Submitted. November 31

Predict series of counts with a lot of zeros.

* The outcome is discrete counts,

Should we require at least one suicide?

## New Approach

1. Fit Y(0)
2. Include multinomial propensity
   * Balancing weights
   * Quantiles?
3. Include prognostic factor
4. Include ST error
5. Predict

This allows for any sort of effect heterogeneity

Other ideas. Fit [Y(0), Y(1)] with expected value from 1 as an offset.

BUT

* There is uncertainty about the expect value.
* Nothing to be gained, unless some homogeneity assumption is introduced.
  + Or meta-distribution?

## TWFE

Wooldridge (2021) discuss TWFE and related method for the linear case.

He establishes the equivalence between TWFE estimators and pooled OLS with two-way Mundlack device (double deameaning ) .

* Adding both the unit-specific time series averages and the period-specific cross-sectional averages in a POLS regression reproduce the two-way FE estimates.
* Thus, including ever-treated (the unit-specific average), time (time-specific average), and the interaction, is also equivalent.

Define

then

All generate the same estimate for the coefficient associated with .

* it shows that and effectively act as sufficient statistics in accounting for any unit-specific heterogeneity and time-specific heterogeneity that is correlated with xit.
* Once and have been included in the regression, adding variables that change only across i or only across t does not affect the coefficients on
* thus

with estimated by GLS. And thus it is the BLUP.

* By the same token, the coefficient on an intervention in

Do not change if we add more variable that vary exclusively in one dimension.

We can relax the common trend assumption if we have some constant (pre-intervention) covariates, and also allow the treatment effects to depend on the covariates.

He establishes the equivalence between pooled OLS and TWFE estimators in the estimation of the ATT for simultaneous adoption (including two-period).

Note that is a time-varying indicator.

### Nonlinear TWFE

Nonlinear TWFE is discussed recently in Wooldridge (2022). He builds on the equivalence between pooled OLS and TWFE for the linear case.

* Poisson fixed effect is less straightforward than the linear counterpart.
* Effects that vary by cohort and by year can be accommodated via interactions of cohort and time.
* Apparently, individual level effect can be replaced with cohort-level effects

POLS

The main difference is a change in the parallel trend assumption, i.e., instead of assuming that (absolute) changes over time are parallel we require that relative changes are parallel.

parallel trends assumption holds for a known, strictly increasing transformation of the conditional mean of the response variable

W drops the i.

### The imputation estimator

One can use an imputation approach, based on first estimating the nonlinear model using the control observations, or a pooled method. In the leading cases mentioned above (canonical link), the two methods are the same, with the pooled method being more convenient for obtaining standard errors and conducting inference.

SE?

Panel bootstrap

Simulation APHA

Placebo Study

* Single regressor generated with random forest
* Note that me in sim, is negative binomial, with time varying trends

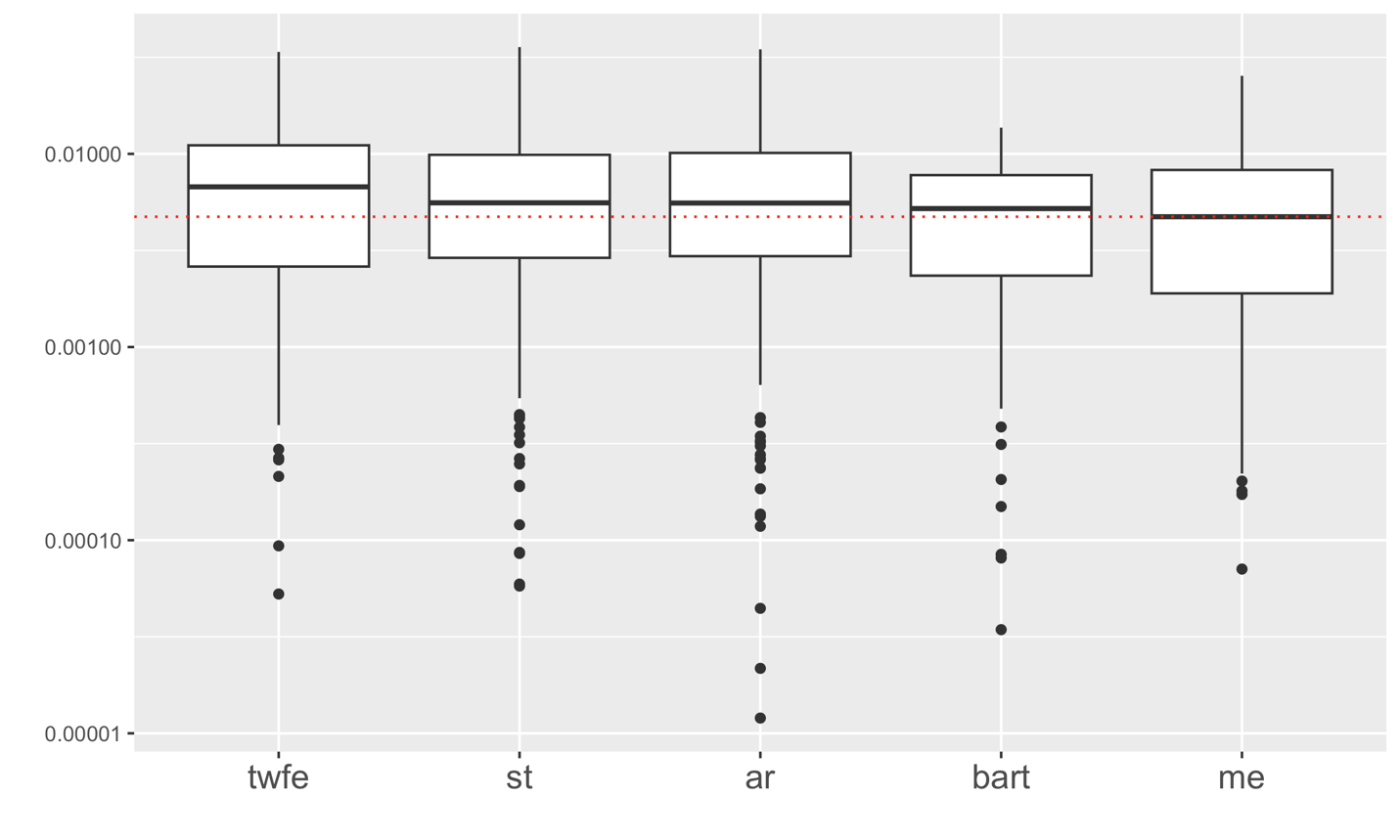
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | File | placebo | sim |
| RI + AR(1) data | Y | hhh4 | nbar | nbar |  |  |
| RI + AR(1) data +nb1 | Y | hhh4 | nb\_ar |  |  | ar |
| RI + RS(t) + nb | Y | glmer.nb | mepois | p\_mepoisfit |  |  |
| RI + RS(t) | Y | glmer | mepois | Mepoisfitfit  gfit? |  | me |
| RI + RS(x) + BYM2 x AR(1) | Y | inla | st | p\_stfit | inla |  |
| RI + RS(x) + AR(1) | Y | inla | ar | p\_arfit | ar |  |
| RI + RS(x) + Besag + AR(1) | Y | inla | ar\_sp | p\_arspfit | ar\_sp |  |
| TWFE on SMR | SMR | lm | twfe | p\_ols\_fits | twfe |  |
| RS(x) + RS(x|time) | Y | glmer.nb | mep | p\_mep\_fits |  |  |
| RS(x) + RS(x|time) | SMR | lmer | me | p\_me\_fits | me |  |
| SMR ~ st + AIAN + time + urb + f\_id + f\_time | SMR |  |  | p\_bart\_fits | bar |  |

X: ranger(SMR ~ st + AIAN + time + urb)

Comparison, read once more CM exponential, I think they compare ratio of RMSE and ANOVA

Compare SMR or compare rate?

