

SPARSE CONTROL SELECTION FOR SPATIAL EPIDEMIOLOGICAL DATA: DOUBLE LASSO AND COUNTY COVID-19 MORTALITY

MAXIMILIAN JAMES GEBAUER

A THESIS

in

STATISTICS AND DATA SCIENCE

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Master of Arts

2026

Supervisor of Thesis

LINDA ZHAO

PROFESSOR OF STATISTICS AND DATA SCIENCE

Graduate Group Chair

DAMEK DAVIS

ASSOCIATE PROFESSOR OF STATISTICS AND DATA SCIENCE

Dedication

To my parents, John and Paula, for their sacrifices and support throughout my academic journey.

Acknowledgments

I thank my advisor, Linda Zhao, for her encouragement and advising without which this thesis would not have been possible, Abraham Wyner whose mentorship has proven invaluable, Paul Sabin who advised an earlier version of the project that improved its trajectory immeasurably, Errol Lord who allowed a philosopher to moonlight as a statistician, and the following additional faculty whose courses were indispensable for my academic journey: Robert Stine, Dylan Small, Michael Carchidi, Wei Wang, Alex Vekker, and Rommel Regis.

Abstract

We investigate whether county-level political ideology—proxied by Republican vote share—is a robust predictor of COVID-19 mortality across U.S. counties, while explicitly avoiding individual-level causal claims. Using cumulative county deaths from January 2020 to May 2022 and a high-dimensional set of 127 county characteristics for ($n \approx 3,096$) counties, we address three methodological obstacles common in this literature: high-dimensional confounding, instability of rare-event rates in small populations, and spatial dependence across adjacent counties. To construct a principled adjustment set, we apply post-double-selection (“Double-LASSO”) using a stabilized log-mortality-rate response for selection, and then carry the resulting union of controls into the final count-based model which incorporates an areal Markov random field smooth based on county adjacency to capture residual spatial dependence not explained by observed covariates.

Results: Republican vote share remains a strong positive predictor of county mortality in the fully adjusted spatial model ($\hat{\beta} = 0.01194$, $SE = 0.000925$, $z = 12.91$, $p < 2 \times 10^{-16}$), corresponding to an estimated rate ratio of $\exp(10\hat{\beta}) \approx 1.13$ per ten percentage-point increase (approximately $\exp(\hat{\beta}) \approx 1.012$ per one point); the 95% confidence interval is $\exp(10(\hat{\beta} \pm 1.96 SE)) \approx [1.11, 1.15]$ on the ten-point scale. Residual spatial autocorrelation is pronounced in the non-spatial specification (Moran’s $I \approx 0.122$, $p < 2.2 \times 10^{-16}$) and is substantially attenuated after incorporating the MRF term (Moran test $p = 0.176$). Simulation-based residual diagnostics show no evidence of residual over/under-dispersion and no excess outliers under a bootstrap outlier test.

Contents

Dedication	ii
Acknowledgments	iii
Abstract	iv
1 Introduction	1
1.1 Existing Work	2
1.2 Contributions	6
2 Methods	8
2.1 Double-Lasso Control Selection	8
2.2 Negative Binomial Likelihood	11
2.3 Modeling Spatial Dependence	12
2.4 Methods Summarized	13
3 Results	15
3.1 Data & EDA	15
3.2 Model Fitting Procedure	19
3.2.1 Post-double Selection Procedure	19
3.2.2 Generalized Additive Model	20
3.3 Key Results	21

4	Discussion	23
4.1	Discussion	23
4.2	Limitations	26
4.3	Future Directions	28
A	Technical Appendix A: Post-Double Selection Algorithm	30
B	Technical Appendix B: Feature Dictionary & Full Results Tables	33
B.1	Feature Dictionary	33
B.2	Full Results Tables	36

List of Tables

B.1	Classification of all candidate control variables considered in the post-double-selection procedure. Factor variables are reported at the parent-variable level rather than the dummy-column level.	34
B.2	Parametric coefficient estimates from the negative binomial GAM: substantive covariates.	36
B.3	State fixed effects from the negative binomial GAM. Coefficients are relative to the omitted reference state.	38

List of Figures

3.1	County-level bivariate association and marginal distributions for partisanship and COVID-19 mortality.	17
3.2	Spatial distributions of county-level COVID-19 mortality and partisanship (contiguous U.S.).	18
3.3	Estimated state fixed effects from the negative binomial GAM, ordered from smallest to largest, with 95% confidence intervals. Effects are on the log-rate scale relative to the reference state (Alabama).	22

Chapter 1

Introduction

The impact of the COVID-19 pandemic, and communities' responses to it, were shaped in part by political ideology. Indeed, prior research suggests that communities in the United States with higher Republican vote share may have experienced higher COVID mortality rates (Chen & Karim, 2022; Kaashoek et al., 2022; Lhila & Alghanem, 2023; Sehgal et al., 2022; Zhuo & Harrigan, 2026). Since data on individual's voting behavior and COVID-19 mortality status are beyond the reach of publicly accessible data, most work in this area treats counties or similarly defined geographic regions as the relevant observational units (Wallace et al., 2023; Wallace et al., 2022). Accordingly, such work is not causal in nature, so as to avoid the ecological fallacy, and instead addresses a question for public health researcher and future pandemic planning: is political ideology robustly predictive of pandemic mortality at the level of U.S. counties? If the answer is affirmative, this pushes public health researchers and policy planners towards further study on what drives this phenomenon and how it might be useful in efficiently responding to future pandemics.

Even given the explicitly non-causal nature of the question, an affirmative answer would be uninteresting if the model's control set were inadequate. Spurious correlations are ubiquitous, and both county Republican vote share and COVID-19 mortality are plausibly related to population density,

age structure, racial and ethnic composition, poverty, healthcare capacity, vaccination coverage, and broader social vulnerability (Cuadros et al., 2023; Foster et al., 2022; Hughes et al., 2021; Patel et al., 2020). Pandemic mortality is likewise a complex phenomenon associated with a relatively large number of controls implying that data-driven feature selection is preferable to *ad hoc* methods. Further complicating the question is the problem of spatial dependence between observational units. Given that most counties in the U.S. share borders with other counties wherein people and goods flow across such borders daily, assuming independence of observations is implausible on its face. Hence, any adequate model for this question ought to explicitly model spatial dependence between counties.

The significance of the framing question only increases as political polarization and public conflict over health policy continue to shape how communities respond to infectious disease threats. While the COVID-19 pandemic represents only one data point, its sheer spatial and temporal scale and its recency demands outsized attention with respect to the question motivating this project. The remainder of this chapter reviews the existing literature, identifies the thesis’s substantive and methodological contributions, and summarizes the data collection and preprocessing pipeline.

1.1 Existing Work

COVID-19 mortality differed sharply across U.S. counties, especially in the first two pandemic years. Standard explanations—age structure, baseline health, poverty, crowded housing, and health-care access—matter, but they do not fully capture the behavioral and policy environments that shape exposure and downstream outcomes. County political orientation—typically measured using Republican vs. Democratic presidential vote share—has therefore been used as a proxy for local “pandemic culture”: risk perception, willingness to adopt non-pharmaceutical interventions (NPIs), trust in public health guidance, and the policy choices made by elected officials.

This literature is overwhelmingly ecological. County partisanship is not an individual trait, and county-level associations cannot be interpreted as individual-level causal effects without strong additional assumptions. Many studies acknowledge this explicitly and interpret political variables as population markers entangled with social vulnerability and governance. The present thesis adopts the same stance: it tests whether Republican vote share is a robust county-level predictor of mortality while avoiding individual-level causal claims.

Across studies, the partisanship–mortality association is not stable. A common trajectory is (i) early concentration of mortality in dense, globally connected metropolitan areas (often more Democratic-leaning), followed by (ii) later periods in which mortality is higher in more Republican-leaning counties. This “flip” is consistent with a shift from exposure driven by initial seeding and urban connectivity to exposure shaped more by sustained preventive behavior and policy polarization.

(Kaashoek et al., 2022) analyze February 2020–February 2021 and organize the year into three phases using clustering of county death trajectories (approximately Feb–May 2020, Jun–Sep 2020, and Oct 2020–Feb 2021). Their central descriptive result is that predictors change across phases: early mortality is concentrated in urban, Democratic-leaning counties, while later mortality becomes more strongly associated with Republican vote share, especially during the winter wave. They also emphasize policy context: in the third period, Republican-leaning counties in states with looser mask mandates experienced substantially higher death rates—reported as “up to 3 times higher” than Democratic-leaning counties—after adjustment for multiple vulnerability measures. Methodologically, they separate introduction (whether a county is seeded) from severity (mortality intensity once seeded), and they use multiple modeling strategies (e.g., regularization and machine learning), with spatial models included in supporting analyses.

(Chen & Karim, 2022) also document a reversal within 2020 using a county panel split into ten 30-day periods (Jan 22–Dec 5, 2020). They define partisanship dichotomously based on the 2016 presidential election majority and model deaths per period with a zero-inflated negative

binomial (ZINB), motivated by both overdispersion and many zero-death observations early in the pandemic. Their estimates show very large early differences (Democratic counties higher mortality in early windows), followed by a late-2020 reversal (Democratic counties lower mortality in the last two windows). Substantively, they interpret this pattern as consistent with politically polarized behaviors and policy environments becoming increasingly consequential as the pandemic progressed. Methodologically, their explicit treatment of “excess zeros” is directly relevant for county-period mortality modeling.

