a large fraction of research in political science. However, the method depends fundamentally on the existence of relatively good individual models; otherwise, the ensemble is empty. Thus, EBMA and other ensemble methods should not discourage the development of individual prediction models, but rather leverage their individual contributions with those from other models in order to achieve more accurate predictions.

Finally, we demonstrated the utility of the EBMA method for improving out-of-sample forecasts in three empirical analyses. In each, the EBMA model outperformed its components and was less sensitive to idiosyncratic data issues than the individual models. The EBMA method was applied to improve the prediction of insurgencies around the Pacific Rim, U.S. presidential election results, and votes of U.S. Supreme Court Justices. However, we believe these applications represent only a portion of the areas to which the EBMA method could be fruitfully applied. Using the software we have developed for this project, it will be possible for researchers to increase the accuracy of forecasts of a wide array of important events.[27]

Funding

Information Processing Technology Office of the Defense Advanced Research Projects Agency through a holding grant to the Lockheed Martin Corporation (FA8650-07-C-7749).

Appendix

Adjustments for Dichotomous Outcomes

Past work on EBMA does not apply directly to the prediction of many political events because the assumed PDFs are normal, Poisson, or gamma. In many settings (e.g., international conflicts), the data are not

[26] In addition, some scholars have advanced the argument that prediction is closely related to the identification of causal processes (e.g., Spirtes, Glymour, and Scheines 2000). However, this is far from a universally accepted position and is not the basis for our advocacy of increased forecasting in political science.

[27] All data used to generate the results in this article will be made available to the public in the journal's dataverse upon publication at http://hdl.handle.net/1902.1/17286 (Montgomery, Hollenbach, and Ward 2012). The package for EBMA, EBMAforecast, is available through the Comprehensive R-Archive Network at http://cran.r-project.org/.

sufficiently fine-grained to justify these distributional assumptions. Usually, the outcomes of interest are dichotomous indicators for whether an event (e.g., civil war) has occurred in a given time period and country. Thus, none of the distributional assumptions used in past work are appropriate in this context. Fortunately, it is a straightforward extension of Sloughter et al. (2007) and Sloughter, Gneiting, and Raftery (2010) to deal appropriately with binary outcomes.[28]

We follow Sloughter et al. (2007) and Hamill, Whitaker, and Wei (2004) in using logistic regression after a power transformation of the forecast to reduce prediction bias. That is, point predictions are raised to a power, $\frac{1}{b} \leqslant 1$. This shrinks predictions downward toward zero. The transformation dampens the effect of extreme observations and helps reduce overfitting that might occur because certain models do slightly better in predicting high-leverage observations. Since the predictions for dichotomous outcomes are necessarily between -1 and 1, our adjustment process is slightly more complex. Nonetheless, the results for bias reduction are the same.

For notational ease, we assume that $\mathbf{f}_k$ is the forecast after the adjustment for bias reduction (we will omit the superscripts for the moment). Therefore, let $\mathbf{f}_k^{'} \in [0, 1]$ be the forecast on the predicted probability scale and

$$\mathbf{f}_k = \left[(1+\text{logit}(\mathbf{f}_k^{'}))^{1/b}-1\right] I\left[\mathbf{f}_k^{'} > \frac{1}{2}\right] - \left[(1+\text{logit}(|\mathbf{f}_k^{'}|))^{1/b}-1\right] I\left[\mathbf{f}_k^{'} < \frac{1}{2}\right], \tag{A1}$$

where $I[.]$ is the general indicator function. Hamill, Whitaker, and Wei (2004) recommend setting $b = 4$, whereas Sloughter et al. (2007) use $b = 3$. We use $b = 3$ in the insurgency example above and $b = 4$ in the courts example. However, we found that this choice makes very little difference for these examples.

