Bayesian Methods of Modeling Residential Rooftop Solar Adoption in Massachusetts

Lucía Vilallonga

ABSTRACT

With rising concerns about the impacts of climate change and the accessibility and effectiveness of solutions both in Massachusetts and nationwide, it will be necessary to examine the factors that most encourage residents and businesses to participate in a just energy transition. This paper reveals that both single and multiple regression Bayesian models agree that MA towns' voting pattern in the 2020 election is the most significant factor explaining the rate of residential rooftop solar adoption in those towns: generally, towns with a higher percentage of votes for Biden in 2020 have more solar installations. However, given the models' relatively poor fits and the known spatial correlation in voting data, this work should be extended to take these spatial relationships into account. As it stands, the correlation between voting and solar adoption is not surprising, but it is clear that there are more potential factors at work.

INTRODUCTION

In March 2021, Massachusetts governor Charlie Baker signed into law An Act Creating A Next-Generation Roadmap for Massachusetts Climate Policy, which sets greenhouse gas (GHG) emissions limits for 2030 and requires interim limits every 5 years. Crucial to the Act is such programs as the Solar Massachusetts Renewable Target (SMART), which inventivises the adoption of photovoltaic (PV) systems by MA residents and businesses, and a Clean Energy and Climate Plan.[1] Clearly, the Massachusetts electric grid is shifting towards more renewable sources of energy: in 2020, solar energy accounted for 19% of Massachusetts' total in-state electricity net generation, and the state ranked ninth in the nation in net generation from all solar in 2020.[2] However, the Act shows that emissions are still not keeping pace with the state's goals. In order to make the transition more effective and more equitable and to reduce emissions more quickly, it will be necessary to determine the factors that encourage stakeholders to adopt renewable energy generation over traditional fossil fuels. This paper specifically focuses on residential rooftop PV systems among homeowners in 350 towns of Massachusetts. The potential factors investigated are: education, income, and voting pattern in the 2020 presidential election. The initial hypothesis was that income would be the most important factor predicting rooftop solar adoption, because of the high upfront costs associated with adoption (even after incentives)[3], but that voting pattern could also be significant.

METHODS

The data were collected from multiple sources:

Table 1: Data sources

Data	**Variable**	**Source**	**Notes**
Existing rooftop PV installations in MA, by town and year	Y, number of adoptions per town	Massachusetts Clean Energy Center (MassCEC)[4]	In this case, town-level distinctions were equivalent to the precincts in the other datasets.
Household income in MA by precinct	x_1, average income per town	NHGIS[5]	Reported as the mean of 2015-2019 incomes as reported in the annual census (2019 USD). "Educational attainment" was reported as one of 24 categories, which was mapped onto a rough "years of education" variable and aggregated up to the town level in data processing.
Educational attainment in MA by block group	x_2, average educational attainment per town	NHGIS[5]	Provided as the mean educational attainment of adults aged 25 and older in 2015-2019, by block group. Data were aggregated up to the town level in data processing.
2020 presidential election results in MA by town	x_3, % of town that voted for Biden in 2020	WBUR[6]	Data presented as a map and a table, which was downloaded to a PDF and manually converted to a CSV file in data processing.

In addition to the covariates of interest, NHGIS provides geographic layers, which can be used in further analyses.

Data processing, to aggregate and format the data, was done in Python (Appendix A). Discrete datasets were joined by their spatial components, since each one then included a term for the town after aggregating. No data were removed during processing.

First, it was necessary to determine which covariates are correlated with the response variable, Y. A pairs plot reveals that x_1, income, and x_2, education, are heavily correlated, and that both are loosely correlated with x_3, voting (Figure 1). x_3 also seemed to have the strongest correlation with Y, and since it is correlated with both x_1 and x_2, it's possible that x_3 already captures the relationship that each of them has to Y. This makes sense: it has been reported that Democratic voters are more likely to be college-educated than those who vote for Republicans,[7] and the ability to attend college is highly dependent on family and personal wealth. This may be particularly true in Massachusetts, which is an overwhelmingly blue state and is also home to several elite universities with high tuition rates.

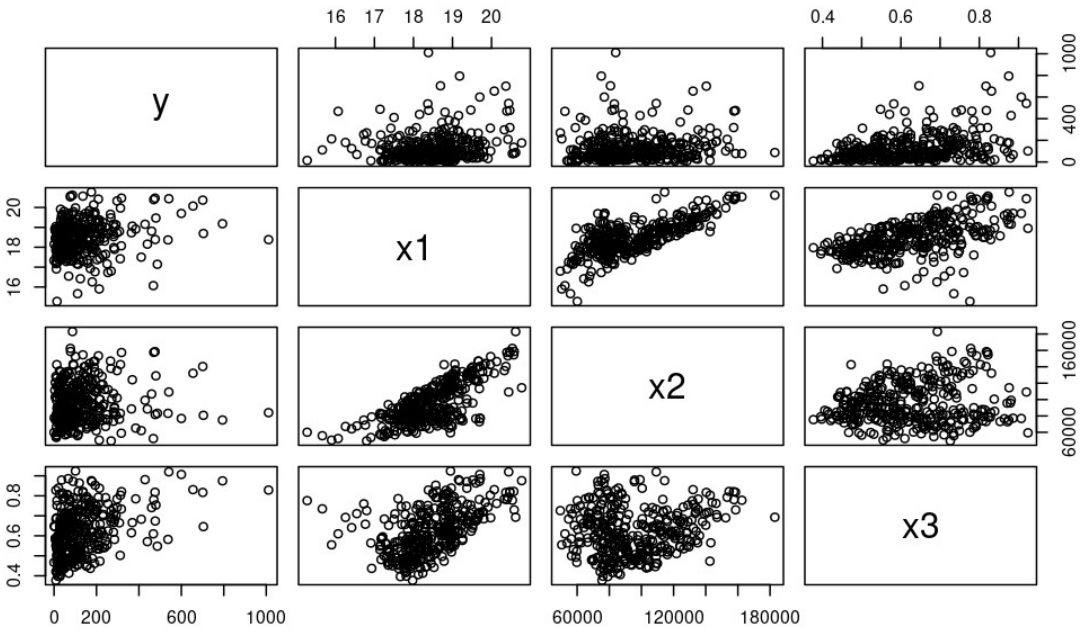


Figure 1: pairs plot of correlations between covariates and response variable.

With this insight in hand, I construct models relating Y to x_3, starting with a Bayesian Poisson regression:

Model 1: Bayesian Poisson regression

Process	$\log(\lambda_i) = \beta_1 + \beta_2 x_i$
Data	$y_i \; Poisson(\lambda_i)$
Parameters	$\beta_i \; Normal(B_{0,i}, V_{b,i})$

A Poisson regression was chosen for this model because Y is a discrete, nonnegative count variable.

Model 2: Bayesian linear regression

Process	$\mu_i = \beta_1 + \beta_2 x_i$
Data	$y_i \; Normal(\mu_i, S)$
Parameters	$\beta_i \; Normal(B_{0,i}, V_{b,i})$

To further explore the relationships between the other covariates and the response, I also analyzed the results of Bayesian multiple regression models, both with and without errors in variables. In these models, β is an *mx1* vector of parameters and X is an *mxn* matrix of covariates.