(Zhuo & Harrigan, 2026) extend the time-varying association framework by explicitly testing whether the vote–mortality relationship differs across pandemic waves. Using 3,108 counties and 60 two-week periods, they estimate multilevel negative binomial models and report little/no association in an early wave (Feb–Aug 2020), with a stronger positive association between Republican vote share and mortality in subsequent waves. They also stress-test interpretation using robustness probes (e.g., dose-response and falsification tests), emphasizing that Republican vote share correlates with other health outcomes—an important reminder that stable ecological associations do not, by themselves, identify a causal political mechanism.

Despite convergence on the broad substantive pattern (weak/reversed early association; stronger positive association later), the county-level literature shares constraints that shape inference and motivate the present thesis.

Outcome modeling and the “zero problem:” County mortality data, especially in short windows or small counties, feature rare events and many zero counts. Studies address this in different ways: (Chen & Karim, 2022) explicitly use ZINB; (Zhuo & Harrigan, 2026) use negative binomial models in short panels; (Kaashoek et al., 2022) separate introduction from severity; (Lhila & Alghanem, 2023) use cumulative rates over a long window. Collectively, these choices reflect a shared recognition that crude rates can be unstable when denominators are small and deaths are

rare. This thesis responds by using both (i) stabilized rate transformations and (ii) count models with population offsets.

Confounding control and model dependence: . All studies adjust for socioeconomic and demographic factors, but adjustment is limited by (a) hand-selected covariate sets and (b) the difficulty of capturing the full bundle of county attributes correlated with both vote share and mortality. (Kaashoek et al., 2022) assemble a large covariate set and use regularization/ML to manage redundancy and nonlinearity, but these tools are primarily used for prediction and feature importance rather than to construct a targeted adjustment set for a focal exposure. (Chen & Karim, 2022) rely on a more conventional adjustment strategy with a binary political measure. (Lhila & Alghanem, 2023)’s decomposition clarifies how much of the gap is compositional versus residual but does not eliminate concerns about unobserved confounding. (Zhuo & Harrigan, 2026) explicitly emphasize that vote share tracks broader health-relevant characteristics, reinforcing why causal language should be constrained in ecological designs. These patterns motivate the present thesis’s use of high-dimensional adjustment procedures designed to reduce omitted-variable bias when the exposure (vote share) is correlated with a wide covariate space.

Spatial dependence and spillovers: Infectious mortality is spatially clustered: neighboring counties share commuting networks, hospital catchment areas, and synchronized waves. Ignoring spatial structure can produce overconfident inference and may leave unmeasured regional exposures in the residual. Among the reviewed studies, (Kaashoek et al., 2022) treat spatial autocorrelation as substantively real and implement adjacency-based spatial models in supporting analyses; (Zhuo & Harrigan, 2026) also treat residual spatial clustering as a methodological issue rather than assuming independence. By contrast, approaches that rely mainly on state fixed effects or state clustering address some shared policy context but do not directly model adjacency-driven dependence. This uneven treatment of spatial structure motivates the present thesis’s explicit spatial smoothing

component. Taken together, the existing literature provides strong descriptive evidence that county-level partisanship and COVID-19 mortality are linked, but leaves open how robust that association remains under richer confounding adjustment, distributionally appropriate count modeling, and explicit spatial correction.

1.2 Contributions

The existing county-level literature supports a consistent descriptive claim: the association between Republican vote share (and related partisan measures) and COVID-19 mortality is weak or reversed early in 2020, then becomes positive and larger in later waves, alongside evidence that policy and behavior plausibly mediate part of this association. The present thesis extends this work in three linked ways:

High-dimensional confounding adjustment: Rather than relying on a limited covariate set (or using regularization primarily for prediction), the thesis uses post-double-selection (Double-LASSO) to build a principled adjustment set predictive of either mortality or Republican vote share before estimating the focal association.

Rare-event stability and measurement: Motivated by the prevalence of zeros and denominator instability, the thesis combines smoothed mortality-rate transformations (with pseudo-counts/pseudo-population offsets, interpretable via a Poisson–Gamma empirical-Bayes lens) with count-based negative binomial models using population offsets.

Explicit spatial structure: Because residual spatial autocorrelation persists even after rich adjustment, the thesis incorporates an adjacency-based areal Markov random field smooth to capture

remaining spatial dependence while modeling deaths via negative binomial regression.

Substantively, this design enables a more stringent test of whether Republican vote share remains a robust county-level predictor of cumulative COVID-19 mortality through May 2022, spanning multiple waves, widespread vaccine availability, and shifting policy regimes that are not covered by studies ending in 2020 or early 2021. Throughout, the thesis maintains an explicitly ecological interpretation aligned with what county data can credibly support.

Chapter 2

Methods

2.1 Double-Lasso Control Selection

Our initially assembled dataset is high-dimensional, containing 127 features describing each of the $n = 3,109$ counties’ age and racial/ethnic demographics, income and poverty characteristics, population density, and related covariates. Given the relatively high candidate-feature-to-observation ratio, some form of principled feature selection is advisable to reduce variance in estimated coefficients and to mitigate overfitting. Existing work (Section 1.1) appears to rely primarily on domain knowledge to select relevant controls. While this is a sensible starting point, it would be impractical to manually screen the full feature set in our setting.

Instead, we use a modified version of the post-double-selection (“Double-LASSO”) procedure introduced by (Belloni et al., 2014). In standard LASSO-based selection, one fits a linear model by minimizing the sum of squared errors subject to an $\mathcal{L}_1$ penalty, with the tuning parameter λ typically chosen by 10-fold cross-validation. The $\mathcal{L}_1$ penalty shrinks many coefficients toward zero and sets some exactly to zero, yielding a sparse set of selected predictors; one common follow-up

is to refit an unpenalized (“relaxed”) regression on the selected variables.

However, selecting controls solely for their predictive power for the outcome can fail to address omitted variable bias in the coefficient of a focal independent variable (the “treatment”). In particular, a confounder may strongly predict the focal independent variable while only weakly predicting the outcome conditional on other covariates; such variables are at risk of exclusion under outcome-only LASSO selection, yet omitting them can bias the estimated treatment effect. Double-LASSO addresses this by performing two separate LASSO selections: (i) regressing the outcome on the full set of candidate controls, and (ii) regressing the focal independent variable on the same set of controls. The final set of controls is taken to be the union of variables selected in these two steps, and inference is then performed by estimating an unpenalized regression of the outcome on the focal independent variable and this union set of controls.

Feature Engineering: Smoothed Log–Mortality Rate

Target definition For county i we model

$$Z_i = \log\left[\frac{\{D_i + 1\}}{\{P_i + 1,000\}} \times 100,000\right],$$

where D_i is cumulative COVID-19 deaths (January 2020–May 2022) and P_i is the 2021 population estimate.

Strict positivity: The logarithm is defined only for positive arguments. 13 U.S. counties had 0 reported COVID fatalities in the coverage period of our data, meaning their raw log-rates would be undefined. Adding one pseudo-death ensures $D_i + 1 > 0$ for *every* county, so the log–rate is always real-valued.

Variance stabilization for small populations: Low-population counties are prone to extreme $\frac{D_i}{P_i}$ variation. Consider Texas’s King County with population $P_i = 256$. Consider two cases of its

unadjusted mortality rate where the raw Death count D_i differs only by 1:

$$\frac{1}{256} \times 100,000 = 391 \quad \text{vs.} \quad \frac{0}{256} \times 100,000 = 0.$$

Now consider the rate difference using the adjustment formula. The +1 / +1,000 offset significantly reduces volatility:

$$\frac{D_i + 1}{P_i + 1,000} \times 100,000 = \begin{cases} 79 & (D_i = 0) \\ 159 & (D_i = 1) \end{cases}$$

Importantly, the adjustment does not introduce a one-sided distortion. Because constants are added to both deaths and population, the adjusted rate can move either upward or downward relative to the raw rate, depending on the original count and denominator. In very small counties, the adjustment shrinks extreme rates toward a common baseline; in large counties, where $P_i \gg 1000$, its effect is negligible.

Empirical–Bayes motivation: The adjustment strategy we employ also enjoys theoretical support from a Bayesian perspective. We can redescribe our adjustment method as Bayesian updating that exploits the Gamma-Poisson conjugacy. If we model the i th county’s death count D_i conditional on rate parameter λ_i and population P_i under a Poisson likelihood, $D_i \mid \lambda_i \sim \text{Poisson}(\lambda_i P_i)$, with a Gamma prior on λ_i , $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, then our adjusted raw rate parameter is just the posterior mean under Bayesian updating:

$$\hat{\lambda}_i = \frac{D_i + \alpha}{P_i + \beta}.$$

Setting $\alpha = 1$, $\beta = 1,000$ yields exactly $\frac{D_i+1}{P_i+1,000}$: a *weak*, Gamma(1,1000) prior suggesting one death per 1,000 people baseline. From an empirical-Bayes perspective, this adjustment is closely akin to the shrinkage principles used in Good–Turing or Jiang–Lahiri smoothing for rare-event rates: it tames extreme estimates by pulling them toward a common baseline, while exerting minimal influence on large-population counties (Good, 1953), (Jiang & Lahiri, 2006). An application

of this method to rare-event rate modeling with counties as observational units can be found in (Gelman et al., 2013). This stabilized log-rate response is used in the selection stage to construct the adjustment set; final inference is then conducted on raw death counts using a negative binomial model with population offset.