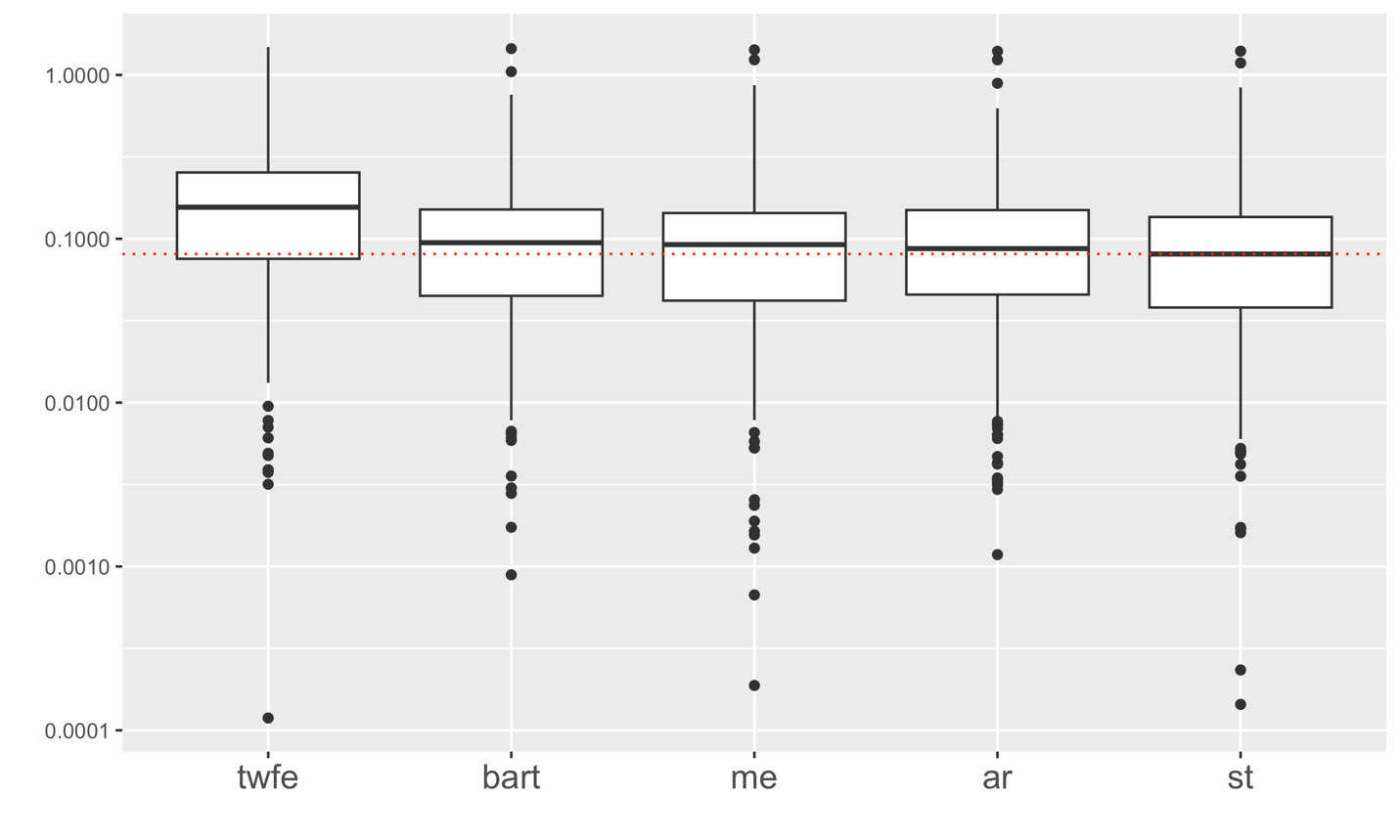
Placebo Study

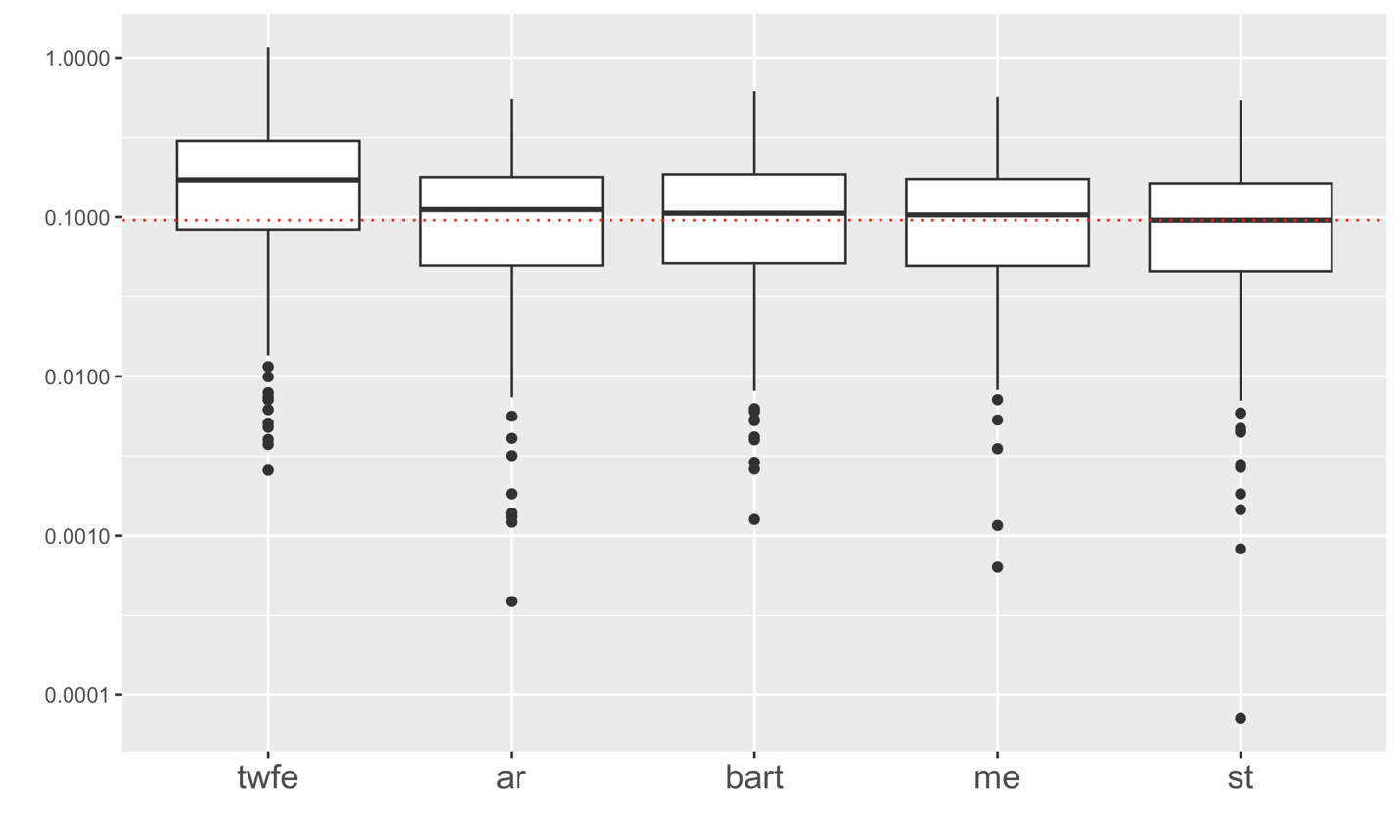


A diagram of a graph

Description automatically generated with medium confidence

Simulation





TWFE has to be Poisson rather than OLS!!

Can we try with a low-rank approximation to spatial info?

• Uncertainty estimation / CI coverage  
- How accurate are the estimates of uncertainty form each approach?

• Incorporate in ongoing GLS impact analysis • Examining effect heterogeneity

## Set up

We consider a setting with N units,[[1]](#footnote-1) for which we observe outcome for units over time periods. Some but not all units adopt the treatment during the panel; once a unit adopt the treatment it stays treated for the reminder of the panel. Let represent the time that the unit receives the treatment, with denoting never-treated units.

We adopt a potential outcomes framework to express causal quantities (Rubin, 1974; Imbens & Rubin, 2015), and assume that stable treatment and no interference between units (i.e., SUTVA). In principle, each unit in each time might have a distinct potential outcome for each potential treatment time , for .[[2]](#footnote-2)

We assume that prior to treatment, a unit’s potential outcomes are equal to unit never-treated potential outcome: for . Thus, the observed outcome is . Define , i.e., an indicator signaling whether unit at time is either treated or not. We may also observe pre-treatment characteristics, denoted by the k-component vector .

## Estimand

For , define the unit-year-level causal effect as

i.e., the difference between the potential outcome at that time under treatment (given adoption at time ) and under never treatment.

Out main interest lies on estimating the average impact across the counties exposed, averaged across counties and posttreatment periods, i.e.,

Other quantities of interest are effects at specified duration after treatment onset -known as event time. Let the event time . The unit-level treatment effect for treated unit at event time , is the difference between the potential outcome at time under treatment at time and under never treatment:

The Average Treatment Effect on the Treated (ATT) k periods after treatment onset:

We can recover the ATT by taking the average of post-treatment effect, i.e.,

Another frequent quantity of interest is the cohort specific effect, i.e., the effect among units with , averaging across follow up periods:

The ATT can then be obtained as weighted average of these cohort-specific estimates with weights proportional to the number of treated units,[[3]](#footnote-3)

## A model for the counterfactual outcome

Because we are interested in treatment effects on treated units -and we observe the potential outcomes under treatment- the challenge for estimating ATT is to impute the average of the missing never-treated potential outcomes, . We can estimate this quantity if we put forward some model for the outcome under the never treated condition. [[4]](#footnote-4)

There is an array of models for count data (discrete, nonnegative) that can accommodate spatial and temporal dependence (see Davis et al., 2021, for a recent review). A frequent framework in disease mapping and spatiotemporal modelling is that of hierarchical mixed-models, where relatively simple models for the observed data are coupled with more complex models for an underlying process. This are also known as parameter-driven models, or state-space models in time-series analysis.[[5]](#footnote-5)

A frequent choice for the observed counts follows independent Poisson distributions,

and then model the log of the expected count,

where is the relative risk in county i, and is the expected count under homogenous risk,, across counties for the entire period.[[6]](#footnote-6)

We chose a flexible mixed-effect model representation for the relative risk, as the addition of three independent component, sometimes referred as large-, small-, and fine-scale variation, i.e.,

In this representation, is a deterministic mean function, such as

with a set of observed county-specific characteristics that may vary over time, (such as number of hospitalizations for a cause unrelated to suicide), a preselected transformation, , the basis chosen to represent the information in , and -dimesntional vector of unknown parameters. The are structured random components, such as,

where again is a preselected transformation of observed covariates, , and an -dimensional vector of random coefficients. [[7]](#footnote-7) Finally, is white noise with . [[8]](#footnote-8)

### Examples

A classical model in policy evaluation, known as two-way fixed effect (TWFE), sets

a unit and time specific fixed effects. This model typically does not include random effect, i.e., . However, deviation from this deterministic function, , are not assumed to be homoscedastic or independent over time, and estimates of uncertainty are obtained that are robust to varying dispersion as well as arbitrary patterns of temporal correlation (but no attempt is made of model the errors explicitly).