Supplementary model a: Bayesian multiple linear regression

Process	$\mu_i = \beta \, X_i$
Data	$y_i \; Normal(\mu_i, S)$
Parameters	$\beta_i \; Normal(B_{0,i}, V_{b,i})$ $S \; Gamma(s_1, s_2)$

Supplementary model b: Bayesian multiple linear regression with errors in income

Process	$\mu_i = \beta \, X_i$
Data	$y_i \; Normal(\mu_i, S)$ $X_{income,obs,i} \; Normal(\alpha_i X_{income,i}, \tau^2)$
Parameters	$\beta_i \; Normal(B_{0,i}, V_{b,i})$ $S \; Gamma(s_1, s_2)$

	$\alpha_i \, Normal\left(a_{0,i}, V_{a,i}\right)$

In this case, I had prior knowledge of the variance in the observed income data, thanks to a report by the US Census Bureau.[8] Instead of sampling tau, I computed the standard error (SE) of the data from reported margins of error (MOE), which give a 95% confidence interval of each reported income value. Inputting this directly into JAGS as a precision greatly simplified the code and model outputs.

Supplementary model c: Bayesian state-space time series with missing observation
This model, which does not have covariates in the same way that the regression models do, attempts to account for how much of the growth in rooftop solar adoption in Massachusetts is attributable to simply the passage of time, perhaps as more people become aware of the option and as prices decrease. Here, the only observed variables are Y, the total number of adoptions across all towns, and t, the observation year (the year in which the installation came online and was reported to MassCEC, 2000-2021).

Process	$N_t = N_{t-1} e^{\epsilon_{t-1} + r}$ $X_t = \log\left(N_t\right) = X_{t-1} + r + \epsilon_{t-1}$
Data	$Y_t \, Normal\left(X_t, \tau^2\right)$
Parameters	$\sigma^2 \, IG\left(s_1, s_2\right)$ $\tau^2 \, IG\left(t_1, t_2\right)$ $r \, Normal\left(r_0, V_r\right)$ $X_0 \, Normal$

Here, X_t are the latent time series, and the model has both a process error (σ^2) and an observation error (τ^2), and a prior on the initial condition (X_0).

RESULTS

Single regression models

The Bayesian Poisson model was initialized using uninformative priors and sampled over 70,000 iterations with 3 MCMC chains. Trace plots revealed satisfactory convergence, and after a burn-in of 5,000 the GBR was below 1.05 (see Appendix B for model convergence plots). The effective size of each beta parameter was around 2,000, which is small. This model could likely be improved upon with better initial conditions and more informative priors. A larger sample size might be possible on another machine; memory problems made the code impossible to run

with more than 70,000 samples. Plotting the credible and predictive intervals for this model suggests that it may not be the best fit: it is very over-confident, and does not capture even a majority of the variance in Y (Figure 2).

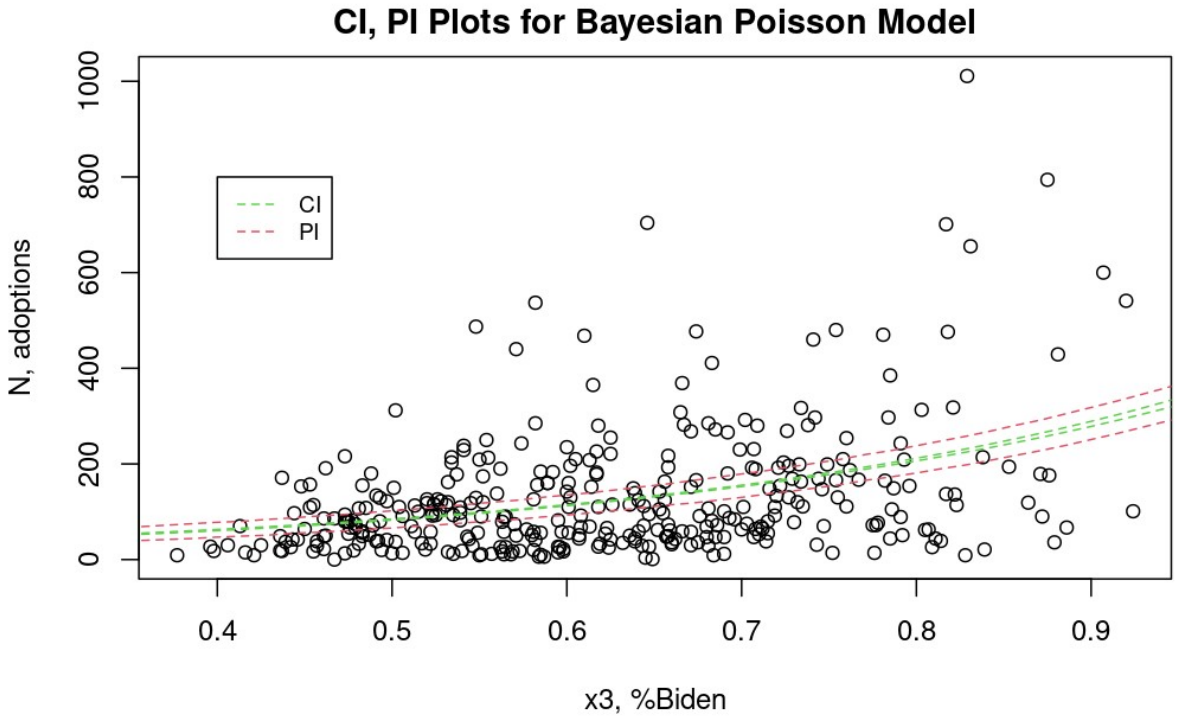


Figure 2: credible and predictive intervals for the Bayesian Poisson model.

The Bayesian linear model was also initialized with uninformative priors and 3 MCMC chains, though this time just 30,000 samples were sufficient for convergence. After a burn-in of 2,000, the effective size of each beta parameter was almost 8,000, and that for S was over 69,000. Unfortunately, this model seems to suffer from the same overconfidence as the Poisson model (Figure 3). However, by both DIC and WAIC, the linear model was a better fit than the Poisson (Table 2).

Table 2: single regression model fit criteria

	DIC	**WAIC**
Model 1	31,990.451	0
Model 2	4,392.125	-446.68

Table 3: model parameter values

Parameter	Model 1 (log scale)	Model 2
β_1	2.904	53.252
β_2	3.048	118.986
S	-	0.609e-4