The logistic model for the outcome variables is

$$\text{logit}\, P(\mathbf{y} = 1|\mathbf{f}_k) \equiv \log \frac{P(\mathbf{y} = 1|\mathbf{f}_k)}{P(\mathbf{y} = 0|\mathbf{f}_k)} = a_{k0} + a_{k_1}\mathbf{f}_k. \tag{A2}$$

The conditional PDF of some within-sample event, given the forecast $f_k^{s|t}$ and the assumption that k is the true model, can be written

$$g_k\left(y|f_k^{s|t}\right) = P\left(y = 1|f_k^{s|t}\right) I\left[y = 1\right] + P\left(y = 0|f_k^{s|t}\right) I\left[y = 0\right]. \tag{A3}$$

Applying this to equation (3), the PDF of the final EBMA model for y is

$$\begin{aligned} p\left(y|f_1^{s|t}, f_2^{s|t}, \ldots, f_K^{s|t}\right) = \sum_{k=1}^{K} w_k \Big\{ & P\left(y = 1|f_k^{s|t}\right) I[y = 1] \\ & + P\left(y = 0|f_k^{s|t}\right) I[y = 0] \Big\}. \end{aligned} \tag{A4}$$

Parameter estimation is conducted using only the data from the validation period T. The parameters a_{0k} and a_{1k} are specific to each individual component model. For model k, these parameters can be estimated as traditional linear models, where $\mathbf{y}$ is the dependent variable and the covariate list includes only $\mathbf{f}_k$ and a constant term.

The difficulty is in estimating the weighting parameters, $w_k \forall k \in [1, 2, \ldots, K]$. For the moment, we have followed Raftery et al. (2005) and Sloughter et al. (2007) in using maximum likelihood methods. In future work, we plan to implement a fully Bayesian analysis by placing priors on all parameters and using Markov chain Monte Carlo techniques to estimate model weights (cf. Vrugt, Diks, and Clark 2008).

The log-likelihood function cannot be maximized analytically, but Raftery et al. (2005) and Sloughter et al. (2007) suggest using the expectation-maximization (EM) algorithm. We introduce the unobserved

[28] The method for dealing with binary outcomes is implicit in Sloughter et al. (2007) and Sloughter, Gneiting, and Raftery (2010), who assume a discrete-continuous distribution for outcomes that includes a logistic component. However, they do not explicitly and fully develop the model for dichotomous outcomes. A related strain of research on dynamic model averaging (cf. Muhlbaier and Polikar 2007; Raftery, Kárný, and Ettler 2010) has recently been extended for direct application to binary outcomes (e.g., McCormick et al. 2011; Tomas 2011).

quantities $z_k^{s|t}$, which represent the posterior probability for model k for observation $s|t$. The E step involves calculating estimates for these unobserved quantities using the formula

$$\hat{z}_k^{(j+1)s|t} = \frac{\hat{w}_k^{(j)} p^{(j)}\left(y|f_k^{s|t}\right)}{\sum_{k=1}^{K} \hat{w}_k^{(j)} p^{(j)}\left(y|f_k^{s|t}\right)}, \tag{A5}$$

where the superscript j refers to the jth iteration of the EM algorithm.

$w_k^{(j)}$ is the estimate of w_k in the jth iteration, and $p^{(j)}(.)$ is shown in equation (A4). Assuming these estimates of $z_k^{s|t}$ are correct, it is then straightforward to derive the maximizing value for the model weights. Thus, the M step estimates these as $\hat{w}_k^{(j+1)} = \frac{1}{n}\sum_{s,t} \hat{z}_k^{(j+1)s|t}$, where n represents the number of observations in the validation data set.[29] The E and M steps are iterated until the improvement in the log-likelihood is no larger than some predefined tolerance.[30]

