Causal Inference for Observational Time-Series with Encoder-Decoder Networks (Online Appendix)

December 10, 2017

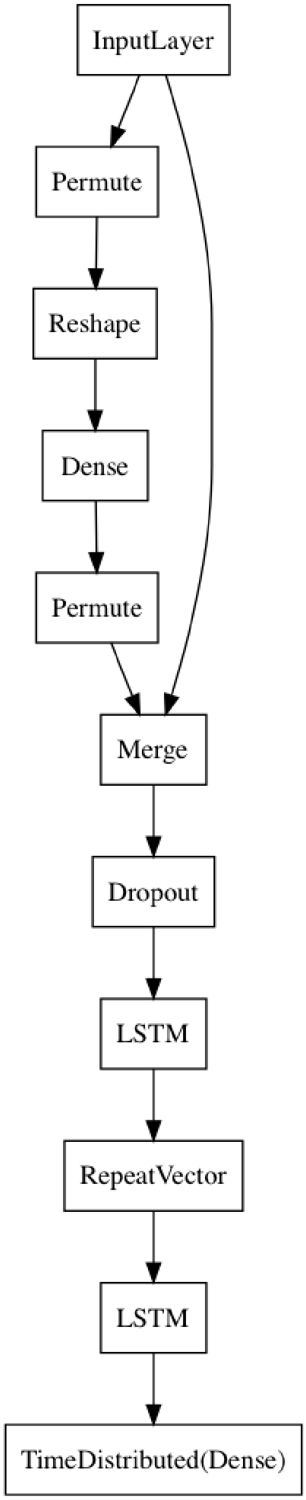

Figure 1: Encoder-decoder networks architecture. The model permutes and reshapes the inputs in order to compute them as importance weights that are automatically optimized by the networks. The output sequence of the first dense layer is permuted in order to perform matrix multiplication on the inputs. Dropout is applied to the LSTM encoder, which drops the temporal dimension of the output. The output of the encoder is repeated and fed to the LSTM decoder, which returns the full output sequence. Lastly, a second dense layer outputs the prediction.

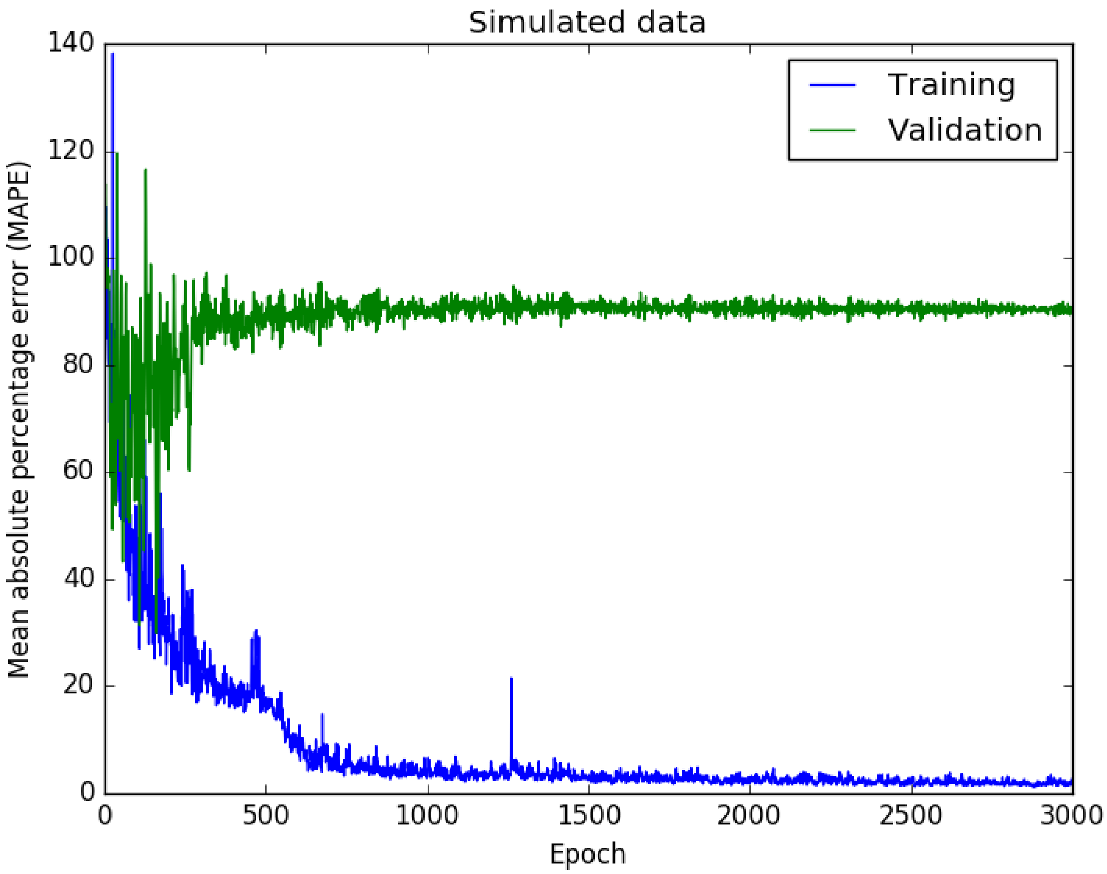

Figure 2: Evolution of encoder-decoder RNN training and validation loss in terms of mean absolute percentage error (MAPE) on ARMA simulated data. The model is trained for 3,000 epochs and check-pointed every 10 epochs.