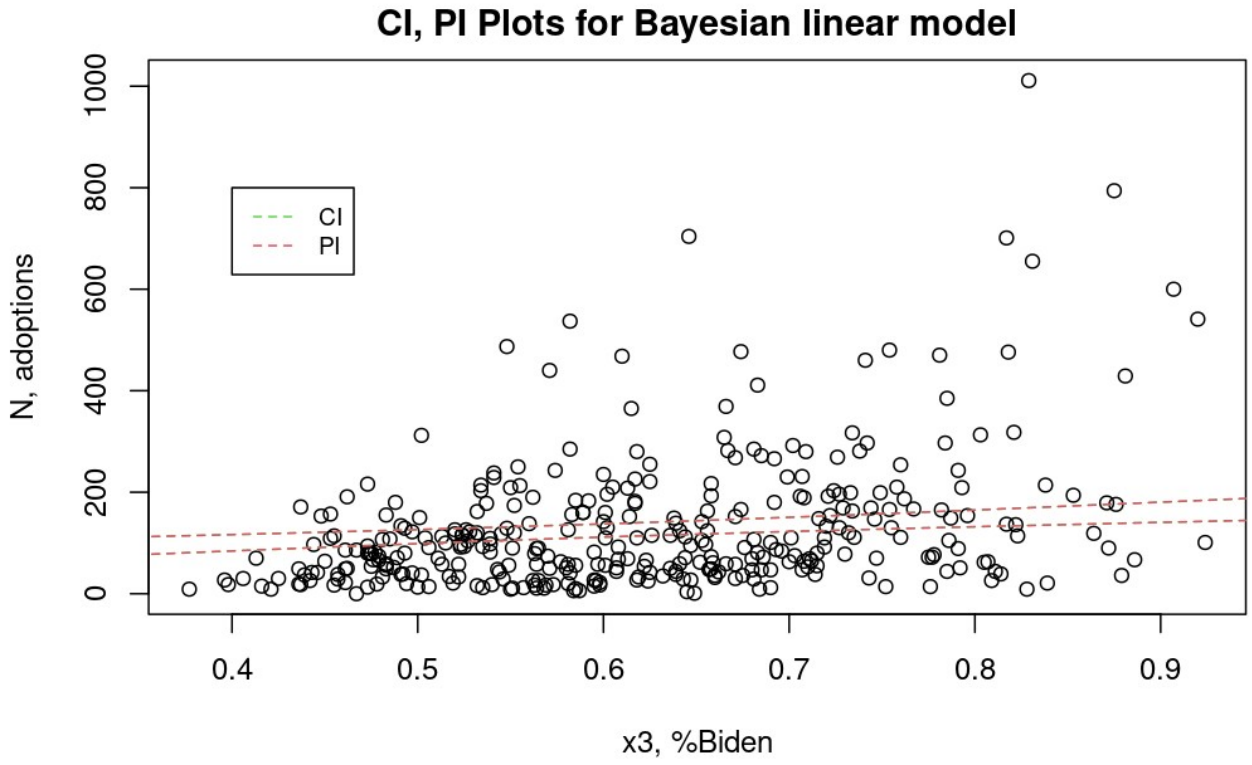


Figure 3: credible and predictive intervals for the Bayesian linear model.
Note: it is unclear why the green CI lines don't appear on the plot.

Multiple regression models

According to DIC, the errors-in-variables model was a better fit (the ΔWAIC is negligible). Interestingly, both models arrived at very similar parameter estimates, including a slight contribution from income. However, voting pattern (β_4) remains the primary model covariate.

The errors-in-variables model may have performed better than simple multivariate regression because it was better able to capture the variability in income values (Figure 4), which are notoriously imprecise in census data.

Table 4: single regression model fit criteria

	DIC	WAIC
Model a	4,372.516	-152.300
Model b	1,095.995	-152.284

Table 5: model parameter values

Parameter	Model a	Model b
β_1	-43.568	-44.797
β_2	-3.172	-3.067
β_3	$0.450e^{-3}$	$0.446e^{-3}$
β_4	304.406	303.776
S	$0.650e^{-4}$	$0.555e^{-4}$

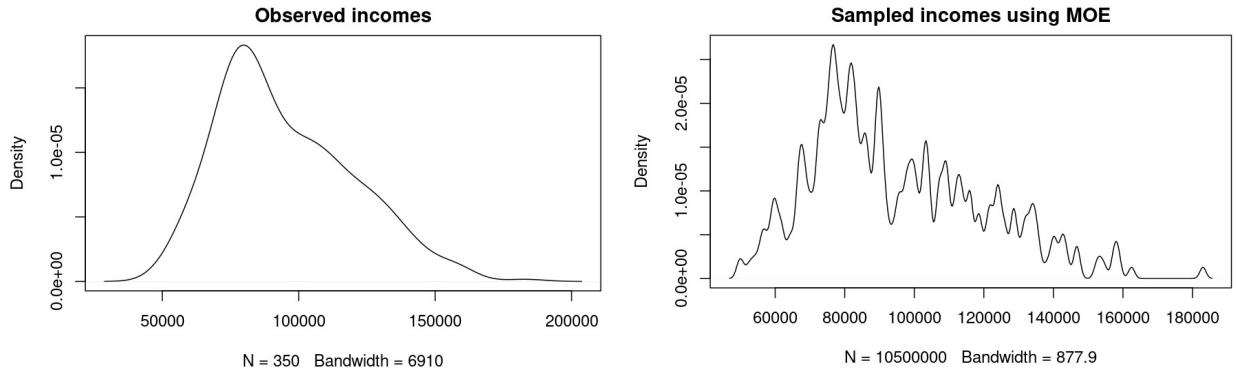


Figure 4: comparison of observed vs sampled incomes.

Since the multiple regression models have additional data, their DIC and WAIC scores cannot be compared to those of the linear models. Still, the best of the two sets of models were: (1) Bayesian single linear regression on %Biden and (2) Bayesian multiple linear regression with errors in income variables. Interestingly, all the linear regression models arrived at similar values for S, the precision on Y, but the addition of the income term in the multiple regression decreased the effect of the %Biden term compared to the single regression case. This may mean that both income and %Biden can describe rooftop adoption, but the two are still very inter-correlated.

Time series model

The initial conditions used were: $r = 0.8, \tau^2 = 100, \sigma^2 = 1$ and the latent X's were initialized to the values of the observed Y's. With 3 chains and 200,000 samples, and after a burn-in of 10,000, the model converged satisfactorily and the effective size for all the parameters, including the missing observation in 2003, were well over 10,000.

The model follows the data pattern well, with wider credible intervals where there are more observations or more uncertainties in the observations (missing value in 2003), and narrower intervals elsewhere (Figure 5). Given how narrow the intervals are overall however, I wonder whether this model is over-confident.

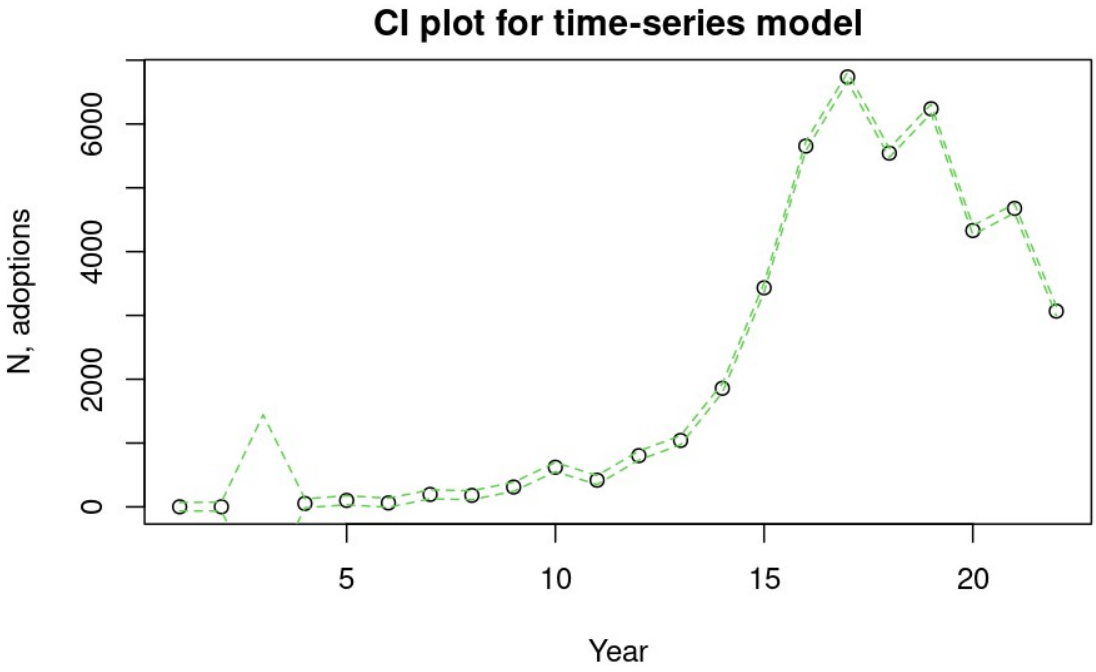


Figure 5: credible interval plot for time-series model with missing observation (2003).