Ensemble Prediction for Dichotomous Outcomes

With these parameter estimates, it is now possible to generate ensemble forecasts. If our forecasts $\mathbf{f}_k^{s|t}$ are each generated from a statistical model, we now generate a new prediction $f_k^{s|t^*}$ from the previously fitted models. For convenience, let $\hat{\mathbf{a}}_k \equiv (\hat{a}_{k0}, \hat{a}_{k1})$. For some dichotomous observation in the test period $t^* \in T^*$, we can see that

$$P\left(y = 1|f_1^{s|t^*}, \ldots, f_K^{s|t^*}; \hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_K; \hat{w}_1, \ldots, \hat{\mathrm{w}}_K\right) = \sum_{k=1}^{K} \hat{w}_k \mathrm{logit}^{-1}\left(\hat{\mathrm{a}}_{k0} + \hat{\mathrm{a}}_{k1} f_k^{s|t^*}\right). \tag{A6}$$

References

Abramowitz, A. I. 2008. Forecasting the 2008 presidential election with the time-for-change model. *PS: Political Science & Politics* 41:691–5.

Andriole, S. J., and R. A. Young. 1977. Toward the development of an integrated crisis warning system. *International Studies Quarterly* 21:107–50.

Armstrong, J. S. 2001. Combining forecasts. In *Principles of forecasting: A handbook for researchers and practitioners,* ed. J. S. Amstrong. Norwell, MA: Kluwer Academic.

Ascher, W. 1978. Forecasting: An appraisal for policy-makers and planners. Baltimore: Johns Hopkins University Press.

Bartels, L. M. 1997. Specification uncertainty and model averaging. *American Journal of Political Science* 41:641–74.

Bartels, L. M., and J. Zaller. 2001. Presidential vote models: A recount. *PS: Political Science and Politics* 34:9–20.

Bates, J., and C. Granger. 1969. The combination of forecasts. *Operations Research* 20:451–68.

Bennett, D. S., and A. C. Stam. 2009. Revisiting predictions of war duration. *Conflict Management and Peace Science* 26:256–67.

Berg, J. E., F. D. Nelson, and T. A. Rietz. 2008. Prediction market accuracy in the long run. *International Journal of Forecasting* 24:285–300.

Berrocal, V. J., A. E. Raftery, T. Gneiting, and R. C. Steed. 2010. Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association* 105:522–37.

Billio, M., R. Casarin, F. Ravazzolo, and H. K. Van Dijk. 2010. Combining predictive densities using Bayesian filtering with applications to U.S. economics data. Norges Bank Working Paper. http://ssrn.com/abstract=1735421 (accessed June 1, 2011).

Billio, M., R. Casarin, F. Ravazzolo, and H. K. Van Dijk. 2011. Bayesian combinations of stock price predictions with an application to the Amsterdam exchange index. Tinbergen Institute Discussion Paper No. 2011-082/4. http://www.tinbergen.nl/discussionpapers/11082.pdf (accessed June 1, 2011).

Brandt, P. T., M. Colaresi, and J. R. Freeman. 2008. The dynamics of reciprocity, accountability, and credibility. *Journal of Conflict Resolution* 52:343–74.

[29] In the case of normally distributed data, $\hat{\sigma}^{2(j+1)} = \frac{1}{n}\sum_{s,t}\sum_{k=1}^{K} \hat{z}_k^{(j+1)s|t}\left(y - f_k^{s|t}\right)^2$.

[30] In the examples above, we begin with the assumption that all models are equally likely, $w_k = \frac{1}{K} \forall k \in [1, \ldots, K]$. Critics of MLE methods and the EM algorithm have raised concerns that convergence to a local rather than a global maximum may occur. We have found no differences in our results based on different starting values, but convergence can be slow if starting values are too dissimilar from the final estimates. Although we feel confident in the results reported here, in future research, we plan to expand the model estimation technique to include Bayesian methods. This will facilitate comparisons of estimates resulting from multiple estimation techniques.

Brandt, P. T., J. R. Freeman, and P. A. Schrodt. 2011a. Racing horses: Constructing and evaluating forecasts in political science. Paper prepared for the 28th Annual Summer Meeting of the Society for Political Methodology. http://polmeth.wustl.edu/media/Paper/RHMethods20110721small_1.pdf (accessed August 20, 2011).