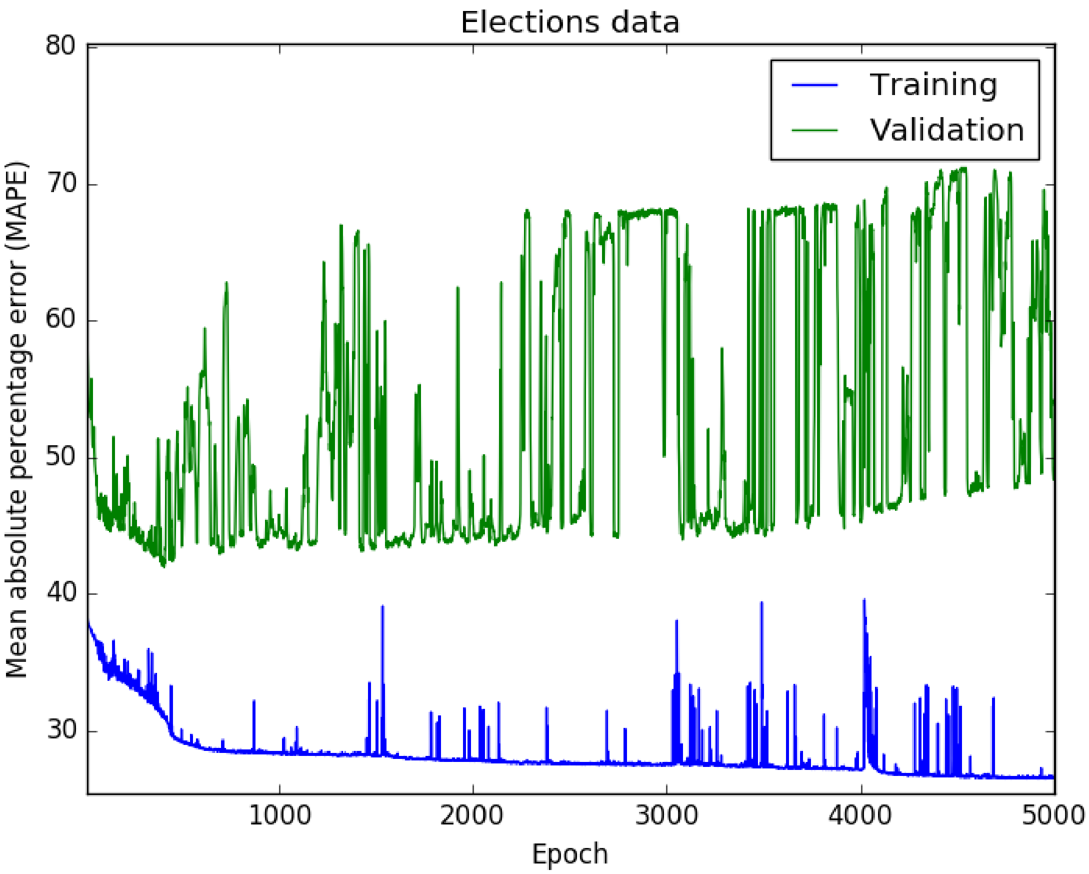

Figure 3: Evolution of encoder-decoder RNN training and validation loss in terms of mean absolute percentage error (MAPE) on mayoral elections data. The model is trained for 5,000 epochs and check-pointed every epoch.