In field like small area estimation (SAE) or disease mapping, the random components are model explicitly in the hopes of improving precision by sharing information across units and over time. The simplest way to model dependence among observations from the same county is by introducing a random intercept, i.e.,

The county-specific random effects allows to “borrow strength” for estimation across counties, similar to the Fay and Herriot (1979) model for a gaussian outcome, by smoothing county estimates toward a global rate. [[9]](#footnote-9)

Rather than smooth small county estimates toward a global rate, disease mapping models, introduce a spatial structure to borrow strength locally rather than globally. A frequent choice is to assume that each follows a conditional autoregressive (CAR) model (Clayton & Kaldor, 1987; Besag et al., 1991),

where each is a user-defined spatial dependence weights determining which counties j are “neighbors” to county i, typically based on adjacency.[[10]](#footnote-10) In other words, the expected value of the error in county i, is an average of the value in the adjacent counties, and the variance is inversely proportional to the number of neighbors. [[11]](#footnote-11)

Both spatial and nonspatial random effects can be combined. For instance, using re-parameterization of the Besag et al. (1991) model by Riebler et al. (2015),

where is the variance, 0 is a mixing parameter determining the amount of variance that is spatially structured, is the identity matrix, and is the precision matrix of the (standardized) spatial component.

Temporal dependence could also be accommodated, for instance, using a first order autoregressive model for , i.e.,

where is the autocorrelation coefficient and are the independent shocks (as in Rao & Yu, 1994, in the context of SAE). [[12]](#footnote-12)

Rather than independent spatial and temporal components, a spatiotemporal variance covariance matrix for the error term can be constructed using the Kronecker product of the spatial and temporal precision matrices, as suggested by Knorr-Held (2000) (see also Martínez-Beneito et al.[2008] for a closely related approach):

(This is also an interaction, but now spatially structured)

## Approach

We propose to learn and using random forest fitted to control data,

The model becomes,

## Assessment

Bootstrap residuals.

Classical measure of performance separately assesses accuracy of the point estimates (mean square or absolute error) and uncertainty estimates (CI coverage and length)[[13]](#footnote-13).

Scoring (strictly proper and proper) attempts to assess the predictive distribution as a whole (Czado et al., 2009). One of the simplest proper scores depeding only on the first two moments of the predictive distribution is

And similarly,

A random forest fit to the observed before treatment. This is a transformation from , a single dimensional regressor, chosen to maximize predictive performance.

Bias can be estimated using direct estimator of true rate, (that is just Y\_it?

(The justification needs to be effect heterogeneity. Otherwise, there are areay unbiased estiamtes and we can not dobetter))

Best predictive SAE Jiang, Nguyen and Rao (2011)

Noting that in SAE the actual target is the prediction of the area means, and the estimation of model parameters is just an intermediate step, the authors propose to estimate the fixed parameters in such a way that the resulting predictors are optimal under some loss function.

(they speack about predictic erro, because it indoves the realization of the. random coeffeicitns, I am interest in prediction because also incol

The DE of the rate is jus O\_it / n\_it (this is assuming O is a random variable)

So the di

Torabi, M. and Rao, J. N. K. (2008). Small area estimation under a two-level model. Survey Methodology 34 11–17.

For that purpose, we will regres

sed a direct estimate of the standardized ratio on a set of covariates, including the county’s historical youth suicide rates (from 1999 to 2006) and demographic characteristics of the county, such as the proportion of the population that is American Indian/Alaska Native, unemployment rate, median income, population without health insurance (from 1999 through the current year), and rates of hospital use among youth and adults for causes unrelated to suicide (from 2008 to the current year). We also included time- and county-specific effects, as well as the state, region, division, and the six-class urban-rural classification. Following recent recommendations, we will further include as a predictor an estimate of the (time-varying) propensity score as a function of these same covariates (Hahn et al. 2020) to avoid confounding induced by regularization. These propensity can itself be estimated using random forest.

## Bias

The crude rate estimator is unbiased, so this aspect cannot be improved upon. However, other, biased estimators may have a smaller variance, which yields a smaller overall mean squared error (the sum of the variance and the squared bias). The so called BLUPs or EBLUPs are only best under the model (they may be better with fixed effects, but that is not guaranteed). [James and Stain 1961]

The principle is referred to as shrinkage, in the sense that the crude rate is moved (shrunk) towards an overall mean, as an inverse function of the inherent variance.

The estimated prediction uncertainty in two-stage mixed effect models plugin estimator in SAE is known to be biased.(Lahiri & Maiti, 2006). Second order unbiased. I think this is due to ignoring the uncetinty in the estimation of variance

BLUP (Anderson)

Best (minimum variance)

Linear (weighted average) <https://icfonline.sharepoint.com/:p:/r/sites/NIHR03/Shared%20Documents/General/III.%20Reporting%20and%20Dissemination/III.%201.%20%20Conference%20presentation/AAS_2024_NIHR03.pptx?d=w84b65e7a37fa4f26994e99908d26c439&csf=1&web=1&e=tcxNCo>

Unbiased

EBLUP, EB, and HB

Are unbiased under the joint distribution of the outcome and the random effects, but conditionally on the realized values of the random effects (for example, if the source of randomness comes from assigments) they are not.

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. P. (1998). Generalized linear models for smallarea estimation. J. Amer. Statist. Assoc.

As stated in the introduction, an important aspect of SAE is the assessment of the accuracy of the predictors. This problem is solved “automatically” under the Bayesian paradigm, which produces realizations of the posterior distribution of the target quantities.

As pointed out by Longford (2007), the ultimate aim in SAE is to make inferences about small area characteristics conditional on the realised (but unknown) values of small area effects, *i.e.*, with respect to (1)

the MSE of real interest is that defined by the area-specific model

Longford (2007)

For example,

with R latent time-varying factors (with R typically small relative to N and T), and N by R unit-specific, time-invariant factor loading.

A time-varying process,

in combination with independence assumptions on the noise component, . In particular, we need to assume that the timing of the treatment is independent of the noise in any period, i.e., , or independent of the noise after the treatment, i.e., for , respectively. And the noise does not have fat tails (sub-gaussian)

Series of counts

https://github.com/JonathanBradley28/CM

1. Multivariate. INAR
2. Copula
3. Parameter-Driven models (state-space
4. Observation-Driven

Bayesian Dynamic GLM. (Gaussian process

1. Level Correlated Models (INLA
2. Hierarchical Multivariate Dynamicm models

# Multivariate Poisson with Latent Multivariate Log Gamma

The model was introduced by Bradley and colleagues in a series of paper (Bradley et al., 2015a, 2018, 2019). The key advantages to this framework are that the multivariate log gamma distribution is conjugate to the Poisson distribution, which permits model flexibility and fast computational performance.

the Poisson data model with a latent Gaussian process model has become the de facto model. However, this model can be difficult to use in high dimensional settings, where the data may be tabulated over different variables, geographic regions, and times.

## Background

Hierarchical GLM with conjugate priors

a type of hierarchical generalized linear model

* Latent conjugate multivariate.
  + providing conjugacy in the nonGaussian dependent data setting
  + more flexible
* Reduced rank methods
  + it can easily be cast within the reduced rank modeling framework

linear combination of independent log-gamma random variables to build their multivariate log-gamma distribution

implement the LCM using a collapsed Gibbs sampler

## The model

Data model

Process model

The r-dimensional random vector is assumed to be mean-zero and have an unknown covariance matrix. The r-dimensional real vectors can belong to any class of areal basis functions.

First order vector autoregressive model, VAR(1),

are known r × r propagator matrices.

ars

the Nt × r matrix is specified to be contained within the orthogonal complement of the column space of . Specifically, define the MI operator as

is a generic weight matrix, such as the adjacency matrix. Notice that the MI operator defines a column space that is orthogonal to XP t. let the spectral representation

Set to the first r columns of

Hughes and Haran (2013) suggests setting r equal to roughly 10% of the positive eigenvalues given

In turn the propagation matrix, is the r eigenvectors of

## The model

### The data

### The process

Addition of three independent component, sometimes termed large-, small-, and fine-scale variation, i.e.,

where is a deterministic mean function (i.e., fixed effects), such as, , with a -dimensional vector of known covariates and -dimesntional vector of unknown parameters. (They may vary over time, , but more commonly not generally allowed)

models small scale variation. A very flexible way to represent is through a basis-function expansion,

with a pre-specified -dimensional vector of basis-functions and an - dimensional vector of random effects.

The random effects maybe allowed to vary overtime and specified to have a first-order vector autoregressive, VAR(1), structure, i.e.,

Where is a prespecified propagator matrix. , and

with and unknown. is independent of .

with

The follow a multivariate gamma rather than multivariate gaussian, which means there are additional parameters to estimate and , i.e.,

The are assumed to follow a white-noise, mean zero process with variance . In the CM proposal a log-gamma process rather than a Gaussian one, with unknown rate and shape,

Flat, conjugate priors are placed in all parameters ).

The set of basis function can accommodate many different processes. A reduced rank spatiotemporal specification is discussed in Bradley et al. (2015b) for a gaussian outcome and was introduced as a dimension reduction mechanism by Hughes & Haran, (2013).

## Moran’s I (MI) basis function and propagation matrix

### Moran’s I basis functions

MI basis functions can be used to model areal data in a reduced dimensional space. allow for nonstationarity in space. they guarantee there are no issues with confounding between fixed and random effects.

Let be the covariate matrix. Define, , the “hat” matrix, or the projection matrix, i.e., computes the orthogonal projection. Then the MI operator (mapping or function) is

for ; where is an covariate matrix, is an identity matrix, is an adjacency matrix. The MI operator defines a column space that is orthogonal to .[[14]](#footnote-14) This can be used to ensure non-confounding between and .

is diagonalizable, i.e.

The proposed basis are the first r eigenvectors, i.e., [[15]](#footnote-15)

is

### Moran’s I propagator matrix

Define the matrix ,[[16]](#footnote-16)

and

is

### Parameter models or priors

We consider specifying as positive semi-definite matrices that are “close” to target precision matrices, ,which includes some of the sources of variability ignore by the MI approximation (by removing some of its principal components).

There are many choices for the “target precision” matrices, such as a CAR model, i.e., . For some generic square matrix , the Frobenius norm is defined as

With both {Kt } and {Mt } specified we can solve for {Wt }, that is, using the VAR(1) model

It is approximated by , where

Appendix B

bradleyjr@missour

### conditional autoregressive smoothing for space and B-splines without penalties for temporal trends

(MacNab & Gustafson, 2007)

we use a tensor product spline model with a Markov random field prior on the coefficients of the basis functions

are fixed effects.

is a set of basis functions (without the intercept) for a R-dimensional space of B-splines of degree 3, denoting L pre-specified inner knots. denoting the rth B-spline basis function evaluated at time t, and

Are random spline coefficients.