The model estimated that there were about 27 installations in 2003, the year of the missing observation, and that the year-over-year growth rate, r, is about 3.3 (Table 6). The missing observation has a very wide confidence interval, however: ranging from -1,371 (not possible) to 1,424. The confidence interval for r is much narrower (Appendix C: Figure 7).

Table 6: model parameter values

Parameter	**Estimate**
r	3.333

σ^2	$0.011e^{-4}$
τ^2	68.564
Y_{2003}	27.134

The observation error (τ^2) is much higher than the process error (σ^2), which suggests that exponential growth over time may be a good predictor of rooftop PV installations.

DISCUSSION

That voting pattern emerged as the most significant factor in rooftop PV adoption in Massachusetts should come as no surprise: the Republican party overall, and especially the 2020 Trump campaign, has been characterized by extreme climate denial, whereas Biden's 2020 campaign ran particularly aggressively on pushing to pass ambitious climate legislation. That is not to say that Massachusetts towns with predominantly Republican voting patterns had no rooftop solar adoptions, but the trend is clear: more votes for Biden in 2020 are correlated with more residential rooftop solar systems.

Still, there may be powerful relationships that neither the single nor multiple Bayesian linear regression models have been able to capture: spatial correlations. Clearly, voting patterns are spatially related, which pockets of either predominantly red or blue towns clustering together, bordered by split towns (Figure 6). Areas with higher populations and more urban centers (eg. the Greater Boston Area) are more likely to have more clusters of blue towns than those in the center of the state. Indeed, population may also play a role in rooftop solar adoption: denser towns tend to have more apartment complexes and fewer individual homes, perhaps reducing residents' influence over the building's energy sources, but these are also more likely to be urban, Democratically-leaning areas, which we now see are more likely to have more rooftop solar installations.

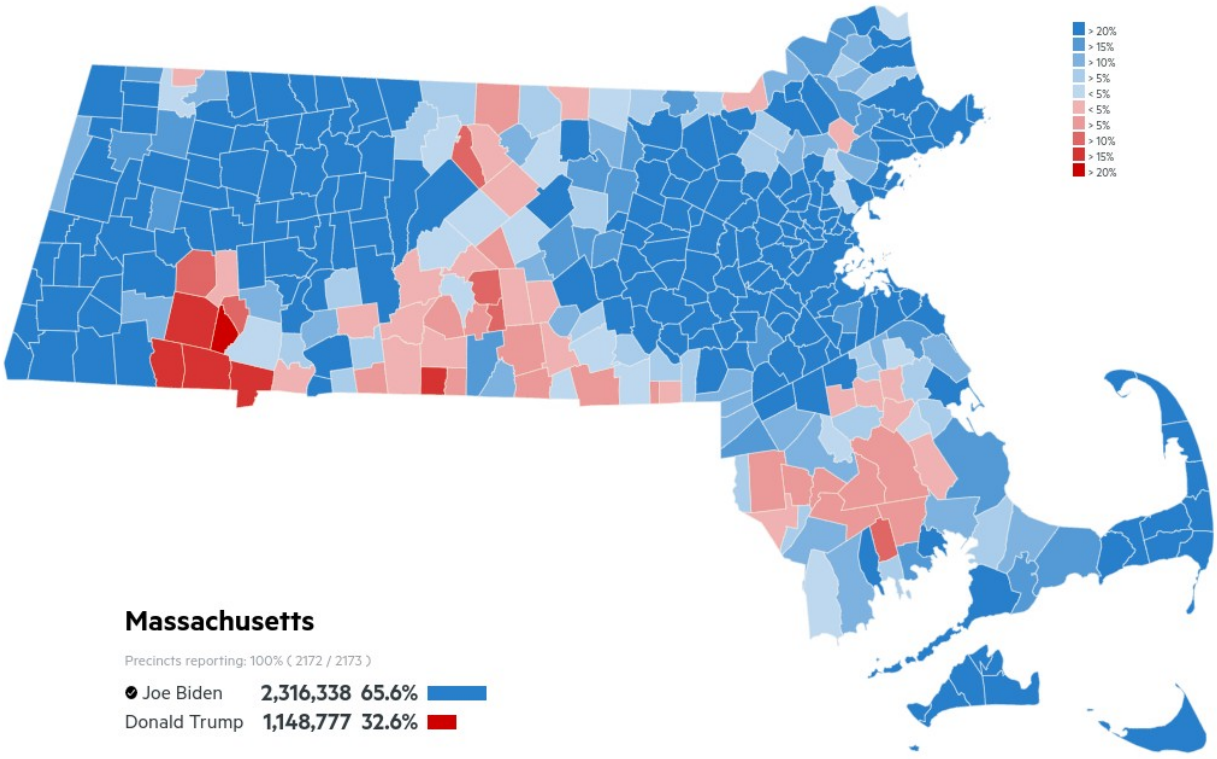


Figure 6: map of 2020 presidential election voting pattern (source: WBUR).

Furthermore, the simple effect of growth over time cannot be ignored. More interesting would be an analysis of the potential reasons for the drop in installations in 2016, perhaps because of incentive policy rollbacks, changes in the overall economy, or even the results of the presidential election in that year. As such, further work should attempt to model these spatial and temporal relationships and their effects on residential rooftop solar adoption. For now, it is perhaps discouraging, if not surprising, that solar adoption among Massachusetts residents follows the polarization of the 2020 presidential election.

ACKNOWLEDGEMENTS

Thanks Professor Dietze, for tolerating my utter confusion in the initial stages of this project and for always expertly and respectfully answering all my questions. Thanks also to Olivia and Alicia, my rocks throughout this course; and to Charlotte, who reviewed my preliminary analysis and pointed out a superficial coding error, which, after I fixed it, made everything work wonderfully again.

Very special thanks to the blue light filter on my laptop for keeping my eyes from burning out of my skull.

No thanks to R, RStudio. Bye, Felicia!

REFERENCES

[1] “Massachusetts Clean Energy and Climate Plan for 2025 and 2030.” *Mass.gov*, Executive Office of Environmental Affairs, https://www.mass.gov/info-details/massachusetts-clean-energy-and-climate-plan-for-2025-and-2030.
[2] “Massachusetts: State Profile and Energy Estimates.” *US Energy Information Administration (EIA)*, US EIA, https://www.eia.gov/state/?sid=MA.
[3] Vilallonga, Lucia. “Policy Analysis of Rooftop Solar Incentives in MA.” *GitHub*, 31 Jan. 2022, https://github.com/ghostpress/ma-solar.
[4] “PV in PTS Public Records Request.” Massachusetts Clean Energy Center, May 2021. https://www.masscec.com/public-records-requests.
[5] Steven Manson, Jonathan Schroeder, David Van Riper, Tracy Kugler, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 16.0 [dataset]. Minneapolis, MN: IPUMS. 2021. http://doi.org/10.18128/D050.V16.0
[6] Smith, William. “Map: See How Your Town or City Voted in the 2020 Election.” *WBUR News*, WBUR, 3 Nov. 2020, https://www.wbur.org/news/2020/11/03/2020-massachusetts-election-map.
[7] Pew Research Center, April, 2015, “A Deep Dive Into Party Affiliation”
[8] U.S. Census Bureau, A Compass for Understanding and Using American Community Survey Data: What Researchers Need to Know U.S. Government Printing Offi ce, Washington, DC, 2009.