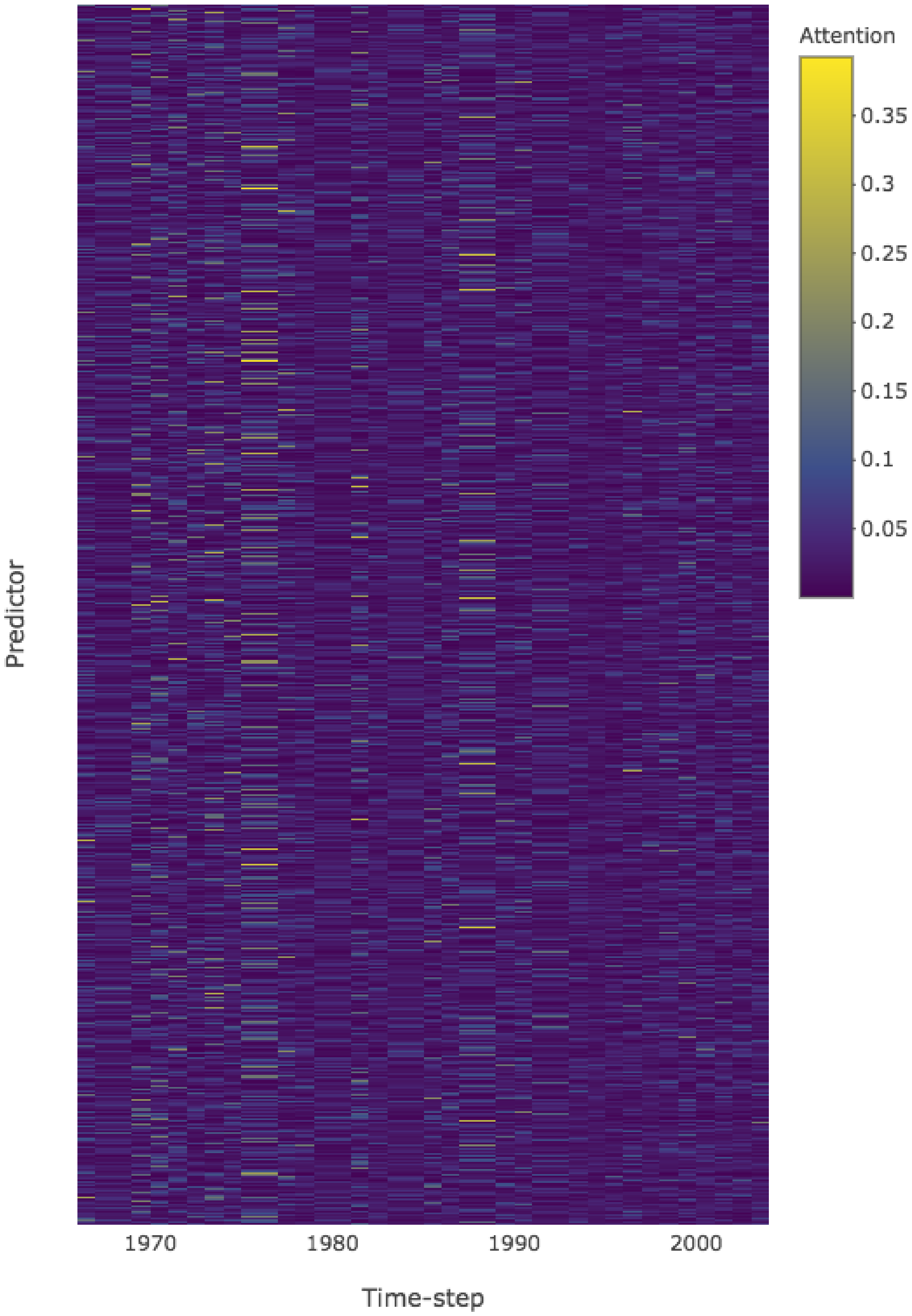

Figure 4: Attention mechanism as a function of predictors and time-steps for encoder-decoder RNN trained on ARMA time-series. Attention is the normalized (softmax) distribution of the importance of each time-step regarding a predictor.