The prior for this random coefficients can be used to accommodate spatial dependence

The prior for this random coefficients can be used to accommodate spatial dependence

##

Rather than introducing the splines, I can use the forest on everything fix

Algorithm

1. Using only control data, use some ML to learn a fixed predictor, say , where x includes time and unit indicators.

“represents an arbitrary smoothing function”

* 1. For learning, we need a transformation of y, such as the SMR.
  2. We could include propensity.
  3. Should we use clustered out of sample performance?

1. Fit a mix-model with a fix and random coefficients for .
   1. Should I include the group mean of ?
2. Consider different priors for the random coefficients: iid (i.e., ri), ar(1), spatial(), spatio-temporal. (knoekner

Baslien Poisson TWFE with cluster bootstrap?

(DASH) colin

MacNab YC, Dean BC. Autoregressive spatial smoothing and temporal smoothing B-splines for mapping rates. Biometrics 2001; 57(3):949–956.

Eilers PHC, Marx BD. Flexible smoothing using B-splines with penalized likelihood (with comments ad rejoinder). Statistical Science 1996; 11(2):89–121

de Boor. A Practical Guide to Splines. Springer: New York, 1978

Ugarte et al. (2012a)

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program. *Journal of the American Statistical Association*, *105*(490), 493–505. https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method: COMPARATIVE POLITICS AND THE SYNTHETIC CONTROL METHOD. *American Journal of Political Science*, *59*(2), 495–510. https://doi.org/10.1111/ajps.12116

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, *93*(1), 113–132. https://doi.org/10.1257/000282803321455188

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist’s companion*. Princeton University Press.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2018). Matrix Completion Methods for Causal Panel Data Models. *arXiv:1710.10251 [Econ, Math, Stat]*. http://arxiv.org/abs/1710.10251

Athey, S., & Imbens, G. (2015). Recursive Partitioning for Heterogeneous Causal Effects. *arXiv:1504.01132 [Econ, Stat]*. http://arxiv.org/abs/1504.01132

Athey, S., & Imbens, G. (2018). Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption. *arXiv:1808.05293 [Cs, Econ, Math, Stat]*. http://arxiv.org/abs/1808.05293

Bai, J., & Ng, S. (2020). Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data. *arXiv:1910.06677 [Econ]*. http://arxiv.org/abs/1910.06677

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How Much Should We Trust Differences-In-Differences Estimates? *The Quarterly Journal of Economics*, *119*(1), 249–275. https://doi.org/10.1162/003355304772839588

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20. https://doi.org/10.1007/BF00116466

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2015a). *Computationally Efficient Distribution Theory for Bayesian Inference of High-Dimensional Dependent Count-Valued Data* (arXiv:1512.07273). arXiv. http://arxiv.org/abs/1512.07273

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2015b). Multivariate spatio-temporal models for high-dimensional areal data with application to Longitudinal Employer-Household Dynamics. *The Annals of Applied Statistics*, *9*(4). https://doi.org/10.1214/15-AOAS862

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2018). Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion). *Bayesian Analysis*, *13*(1). https://doi.org/10.1214/17-BA1069

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2019). *Bayesian Hierarchical Models with Conjugate Full-Conditional Distributions for Dependent Data from the Natural Exponential Family* (arXiv:1701.07506). arXiv. http://arxiv.org/abs/1701.07506

Breiman, L. (Ed.). (1998). *Classification and regression trees* (Repr). Chapman & Hall [u.a.].

Carvalho, C., Feller, A., Murray, J., Woody, S., & Yeager, D. (2019). Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge. *arXiv:1907.07592 [Stat]*. http://arxiv.org/abs/1907.07592

Chernozhukov, V., Wuthrich, K., & Zhu, Y. (2019). An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls. *arXiv:1712.09089 [Econ, Stat]*. http://arxiv.org/abs/1712.09089

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-AOAS285

Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, *43*(3), 671–681.

Condron, S., Garraza, L. G., Walrath, C. M., McKeon, R., Goldston, D. B., & Heilbron, N. S. (2015). Identifying and Referring Youths at Risk for Suicide Following Participation in School-Based Gatekeeper Training. *Suicide and Life-Threatening Behavior*, *45*(4), 461–476. https://doi.org/10.1111/sltb.12142

Condron, S., Godoy Garraza, L., Kuiper, N., Sukumar, B., Walrath, C., & McKeon, R. (2019). Comparing the Effectiveness of Brief Versus In-Depth Gatekeeper Training on Behavioral Outcomes for Trainees. *Crisis*, *40*(2), 115–124. https://doi.org/10.1027/0227-5910/a000539

Czado, C., Gneiting, T., & Held, L. (2009). Predictive Model Assessment for Count Data. *Biometrics*, *65*(4), 1254–1261. https://doi.org/10.1111/j.1541-0420.2009.01191.x

Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., Pipiras, V., & Ravishanker, N. (2021). Count Time Series: A Methodological Review. *Journal of the American Statistical Association*, *116*(535), 1533–1547. https://doi.org/10.1080/01621459.2021.1904957

DISEASE MAPPING COLLABORATIVE GROUP. (2000). Disease mapping models: An empirical evaluation. *Statistics in Medicine*, *19*(17–18), 2217–2241. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18<2217::AID-SIM565>3.0.CO;2-E

Doudchenko, N., & Imbens, G. W. (2017). Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. *arXiv:1610.07748 [Stat]*. http://arxiv.org/abs/1610.07748

Fay, R. E., & Herriot, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, *74*(366), 269. https://doi.org/10.2307/2286322

Hahn, P. R., Murray, J. S., & Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects. *Bayesian Analysis*. https://doi.org/10.1214/19-BA1195

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217–240. https://doi.org/10.1198/jcgs.2010.08162

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933

Hughes, J., & Haran, M. (2013). Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *75*(1), 139–159. https://doi.org/10.1111/j.1467-9868.2012.01041.x

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, *19*(17‐18), 2555–2567. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18<2555::AID-SIM587>3.0.CO;2-#

Kolak, M., & Anselin, L. (2020). A Spatial Perspective on the Econometrics of Program Evaluation. *International Regional Science Review*, *43*(1–2), 128–153. https://doi.org/10.1177/0160017619869781

Lahiri, S. N., & Maiti, T. (2006). *Nonnegative mean squared prediction error estimation in small area estimation* (arXiv:math/0604075). arXiv. http://arxiv.org/abs/math/0604075

MacNab, Y. C., & Gustafson, P. (2007). Regression B‐spline smoothing in Bayesian disease mapping: With an application to patient safety surveillance. *Statistics in Medicine*, *26*(24), 4455–4474. https://doi.org/10.1002/sim.2868

Martínez-Beneito, M. A., López-Quilez, A., & Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, *27*(15), 2874–2889. https://doi.org/10.1002/sim.3103

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Paul, M., & Held, L. (2011). Predictive assessment of a non‐linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, *30*(10), 1118–1136. https://doi.org/10.1002/sim.4177

Rao, J. N. K., & Yu, M. (1994). Small-area estimation by combining time-series and cross-sectional data. *Canadian Journal of Statistics*, *22*(4), 511–528. https://doi.org/10.2307/3315407

Riebler, A., Sørbye, S. H., Simpson, D., & Havard Rue. (2015). *An intuitive Bayesian spatial model for disease mapping that accounts for scaling*. https://doi.org/10.13140/RG.2.1.1899.2080

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Sparapani, R. A., Logan, B. R., McCulloch, R. E., & Laud, P. W. (2016). Nonparametric survival analysis using Bayesian Additive Regression Trees (BART). *Statistics in Medicine*, *35*(16), 2741–2753. https://doi.org/10.1002/sim.6893

Wager, S., & Athey, S. (2017). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *arXiv:1510.04342 [Math, Stat]*. http://arxiv.org/abs/1510.04342

Wooldridge, J. M. (2021). Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3906345

Wooldridge, J. M. (2022). Simple Approaches to Nonlinear Difference-in-Differences with Panel Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4183726

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514. https://doi.org/10.1198/106186008X319331

Research Strategy

Significance

Suicide rates for youth aged 10–24 years in the United States increased 56% from 2007 to 2017, surpassing homicide as the second-leading cause of death for this group (Curtin & Heron, 2019). As in other age groups, suicide among youth occurs at much higher rates in smaller, more rural areas compared with larger, more urban areas, a gap that has widened in recent years (Ivey-Stephenson et al., 2017; Kegler et al., 2017). During the years 2016 through 2018, the youth suicide rate in nonmetropolitan counties was 13.7 per 100,000, compared to 8.9 per 100,000 in large metropolitan counties (Centers for Disease Control and Prevention [CDC], 2020). Suicide disproportionally burdens AI/AN youth (Leavitt et al., 2018). During the years 2016 through 2018, the suicide rate among AI/AN youth in nonmetropolitan counties reached 32.1 per 100,000 (CDC, 2020).

Since 2005, the GLS youth suicide prevention program administered by the Substance Abuse and Mental Health Services Administration (SAMHSA) has funded U.S. states and territories, tribes, and college campuses to implement comprehensive, community-based youth suicide prevention programs. GLS program grantees utilize funding for a variety of suicide prevention interventions, including gatekeeper trainings; outreach and awareness strategies; early identification screening programs; partnership development for early intervention, linkages to community providers, and appropriate treatment; care transitions; culture-based prevention activities; and means restriction (Goldston et al., 2010).