2.2 Negative Binomial Likelihood

Traditionally, post-double-selection (“Double-LASSO”) proceeds by estimating the final, unpenalized model using ordinary least squares—implicitly invoking a Gaussian error model for the outcome regression used for inference. When we implemented this approach, even after applying a log transformation to the outcome, the residuals remained strongly non-normal, with pronounced skewness and heavy tails. This diagnostic pattern indicates that the transformed-outcome linear model is a poor approximation to the conditional distribution of the data and that standard OLS-based standard errors and tests may be unreliable for inference about the focal independent variable.

A more basic issue is that the outcome here is a count generated by an underlying event process: deaths are nonnegative integers whose variance typically increases with the mean. In such settings, modeling $\log(\text{deaths})$ with Gaussian errors is not merely a convenience—it is a substantive approximation that can fail when counts are small, when heteroskedasticity is severe, or when there is excess variability beyond what a Poisson process would predict. The observed heavy-tailed residuals are consistent with this mismatch: the linear-Gaussian model struggles to accommodate the combination of discreteness, strong mean–variance coupling, and heterogeneous county-level risk.

For these reasons, we instead model deaths on their natural scale using a negative binomial generalized linear model. Specifically, we model $Y_i = \text{TotalDeaths}_i$ with a log link and include

$\log(\text{Population}_i)$ as an offset, so that the linear predictor governs the log death rate rather than the log count: $\log(\mathbb{E}[Y_i | X_i]) = \log(\text{Population}_i) + \eta_i$, equivalently $\log\left(\frac{\mathbb{E}[Y_i | X_i]}{\text{Population}_i}\right) = \eta_i$. This specification preserves the original inferential target—a multiplicative effect on mortality rates—while using a likelihood appropriate for count outcomes. The negative binomial model also explicitly accommodates overdispersion through an additional dispersion (shape) parameter, allowing $\text{Var}(Y_i | X_i)$ to exceed $\mathbb{E}[Y_i | X_i]$. In other words, rather than forcing the data into a Gaussian framework via transformation, we adopt a distributional model that aligns with the measurement scale and variance structure of county death counts, yielding more defensible uncertainty quantification for the effect of the focal independent variable. In the following section, we conclude the methods section by expanding the negative binomial model to explicitly capture spatial dependence amongst the counties.

2.3 Modeling Spatial Dependence

In the standard post-double-selection (“Double-LASSO”) workflow, one would then fit an unpenalized outcome regression on the focal independent variable and the union of selected controls. While this addresses both selection pathways that can generate omitted-variable bias, it does not explicitly represent the a priori expectation that outcomes are spatially dependent across counties. For an infectious disease such as COVID-19, transmission and shared exposures routinely cross county borders, so a model that conditions only on observed covariates can miss an important component of the data-generating process. Likewise, clustering standard errors at the state level does not resolve dependence at the finer geographic scale at which spillovers plausibly operate.

To capture this dependence directly, we incorporate an areal spatial effect that partially pools information across adjacent counties via a Markov random field (MRF) smooth. We construct a county-to-county adjacency matrix from a U.S. county shapefile and use it to define the neigh-

neighborhood graph underlying the MRF penalty. The resulting regularization shrinks neighboring counties’ latent effects toward one another—allowing localized departures when supported by the data—while standard identifiability constraints (e.g., centering the spatial effects) prevent the spatial term from being confounded with the intercept and other additive components of the model.

2.4 Methods Summarized

We analyze county-level COVID-19 mortality using a three-stage modeling strategy designed for high-dimensional covariate adjustment, rare-event rate stabilization, and spatial dependence. First, because the dataset contains 147 candidate controls for $n = 3,096$ counties, we use the post-double-selection (“Double-LASSO”) procedure to construct a parsimonious but defensible adjustment set. This method performs two separate LASSO selections—one predicting the outcome and one predicting the focal independent variable—and then takes the union of selected covariates, thereby reducing the risk of omitted-variable bias from confounders that are strongly related to the treatment but only weakly related to the outcome.

Second, we define a smoothed log–mortality-rate target that is well-defined for all counties and less volatile for very small populations. The target replaces the raw rate D_i/P_i with $(D_i + 1)/(P_i + 1000)$, scaled per 100,000 residents and log-transformed. This adjustment guarantees strict positivity even for counties with zero reported deaths, dampens extreme rate fluctuations driven by tiny denominators, and admits an empirical–Bayes interpretation as the posterior mean of a Poisson–Gamma model with a weak prior centered on one death per 1,000 residents.

Third, for inference we move beyond the Gaussian residual assumptions implicit in an OLS final stage. Diagnostics indicated that even log-transformed outcomes yielded heavy-tailed, markedly non-normal residuals, consistent with the discreteness and strong mean–variance coupling of count data. We therefore model death counts directly using a negative binomial generalized linear model

with a log link and a population offset, which preserves the interpretation in terms of mortality rates while explicitly allowing overdispersion. Finally, because infectious-disease outcomes plausibly exhibit spatial spillovers across county borders, we augment the model with an areal spatial effect implemented as a Markov random field smooth based on a county adjacency matrix derived from a U.S. county shapefile. This term partially pools neighboring counties' latent effects through a graph-based penalty, capturing residual spatial dependence at the appropriate geographic scale while maintaining identifiability via standard centering constraints.

Chapter 3

Results

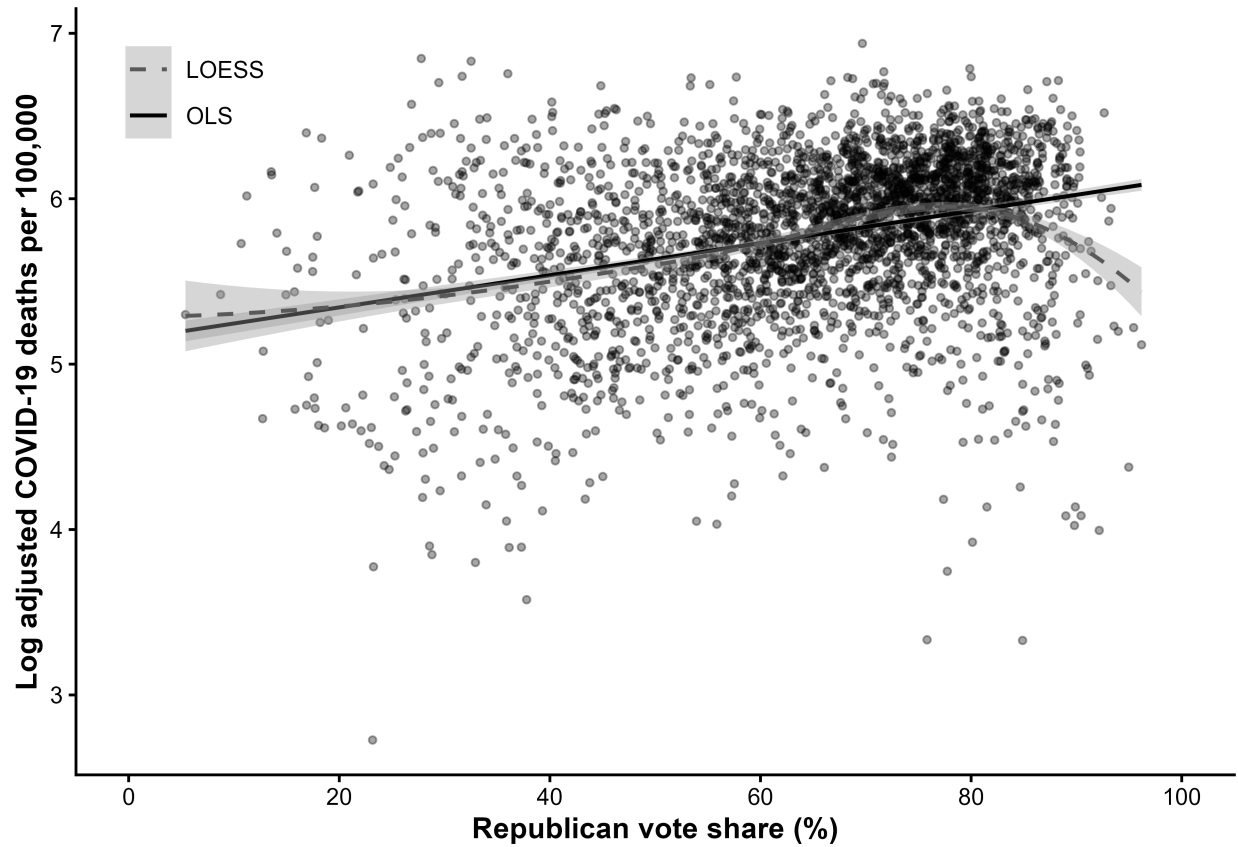
3.1 Data & EDA

The county-level data for this project were drawn from several original sources. Mortality data was sourced from the New York Times COVID mortality database which concatenated and cross-referenced local, state, and federal data reports (The New York Times, 2025). Data on age-adjusted incidence rates of comorbidities and Social Determinants of Health were extracted from the PLACES project by the U.S. Center for Disease Control and Prevention (CDC) (Centers for Disease Control and Prevention, 2025). Election data was sourced from (Algara & Amlani, 2021) and AP (POLITICO LLC, 2020). Data on county-populations and age distributions were drawn from the U.S. Census Bureau’s annual reported estimates (National Cancer Institute SEER Program, 2025). Hospital infrastructure and physician data were extracted from the U.S. Department of Health and Human Services Area Health Resource Files (AHRF) (Health Resources and Services Administration, 2025). Data on vaccination coverage was sourced from the CDC (Hughes et al., 2021). The remaining data on county-level characteristics is drawn from the U.S. Department of Agriculture’s Economic Research Service (U.S. Department of Agriculture, Economic Research