———. 2011b. Real-time, time-series forecasting of inter- and intra-state political conflict. *Conflict Management and Peace Science* 28:41–64.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78:1–3.

Brock, W. A., S. N. Durlauf, and K. D. West. 2007. Model uncertainty and policy evaluation: Some theory and empirics. *Journal of Econometrics* 136:629–64.

Brown, L. B., and H. W. Chappell. 1999. Forecasting presidential elections using history and polls. *International Journal of Forecasting* 15:127–35.

Bueno de Mesquita, B. 2002. *Predicting politics*. Columbus: Ohio State University Press.

———. 2011. A new model for predicting policy choices: Preliminary tests. *Conflict Management and Peace Science* 28:65–85.

Campbell, J. E. 1992. Forecasting the presidential vote in the states. *American Journal of Political Science* 36:386–407.

———. 2008. The trial-heat forecast of the 2008 presidential vote: Performance and value considerations in an open-seat election. *PS: Political Science & Politics* 41:697–701.

Campbell, J. E., and K. A. Wink. 1990. Trial-heat forecasts of the presidential vote. *American Politics Research* 18:251–69.

Chmielecki, R. M., and A. E. Raftery. 2010. Probabilistic visibility forecasting using Bayesian model averaging. *Monthly Weather Review* 139:1626–36.

Choucri, N., and T. W. Robinson, eds. 1978. *Forecasting in international relations: Theory, methods, problems, prospects*. San Francisco, CA: W. H. Freeman.

Clyde, M. 2003. Model averaging. In *Subjective and objective Bayesian statistics: Principles, models, and applications,* ed. S. J. Press, 320–35. Hoboken, NJ: Wiley-Interscience.

Clyde, M., and E. I. George. 2004. Model uncertainty. *Statistical Science* 19:81–94.

Cuzàn, A. G., and C. M. Bundrick. 2008. Forecasting the 2008 presidential election: A challenge for the fiscal model. *PS: Political Science & Politics* 41:717–22.

Davies, J. L., and T. R. Gurr. 1998. *Preventive measures: Building risk assessment and crisis early warning systems*. Lanham, MD: Rowman & Littlefield.

Dawid, A. P. 1982. The well-calibrated Bayesian (with discussion). *Journal of the American Statistical Association* 77:605–13.

———. 1984. Present position and potential developments: Some personal views. Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society Series A (Statistics in Society)* 147:278–92.

de Marchi, S., C. Gelpi, and J. D. Grynaviski. 2004. Untangling neural nets. *American Political Science Review* 98:371–8.

de Sola Pool, I., R. P. Abelson, and S. L. Popkin. 1964. *Candidates, issues, and strategies: A computer simulation of the 1960 and 1964 presidential elections*. Cambridge, MA: MIT Press.

Draper, D. 1995. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society Series B (Methodological)* 57:45–97.

Enders, W., and T. M. Sandler. 2005. After 9/11: Is it all different now? *Journal of Conflict Resolution* 49:259–77.

Erikson, R. S., and C. Wlezien. 2008. Leading economic indicators, the polls, and the presidential vote. *PS: Political Science & Politics* 41:703–7.

Fair, R. C. 1978. The effect of economic events on votes for president. *Review of Economics and Statistics* 60:159–73.

———. 2010. Presidential and congressional vote-share equations: November 2010 update. Working Paper, Yale University. http://fairmodel.econ.yale.edu/RAYFAIR/PDF/2010C.pdf (accessed June 7, 2011).

Fearon, J. D., and D. D. Laitin. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97:75–90.

Feder, S. A. 2002. Forecasting for policy-making in the post-Cold War period. *Annual Review of Political Science* 5:111–25.