Using a combination of propensity score matching and inverse probability weighting, Walrath et al. (2015) estimated a positive impact on suicide mortality in close to 500 communities exposed to the program before 2010. Subsequently, Godoy Garraza et al. (2015) found a similar pattern with respect to nonfatal suicide behavior. A more recent study examined the long-term impact of the intervention on these same communities and the role of length of exposure to the program (Godoy Garraza et al., 2019). While the mechanisms of action were not directly examined, it is possible that program effort had altered the sense of belonging, particularly in marginalized communities (Joiner, 2007). Indeed, Godoy Garraza et al. (2019) found evidence suggestive of a greater positive impact of GLS in rural counties. In these studies, however, the sample deliberately excluded very small communities (specifically, counties with fewer than 3,000 youth on average between 1999 and 2006) because yearly rate estimates in these areas was deemed unreliable. This exclusion affected over 1,000 counties, about 80% of them noncore (i.e., the most rural counties based on the National Center for Health Statistics [NCHS] six-level urban-rural classification scheme [Ingram & Franco, 2014]).

Assessing the impact of an intervention (such as a project, program, or policy) ultimately depends on the ability to estimate what would have happened had the intervention not taken place. While randomized studies are the gold-standard approach to estimating counterfactual outcomes, many situations make randomization unfeasible. The problem of estimating the counterfactual outcome in such situations (i.e., by using observational data) has received considerable attention in the last few decades, particularly since the formalization of the Rubin causal model (Holland, 1986) and the introduction of propensity score matching (Rosenbaum & Rubin, 1983) as a frequently convincing solution. A notable example is the introduction of the synthetic control methodology for comparative case studies (Abadie et al., 2003, 2010, 2015). For the most part, this work has focused on how to deal with confounders in the absence of randomization, an essential issue regardless of sample size.

The challenges associated with counterfactual estimation with observational data are compounded in the context of small areas or communities where reliable estimates of the outcome of interest are difficult to obtain. Indeed, an area or domain of estimation is considered to be small precisely when direct estimations are extremely unreliable or entirely unfeasible given the sample size (Pfeffermann, 2013). This situation often arises with relatively rare outcomes (such as suicide), as well as with more frequent outcomes (such as suicide-related hospitalizations), when the interest lies on specific segments of the population.

Closely related with small-area estimation, the field of disease mapping is concerned with estimating the risk of a disease or health outcome using case counts within small administrative districts or regions (Waller & Carlin, 2010; Martinez-Beneito & Botella-Rocamora, 2019). Building on hierarchical models originally proposed for small-area estimation, disease-mapping models further incorporate spatial dependence. While the initial models focused on a cross-sectional estimation (Besag et al., 1991), spatiotemporal models that take advantage of multiple time periods were soon developed (Bernardinelli et al., 1995; Knorr-Held, 2000). The hierarchical structure of these models has made either empirical or full Bayesian estimation the predominant approach. The recent introduction of more efficient computational methods for Bayesian estimation have made these models even more attractive (Rue et al., 2009). Bauer et al. (2016), for example, recently took advantage of this development to propose a flexible but computationally efficient spatiotemporal model based on the use of splines. Neither of these developments, however, has been applied to program impact evaluation.

Innovation

Doudchenko & Imbens (2016) explicitly articulated the link among a range of methods to estimate impact, including difference in differences, propensity score matching, and synthetic control. Athey et al. (2017) further examined this connection, emphasizing the nature of the impact-evaluation problem as an imputation problem—that is, the imputation of the missing counterfactual outcome. Previous proposals, including Brodersen et al. (2015) and Amjad et al. (2017), have used Bayesian approaches to solve this imputation problem. While these models incorporate information on time dependence to aid estimation, they do not take advantage of the potential spatial structure. To the best of our knowledge, Bayesian spatiotemporal models developed for disease mapping—such as that proposed by Bauer et al. (2016), which explicitly takes advantage of spatial structure together with temporal dependencies to aid the estimation—have not been utilized for impact evaluation. Applying this approach to suicide prevention impact evaluation both extends the utility of the approach beyond disease mapping and significantly advances the ability to understand the impact of suicide prevention programming in some of the highest-risk populations. While some of the insights from disease mapping are not necessarily new, their incorporation into impact evaluation is innovative and can represent a significant methodological improvement as well as spark further research and developments in the field. As such, we are proposing a methodological innovation that can be applied immediately to real-world practice.

Approach

1. Design

We will use both simulated and real data to examine the performance of the proposed approach and compare it with alternative methods to estimate impact. In particular, we will compare the following algorithms:

1. Synthetic control (SC) (Abadie et al., 2003, 2010, 2015)
2. Elastic net (EN) (Doudchenko & Imbens, 2016)
3. Matrix completion (MC) (Athey et al., 2017)
4. Disease mapping (DM) (Bauer et al., 2016)

These four procedures use information on the outcome variables (e.g., youth suicide mortality) from unexposed communities and periods (i.e., from communities never exposed and from communities exposed before the start of the intervention) to estimate the counterfactual outcome. Only the DM approach incorporates information regarding the spatial proximity of communities.

1. Data

The algorithms will be compared using both simulated and real data.

2.1. The simulation

We will use simulated health outcome counts for a sample of areas under different hypotheses regarding underlying spatial and temporal dependence (and other relevant parameters). With simulated data, we know the true quantities of interest and computation of bias, variance, and overall mean square error (MSE) are straightforward.

Data-generating model

We will simulate data using a generalized linear mixed model with spatially and serially correlated random effects. In particular, we will use the formulation proposed by Rushworth et al. (2014), which is closely related to that proposed earlier by Martínez-Beneito et al. (2008). Specifically, we will draw the counts in each area and period from independent Poisson distributions with expected value determined by the location- and time-specific relative risk. In turn, the log relative risk within each area will follow a first-order autoregressive process with a specified correlation parameter. We will independently draw the initial relative risk and the subsequent random shocks or error terms for each period from a multivariate normal distribution with a unit variance and a spatially structured covariance matrix.[[17]](#footnote-17) Importantly, we will take the neighborhood structure of the areas and their relative size from the real application. We will also take the total number of available periods from the real application (i.e., nine years). After we generate each dataset, we will select a sample of areas that will be considered exposed to a hypothetical intervention and a random initial time within each area representing a staggered adoption.

Parameters of the data-generating process

We will implement simulation under different values for the following parameters:

1. Two levels of the number of areas ever exposed ()
2. Two levels of first initial time of exposure () (individual initial time will be equally spaced after that year)
3. Two levels of baseline risk (corresponding approximately with suicide to self-harm hospitalizations among youth) ()
4. Three levels of spatial association (
5. Three levels of serial dependence ()
6. Additional noise from a t distribution with five degrees of freedom ()

For each cell, we will generate 25 datasets, resulting in 25\*2\*2\*2\*3\*3\*2=3,600 simulated datasets. This number of simulations will ensure high precision in the estimation of performance of each algorithm (a standard error due to the simulation of less than 2% of the typical variation in the measure of performance).

2.2. The case study

Since 2005, the GLS program administered by SAMHSA has funded U.S. states and territories, tribes, and college campuses to implement comprehensive, community-based youth suicide prevention programs (Goldston et al., 2010). We will use the proposed approach to estimate the impact of the GLS program in micropolitan and noncore communities in a sample of states. We expect GLS was particularly impactful in these smaller, often marginalized, communities, possibly by altering the sense of belonging (Joiner, 2015).

Sample

The sample encompasses 516 counties and county equivalents in ten states: Arizona, Kentucky, Massachusetts, Michigan, Nebraska, New Mexico, Nevada, Oregon, South Dakota, and Washington. A total of 204 nonmetropolitan counties in these states (87 micropolitan and 117 noncore) were exposed to GLS suicide prevention activities between 2007 and 2016; we will focus our analysis on these counties. We will use information from the remaining 312 counties to estimate the counterfactual.

Measures and sources

Our outcome variables will encompass suicide mortality rates and self-harm hospitalizations among youth aged 10–24 years from 2008 to 2016. We will obtain the suicide mortality rate from CDC’s Compressed Mortality File. We will derive hospitalization rates from the Healthcare Cost and Utilization Project’s (HCUP’s) State Inpatient Databases. We will also obtain emergency department visits due to self-harm for communities in a subsample of states (five of the 10 states in our sample) participating in HCUP’s State Emergency Department Databases.

The independent variable is the exposure to the GLS program. The National Outcome Evaluation (NOE) of the GLS program will serve as the source for this information. NOE systematically collects information on the location of program activities, particularly training activities. As the national evaluator for the program, ICF currently holds that information and has SAMHSA’s permission to use the data in this project.

We will consider multiple covariates, including each community’s total population and youth population, race-ethnic composition (percentages of Hispanic and non-Hispanic white, African American, AI/AN, and other races), median household income, poverty rates, unemployment rates, percentage of population with no health insurance, and rurality (using the NCHS six-level urban-rural classification scheme [Ingram & Franco, 2014]). We will acquire most of this information via the U.S. Census Bureau; however, we will obtain unemployment estimates from the U.S. Department of Labor and bridged race-ethnic counts and the urban-rural classification from NCHS.

1. Analysis

The goal of the analysis is to compare the accuracy of the four approaches (i.e., SC, EN, MC, DM) in estimating the outcome(s) that would have been observed in communities exposed to an intervention had the intervention not taken place. We can directly compute the predictive accuracy of the different algorithms with the simulated data but need to estimate this accuracy in the case of the real application.

3.1. Simulation

For each simulated dataset, we will estimate the counterfactual outcome for the communities and periods exposed to the hypothetical intervention using each of the four approaches (SC, EN, MC, or DM). Because the true counterfactual value will be available, we can directly compute measures of discrepancy. We will particularly focus on the correlation coefficients representing the linear association between the true and predicted counterfactual value.