APPENDIX A: Data processing

The overall steps of data processing were as follows:

1. Merging each disparate data set (education, income, installations, voting) to the proper geographic levels:
    a. Aggregate education data from the block group level to the town level
    b. Aggregate income data from the block group level to the town level
    c. Aggregate PV installations data from the installation level (single system, household) to the town level
2. For the time series data, different steps were required:
    a. Isolate the years over which the data were collected (2000-2021)
    b. Aggregate the installations to the year level, rather than installation level or town level

Education data

Education data were given in categories, which roughly but not exactly mapped onto years of education. For example: category 1, no schooling completed, is equivalent to 0 years of education; category 24, doctorate degree, is roughly equivalent to 25 years of education. Note: education level includes nursery school and kindergarten.

Income data

Aggregating income data was simpler, since values were given directly as average income from 2015-2019 per block group. The aggregated income per town was a mean of each block group in that town, and the MOE's were also similarly aggregated.

Installations data

To get the response variable, a simple count of the number of installations per town was sufficient. For the time series, the count was instead done per year.

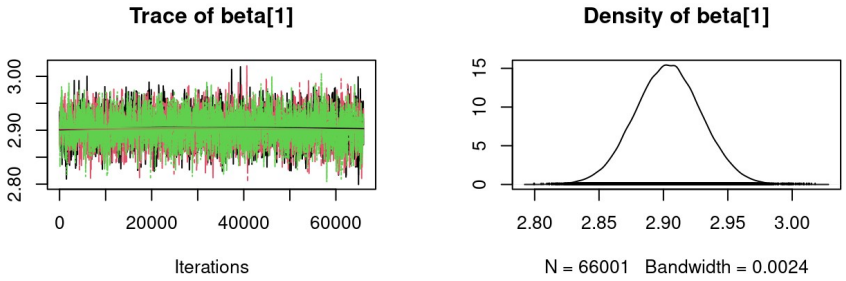
Please see code here (GitHub).

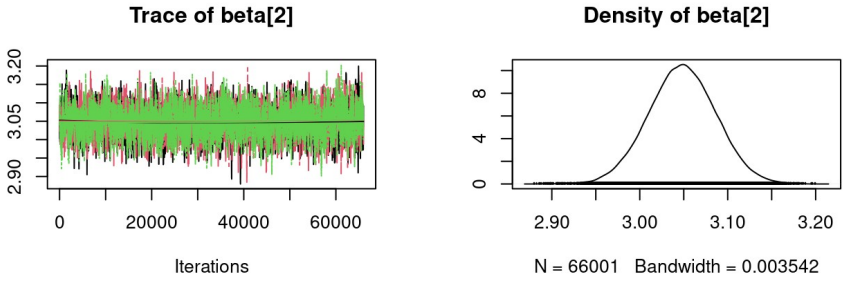
Voting data

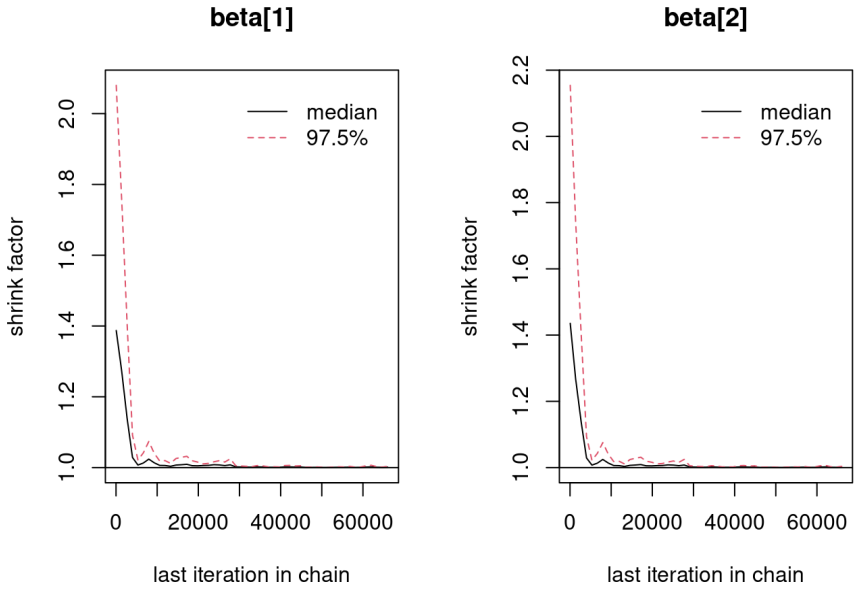
Voting data were provided in PDF form, which was downloaded directly from the WBUR source article. From there, I manually transferred each row of the PDF to a CSV file.