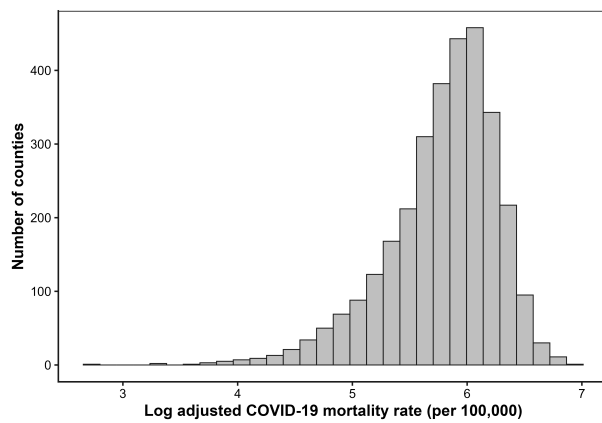
Service, 2023).¹ Post data cleaning, we have 3,096 complete cases (counties) covering all States in the contiguous U.S. (treating D.C. as a state) which collectively contain 3,109 counties—a coverage rate of $\approx 99\%$. The shape file used to create a county adjacency matrix was sourced from the U.S. Census Bureau (U.S. Census Bureau, 2021).

All data sets were merged by county FIPS code, the unique identifier for every U.S. county. Of the nation’s 3,144 counties, we dropped Alaska’s 30 county-equivalents because the state does not publish county-level election returns—our focal independent variable—and removed Hawaii’s 5 counties owing to systematic missingness in key covariates. The merge therefore covered 3,109 counties. A final row-wise deletion of sporadic blanks in socioeconomic or health variables trimmed the sample to $n = 3,096$ or approximately 99% of all continental-U.S. counties plus the District of Columbia. Given this near-complete coverage and the geographic isolation of the excluded states, we judged imputation unnecessary and theoretically unwarranted. Figure 3.1 gives the marginal distributions of the adjusted log COVID death-rate per 100k and county-level Republican vote-share and their scatterplot. Raw cumulative county-level deaths range from 0 to 32,022 (Median = 101, Mean = 304.9) whereas Republican vote-share ranges from 5.4%-96.2% (Median = 68.4%, Mean = 65.1%). The adjusted log death rate has a moderate positive raw correlation with Republican vote share ($r = 0.32$), consistent with the upward trend visible in Figure 3.1 but not yet informative about confounding or spatial dependence. Finally, Figure 3.2 plots two county-level choropleth maps of the contiguous U.S., one showing the distribution of Republican vote-share and the other of cumulative COVID deaths/100k. In both we observe noticeable spatial clustering, with counties tending to take similar values to their neighbors with respect to both variables. This EDA motivates our model which includes a smoothed Markov Random Field over the spatial structure of the data.

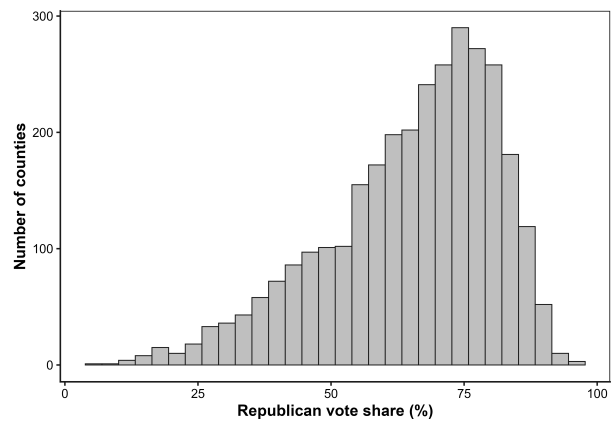
¹Our thanks to Dr. Linda Zhao and her team of RAs for extracting these covariates from the ERS database.



(a) County-level association between Republican vote share and adjusted log deaths per 100,000.

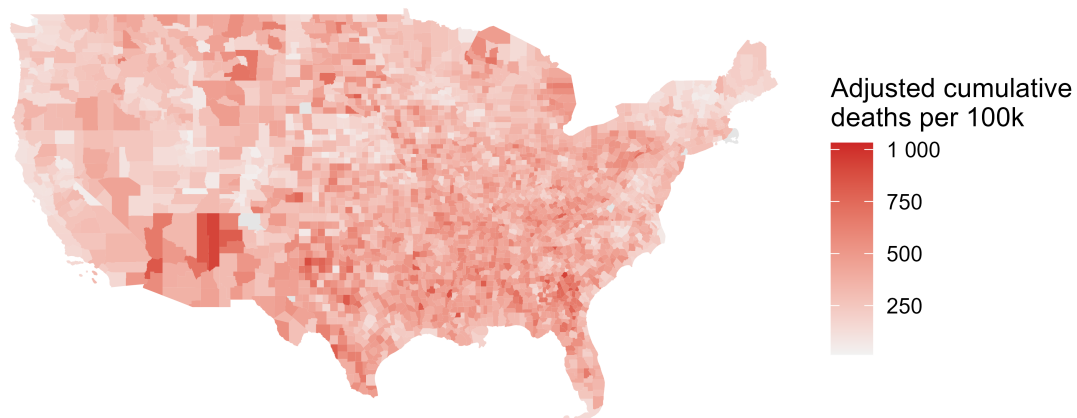


(b) Adjusted log deaths per 100,000 (distribution).

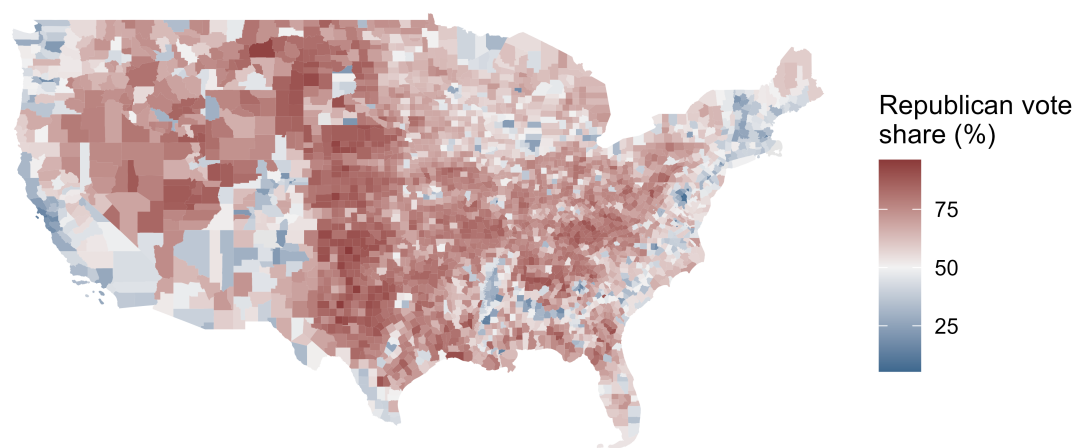


(c) Republican vote share (%) (distribution).

Figure 3.1: County-level bivariate association and marginal distributions for partisanship and COVID-19 mortality.



(a) Adjusted COVID-19 mortality (per 100k).



(b) Republican vote share.

Figure 3.2: Spatial distributions of county-level COVID-19 mortality and partisanship (contiguous U.S.).

3.2 Model Fitting Procedure

3.2.1 Post-double Selection Procedure

With $p = 127$ candidate features and $n = 3,096$, we ran the post-double selection procedure laid out in detail in Technical Appendix A. Each LASSO procedure was fit using the `cv.glmnet()` function from the `glmnet` package in R. We tuned λ using 10-fold cross-validation with the objective of minimizing mean-squared error (1 SE rule for λ). We chose to force-in the following features for theoretical reasons: State, Rural Urban Census Code, Average Vaccination Coverage, $\$ > \65 Vaccination Coverage, Population % 65–74, Population % 75+, White Population %, Black Population %, Hispanic Population %, Poverty Rate, Log Median Household Income, % Population with 4+ Year Degree, Obesity Prevalence, Diabetes Prevalence, Hospital Beds/1k, General Practitioners/1k, and Log Population Density. We forced in these features by passing a penalty vector to both `cv.glmnet()` calls containing 0 for the relevant indices in the model matrix and 1 otherwise. We elected to force in features that prior work suggests are predictive of our response while still benefiting from an overall data-driven feature-selection procedure. The union of the two selection steps contained 48 features drawn from multiple substantively relevant domains, including age and racial/ethnic composition, vaccination coverage, comorbidity prevalence, socioeconomic conditions, healthcare capacity, employment structure, and geographic controls. This breadth is consistent with the multifactorial character of pandemic mortality and provides some reassurance that the post-double-selection procedure recovered a rich conditioning set rather than a narrowly predictive collection of covariates.