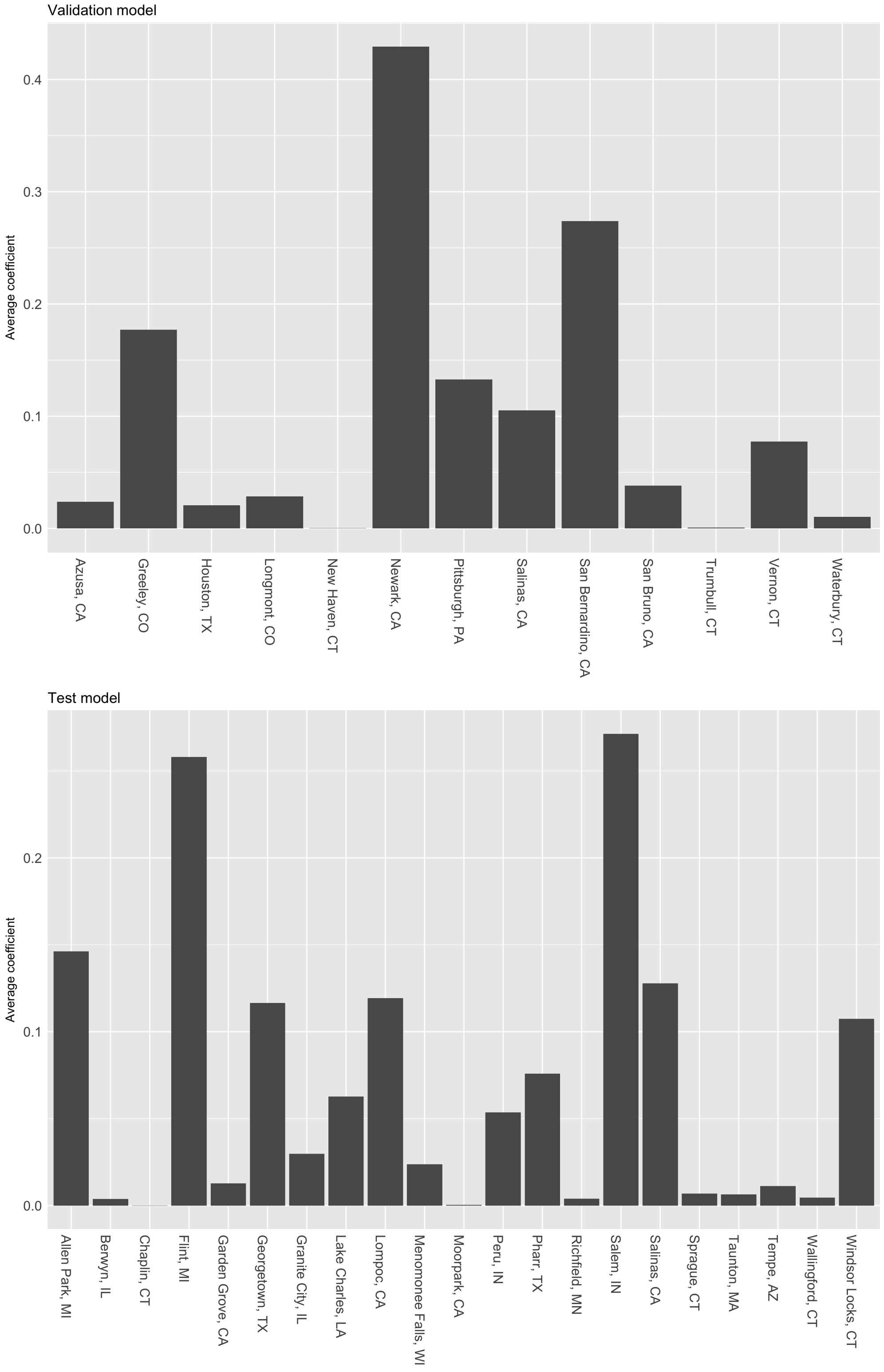

Figure 5: Average coefficients when predictors were selected by BSTS model trained on training set predictors (1948 to 1999) (*top*) and a model trained on both training and validation set predictors (1948 to 2004) (*bottom*). Only predictors with non-zero values are shown. BSTS model with semilocal linear trend and seasonal components and spike-and-slab priors is trained with 10,000 MCMC samples with the first 1,000 samples discarded as burn-in. Each model predicts the time-series of the mean winner margin in Panagopoulos and Green treated cities using only winner margins in non-treated cities as predictors.