APPENDIX B: Model convergence plots

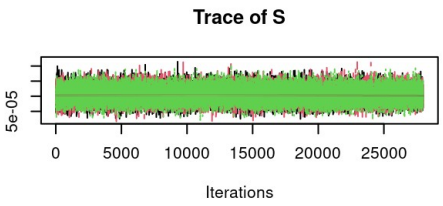
Model 1

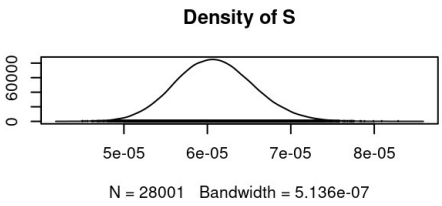


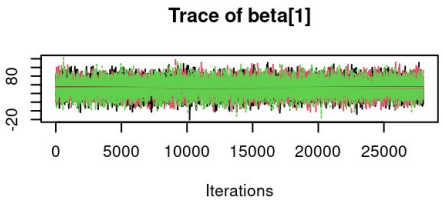


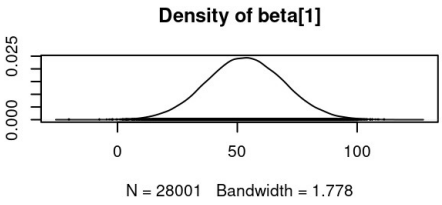


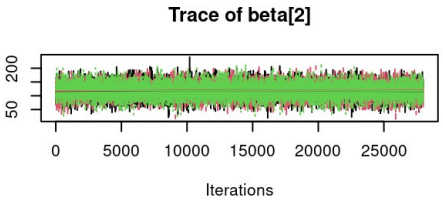
Model 2

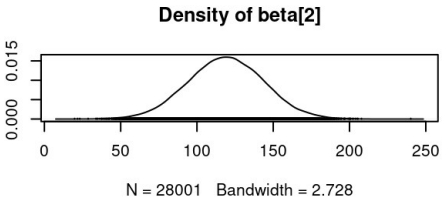

Trace of S
5e-05
0
5000
10000
15000
20000
25000
Iterations

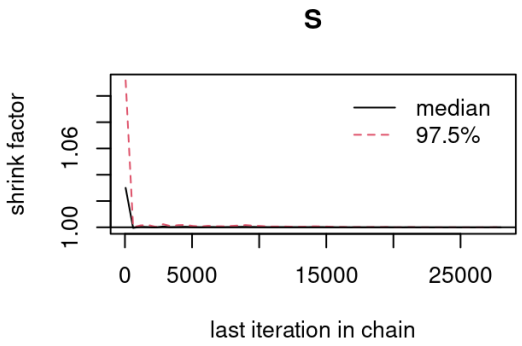

Density of S
0
60000
5e-05
6e-05
7e-05
8e-05
N = 28001 Bandwidth = 5.136e-07

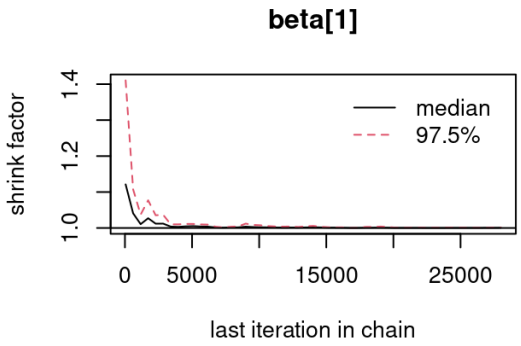

Trace of beta[1]
-20
80
0
5000
10000
15000
20000
25000
Iterations

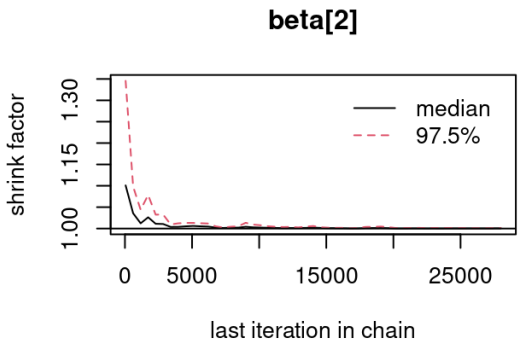

Density of beta[1]
0.000
0.025
0
50
100
N = 28001 Bandwidth = 1.778


Trace of beta[2]
50
200
0
5000
10000
15000
20000
25000
Iterations

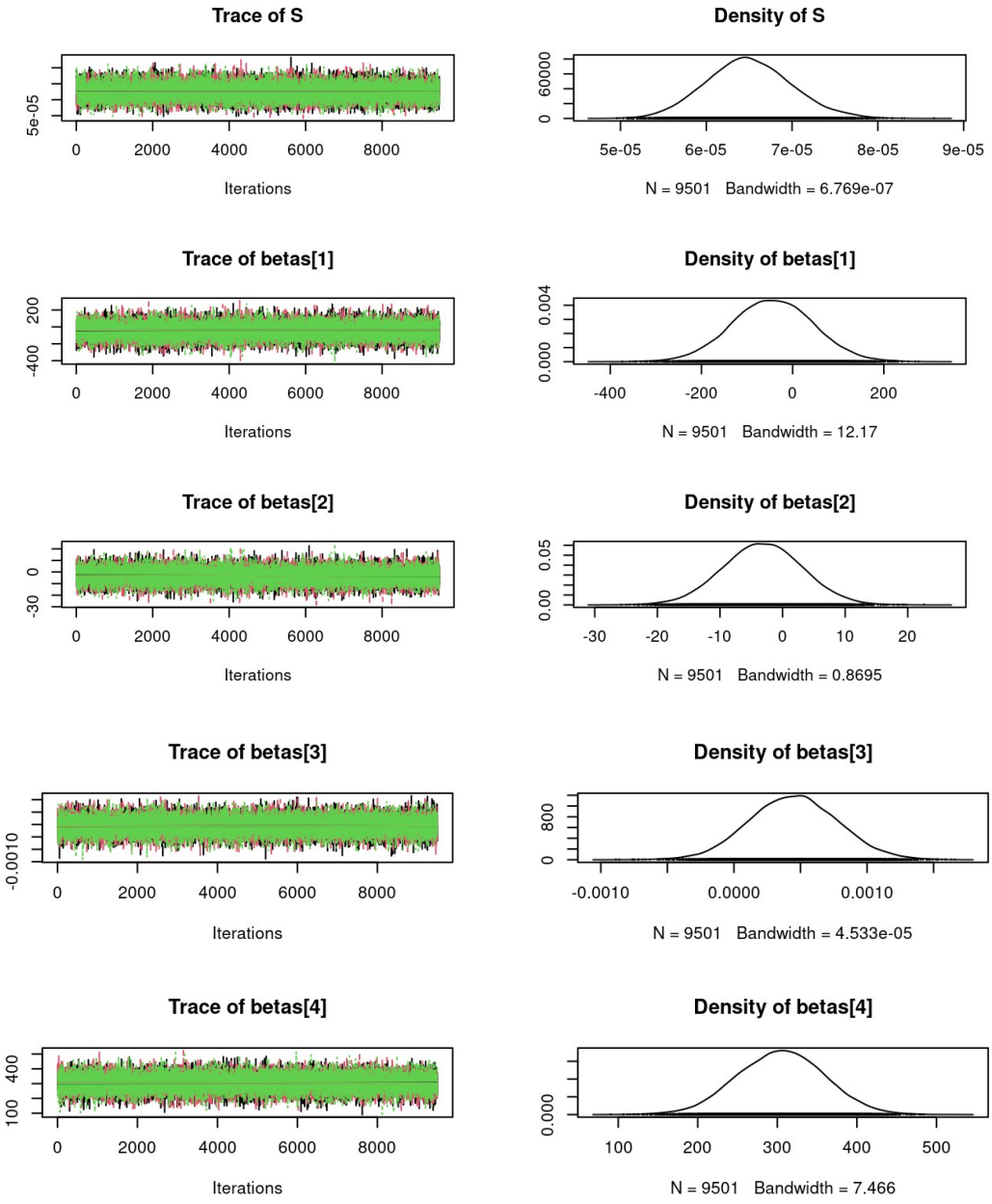

Density of beta[2]
0.000
0.015
0
50
100
150
200
250
N = 28001 Bandwidth = 2.728

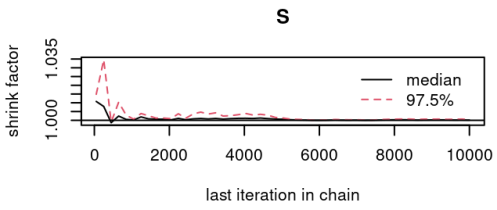

S
shrink factor
1.00
1.06
median
97.5%
0
5000
15000
25000
last iteration in chain

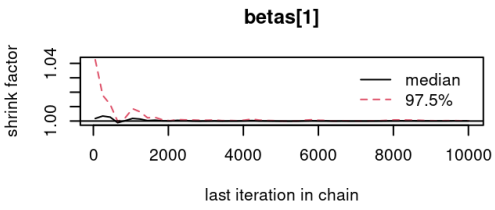

beta[1]
shrink factor
1.0
1.2
1.4
median
97.5%
0
5000
15000
25000
last iteration in chain

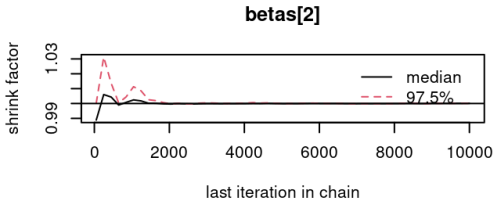

beta[2]
shrink factor
1.00
1.15
1.30
median
97.5%
0
5000
15000
25000
last iteration in chain