Feldkircher, M. Forthcoming 2012. Forecast combination and Bayesian model averaging: A prior sensitivity analysis. *Journal of Forecasting.*

Fraley, C., A. E. Raftery, and T. Gneiting. 2010. Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Monthly Weather Review* 138:190–202.

Fraley, C., A. E. Raftery, T. Gneiting, J. M. Sloughter, and V. J. Berrocal. 2011. Probabilistic weather forecasting in R. *R Journal* 3:55–63.

Fraley, C., A. E. Raftery, J. M. Sloughter, and T. Gneiting. 2010. *EnsembleBMA: Probabilistic forecasting using ensembles and Bayesian model averaging. R package version 4.5.* http://CRAN.R-project.org/package=ensembleBMA.

Freeman, J. R., and B. L. Job. 1979. Scientific forecasts in international relations: Problems of definition and epistemology. *International Studies Quarterly* 23:113–43.

Geer, J., and R. R. Lau. 2006. Filling in the blanks: A new method for estimating campaign effects. *British Journal of Political Science* 36:269–90.

Gill, J. 2004. Introduction to the special issue. *Political Analysis* 12:647–74.

Gleditsch, K. S., and M. D. Ward. 2010. Contentious issues and forecasting interstate disputes. Presented at the 2010 Annual Meeting of the International Studies Association, New Orleans, LA.

Gneiting, T., and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102:359–78.

Gneiting, T., and T. L. Thorarinsdottir. 2010. Predicting inflation: Professional experts versus no-change forecasts. Working Paper. http://arxiv.org/abs/1010.2318v1http://arxiv.org/abs/1010.2318v1 (accessed June 15, 2011).

Graefe, A., A. G. Cuzan, R. J. Jones, and J. S. Armstrong. 2010. Combining forecasts for U.S. presidential elections: The PollyVote. Working Paper. http://dl.dropbox.com/u/3662406/Articles/Graefe_et_al_Combining.pdf (accessed May 15, 2011).