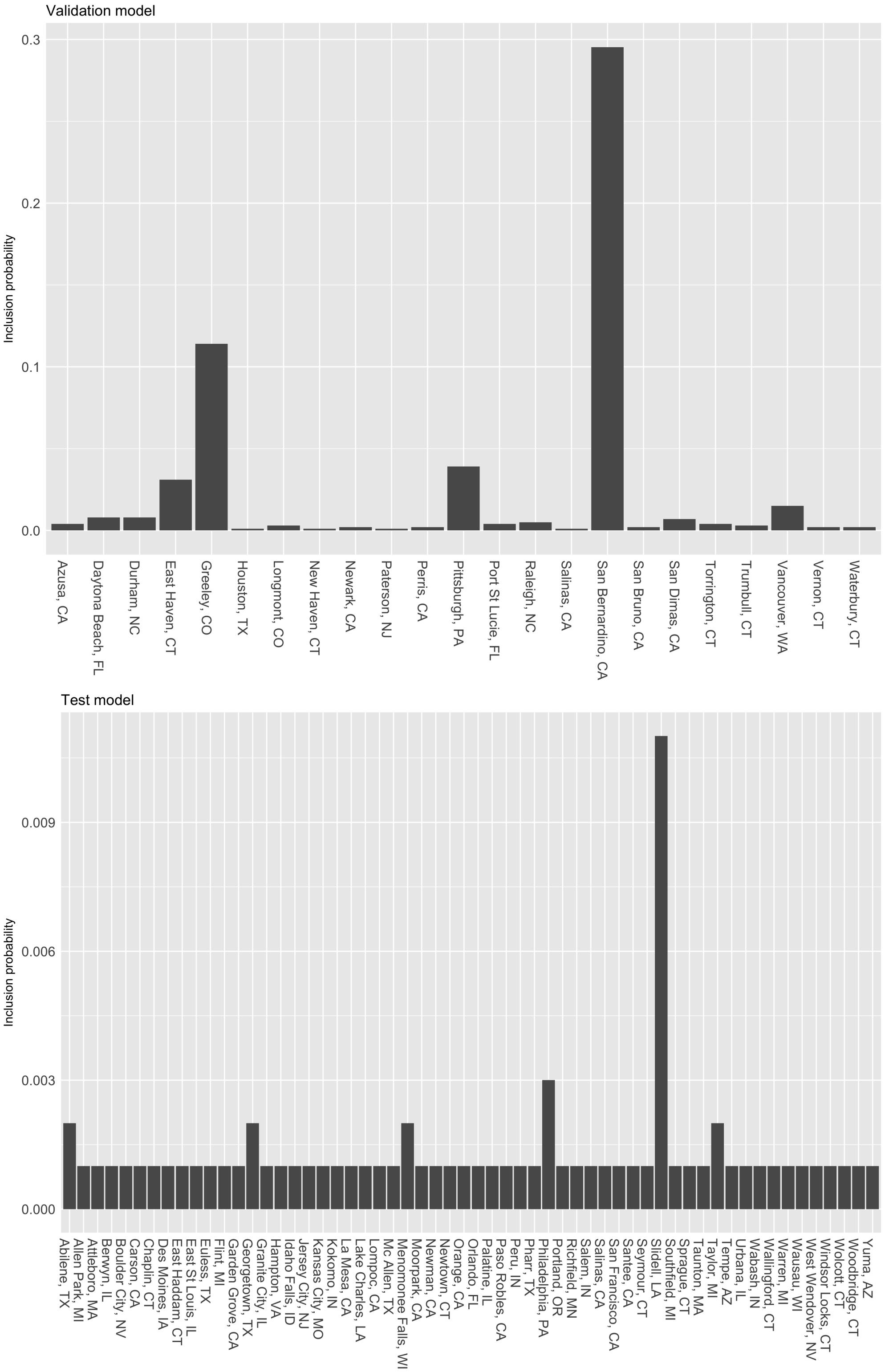

Figure 6: BSTS model inclusion probabilities (i.e., how often the predictors were selected). See notes to Fig. 5.