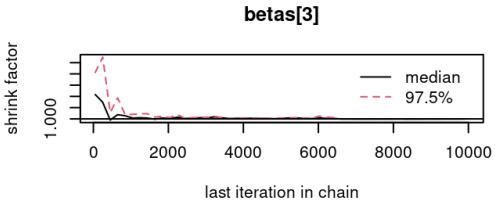
Model Sa

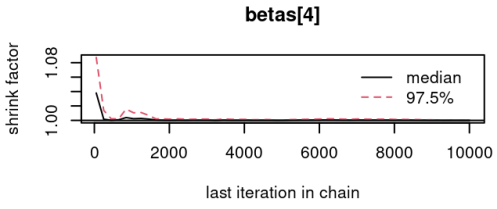

Trace of S
5e-05
0
2000
4000
6000
8000
Iterations
Density of S
0
60000
5e-05
6e-05
7e-05
8e-05
9e-05
N = 9501 Bandwidth = 6.769e-07
Trace of betas[1]
-400
200
0
2000
4000
6000
8000
Iterations
Density of betas[1]
0.000
0.004
-400
-200
0
200
N = 9501 Bandwidth = 12.17
Trace of betas[2]
-30
0
0
2000
4000
6000
8000
Iterations
Density of betas[2]
0.00
0.05
-30
-20
-10
0
10
20
N = 9501 Bandwidth = 0.8695
Trace of betas[3]
-0.0010
0
2000
4000
6000
8000
Iterations
Density of betas[3]
0
800
-0.0010
0.0000
0.0010
N = 9501 Bandwidth = 4.533e-05
Trace of betas[4]
100
400
0
2000
4000
6000
8000
Iterations
Density of betas[4]
0.000
100
200
300
400
500
N = 9501 Bandwidth = 7.466


S
shrink factor
1.000
1.035
median
97.5%
0
2000
4000
6000
8000
10000
last iteration in chain

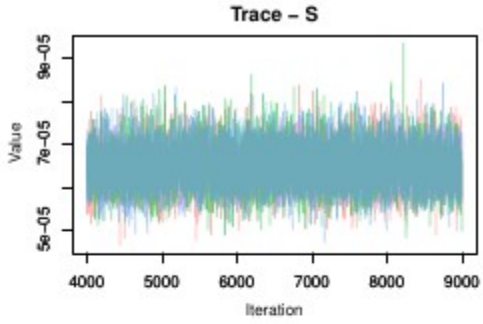

betas[1]
shrink factor
1.00
1.04
median
97.5%
0
2000
4000
6000
8000
10000
last iteration in chain

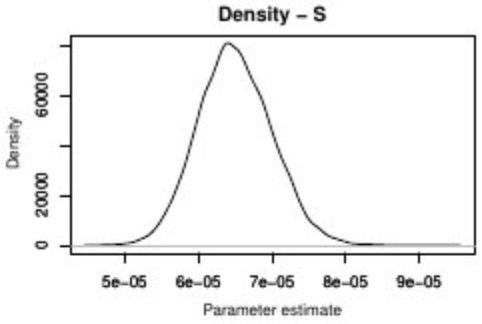

betas[2]
shrink factor
0.99
1.03
median
97.5%
0
2000
4000
6000
8000
10000
last iteration in chain

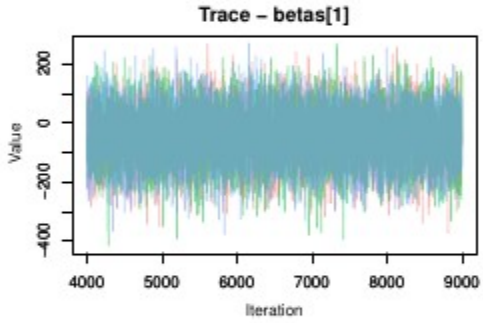

betas[3]
shrink factor
1.000
median
97.5%
0
2000
4000
6000
8000
10000
last iteration in chain

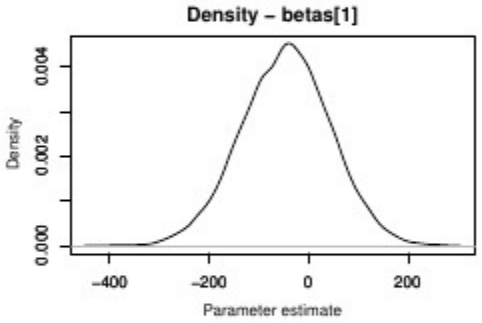

betas[4]
shrink factor
1.00
1.08
median
97.5%
0
2000
4000
6000
8000
10000
last iteration in chain