3.2.2 Generalized Additive Model

To deal with the non-normality of errors in the Gaussian model discussed in the Introduction, we employed a negative binomial generalized additive model with log-link wherein we regress county-level cumulative deaths with log population offset on the set of features selected in the post-double selection procedure. To account for spatial-autocorrelation amongst counties, we included a smoothed Markov random field (MRF) over a queen adjacency matrix built from a Census year 2020 U.S. county shapefile. This enforces partial pooling amongst neighboring counties. We fit the model using the `gam()` function from the `mgcv` package in R.

The algorithm converged successfully and estimated the dispersion parameter $\theta = 33.705$. A dispersion test from the `DHARMA` package was not significant ($p = 0.586$) and a bootstrap test for outliers was likewise non-significant ($p = 0.56$). Crucially, we employed a Moran's I test from the `spdep` package for spatial autocorrelation in the residuals. This test takes as input an adjacency matrix of the spatial units and the residuals from the fitted model to determine if there is non-zero correlation between neighboring units that was not controlled for by the covariates in the model. We failed to reject the null of no-spatial autocorrelation ($p = 0.1761$), providing evidence that the MRF smooth successfully controlled for spatial autocorrelation. Importantly, we re-fit the model specified above but without the MRF term and a Moran's I test on this no-spatial model was highly significant ($p < 2.2 \times 10^{-16}$) suggesting that our *a priori* concerns about spatial autocorrelation are vindicated by the data. In the preferred specification, the adjacency-based MRF smooth had an effective degrees of freedom of 298.3, indicating that the model retained substantial residual spatial structure while remaining far less flexible than the near-saturated spatial fits obtained under weaker smoothing.

3.3 Key Results

In the preferred negative binomial model with a log link and population offset, the coefficient on Republican vote share is positive and precisely estimated ($\hat{\beta} = 0.01194$, $SE = 0.000925$, $z = 12.91$, $p < 2 \times 10^{-16}$). Because the model is log-linear in the mean, this coefficient has a multiplicative interpretation on the death rate: holding constant the selected covariates, state fixed effects, and the spatial MRF term, a one–percentage–point increase in county Republican vote share is associated with a factor of $\exp(0.01194) \approx 1.012$, i.e., roughly 1.2% higher expected COVID-19 mortality rate. Put differently, a 10–point increase in Republican vote share corresponds to about a 12.7% higher expected mortality rate, all else equal. A 95% confidence interval for the per–point rate ratio is (1.010, 1.014) which remains substantively and statistically positive.

We include a table of the remaining coefficients in Appendix B, but caution that because the post–double–selection procedure is designed to improve inference for the focal coefficient on Republican vote share rather than to produce a fully theory–driven structural model for every covariate, the remaining coefficients should be interpreted descriptively and with caution. Figure 3.3 plots the fixed effects and 95% intervals estimated for State with Alabama as the reference level. The state fixed effects suggest residual between–state heterogeneity, but the focal political association operates at a finer county–level scale than a pure state–level summary would capture.

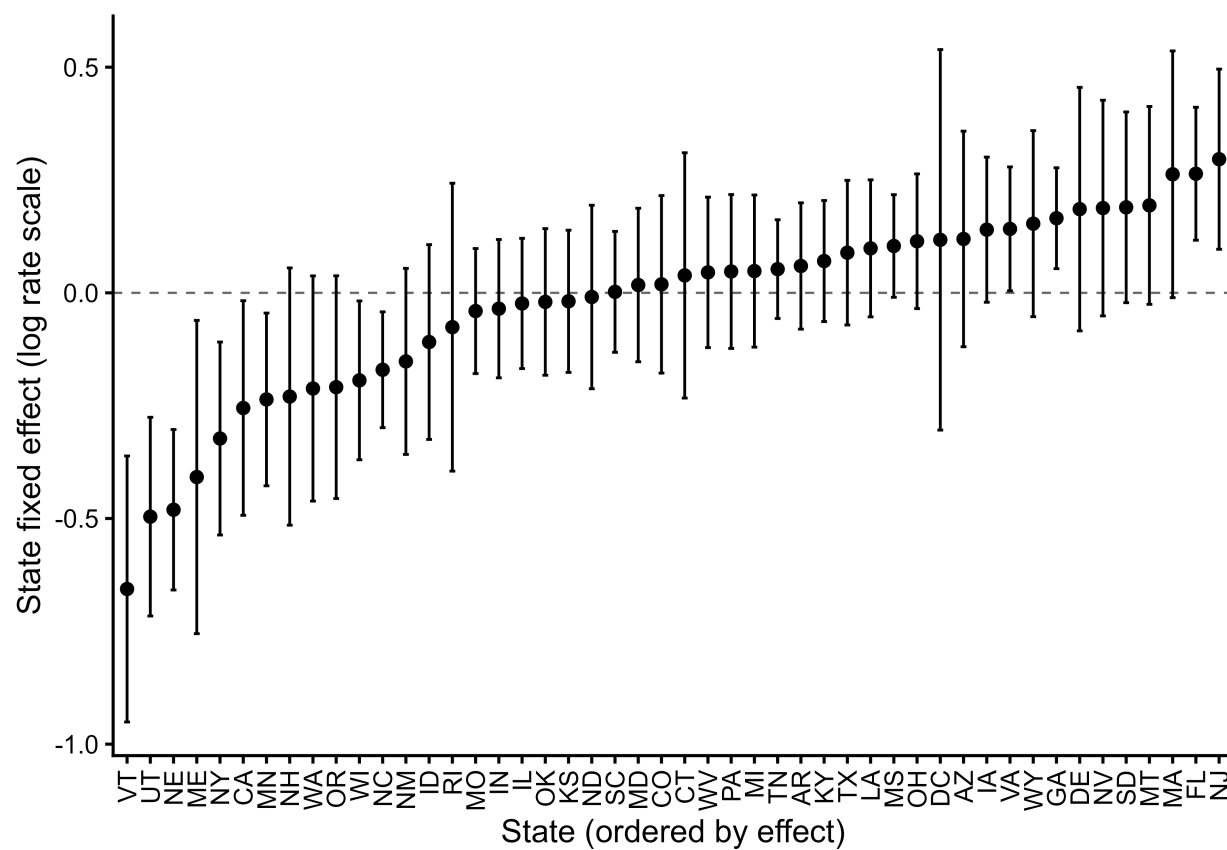


Figure 3.3: Estimated state fixed effects from the negative binomial GAM, ordered from smallest to largest, with 95% confidence intervals. Effects are on the log-rate scale relative to the reference state (Alabama).

Chapter 4

Discussion

4.1 Discussion

The central finding of this thesis is that county-level Republican vote share remains a strong positive predictor of cumulative COVID-19 mortality through May 2022 even after imposing a substantially more demanding adjustment strategy than is typical in the existing literature. In the preferred negative binomial model with population offset, high-dimensional post-double-selection controls, state fixed effects, and an adjacency-based Markov random field smooth, the estimated coefficient on Republican vote share is positive, precisely estimated, and substantively meaningful: a ten percentage-point increase in Republican vote share is associated with an estimated 12–13% increase in expected COVID-19 mortality, holding constant the selected county-level covariates. The result is both statistically precise and methodologically robust: it survives a modeling strategy explicitly designed to address three major threats to naive ecological inference in this setting: high-dimensional confounding, rare-event instability, and residual spatial dependence.

Substantively, this finding reinforces the broader pattern documented in recent county-level and

voter-level work on the politics of the pandemic. Prior studies suggest that the association between partisanship and COVID-19 mortality was weak, reversed, or unstable in the earliest phase of the pandemic, but became increasingly positive as the pandemic evolved and as policy responses, behavioral adaptation, vaccine uptake, and information environments became more politically differentiated. The present thesis is broadly consistent with that literature, but extends it in an important way: the positive association persists in a cumulative county-level analysis spanning January 2020 through May 2022 even after richer adjustment for county characteristics and explicit modeling of spatial dependence. In that sense, the result strengthens the descriptive case that political ideology, or social and institutional features tightly correlated with it, was systematically related to pandemic mortality at the county level.

At the same time, the result should be interpreted with care. This thesis does not estimate the causal effect of individual Republican voting on individual mortality risk, nor can it isolate a single mechanism linking county political composition to county mortality. County Republican vote share is best understood here as an ecological proxy for a bundle of correlated factors that may include local policy responses, risk perceptions, vaccine uptake, social norms, media ecosystems, trust in public health institutions, and broader demographic or institutional differences not fully captured by the observed covariates. The strength of the present analysis lies not in resolving those mechanisms, but in showing that the association remains robust even after a deliberately stringent attempt to account for observed confounding and residual geographic structure.