To assess the effects of the data-generating parameters on predictive accuracy, we will perform analyses of variance (ANOVA) with algorithm type (SC, EN, MC, or DM), county size quintiles, and the parameters of the data-generating process as independent variables. We expect the DM approach to outperform the alternatives among smaller counties. This relationship will likely be affected by the variation in the parameters of the data-generating process. Thanks to the factorial design with multiple replications, the relative importance of these three-way and higher-order interactions as sources of variation can be quantified. We will graphically examine these relationships using trellis plots, particularly for factors and interactions found to explain substantial variation in performance. (See Rubin [1978] for an early example of using ANOVA to summarize simulation results and Morris et al. [2019] for current recommendations).

3.2. The case study

We will estimate the counterfactual youth suicide rate and rate of hospitalization due to self-harm in nonmetropolitan communities exposed to the GLS program following the start of program activities using each of the four methods. Secondary analysis will focus on sex- and race/ethnic-specific rates. In this case, the true value of the quantity of interest is unknown; thus, it is impossible to directly assess the accuracy of the approaches. However, we can estimate the predictive error using only unexposed counties and periods and a procedure akin to leave-one-out cross-validation proposed by Doudchenko & Imbens (2016). Specifically, for each observation from a county never exposed to GLS as well as for each observation from before the start of the exposure among counties eventually exposed, the value of the outcome under no intervention is known (i.e., it is not a counterfactual). If the value of that outcome is ignored momentarily and the rest of the information is used to estimate that value, the discrepancy between observed and estimated values cannot be the impact of the intervention (that has not started). Instead, the discrepancy is an estimate of the predictive error. We will further compare the estimated impact—that is, the difference between the estimated counterfactual and actual observed outcomes after the start of the intervention among the exposed counties.

3.2.1. Heterogeneous effects

Our proposed procedures result in an estimated effect for each county and each year following the start of the intervention. We anticipate these effects will be heterogeneous. We will examine that heterogeneity as a function of both county and intervention characteristics. [[18]](#footnote-18),[[19]](#footnote-19) Some authors have proposed using supervised machine-learning algorithms to carry out this exploration systematically (e.g., Athey & Imbens, 2015). We propose, in particular, the use of two recursive partitioning algorithms (Zeileis et al., 2008 and Hothorn et al., 2006) to identify subgroups of counties with distinct patterns of GLS effect over time since start of exposure and to examine which intervention characteristics are associated with membership in each the identified subgroups. Understanding the circumstances and/or intervention features associated with better outcomes will greatly enhanced the utility of the results for suicide prevention policy and practice. Furthermore, while no intervention was intended to exclusively alter the sense of belonging, certain emphasis (such as the prominence of tribally sponsored activities or peer trainings) might provide further support to that hypothesis.

1. Challenges and remedies

As of the last quarter of 2015, the coding scheme used for diagnostic information for hospital discharge records transitioned from ICD-9 to ICD-10. We will examine quarterly rates of the outcome of interest (self-harm and/or suicide ideation among youth) to determine if there is a discontinuity at the time of transition. We will report the results of this examination. Even if a discontinuity is found, there is no reason to expect that it would be associated with the intervention. In other words, it is expected to affect both exposed and unexposed communities in the same manner.

Inferential methods (e.g., interval estimation, hypothesis testing) for some recently developed methods-of-impact estimation are not yet entirely understood and, as such, are currently an active area of research (e.g., Chernozhuko et al., 2018). Traditional arguments based on large samples (large number of units, periods, or both) can be unappealing in the typical settings for which these approaches were developed. That said, we will focus on the empirical assessment of the predictive accuracy of the counterfactual by the different algorithms, which does not depend on the inferential procedure. Our proposed approach offers well-known methods to quantify uncertainty, albeit under a Bayesian framework.

Response

1. Novelty

* “The method is not new and now often used, which undermines potential impact” (Reviewer 1)
* “The methods themselves are not new.” (Reviewer 1)
* “The use of the proposed Bayesian disease mapping method in the considered application is quite standard. The advantages of the proposed method over competing causal methods are not obvious.” (Reviewer 2)
* “Both method development and data application are standard.” (Reviewer 2)
* “Compared with the competing causal methods, such as SC, EN and MC, the proposed DM method may lack a clear causal interpretation" (Reviewer 2)

We do not claim that disease mapping or Bayesian spatiotemporal models are an innovation (and much less, that we are responsible for such innovation). While spatiotemporal models are still a very active area of research, the use of these models to improve small are estimation of relatively infrequent disease or event rates (including, in particular, suicide rates) is well stablished. The purpose of this standard use is mostly descriptive in nature, e.g., to characterize the distribution of disease or event spatially and over time, to detect ‘hotspots’, or even to examine environmental correlates.

Even if it might appear as a ‘natural’ extension, the use of spatiotemporal models for impact evaluation/causal inference is not standard (neither in the area of suicide prevention nor beyond this field). A very recent article by recognized authorities in spatial econometric (Kolak & Anselin, 2020), highlights the potential fruitfulness of this connection (i.e., between spatiotemporal modelling and causal inference). The article states that “while this has recently received increased attention in the mainstream causal literature […] an explicit spatial focus is still largely absent” (page 129).

We agree with the Reviewer 2 statement that “compared with the competing causal methods, such as SC, EN and MC, the proposed DM method may lack a clear causal interpretation”. The lack of a clear causal interpretation is precisely due to the fact that the proposed method have been largely used for descriptive purposes and not for impact evaluation/causal inference. Embedding DM in a clear causal framework is precisely our innovation.

While the causal framework we use is also not novel, it was only recently articulated explicitly for panel data (Athey et al., 2018). In this framework, effect estimation reduces to estimating the missing entries in an incomplete matrix. While Athey and colleagues proposed a particular method to implement this estimation, once the problem is conceived in this way, other alternatives are possible, including, we argue, the use spatiotemporal models. It is worth mentioning that Athey and colleagues did anticipate the possibility of exploiting the dependence on the time dimension to improve estimation. Out proposal, takes advantage of both temporal and spatial dependence.

Below we present the casual framework more formally, and show our proposed approach fit in this framework.

Potential outcomes framework

We were interested in estimating the effect of exposure to the GLS program on an outcome, e.g. suicide-related hospitalizations. We had data on counties i = 1,..., N, observed during years t = 1,..T. Each year, we observed whether a county was exposed to GLS or not, a circumstance we denoted by Wit ∈{0,1}. Indeed, a subset of the counties, i = N0 +1,…, N, was exposed to GLS starting from year t = t0i +1 onwards. [[20]](#footnote-20) Using the so-called “potential outcome” perspective (Rubin, 1974; Imbens & Rubin, 2015), we defined the effect of GLS in an exposed county as the difference between observed suicide-related hospitalizations and the hospitalizations that would have been observed in the absence of the intervention. Formally, we proposed that for the counties exposed to GLS after exposure started (but not necessarily for all counties and years), there were a pair of potential outcomes, Yit(0) and Yit(1), corresponding to each possible exposure status. We only observed, however, the outcome corresponding to the actual exposure, i.e., Yit = Yit(Wit). The effect of the intervention in a county exposed to GLS during a year after the start of exposure can be defined as

for i > N0 and t > t0i. We hoped to estimate the unobserved outcome Yit(0) using information from unexposed county-years; i.e., suicide-related hospitalizations among exposed counties before the start of exposure, and information among unexposed counties during the entire period. In other words, we wanted to use **Y**(0), the N by T matrix with elements Yit = Yit(0) whenever Wit = 0, and missing entries otherwise.

From this perspective, traditional and more recent proposed approaches can be seen as different ways to combine the available information in **Y**(0), and possibly additional covariates, to estimate the missing entries in this matrix. For example, Athey et al. (2018) proposed that if **Y**(0) can be thought as a low rank matrix plus some noise, i.e.,

,

singular value decomposition coupled with regularization can be used to estimate **L** and to impute the missing entries in **Y**(0). On the other hand, a traditional difference-in-differences (DID) approach (e.g., Bertrand et al., 2004; Angrist & Pischke, 2009) can be thought of as advancing a very specific factor structure for the underlying matrix **L**,

,

i.e., a sum of county- and year-specific effects. Athey et al. (2018) discussed the relationship with other approaches such as interactive fixed-effects or synthetic control. Regardless of the approach, the causal interpretation relies on some form of the ignorability or unconfoundedness assumption (Rosenbaum & Rubin, 1983); i.e., after taking into account the information on **Y**(0) and possibly additional county characteristics or features, assignment to the intervention (both whether the county is exposed and the year the exposure started) is as good as random. Athey & Imbens (2018) discussed more detailed assumptions for staggered adoption settings.

The spatiotemporal approach

Our proposal involved using a spatiotemporal model to impute the missing entries **Y**(0). Using a conventional choice in disease mapping, we assumed that the number of observed hospitalizations was distributed as

,

where was the relative risk in county i, and was the expected count under homogenous risk,, across counties for the entire period.[[21]](#footnote-21) The log of the relative risk can be modeled as a function of a linear predictor, i.e.,

where represents a set of observed county-specific characteristics that may vary over time (such as number of hospitalizations for a cause unrelated to suicide). In disease mapping, it is frequently assumed that , i.e.,a parametric function linear on a set of parameters, , is used to model the relationship between the relative risk and the covariates.This does not need to be the case, and much more flexible approximations are common in impact evaluations. Regardless, spatiotemporal models are primarily distinguished by the way in which the error term is conceived. For instance, if the errors were taken to be independent across counties and over time, the model would correspond to an overdispersed Poisson regression. Classical models for small area estimation (SAE), on the other hand, introduced count-specific random effects, setting , with and representing the residual noise, to borrow strength for estimation across counties (Fay & Herriot, 1979). These SAE models smooth small county estimates toward a global rate. Disease mapping models, in contrast, introduce a spatial structure to borrow strength locally rather than globally. A frequent choice is to assume that each follows a conditional autoregressive (CAR) model (Clayton & Kaldor, 1987; Besag et al., 1991),

where each wij is a user-defined spatial dependence weights determining which counties j are “neighbors” to county i, typically based on adjacency. In other words, the expected value of the error in county i, is an average of the value in the adjacent counties, and the variance is inversely proportional to the number of neighbors. Both spatial and nonspatial random effects can be combined. For instance, using the recently proposed parameterization of the Besag et al. (1991) model by Riebler et al. (2015),

,

where is the variance, 0 is a mixing parameter determining the amount of variance that is spatially structured, is the identity matrix, and is the precision matrix of the (standardized) spatial component.