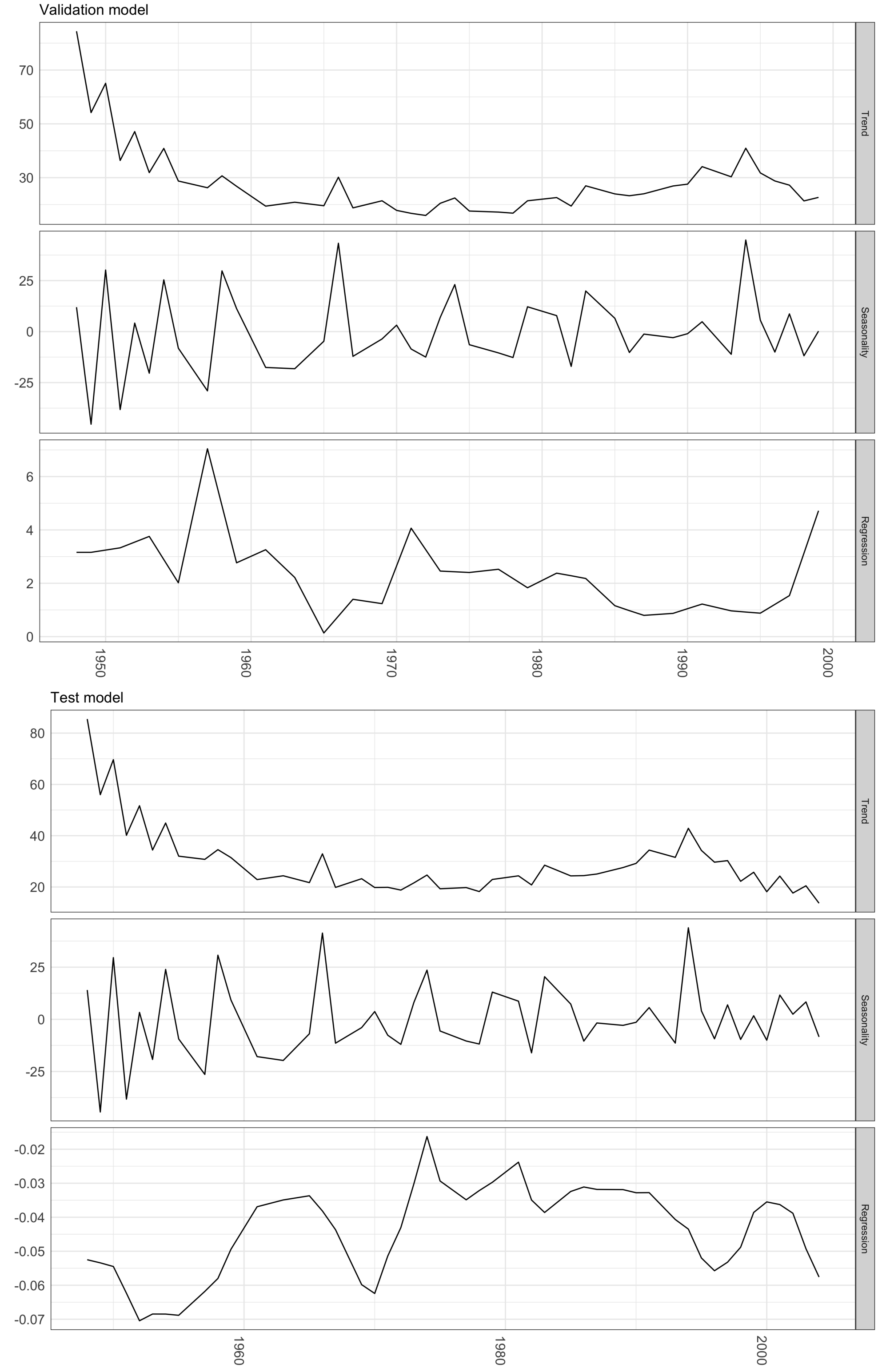

Figure 7: Semilocal linear trend component, monthly seasonal component, and regression component of BSTS model fit on mayoral elections training data. See notes to Fig. 5.

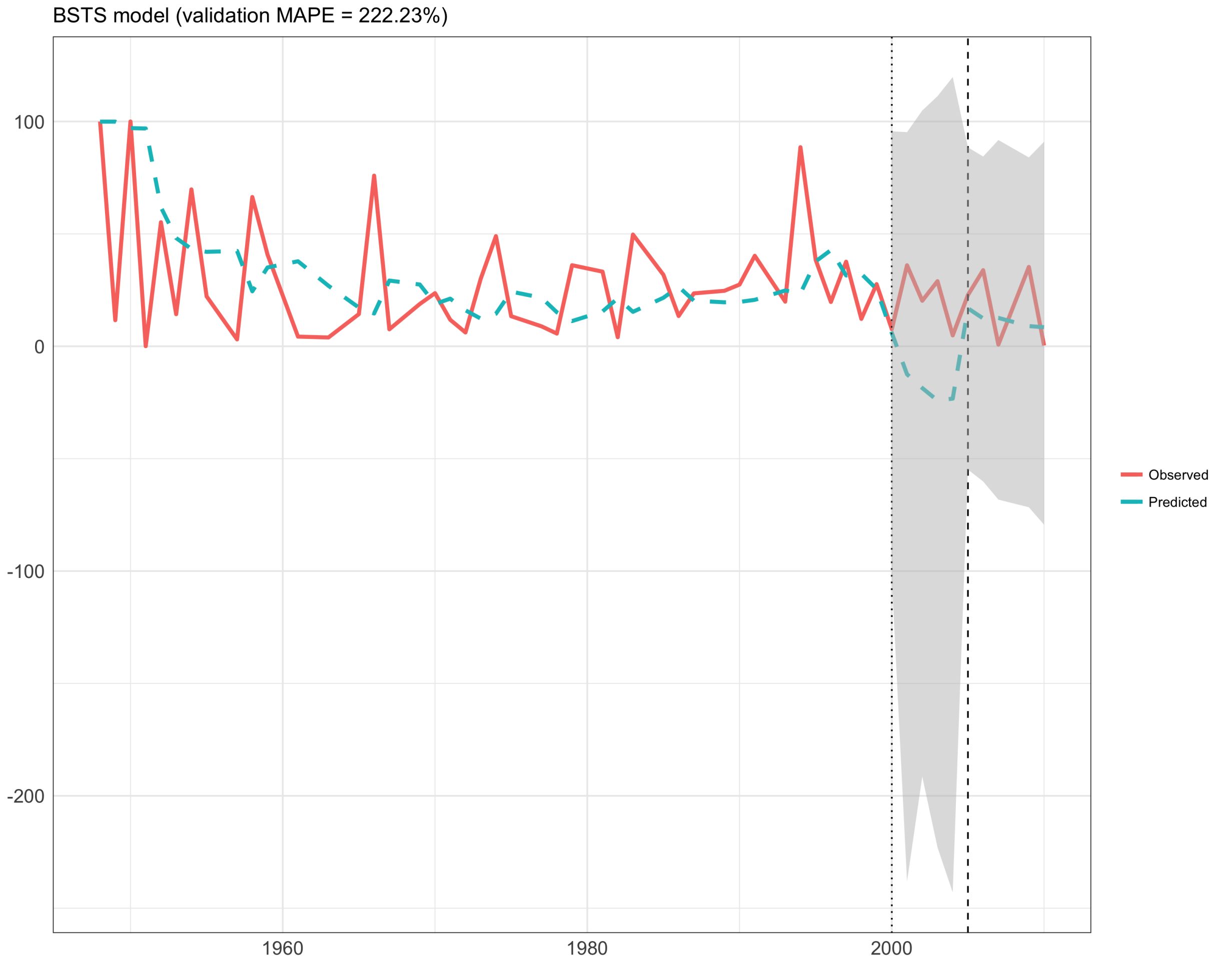

Figure 8: Observed and BSTS predicted winner margins in Panagopoulos and Green treated cities, 1948 to 2010. Predictions are obtained by averaging across MCMC draws and 95% credible intervals (shaded region) are obtained from the distribution of MCMC draws. The dotted vertical line represents the start of the validation set (2000) and the dashed vertical line is the start of the test set (2005). See notes to Fig. 5.