Greenhill, B. D., M. D. Ward, and A. Sacks. 2011. The separation plot: A new visual method for evaluating the fit of binary data. *American Journal of Political Science* 55:990–1002.
Gurr, T. R., and M. I. Lichbach. 1986. Forecasting internal conflict: A competitive evaluation of empirical theories. *Comparative Political Studies* 19:3–38.
Hamill, T. S., J. S. Whitaker, and X. Wei. 2004. Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Monthly Weather Review* 132:1434–47.
Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
Hausegger, L., and L. Baum. 1999. Inviting congressional action: A study of Supreme Court motivations in statutory interpretation. *American Journal of Political Science* 43:162–85.
Hibbs, D. A. 2000. Bread and peace voting in U.S. presidential elections. *Public Choice* 104:149–80.
Hildebrand, D. K., J. D. Laing, and H. Rosenthal. 1976. Prediction analysis in political research. *American Political Science Review* 70:509–35.
Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14:382–417.
Holbrook, T. M. 2008. Incumbency, national conditions, and the 2008 presidential election. *PS: Political Science & Politics* 41:709–12.
Huisman, J., L. Breuer, H. Bormann, A. Bronstert, B. Croke, H.-G. Frede, T. Gräff, L. Hubrechts, A. Jakeman, G. Kite, et al. 2009. Assessing the impact of land-use change on hydrology by ensemble modeling (LUCHEM) II: Ensemble combinations and predictions. *Advances in Water Resources* 32:147–58.
Imai, K., and G. King. 2004. Did illegal overseas absentee ballots decide the 2000 U.S. presidential election? *Perspectives on Politics* 2:537–49.
Jerome, B., V. Jerome, and M. S. Lewis-Beck. 1999. Polls fail in France: Forecasts of the 1997 legislative election. *International Journal of Forecasting* 15:163–74.
King, G., and L. Zeng. 2001. Improving forecasts of state failure. *World Politics* 53:623–58.
Klein, D. E., and R. J. Hume. 2003. Fear of reversal as an explanation of lower court compliance. *Law & Society Review* 37:579–606.
Koop, G., and D. Korobilis. 2009. Forecasting inflation using dynamic model averaging. Working Paper. http://personal.strath.ac.uk/gary.koop/koop_korobilis_forecasting_inflation_using_DMA.pdf (accessed May 25, 2011).
Krause, G. A. 1997. Voters, information heterogeneity, and the dynamics of aggregate economic expectations. *American Journal of Political Science* 41:1170–200.
Leblang, D., and S. Satyanath. 2006. Institutions, expectations, and currency crises. *International Organization* 60:245–62.
Lewis-Beck, M. S. 2005. Election forecasting: Principles and practice. *British Journal of Politics & International Relations* 7:145–64.
Lewis-Beck, M. S., and C. Tien. 2008. The job of president and the jobs model forecast: Obama for '08? *PS: Political Science & Politics* 41:687–90.
Lock, K., and A. Gelman. 2010. Bayesian combination of state polls and election forecasts. *Political Analysis* 18:337–48.
Lockerbie, B. 2008. Election forecasting: The future of the presidency and the house. *PS: Political Science & Politics* 41:713–6.
Madigan, D., and A. E. Raftery. 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 89:1535–46.
Marshall, M. G., K. Jaggers, and T. R. Gurr. 2009. *Polity IV project: Political regime characteristics and transition 1800–2007*. College Park, MD: CIDCM, University of Maryland.
Martin, A. D., K. M. Quinn, T. W. Ruger, and P. T. Kim. 2004. Competing approaches to predicting Supreme Court decision-making. *Perspectives on Politics* 2:761–7.
McCandless, T. C., S. E. Haupt, and G. S. Young. 2011. The effects of imputing missing data on ensemble temperature forecasts. *Journal of Computers* 6:162–71.
McCormick, T. H., A. E. Raftery, D. Madigan, and R. S. Burd. 2011. Dynamic logistic regression and dynamic model averaging for binary classification. Working Paper. http://www.stat.columbia.edu/ madigan/PAPERS/ldbma27.pdf (accessed March 26, 2011).
Min, S.-K., and A. Hense. 2006. A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophysical Research Letters* 33:L08708.
Min, S.-K., D. Simonis, and A. Hense. 2007. Probabilistic climate change predictions applying Bayesian model averaging. *Philosophical Transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences* 365:2103–16.
Montgomery, J. M., F. Hollenbach, and M. D. Ward. 2012. Replication data for: Improving predictions using ensemble Bayesian model averaging. IQSS Dataverse Network. http://hdl.handle.net/1902.1/17286.
Montgomery, J. M., and B. Nyhan. 2010. Bayesian model averaging: Theoretical developments and practical applications. *Political Analysis* 18:245–70.
Muhlbaier, M. D., and R. Polikar. 2007. An ensemble approach for incremental learning in nonstationary environments. *Multiple Classifier Systems* 4472:490–500.
Norpoth, H. 2008. On the razor's edge: The forecast of the primary model. *PS: Political Science & Politics* 41:683–6.
O'Brien, S. P. 2002. Anticipating the good, the bad, and the ugly: An early warning approach to conflict and instability analysis. *Journal of Conflict Resolution* 46:791–811.
———. 2010. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review* 12:87–104.
Page, S. E. 2008. Uncertainty, difficulty, and complexity. *Journal of Theoretical Politics* 20:115–49.
———. 2011. *Diversity and complexity.* Princeton, NJ: Princeton University Press.