Temporal dependence could also be accommodated, for instance, using a first order autoregressive model for , i.e., , where is the autocorrelation coefficient and s are the independent shocks (as in Rao & Yu, 1994, in the context of SAE).

Rather than independent spatial and temporal components, a spatiotemporal variance covariance matrix for the error term can be constructed using the Kronecker product of the spatial and temporal precision matrices, as suggested by Knorr-Held (2000) (see also Martínez-Beneito et al.[2008] for a closely related approach):

1. Use of covariates

* “It is not clear how covariates and program intervention indicators are incorporated into the proposed DM model given in Section 2.1” (Reviewer 2)

Here we discuss the use of covariates in the main counterfactual estimation. We will discuss an additional use of covariates (i.e., to examine effect heterogeneity) in the next title.

The program indicator will not be used as ordinary covariate (to be included in a regression, for instance). Instead, the program indicator will be used to create the incomplete matrix **Y**(0) described above in our casual framework. The elements of this matrix are equal to the observed values of the outcome of interest when the program indicator is zero and are set to missing otherwise. This equates to making no assumptions (advancing no priors) regarding the shape of the program effect itself (e.g., whether it is homogenous across units or over time).

Our approach, like the competing causal methods with which it will be compared, gives preponderance to one particular covariate, i.e., the pre-intervention values of the outcome (i.e., the non-missing elements of **Y**(0)). As noted by Doudchenko and Imbens (2017, page 19) “in terms of predictive power the lagged outcomes tend to be substantially more important, and as a result the decision how to treat these other pre-treatment variables need not be a very important one”. That said, our approach can easily incorporate additional covariates.

In our case study, we propose to start by flexibly estimating , for example using Bayesian Additive Regression Trees (BART, Chipman et al., 2010; Sparapani et al., 2016). We then estimated the full spatiotemporal model, plugging in the estimated values from obtained in the first step as a single regressor. For that purpose, we will regressed a direct estimate of the standardized ratio on a set of covariates, including the county’s historical youth suicide rates (from 1999 to 2006) and demographic characteristics of the county, such as the proportion of the population that is American Indian/Alaska Native, unemployment rate, median income, population without health insurance (from 1999 through the current year), and rates of hospital use among youth and adults for causes unrelated to suicide (from 2008 to the current year). We also included time- and county-specific effects, as well as the state, region, division, and the six-class urban-rural classification. Even though it is not theoretically required, recent work recommends including as a predictor an estimate of the (time-varying) propensity score as a function of these same covariates (Hahn et al. 2020) to avoid confounding induced by regularization. These propensity can itself be estimated using BART as well.

1. Effect heterogeneity and machine learning

* “Effect heterogeneity is not well developed.” (Reviewer 1)
* “Machine learning and effect heterogeneity are key here, but underdeveloped.” (Reviewer 1)

Our exposition of how effect heterogeneity will be analyze in the proposal is rather compact. We present a more detail description bellow.

The exploration of effect heterogeneity has traditional been limited to a few suspected ‘effect modifiers’, hopefully specified at the design stage. Recently, there has been increased interest in carrying out this examination more systematically using supervised machine-learning algorithms (e.g.,Athey and Imbens 2015; Wager and Athey 2017). Our specific proposed approach to examine effect heterogeneity, builds upon the following two advantages of our general approach to counterfactual estimation,

1. our approach will result on ‘individual-level’ estimates (i.e., an estimate of the GLS effect for each county and each year following the start of the intervention), however noisy or instable, together with associated measures of uncertainty, rather than just an estimates of an average effect among some population;
2. we can summarize these individual estimates in any way, including by using any machine learning algorithm to learn patterns, and afterwards obtain valid Bayesian inference since, as argued by Carvalho et al. (2019), this simply entails summarizing the posterior distribution after conditioning on the data only once.

After fitting the spatiotemporal model described above, we obtained estimates of counterfactual outcomes (and, therefore, estimates of the GLS effect) for each county and year following the start of the intervention in the county together with their posterior distribution. To explore and summarize these results, we used two recursive partitioning algorithms.

First, we used model-based recursive partitioning (Zeileis et al., 2008) to identify groups of counties with distinct patterns of GLS effects over time. In this approach, different coefficients for a regression model are considered in each partition rather than just a different average value for the outcome, as in more traditional regression trees. In our application, the regression model considered will be just the estimated effect in each county-year regressed on the number of years since the start of exposure as well as the square of the number of years. Therefore, groups of counties with distinct patterns (specifically, second order polynomial curves) of GLS effects over time were distinguished.

At these stage, we will considered multiple contextual characteristics, such as

1. the community’s size in terms of the youth population;
2. racial/ethnic composition (percentages of Hispanic and non-Hispanic White, African American, AI/AN, and other races);
3. median household income, poverty rates, unemployment rates, and percentage of the population with no health insurance; and
4. rurality (using the NCHS six-level urban-rural classification scheme [Ingram & Franco, 2014]).

Most of this information will come from the U.S. Census Bureau; however, we will obtain unemployment estimates from the U.S. Department of Labor and racial/ethnic composition and the urban-rural classification from NCHS.

Once the groups are identified, we will use conditional inference regression trees (Hothorn et al., 2006) to examine which intervention characteristics were associated with membership in each group. This algorithm is closer to the original recursive partitioning trees from Breiman (1998) but uses a statistical motivated stopping rule rather than cross-validation and pruning to find the size of the tree.

We will use the following indicators for programmatic characteristics as potential predictors for differences in impact,

1. the proportion of trainees per youth population (an indication of program reach);
2. the number of years in which trainings occurred (an indication of program continuity over time);
3. the proportion of short (less than 2 hours) and long (more than 8 hours) trainings (training characteristics were previously suggested to be associated with training outcomes [Condron et al., 2015, 2019]);
4. the proportion of trainings implemented by tribal GLS grantees (a possible indication of cultural appropriateness);
5. the proportion of trainees under 18 years old (an indication of life skills and peer support emphasis vis-à-vis the adult gatekeeper model); and
6. the proportion of trainees who have roles in education and mental health settings.

The source for all these indicators will GLS National Outcome Evaluation. The number and type of indicators is limited by the availability at the level of spatial granularity needed to conduct this analysis. Consequently, most programmatic characteristics were based on training activities implemented in each county.

1. Parametric assumption and Inference for program effects

* “The proposed Bayesian disease mapping method requires more restrictive parametric assumptions than those competing causal methods, such as SC, EN and MC.” (Reviewer 2)
* “In addition, its inference including confidence interval and hypothesis test for assessing program effects may be hard to study.” (Reviewer 2)

Whether the difference between observed and estimated rates in absence of the intervention can be given a causal interpretation depends on some form of the ignorability or unconfoundedness assumption (Rosenbaum & Rubin, 1983). In other words, after taking into account the information on Y(0) and possibly additional county characteristics or features, assignment to the intervention (both whether the county is exposed and the year the exposure started) is as good as random. While this is admittedly a strong assumption, but one that is equally required for all the methods considered (i.e., SC, EN, MC, and DM).

Quantification of uncertainty surrounding the counterfactual estimates does generally involve additional distributional assumptions. This requirement is not exclusive of our approach either. For example, while Athey and colleagues did not discuss inference for MC, it was recently discussed for large samples of both cases and time points (Bai & Ng, 2020). It could be argue that SC does not require distributional assumptions for inference, but only if inference is restricted to the limited placebo studies originally proposed by the authors (Abadie & Gardeazabal, 2003; Abadie et al., 2010, 2015). Finally, there has been recent work on inference for counterfactual estimation with reduced dependence on distributional assumptions (Chernozhukov et al., 2019). This proposals, however, are equally applicable to any of the methods considered (i.e., SC, EN, MC, and DM).

If the distributional assumptions are made, on the other hand, then ‘inference including confidence interval and hypothesis test for assessing program effects’ is particular straightforward in our proposed approach. This ‘straightforwardness’, in fact, has been highlighted as one of the advantage of a Bayesian approaches over frequentist alternatives (e.g., Carvalho et al. 2019; Hill 2011).

All that said, the comparison of method based on predictive accuracy, as proposed, is agnostic with respect to each method’s distributional assumptions or lack thereof. Examples of similar ‘agnostic’ comparisons can be found in Doudchenko and Imbens (2017) and Athey et al. (2018) when introducing the EN and MC methods, respectively.

1. Use of ANOVA to analyze simulation

* “On page 45, there is new text added to justify analysis techniques, but it still not fully developed with a reference to Morris et al for current recommendations. I would have preferred to see that information included here, as the there is room in the grant (minor concern).” (Reviewer 3)

Our reference to Morris, White, and Crowther (2019) seems to suggest that there are additional analytic techniques in their article that we will use but were not detailed in the proposal. This is not the case. We simply meant to state that our choice of analytic approach (i.e., the use of ANOVA) was consistent with current recommendations. Indeed, Morris and colleagues argue simulations should be designed and analyzed more like regular experiments. We also reference an early example of the use of ANOVA to analyzed simulation results.

## A model for the counterfactual outcome

Because we are interested in treatment effects on treated units -and we observe the potential outcomes under treatment- the challenge for estimating ATT is to impute the average of the missing never-treated potential outcomes, . We can estimate this quantity if we put forward some model for the outcome under the never treated condition. A frequent choice for mortality outcome is to assume a Poisson distribution for the observed counts coupled with more complex model for the underlying risk, i.e.,

,

where is the relative risk in county i, and is the expected count under homogenous risk,, across counties for the entire period.[[22]](#footnote-22)

We chose a flexible mixed effect model representation for the relative risk, as the addition of three independent component (sometimes referred as large-, small-, and fine-scale variation), i.e.,

where is a deterministic mean function, such as , with a set of observed county-specific characteristics that may vary over time, (such as number of hospitalizations for a cause unrelated to suicide), a preselected transformation, , the basis chosen to represent the information in , and -dimesntional vector of unknown parameters.