One of the more important methodological findings is that spatial dependence is not a cosmetic issue in these data. The non-spatial negative binomial model leaves strong residual spatial autocorrelation, whereas the model including the MRF spatial term substantially attenuates this pattern. This suggests that county mortality outcomes cannot be adequately treated as conditionally independent once one accounts only for observed county covariates and state fixed effects. Given the geography of infectious disease transmission, regional health-system capacity, commuting patterns,

and shared policy environments, that result is substantively plausible. More importantly, it indicates that models which omit spatial structure in this setting risk overstating precision or attributing geographically patterned residual variation to measured covariates. The present results therefore support not only the substantive association of interest, but also the methodological claim that adjacency-based spatial adjustment is warranted in analyses of county-level pandemic mortality data with similar geographic structure.

A second important methodological implication concerns the construction of the adjustment set. Existing county-level studies typically rely on hand-selected controls motivated by substantive judgment, which is reasonable but difficult to scale in high-dimensional settings. Here, the post-double-selection procedure provides a more principled alternative by selecting covariates predictive of either the outcome or the focal independent variable and then carrying their union into the final model. This matters because omitted-variable bias can arise not only from variables strongly predictive of mortality, but also from variables strongly predictive of county partisanship that only weakly predict mortality conditional on other covariates. The present analysis does not claim that post-double-selection eliminates all confounding, but it does substantially improve the credibility of the conditioning set relative to ad hoc manual screening.

Finally, the negative binomial specification appears substantively and diagnostically appropriate for the outcome at hand. County death counts are discrete, highly skewed, and strongly overdispersed relative to a Poisson benchmark, and diagnostics from the transformed-outcome Gaussian model indicated substantial mismatch. The negative binomial model with population offset therefore offers a better-aligned inferential framework for the mortality-rate question of interest. This interpretation is supported by simulation-based residual diagnostics, which showed no evidence of residual dispersion problems once the negative binomial specification and spatial term were included. In this respect, the thesis also contributes a modeling lesson: in county-level mortality analyses with large variation in population size, it is often preferable to model counts directly with an exposure

offset and explicit overdispersion rather than rely on transformed rates alone.

4.2 Limitations

Several limitations qualify the interpretation of these results. First, the design is ecological. All inferences operate at the county level, and county Republican vote share cannot be interpreted as a proxy for the causal effect of individual partisan identity. Counties are heterogeneous social units, and the mechanisms linking county political composition to county mortality may operate through institutions, policies, social norms, health behaviors, or composition effects that cannot be disentangled with publicly available county aggregates alone. Accordingly, the findings should be interpreted as robust county-level associations, not individual-level causal effects.

Second, the analysis uses cumulative mortality through May 2022 rather than a dynamic panel of wave-specific outcomes. This choice has advantages: cumulative mortality is less sparse, less sensitive to short-window zero inflation, and better aligned with the thesis’s emphasis on broad county-level mortality burdens over the pandemic as a whole. However, cumulative outcomes also compress important temporal variation. They cannot reveal whether the association between partisanship and mortality differed sharply across early, middle, and late pandemic phases, nor can they isolate whether particular shifts in policy, vaccination, or public behavior account for the aggregate relationship observed here. The result should therefore be understood as a summary association over a long and heterogeneous pandemic period.

Third, although the covariate set is rich and the post-double-selection procedure is specifically designed to improve confounding adjustment, unobserved confounding remains possible. County political composition is correlated with many difficult-to-measure features, including local institutional capacity, health-seeking behavior, media environments, social trust, and informal norms regarding mitigation and vaccination. The present modeling strategy reduces the risk that the main

result is driven by an impoverished or ad hoc control set, but it cannot rule out all omitted-variable bias. This is especially important because the focal variable is socially and politically loaded, making over-interpretation particularly tempting.

Fourth, the spatial term should not be interpreted too strongly. The MRF smooth captures residual spatial dependence and improves model adequacy, but it does not by itself identify the underlying source of that dependence. Residual geographic clustering may reflect contagion, regional policy regimes, commuting networks, hospital referral patterns, shared data-reporting practices, or unmeasured covariates with spatial structure. The thesis therefore treats the MRF primarily as a correction for residual spatial dependence, not as a direct estimate of a specific geographic mechanism.

Fifth, the post-double-selection procedure is used here as a principled screening device, while the final inference stage is conducted under a negative binomial likelihood with spatial smoothing. This workflow is defensible and substantively motivated, but it is not identical to the simplest linear-model presentation of post-double-selection in (Belloni et al., 2014). The logic of the method still applies at the level of confounder screening, but formal properties for this exact hybrid implementation are more limited than in the canonical linear setting. For that reason, the results are best understood as strong and carefully stress-tested evidence of robustness rather than as an exact semiparametric identification result.

Finally, the geographic scope is nearly complete but not literally exhaustive. Alaska and Hawaii are excluded for data reasons, and a small number of counties are lost to missing covariates. Given the near-complete coverage of the contiguous United States plus the District of Columbia, this omission is unlikely to alter the broad pattern, but it does mean the thesis is not a census of all U.S. counties in the strictest sense.

4.3 Future Directions

Several directions for future work follow naturally from the present analysis. The most immediate extension is to move from cumulative mortality to a county-by-time panel that estimates how the association between county political composition and mortality evolves across pandemic waves. Such a design would permit a more direct comparison with studies that report early reversals and later positive associations, and could help determine whether the cumulative relationship estimated here is driven primarily by specific phases of the pandemic.

A second priority is to move from association to mechanism. The present results are consistent with several candidate pathways, including differences in vaccination uptake, mitigation behavior, policy responsiveness, healthcare access, and media or information environments. Future work could assess these mechanisms more directly by combining county-level mortality with richer panel data on vaccine uptake, mask usage, mobility, hospitalization capacity, and policy timing, or by embedding the present county-level analysis within a mediation framework. Even then, strong causal claims would remain difficult, but the field would be better positioned to distinguish competing explanations for the association documented here.

Third, future work could explore alternative spatial specifications. The MRF smooth used here is appropriate for areal county data and performs well diagnostically, but it would be informative to compare it with other approaches, including conditional autoregressive models, spatial lag/error models, or Bayesian hierarchical specifications that jointly model structured and unstructured spatial heterogeneity. Such comparisons would help clarify the extent to which the substantive conclusion depends on a particular representation of spatial dependence.

Fourth, the present thesis suggests a broader methodological agenda for high-dimensional ecological health research. Many questions in public health and political epidemiology involve outcomes shaped by a large set of correlated county or regional covariates, along with plausible spatial

dependence. The combination of post-double-selection, count-based generalized modeling, and explicit spatial smoothing may therefore be useful beyond the specific COVID-19 setting studied here. Similar workflows may be especially useful in ecological studies of hospitalization, excess mortality, chronic disease burden, and other county-level health outcomes where researchers face a large and correlated covariate space together with plausible residual spatial dependence.

More broadly, the public-health relevance of the present findings extends beyond the COVID-19 pandemic itself. If county political composition remains strongly associated with mortality even after rich adjustment, then ideology is not merely a background cultural variable; it is part of the practical landscape within which health communication, mitigation policy, and institutional trust operate. Future pandemic planning may therefore benefit from taking political context seriously as a predictor of vulnerability, even when the underlying mechanisms remain incompletely understood.

Appendix A

Technical Appendix A: Post-Double Selection Algorithm

Notation. Let $n = 3,096$ denote the number of counties and $p =$ the number of potential controls.

- Z_i – engineered outcome

$$Z_i = \log\left\{\left(\frac{\text{Deaths}_i + 1}{\text{Population}_i + 1,000}\right) \times 100,000\right\};$$

- T_i – focal IV: 2020 Republican vote share in county i ;
- $\mathbf{X}_i \in \mathbb{R}^p$ – remaining covariate vector;
- $\mathbf{S} \subseteq \{1, \dots, p\}$ – index set of controls retained for inference;
- $\mathcal{F}$ – index set of *theory-driven* confounders (vaccination, age 75+, race, poverty & income, obesity, diabetes, health-care capacity, population density, and the full block of State and RuralUrbanContinuumCode2013 dummies).

All variables in $\mathcal{F}$ are *forced in*—i.e. assigned a zero ℓ_1 penalty in both LASSO stages. State fixed

effects δ_{state} are therefore present in every equation.

Algorithm. Let $\mathbf{W}_i = (\mathbf{X}_{i,-T}, \delta_{\text{state}})$. Define a penalty vector $\omega \in \{0, 1\}^p$ with $\omega_j = 0$ for $j \in \mathcal{F}$ and $\omega_j = 1$ otherwise.

1. *Outcome-selection LASSO.*

$$\hat{\pi}(\lambda_Z) = \arg \min_{\pi} \frac{1}{2n} \sum_i (Z_i - \mathbf{W}_i^\top \pi)^2 + \lambda_Z \sum_j \omega_j |\pi_j|.$$

Choose $\hat{\lambda}_Z$ by 10-fold CV (1SE rule). Let $\mathbf{S}_Z = \{j : \hat{\pi}_j(\hat{\lambda}_Z) \neq 0\}$.