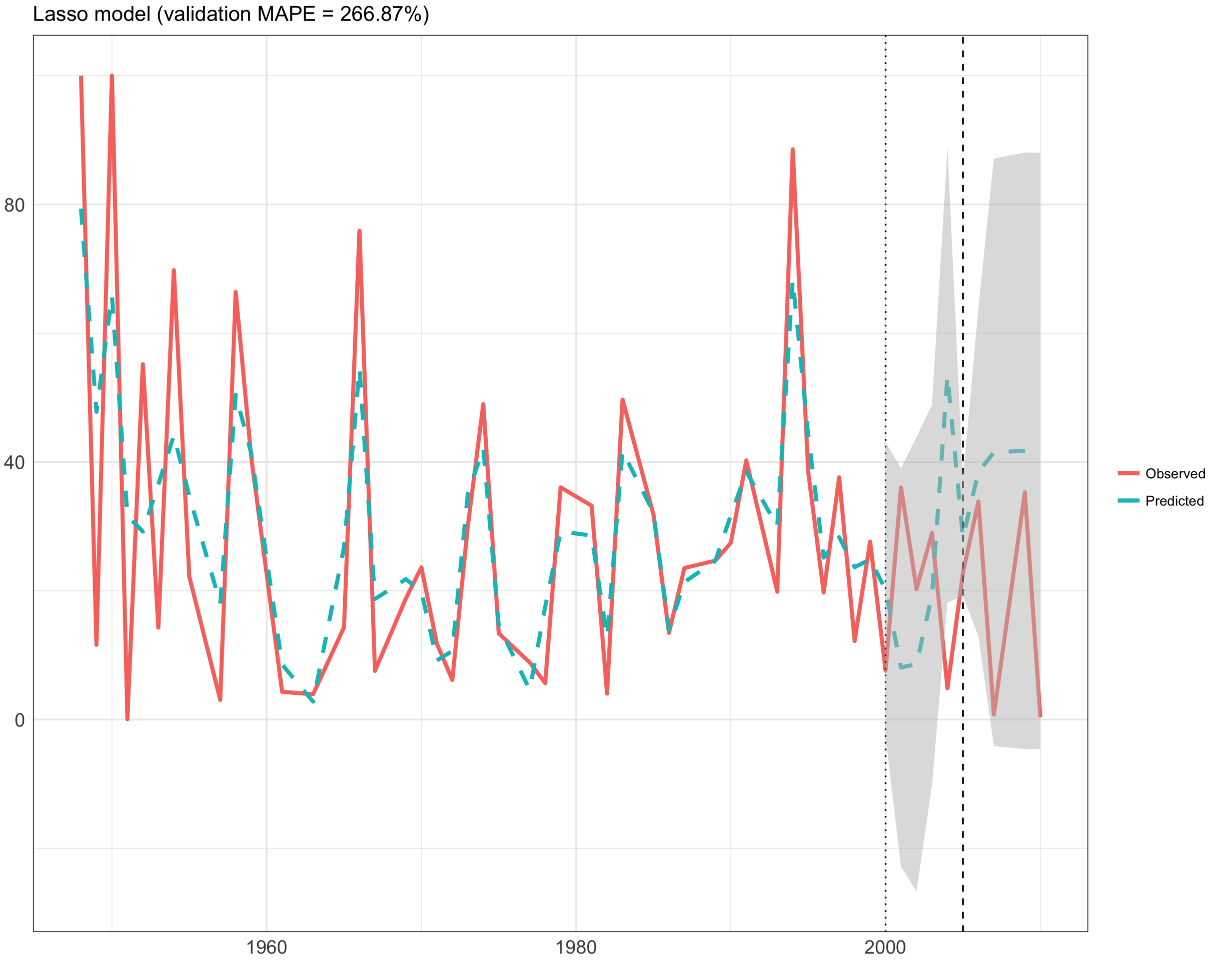

Figure 9: Observed and lasso predicted winner margins. Eq. 1 is estimated by ℓ_1-penalized linear regression via lasso. Predicted winner margins are obtained by averaging across predictions generated by 25,000 different values of the penalty parameter λ and 95% prediction intervals (shaded region) are obtained from the prediction distribution and estimated via Eq. 7. The dotted vertical line represents the start of the validation set (2000) and the dashed vertical line is the start of the test set (2005).