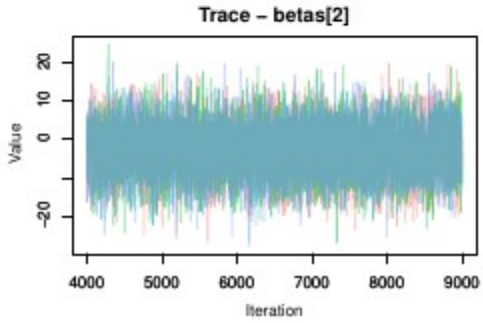
Model Sb

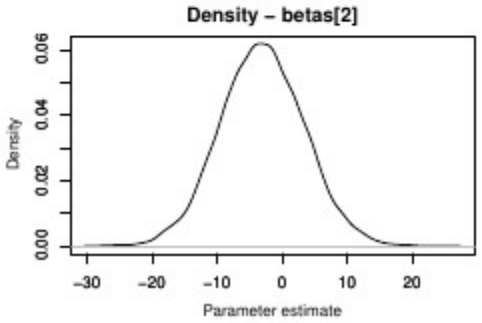

Trace – S
Value
Iteration

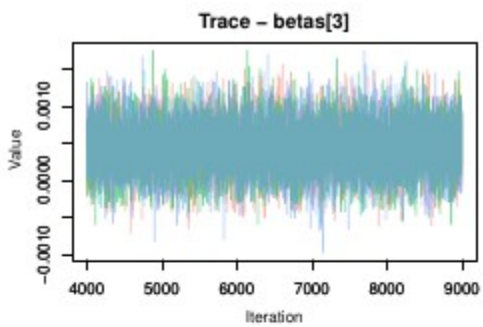

Density – S
Density
Parameter estimate

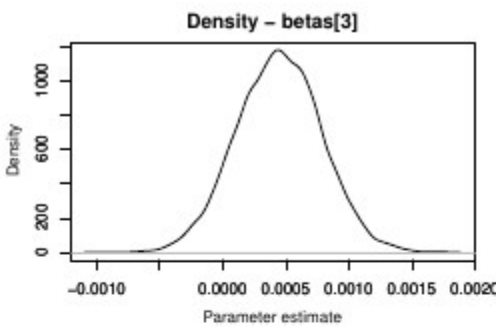

Trace – betas[1]
Value
Iteration

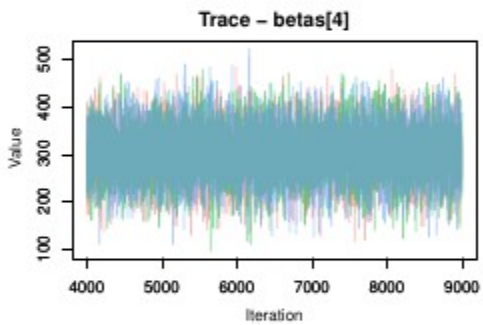

Density – betas[1]
Density
Parameter estimate

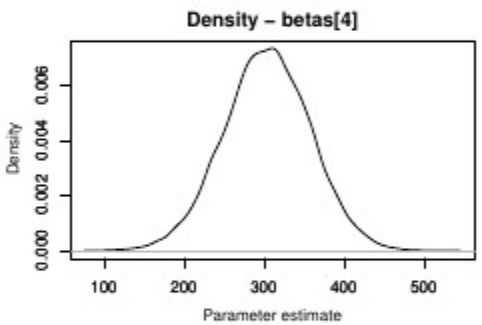

Trace – betas[2]
Value
Iteration


Density – betas[2]
Density
Parameter estimate

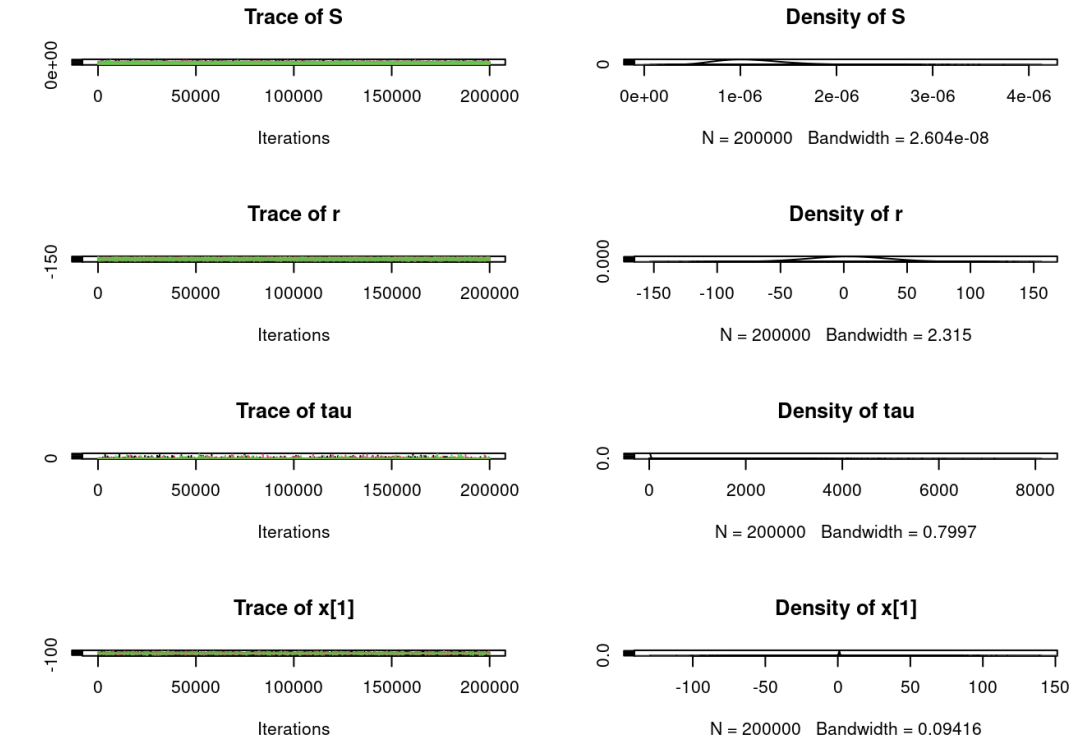

Trace – betas[3]
Value
Iteration

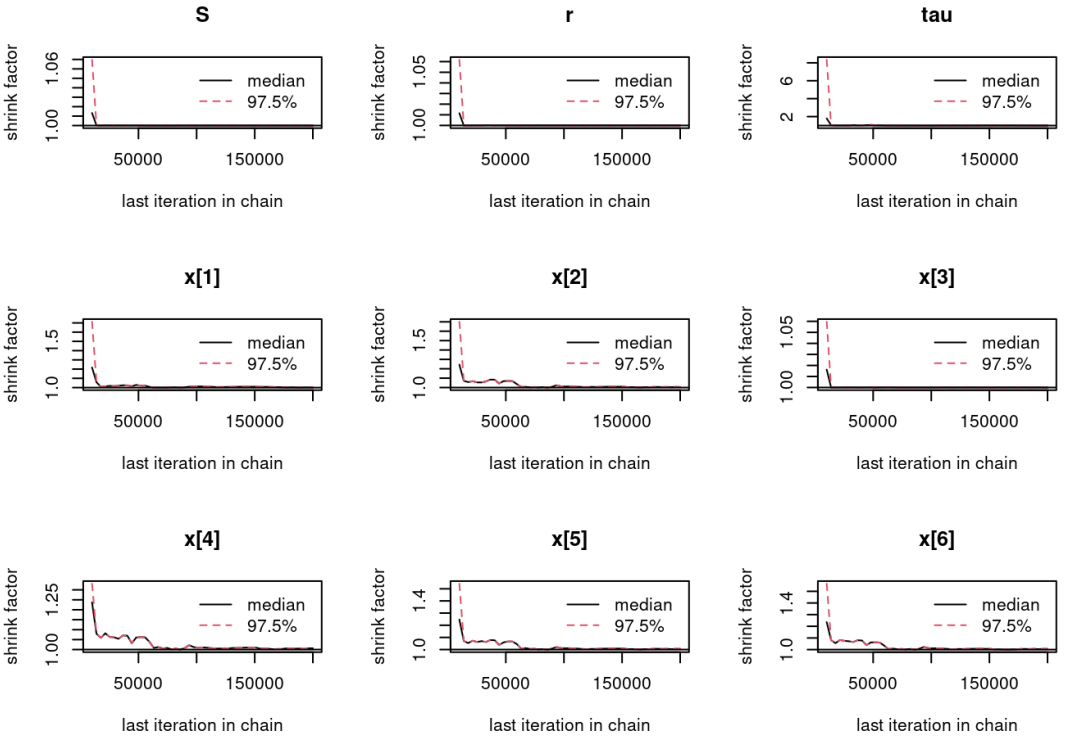

Density – betas[3]
Density
Parameter estimate


Trace – betas[4]
Value
Iteration


Density – betas[4]
Density
Parameter estimate

Model Sc


Trace of S
Iterations
Density of S
N = 200000 Bandwidth = 2.604e-08
Trace of r
Iterations
Density of r
N = 200000 Bandwidth = 2.315
Trace of tau
Iterations
Density of tau
N = 200000 Bandwidth = 0.7997
Trace of x[1]
Iterations
Density of x[1]
N = 200000 Bandwidth = 0.09416


S
r
tau
x[1]
x[2]
x[3]
x[4]
x[5]
x[6]
shrink factor
median
97.5%
last iteration in chain

APPENDIX C: Supplementary figures

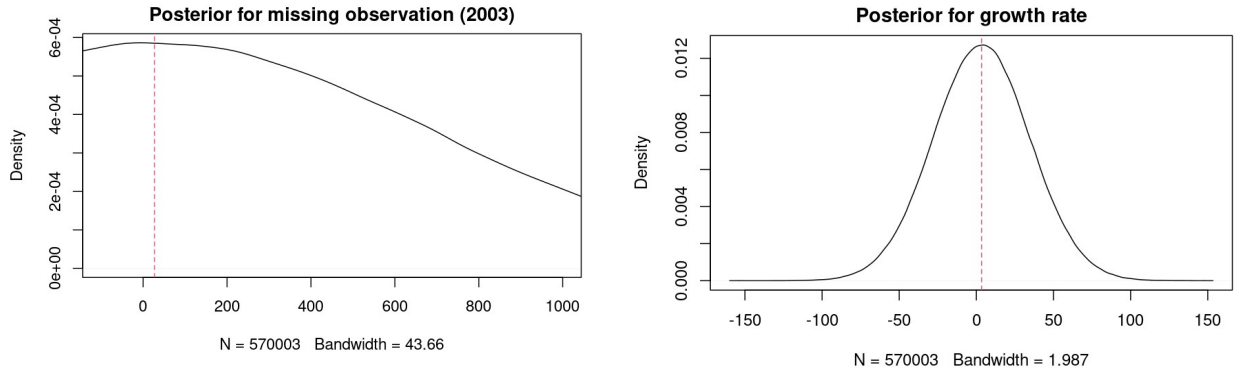


Figure 7: posteriors for missing observation (left) and growth rate, r (right) in time series model.