The are structured random components. We can select a similarly flexible way to represent through a basis-function expansion of X,

again is a preselected transformation of observed covariates, , and an -dimensional vector of random coefficients. Finally, is white noise with .

A classical approach in policy evaluation, known as two-way fixed effect (TWFE), sets

a unit and time specific fixed effects. In this approach, deviation from this deterministic function, , are not assumed to be homoscedastic or independent over time, and estimates are obtained that are robust to heteroscedasticity and arbitrary patterns of temporal correlation (but no attempt is made of model the errors explicitly).

Classical models for small area estimation (SAE), introduce county-specific random effects,

to “borrow strength” for estimation across counties (Fay & Herriot, 1979).

These SAE models smooth small county estimates toward a global rate. Disease mapping models, in contrast, introduce a spatial structure to borrow strength locally rather than globally. A frequent choice is to assume that each follows a conditional autoregressive (CAR) model (Clayton & Kaldor, 1987; Besag et al., 1991),

where each is a user-defined spatial dependence weights determining which counties j are “neighbors” to county i, typically based on adjacency. In other words, the expected value of the error in county i, is an average of the value in the adjacent counties, and the variance is inversely proportional to the number of neighbors. Both spatial and nonspatial random effects can be combined. For instance, using re-parameterization of the Besag et al. (1991) model by Riebler et al. (2015),

where is the variance, 0 is a mixing parameter determining the amount of variance that is spatially structured, is the identity matrix, and is the precision matrix of the (standardized) spatial component.

Temporal dependence could also be accommodated, for instance, using a first order autoregressive model for , i.e.,

where is the autocorrelation coefficient and are the independent shocks (as in Rao & Yu, 1994, in the context of SAE).

Rather than independent spatial and temporal components, a spatiotemporal variance covariance matrix for the error term can be constructed using the Kronecker product of the spatial and temporal precision matrices, as suggested by Knorr-Held (2000) (see also Martínez-Beneito et al.[2008] for a closely related approach):

Parametric Bootstrap

Specifically, in the case of counts of events, a bootstrap “sample” of the Oi can be obtained by generating n independent Poisson random variates, each with mean  or . These “data” are then used to obtain EB estimates for the risk or relative risk parameters. The process is repeated multiple times, such that the empirical distribution of the EB estimates from the bootstrap samples can form the basis for an estimate of variance.

The parametric bootstrap method is also increasingly used to obtain estimates of precision in more complex models for risk or relative risk, such as non-parametric mixture models (see, for example Ugarte et al. 2003, and Section 6.7), and fully Bayesian CAR models (e.g., MacNab and Dean 2000, MacNab et al. 2004)

1. This N is frequently thought as a random sample from a large, well-defined population. This conceptualization is quite reasonable for cohort studies (particularly if enrollment was the result of probabilistic sampling as it is sometimes the case). Alternative, N can be though as a finite population, which maybe more compelling when the units are all the US counties. In such case the randomness on observed outcome arises from the treatment assignment. [↑](#footnote-ref-1)
2. Even though the treatment is binary, the different timing of the treatment may result on different effects. That is why there are not just 2 potential outcomes at each time point but entire treatment paths. This is analogous to different dosages of a drug. We could have some response model, [↑](#footnote-ref-2)
3. Ben-Michael et al. (2021) suggests taking some measures to ensured that the sample of units remains constant when averaging across K. In that case, the targeted ATT is not the same as in BART or SDID. [↑](#footnote-ref-3)
4. The number of cases (for an untreated unit-period) is a random variable. The observed count (for past untreated unit-period) is a realization of this random variable.

   We can conceive this random count as arising from a population of size exposed to a certain risk, say , then,

   with mean and variance . This may seem as a very particular parametric assumption, but allows for quite flexible models, since we have not stated how the risk in each unit and period comes to be (there are, in fact as many parameters as observations).

   The observed rate

   is an unbiased estimator of the underlying risk. The variance of the estimator is

   which is inversingly proportional to the size of the population. A Poisson assumption leads to a similar expression. In that case, , and

   This does not change if we replace , by , i.e., indirect standardization. The rate is the ML estimator of the underlying risk, and plugin it in the variance expression results on the ML estimator of the variance. The variance is proportional to the mean, inversingly proportional to the population and estimated to be zero if there happens to be no events. [↑](#footnote-ref-4)
5. The main alternative is so called observation-driven models, e.g.,

   with . This representation may be more compelling for modelling contagious disease (than cancer or suicide). Hybrid models combining observation- and parameter- driven components are also possible (e.g., Paul & Held (2011) allow and to vary over units, and includes random effects in and ) [↑](#footnote-ref-5)
6. A direct, maximum likelihood estimate of the relative risk is the ratio of observed to expected cases. The variance of the direct estimate is inversely proportional to the expected number of cases and can become very large as the expected number of cases decreases. When the outcome of interest is the number of deaths, the ratio is usually referred to as the standardized mortality ratio. For our application, we chose the term standardized hospital use ratio. [↑](#footnote-ref-6)
7. In spatio-temporal models, a basis expansion that has received particular attention are splines, such as B-splines, to obtain a lower rank representation of the matrix of locations. [↑](#footnote-ref-7)
8. While it is barely written in this way, we could consider in terms of Knorr-Held (2000) a spatiotemporal interaction with no spatial or temporal structure, i.e., with , i.e., is just a N.T by N.T diagonal matrix. [↑](#footnote-ref-8)
9. in matrix notation, where is a T-vector of 1’s, and is the identity matrix. [↑](#footnote-ref-9)
10. Alternative neighbors could be defined in terms of distance between centroids. [↑](#footnote-ref-10)
11. The joint (instead of the conditional distribution) is gaussian and proportional to

    with

    This matrix is not full rank; to make RE identifiable we need sum-to-zero constraint. [↑](#footnote-ref-11)
12. This is a type of space-time interaction in terms of Knorr-Held (2000), but one that is not spatially structured, , where ; where is a T by T temporally structured matrix. For example, in the case of AR(1), the element of the matrix are . [↑](#footnote-ref-12)
13. An example of an old comparison focusing on model fit (DISEASE MAPPING COLLABORATIVE GROUP, 2000). Ma be the extent to which the underlying rate is recovered is the most relevant. [↑](#footnote-ref-13)
14. PsiOper = (diag(n) - X %\*% solve(t(X) %\*% X) %\*% t(X)) %\*%

    A %\*% (diag(n) - X %\*% solve(t(X) %\*% X) %\*% t(X)) [↑](#footnote-ref-14)
15. If A is a square matrix

    Where Q is an orthogonal matrix, and R is upper triangular. If A has n linearly independent columns, then the first n columns of Q form an orthonormal basis for the column space of A, they are all unit vectors and orthogonal to each other [↑](#footnote-ref-15)
16. [↑](#footnote-ref-16)
17. The data generation model can be written as follows. The count of cases for area i and time t,

    where N is the total number of areas, T is the total number of periods, is the relative risk in the area i at time t, and is the expected count under homogenous risk (i.e., , where is the population at risk and is some baseline risk). In turn, the log relative risk, from t=2 onwards,

    follows a first order autoregressive process, where is the serial correlation parameter and is a random shock (at t=1,). Importantly, has a spatial structure—that is, for each period t,

    where denotes spatial covariance matrix. Following Leroux et al. (1999), and **,** where is the identity matrix, is a variance parameter, is a spatial dependence parameter, and Q is a matrix with the neighborhood structure. In particular, Q contains the number of neighbors for each region along the diagonal, and off-diagonal elements are if i and j are neighbors and is zero otherwise. Importantly, the total number of areas (N), the population at risk in each area and period (), and the neighborhood structure (Q) are not simulated but taken from the real data. [↑](#footnote-ref-17)
18. County-level characteristics include racial/ethnic composition (percentages of Hispanic and non-Hispanic White, African American, AI/AN, and other races); median household income, poverty rates, unemployment rates, and percentage of the population with no health insurance; and rurality (using the NCHS six-level urban-rural classification scheme [Ingram & Franco, 2014]). [↑](#footnote-ref-18)
19. Intervention-level characteristics include the proportion of trainees per youth population (an indication of program reach); the number of years in which trainings occurred (an indication of program continuity over time); the proportion of short (less than two hours) and long (more than eight hours) trainings (training characteristics were previously suggested to be associated with training outcomes [Condron et al., 2015, 2019]); the proportion of trainings implemented by tribal GLS grantees (a possible indication of cultural appropriateness); the proportion of trainees under age 18 (an indication of life skills and peer support emphasis vis-à-vis the adult gatekeeper model); and the proportion of trainees who have roles in education and mental health settings. [↑](#footnote-ref-19)
20. This so called “staggered adoption” representation is, admittedly, a simplification; we will not consider that exposure may have been intermittent after that initial year. As a result, we may underestimate the effect of GLS [↑](#footnote-ref-20)
21. A direct, maximum likelihood estimate of the relative risk is the ratio of observed to expected cases. The variance of the direct estimate is inversely proportional to the expected number of cases and can become very large as the expected number of cases decreases. When the outcome of interest is the number of deaths, the ratio is usually referred to as the standardized mortality ratio. For our application, we chose the term standardized hospital use ratio. [↑](#footnote-ref-21)
22. A direct, maximum likelihood estimate of the relative risk is the ratio of observed to expected cases. The variance of the direct estimate is inversely proportional to the expected number of cases and can become very large as the expected number of cases decreases. When the outcome of interest is the number of deaths, the ratio is usually referred to as the standardized mortality ratio. For our application, we chose the term standardized hospital use ratio. [↑](#footnote-ref-22)