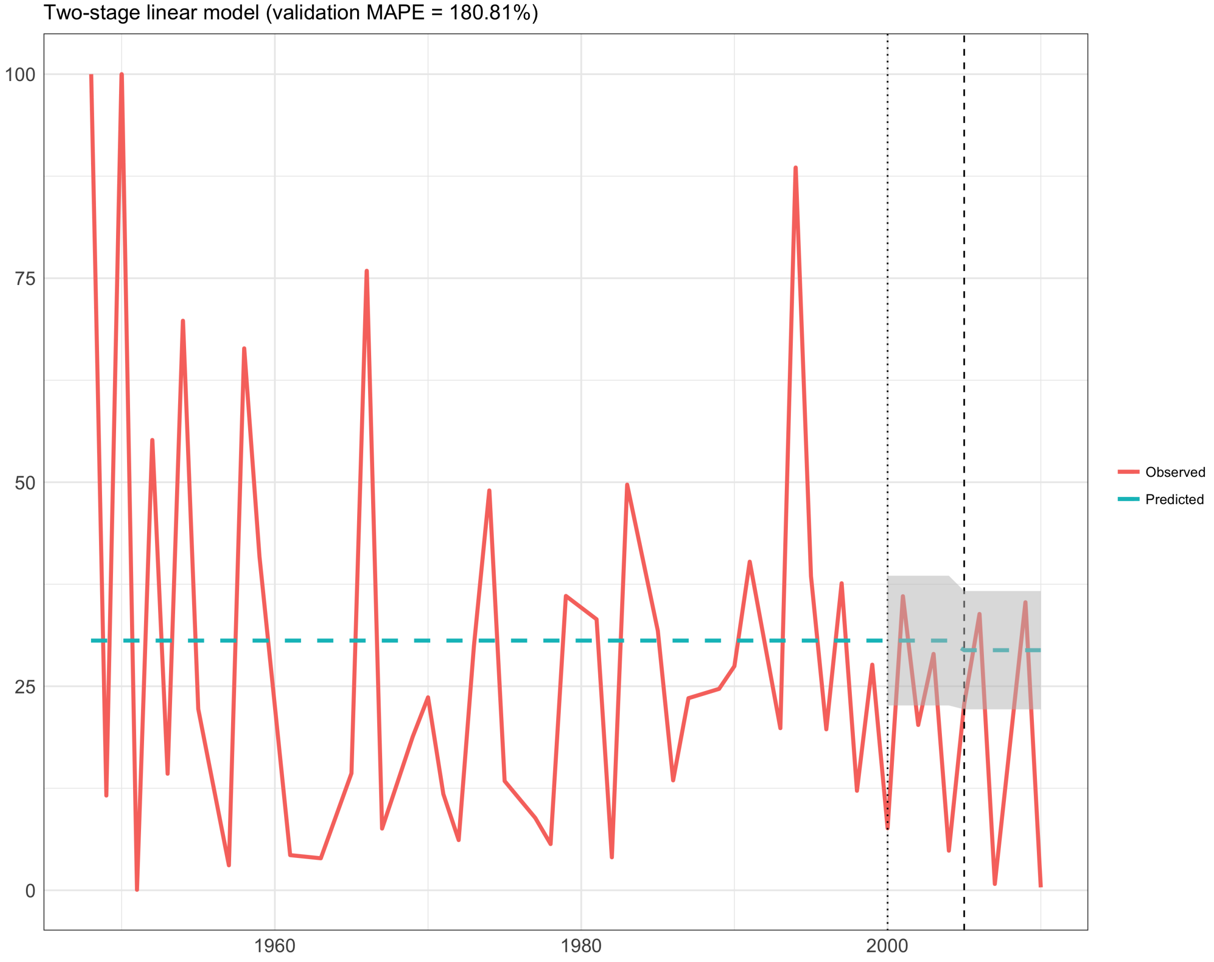

Figure 10: Observed and predicted winner margins from two-step linear model. In the first step, a lasso regression of treated winner margins on control winner margins is run for the purpose of selecting predictors with non-zero coefficients. The penalty parameter λ is selected by 10-fold cross-validation. In the second step, a linear model is fit on the same data using the predictors selected in the first step. 95% prediction intervals (shaded region) are obtained from the prediction distribution and estimated via Eq. 7. The dotted vertical line represents the start of the validation set (2000) and the dashed vertical line is the start of the test set (2005).

References

Brodersen, Kay H, Fabian Gallusser, Jim Koehler, Nicolas Remy, and Steven L Scott. 2015. "Inferring Causal Impact Using Bayesian Structural Time-series Models." *The Annals of Applied Statistics* 9 (1): 247–274.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* Cambridge, MA: MIT press.

Kerman, Jouni, Peng Wang, and Jon Vaver. 2017. *Estimating Ad Effectiveness using Geo Experiments in a Time-Based Regression Framework.* Technical report. Google, Inc.