2. *Treatment-selection LASSO.*

$$\hat{\kappa}(\lambda_T) = \arg \min_{\kappa} \frac{1}{2n} \sum_i (T_i - \mathbf{W}_i^\top \kappa)^2 + \lambda_T \sum_j \omega_j |\kappa_j|.$$

Choose $\hat{\lambda}_T$ by 10-fold CV (1SE rule). Let $\mathbf{S}_T = \{j : \hat{\kappa}_j(\hat{\lambda}_T) \neq 0\}$.

3. *Union of controls.* $\mathbf{S} = \mathbf{S}_Z \cup \mathbf{S}_T$.

4. *Post-selection negative binomial model with spatial MRF.* Form the selected control vector $\mathbf{W}_{i,\mathbf{S}} = (1, T_i, \mathbf{W}_{i,\mathbf{S}})$, i.e., an intercept, the focal treatment T_i , and the subset of controls indexed by $\mathbf{S}$. Then estimate β_T by fitting the negative binomial log-link model with population offset and an areal MRF smooth:

$$Y_i \mid \mu_i \sim \text{NegBin}(\mu_i, \theta), \quad \log(\mu_i) = \log(P_i) + \beta_T T_i + \mathbf{W}_{i,\mathbf{S}}^\top \boldsymbol{\beta}_{\mathbf{S}} + u_{g(i)},$$

where P_i is the county population (exposure), $g(i)$ indexes county i 's GEOID, and $u_{g(i)}$ is a Markov random field effect defined on the county adjacency graph. The MRF term is

estimated via a quadratic difference penalty over neighboring counties,

$$\mathcal{P}(u) = \lambda_u \sum_{(a,b) \in E} (u_a - u_b)^2,$$

with an identifiability constraint (e.g., $\sum_a u_a = 0$) to prevent confounding with the intercept.

Inference for β_T is then conducted in this unpenalized final stage, conditional on the selected set $\mathbf{S}$, with θ and the smoothing parameter λ_u estimated (e.g., by REML).

Appendix B

Technical Appendix B: Feature Dictionary & Full Results Tables

B.1 Feature Dictionary

Table B.1: Classification of all candidate control variables considered in the post-double-selection procedure. Factor variables are reported at the parent-variable level rather than the dummy-column level.

Feature	Outcome LASSO	Treatment LASSO	Forced in	Final union	Status
Average vaccination coverage	Yes	Yes	Yes	Yes	Both
Black population (%)	Yes	Yes	Yes	Yes	Both
College degree or higher (%)	Yes	Yes	Yes	Yes	Both
Diabetes prevalence	Yes	Yes	Yes	Yes	Both
General practitioners per 1,000	Yes	Yes	Yes	Yes	Both
Hispanic population (%)	Yes	Yes	Yes	Yes	Both
Hospital beds per 1,000	Yes	Yes	Yes	Yes	Both
Log median household income	Yes	Yes	Yes	Yes	Both
Log population density	Yes	Yes	Yes	Yes	Both
Obesity prevalence	Yes	Yes	Yes	Yes	Both
Population aged 60–74 (%)	Yes	Yes	Yes	Yes	Both
Population aged 75+ (%)	Yes	Yes	Yes	Yes	Both
Poverty rate	Yes	Yes	Yes	Yes	Both
Rural–Urban Continuum Code	Yes	Yes	Yes	Yes	Both
State fixed effects	Yes	Yes	Yes	Yes	Both
Vaccination coverage, age 65+	Yes	Yes	Yes	Yes	Both
White population (%)	Yes	Yes	Yes	Yes	Both
Asian population (%)	No	Yes	No	Yes	Treatment only
Associate degree (%)	No	Yes	No	Yes	Treatment only
Asthma prevalence	No	Yes	No	Yes	Treatment only
Average household size	No	Yes	No	Yes	Treatment only
COPD prevalence	No	Yes	No	Yes	Treatment only
Coronary heart disease prevalence	No	Yes	No	Yes	Treatment only
Deep poverty rate	No	Yes	No	Yes	Treatment only
Employment in agriculture (%)	No	Yes	No	Yes	Treatment only
Employment in construction (%)	No	Yes	No	Yes	Treatment only
Employment in government (%)	No	Yes	No	Yes	Treatment only
Employment in information (%)	No	Yes	No	Yes	Treatment only
Employment in manufacturing (%)	No	Yes	No	Yes	Treatment only
Employment in services (%)	No	Yes	No	Yes	Treatment only
Employment in trade (%)	No	Yes	No	Yes	Treatment only
Employment in transportation (%)	No	Yes	No	Yes	Treatment only
Female-headed households (%)	No	Yes	No	Yes	Treatment only
Foreign-born population (%)	No	Yes	No	Yes	Treatment only
High cholesterol prevalence	No	Yes	No	Yes	Treatment only

Continued on next page

Table B.1: Classification of all candidate control variables considered in the post-double-selection procedure (*continued*).

Feature	Outcome LASSO	Treatment LASSO	Forced in union	Final union	Status
High school diploma only (%)	No	Yes	No	Yes	Treatment only
Less than high school (%)	No	Yes	No	Yes	Treatment only
Multiple-race population (%)	No	Yes	No	Yes	Treatment only
Net international migration rate, 2010–2019	No	Yes	No	Yes	Treatment only
Non-English households (%)	No	Yes	No	Yes	Treatment only
Owner-occupied housing (%)	No	Yes	No	Yes	Treatment only
Poor physical health prevalence	No	Yes	No	Yes	Treatment only
Population aged 45–60 (%)	No	Yes	No	Yes	Treatment only
SVI Theme 2	No	Yes	No	Yes	Treatment only
SVI Theme 3	No	Yes	No	Yes	Treatment only
Unemployment rate (2019)	No	Yes	No	Yes	Treatment only
Blood pressure medication prevalence	No	No	No	No	Excluded
Cancer prevalence	No	No	No	No	Excluded
Current smoking prevalence	No	No	No	No	Excluded
Employment in finance, insurance, and real estate (%)	No	No	No	No	Excluded
High blood pressure prevalence	No	No	No	No	Excluded
Kidney disease prevalence	No	No	No	No	Excluded
Log per capita income	No	No	No	No	Excluded
Native American population (%)	No	No	No	No	Excluded
No health insurance prevalence	No	No	No	No	Excluded
Poor mental health prevalence	No	No	No	No	Excluded
Routine checkup prevalence	No	No	No	No	Excluded
SVI Theme 1	No	No	No	No	Excluded
SVI Theme 4	No	No	No	No	Excluded
Stroke prevalence	No	No	No	No	Excluded

B.2 Full Results Tables

Table B.2: Parametric coefficient estimates from the negative binomial GAM: substantive covariates.

Feature	Estimate	Std. Error	p-value
Intercept	−2.3230	0.9732	0.017
Republican vote share	0.0119	0.0009	< 0.001
Population aged 60–74 (%)	−0.0163	0.0032	< 0.001
Population aged 75+ (%)	0.0702	0.0052	< 0.001
Average vaccination coverage	−0.0057	0.0019	0.003
Vaccination coverage, age 65+	0.0039	0.0011	< 0.001
Diabetes prevalence	0.0458	0.0142	0.001
Obesity prevalence	−0.0011	0.0022	0.621
Poverty rate	0.0030	0.0025	0.227
White population (%)	−0.0091	0.0018	< 0.001
Black population (%)	−0.0079	0.0015	< 0.001
Hispanic population (%)	−0.0030	0.0019	0.118
RUCC: 2	−0.0377	0.0175	0.031
RUCC: 3	−0.0184	0.0191	0.336
RUCC: 4	−0.0079	0.0216	0.716
RUCC: 5	−0.0025	0.0288	0.931
RUCC: 6	−0.0067	0.0203	0.740
RUCC: 7	−0.0249	0.0235	0.289
RUCC: 8	−0.0325	0.0270	0.229
RUCC: 9	−0.0357	0.0280	0.203
College degree or higher (%)	−0.0059	0.0022	0.009
Hospital beds per 1,000	0.0008	0.0011	0.482
General practitioners per 1,000	0.0771	0.0173	< 0.001
Log median household income	−0.4248	0.0726	< 0.001
Log population density	0.0588	0.0075	< 0.001
Population aged 45–60 (%)	0.0153	0.0042	< 0.001

Continued on next page

Table B.2: Parametric coefficient estimates from the negative binomial GAM: substantive covariates (*continued*).

Feature	Estimate	Std. Error	p-value
Asthma prevalence	0.0830	0.0239	< 0.001
Coronary heart disease prevalence	−0.0847	0.0349	0.015
COPD prevalence	−0.0467	0.0264	0.077
Chronic kidney disease prevalence	−0.0066	0.0035	0.060
High cholesterol prevalence	0.0089	0.0068	0.192
SVI Theme 2	0.0315	0.0234	0.178
SVI Theme 3	0.0209	0.0550	0.704
Unemployment rate (2019)	0.0120	0.0053	0.024
Deep poverty rate	−0.0013	0.0028	0.642
Asian population (%)	0.0052	0.0042	0.214
Multiple-race population (%)	−0.0457	0.0086	< 0.001
Average household size	−0.1652	0.0344	< 0.001
Female-headed households (%)	0.0148	0.0026	< 0.001
Employment in agriculture (%)	−0.0030	0.0021	0.156
Employment in construction (%)	−0.0077	0.0028	0.006
Employment in transportation (%)	0.0059	0.0031	0.057
Employment in government (%)	−0.0012	0.0024	0.617
Employment in information (%)	0.0081	0.0069	0.241
Employment in manufacturing (%)	−0.0003	0.0017	0.842
Employment in trade (%)	−0.0014	0.0025	0.578
Employment in services (%)	−0.0004	0.0019	0.842
Less than high school (%)	0.0057	0.0028	0.046
High school diploma only (%)	0.0052	0.0021	0.014
Associate degree (%)	0.0025	0.0035	0.475
Foreign-born population (%)	−0.0053	0.0026	0.043
Net international migration rate, 2010–2019	0.0005	0.0050	0.917
Non-English households (%)	−0.0007	0.0038	0.860
Owner-occupied housing (%)	0.0027	0.0011	0.020

Note. RUCC = Rural–Urban Continuum Code. The omitted reference category is RUCC 1.