Page, S. E., L. M. Sander, and C. M. Schneider-Mizell. 2007. Conformity and dissonance in generalized voter models. *Journal of Statistical Physics* 128:1279–87.
Pevehouse, J. C., and J. S. Goldstein. 1999. Serbian compliance or defiance in Kosovo? Statistical analysis and real-time predictions. *Journal of Conflict Resolution* 43:538–46.
Raftery, A. E. 1995. Bayesian model selection in social research. *Sociological Methodology* 25:111–63.
Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* 133:1155–74.
Raftery, A. E., M. Kárný, and P. Ettler. 2010. Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics* 52:52–66.
Raftery, A. E., and Y. Zheng. 2003. Long-run performance of Bayesian model averaging. *Journal of the American Statistical Association* 98:931–8.
Richards, M. J., and H. M. Kritzer. 2002. Jurisprudential regimes in Supreme Court decision-making. *American Political Science Review* 96:305–20.
Rosenstone, S. J. 1983. *Forecasting presidential elections*. New Haven, CT: Yale University Press.
Ruger, T. W., P. T. Kim, A. D. Martin, and K. M. Quinn. 2004. The Supreme Court Forecasting Project: Legal and political science approaches to predicting Supreme Court decision-making. *Columbia Law Review* 104:1150–210.
Schneider, G., N. P. Gleditsch, and S. Carey. 2011. Forecasting in international relations: One quest, three approaches. *Conflict Management and Peace Science* 28:5–14.
Schrodt, P. A., and D. J. Gerner. 2000. Using cluster analysis to derive early warning indicators for political change in the Middle East, 1979–1996. *American Political Science Review* 94:803–18.
Segal, J. A., and A. D. Cover. 1989. Ideological values and the votes of U.S. Supreme Court Justices. *American Political Science Review* 83:557–65.
Singer, J. D., and M. D. Wallace. 1979. *To augur well: Early warning indicators in world politics*. Beverly Hills, CA: Sage.
Sloughter, J. M., T. Gneiting, and A. E. Raftery. 2010. Probabilistic wind-speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association* 105:25–35.
Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley. 2007. Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review* 135:3209–20.
Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns. 2009. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association* 104:97–116.
Songer, D. R., J. A. Segal, and C. M. Cameron. 1994. The hierarchy of justice: Testing a principal-agent model of Supreme Court-circuit court interactions. *American Journal of Political Science* 38:673–96.
Spirtes, P., C. N. Glymour, and R. Scheines. 2000. *Causation, prediction, and search*. Vol. 81. Cambridge, MA: MIT Press.
Tomas, A. 2011. A dynamic logistic multiple classifier system for online classification. Working Paper. http://www.stats.ox.ac.uk/tomas/html_links/T2011.pdf (accessed June 1, 2011).
Vincent, J. E. 1980. Scientific prediction versus crystal ball gazing: Can the unknown be known? *International Studies Quarterly* 24:450–4.
Vrugt, J. A., M. P. Clark, C. G. Diks, Q. Duan, and B. A. Robinson. 2006. Multi-objective calibration of forecast ensembles using Bayesian model averaging. *Geophysical Research Letters* 33:L19817.
Vrugt, J. A., C. G. Diks, and M. P. Clark. 2008. Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling. *Environmental Fluid Mechanics* 8:579–95.
Ward, M. D., B. D. Greenhill, and K. M. Bakke. 2010. The perils of policy by *p*-value: Predicting civil conflict. *Journal of Peace Research* 47:363–75.
Ward, M. D., R. M. Siverson, and X. Cao. 2007. Disputes, democracies, and dependencies: A re-examination of the Kantian peace. *American Journal of Political Science* 51:583–601.
Whiteley, P. F. 2005. Forecasting seats from votes in British general elections. *British Journal of Politics & International Relations* 7:165–73.
Wright, J. H. 2008. Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics* 146:329–41.
———. 2009. Forecasting U.S. inflation by Bayesian model averaging. *Journal of Forecasting* 28:131–44.
Zhang, X., R. Srinivasan, and D. Bosch. 2009. Calibration and uncertainty analysis of the SWAT model using genetic algorithms and Bayesian model averaging. *Journal of Hydrology* 374:307–17.