Table B.3: State fixed effects from the negative binomial GAM. Coefficients are relative to the omitted reference state.

State	Estimate	Std. Error	p-value
State: Arizona	0.1194	0.1219	0.327
State: Arkansas	0.0595	0.0714	0.405
State: California	−0.2552	0.1213	0.035
State: Colorado	0.0188	0.1003	0.851
State: Connecticut	0.0386	0.1387	0.781
State: Delaware	0.1855	0.1376	0.178
State: District of Columbia	0.1175	0.2151	0.585
State: Florida	0.2639	0.0752	< 0.001
State: Georgia	0.1653	0.0570	0.004
State: Idaho	−0.1090	0.1101	0.322
State: Illinois	−0.0236	0.0737	0.749
State: Indiana	−0.0351	0.0782	0.654
State: Iowa	0.1400	0.0820	0.088
State: Kansas	−0.0187	0.0804	0.816
State: Kentucky	0.0705	0.0684	0.303
State: Louisiana	0.0986	0.0774	0.203
State: Maine	−0.4081	0.1771	0.021
State: Maryland	0.0174	0.0868	0.841
State: Massachusetts	0.2627	0.1395	0.060
State: Michigan	0.0482	0.0860	0.575
State: Minnesota	−0.2362	0.0976	0.016
State: Mississippi	0.1038	0.0580	0.073
State: Missouri	−0.0404	0.0707	0.568
State: Montana	0.1937	0.1118	0.083
State: Nebraska	−0.4806	0.0907	< 0.001
State: Nevada	0.1878	0.1219	0.123

Continued on next page

Table B.3: State fixed effects from the negative binomial GAM (*continued*).

State	Estimate	Std. Error	p-value
State: New Hampshire	−0.2298	0.1454	0.114
State: New Jersey	0.2961	0.1018	0.004
State: New Mexico	−0.1519	0.1052	0.149
State: New York	−0.3228	0.1091	0.003
State: North Carolina	−0.1704	0.0655	0.009
State: North Dakota	−0.0092	0.1037	0.929
State: Ohio	0.1143	0.0761	0.133
State: Oklahoma	−0.0201	0.0829	0.808
State: Oregon	−0.2090	0.1259	0.097
State: Pennsylvania	0.0473	0.0870	0.587
State: Rhode Island	−0.0760	0.1627	0.640
State: South Carolina	0.0022	0.0683	0.974
State: South Dakota	0.1895	0.1078	0.079
State: Tennessee	0.0525	0.0558	0.347
State: Texas	0.0890	0.0817	0.276
State: Utah	−0.4959	0.1123	< 0.001
State: Virginia	0.1417	0.0700	0.043
State: Vermont	−0.6562	0.1504	< 0.001
State: Washington	−0.2119	0.1273	0.096
State: West Virginia	0.0453	0.0851	0.594
State: Wisconsin	−0.1939	0.0898	0.031
State: Wyoming	0.1533	0.1052	0.145

Bibliography

- Algara, C., & Amlani, S. (2021). Replication data for: Partisanship & nationalization in american elections: Evidence from presidential, senatorial, & gubernatorial elections in the u.s. counties, 1872–2020 [V1; UNF:6:glfQoiLzpXDGTfErebfBIQ==]. <https://doi.org/10.7910/DVN/DGUMFI>
- Belloni, A., Chernozhukov, V., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, 28(2), 29–50. <https://doi.org/10.1257/jep.28.2.29>
- Centers for Disease Control and Prevention. (2025). *Places* [Accessed: 2025-04-22]. <https://www.cdc.gov/places>
- Chen, H.-F., & Karim, S. A. (2022). Relationship between political partisanship and COVID-19 deaths: Future implications for public health. *Journal of Public Health*, 44(3), 716–723. <https://doi.org/10.1093/pubmed/fdab136>
- Cuadros, D. F., Gutierrez, J. D., Moreno, C. M., Escobar, S., Miller, F. D., Musuka, G., Omori, R., Coule, P., & MacKinnon, N. J. (2023). Impact of healthcare capacity disparities on the covid-19 vaccination coverage in the united states: A cross-sectional study. *The Lancet Regional Health – Americas*, 18, 100409. <https://doi.org/10.1016/j.lana.2022.100409>
- Foster, T. B., Fernandez, L., Porter, S. R., & Pharris-Ciurej, N. (2022). *Age, sex, and racial/ethnic disparities and temporal-spatial variation in excess all-cause mortality during the covid-19*

- pandemic: Evidence from linked administrative and census bureau data* (tech. rep. No. CES-WP-22-18). U.S. Census Bureau, Center for Economic Studies. <https://www.census.gov/library/working-papers/2022/adrm/CES-WP-22-18.html>
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd). Chapman & Hall/CRC.
- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3–4), 237–264.
- Health Resources and Services Administration. (2025). *Area health resources files* [Rockville (MD): HRSA; accessed: 2025-04-07]. <https://data.hrsa.gov/topics/health-workforce/ahrf>
- Hughes, M. M., Wang, A., Grossman, M. K., & et al. (2021). County-level covid-19 vaccination coverage and social vulnerability — united states, december 14, 2020–march 1, 2021. *MMWR Morbidity and Mortality Weekly Report*, 70, 431–436. <https://doi.org/10.15585/mmwr.mm7012e1>
- Jiang, J., & Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15(1), 1–96. <https://doi.org/10.1007/BF02595419>
- Kaashoek, J., Testa, C., Chen, J. T., Stolerman, L. M., Krieger, N., Hanage, W. P., & Santillana, M. (2022). The evolving roles of US political partisanship and social vulnerability in the COVID-19 pandemic from february 2020–february 2021. *PLOS Global Public Health*, 2(12), e0000557. <https://doi.org/10.1371/journal.pgph.0000557>
- Lhila, A., & Alghanem, F. (2023). Along party lines: Examining the gubernatorial party difference in COVID-19 mortality rates in U.S. counties. *Preventive Medicine Reports*, 32, Article 102142, 102142. <https://doi.org/10.1016/j.pmedr.2023.102142>
- National Cancer Institute SEER Program. (2025). U.s. county population data [Originally from U.S. Census; downloaded from [urlhttps://seer.cancer.gov/popdata/download.html](https://seer.cancer.gov/popdata/download.html)].

- Patel, J. A., Nielsen, F. B. H., Badiani, A. A., Assi, S., Unadkat, V. A., Patel, B., Ravindrane, R., & Wardle, H. (2020). Poverty, inequality and covid-19: The forgotten vulnerable. *Public Health, 183*, 110–111. <https://doi.org/10.1016/j.puhe.2020.05.006>
- POLITICO LLC. (2020). *Connecticut election results 2020: Live map updates, voting by county & district* [Accessed: 2025-04-30]. <https://www.politico.com/2020-election/results/connecticut/>
- Sehgal, N. J., Yue, D., Pope, E., Wang, R. H., & Roby, D. H. (2022). The association between covid-19 mortality and the county-level partisan divide in the united states. *Health Affairs, 41*(6), 853–863.
- The New York Times. (2025). *Covid-19 data for u.s. counties* [Accessed 2025-04-26]. <https://github.com/nytimes/covid-19-data/blob/master/us-counties.csv>
- U.S. Census Bureau. (2021). *2020 TIGER/Line shapefiles* [All legal boundaries and names are as of January 1, 2020; released February 2, 2021]. Retrieved March 15, 2026, from <https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2020.html>
- U.S. Department of Agriculture, Economic Research Service. (2023). *Atlas of rural and small-town america*. USDA ERS.
- Wallace, J., Goldsmith-Pinkham, P., & Schwartz, J. L. (2023). Excess death rates for republican and democratic registered voters in florida and ohio during the covid-19 pandemic. *JAMA Internal Medicine, 183*(9), 916–923. <https://doi.org/10.1001/jamainternmed.2023.1154>
- Wallace, J., Goldsmith-Pinkham, P., & Schwartz, J. L. (2022, September). *Excess death rates for republicans and democrats during the covid-19 pandemic* (tech. rep. No. 30512). National Bureau of Economic Research. <https://doi.org/10.3386/w30512>
- Zhuo, J., & Harrigan, N. (2026). Political orientation and public health: The evolving relationship between republican vote share on COVID-19 mortality across pandemic waves. *Social Indicators Research, 181*, Article 31, 31. <https://doi.org/10.1007/s11205-025-03735-2>