Violent Crime in Baltimore City

J.C. Lazzaro

Due 05/03/2018 by 11:59 pm

** Disclaimer **: The BPD Part 1 Victim Based Crime dataset was updated the morning of May 3, 2018 and consequently the number of observations available via the API request at the time of this project's submission has been truncated. Only observations after 12/15/2016 are now available whereas data from 01/01/2012 was available prior to the most recent update. We still include the written analysis of all parts dependent on this data, although all code chunks with supporting material and figures have been suppressed. The API appears to be refreshed automatically, so we encourage the reader to run all of the supressed chunks again to see if the dataset has been updated to include all data from 01/01/2012 to 12/31/2017 which was used for the time series analysis.

Abstract

Colloquially refered to as a city of neighborhoods, Baltimore City is perhaps infamous for its relatively high levels of violent crime. Using data from the Open Baltimore project and the Baltimore Neighborhood Indicators Alliance, this project aims to visualize the makeup of violent crime in Baltimore City from 2012 - 2017 and to explore if there is any spatial dependence in the level of violent crime in each neighborhood upon controlling for relevant covariates. We model the trends in aggregate violent crime over time using a SARIMA model and investigate the spaital dependence of violent crime using CAR and SAR models.

Violent Crime in Baltimore over Time

As part of the Open Baltimore Project, the Baltimore City Police Department (BPD) Part 1 Victim Based Crime Data provides daily crime data accessible via an API request. The data includes various variables, the most important of which for our purposes is crime date, description, and location.

As per the 2015 Maryland Uniform Crime Report, we define violent crime as incidents involving: homicide, shooting, rape, robbery, and aggravated assault. We filter the BDP data to only include such incidents and aggregate the data so that the frequency is monthly. The makeup of violent crime in Baltimore City over time is displayed below:

While the aggregate level of violent crime appears to increase over time, with an evident seasonal pattern, the relative incidence of each of the consituent crimes is roughly constant. That is, any increase in overall levels of violent crime appears to be spread commensurately throughout each type of crime and is not due to any singular increase in a particular violent crime. However, there does appear to be a seasonal uptick in the levels of homicide/shootings from the months of April to September. Nonetheless, the overwhelming majority of violent crime in Baltimore City consists of robbery and aggravated assault.

Developing Univariate Time Series Model

There is an apparent trend in the monthly level of aggregate violent crime in Baltimore City over time, implying that the series is not stationary. While one can argue that there is a linear trend in the level of violent crime, the time series can be naively partitioned into two sections - before and after April 2015 - where the mean in the latter partition is markedly higher than that in the former.

To remedy this apparent non-stationarity, we first difference the data; the resulting series does appear to have a constant mean over time. However, there is a persistent seasonal trend which suggests we should seasonally difference the data with a period of 12 months.

Using model selection criteria, we confirm that seasonally differencing in addition to first differencing is indeed warranted. The ACF and PACF of the residuals of the resulting $ARIMA(0,1,0)(0,1,0)_{12}$ model indicates that there is a remaining seasonal structure to the data not accounted for by the model.

The ACF and PACF of the residuals suggests that incorporating a seasonal MA(1) term in the model is needed. Again, we explore the parameter space around the MA(1) term, using model selection criteria to determine the appropriate addition to the model. The residuals of the resulting $ARIMA(0,1,0)(0,1,1)_{12}$ model indicate that there is still structure unaccounted for by the model.

The ACF and PACF of the residuals suggests that adding an MA(1) component to the model is appropriate. Employing the same procedure as above, the residuals of the resulting $ARIMA(0,1,1)(0,1,1)_{12}$ model indicate that there is no significant remaining temporal structure that has not been captured by the model.

Model fit and forecasts

The model fit, overlayed with the time series of aggregate violent crime in Baltimore City, is given below:

Twelve month forecasts using the $ARIMA(0,1,1)(0,1,1)_{12}$ model indicate that the violent crime will continue to follow the seasonal trend of decreasing in the early months of the year but picking up from late spring until early fall until diminishing in the later months of the year.

However, observing the 24 month forecast displays that these forecasts revert to a mean structure and appear to do so rather quickly, perhaps within the 12 month forecast. As expected, the prediction intervals widen with larger forecast horizons.

Violent Crime in Baltimore in 2015

From the above analysis, we see that there was a marked increase in the aggregate level in violent crime in Baltimore City in 2015. The Baltimore Neighborhood Indicators Alliance (BNIA) as part of its Vital Signs series, provides data on a variety of aspects for various Baltimore City localities. More specifically, they aggregate data pertaining to education, employment, crime, health, etc. for the Community and Statistical Areas (CSA) in Baltimore City, as defined by the 2010 Census.

Using this data, we use areal data models to both predict the levels of violent crime in each CSA in 2015 and to investigate whether there is significant spatial dependence in the level of violent crime after controlling for relevant covariates. We choose a larger set of covariates a priori from the BNIA Vital Signs series which may be relevant predictors for violent crime in each CSA; a sample of these predictors include: unemployment rate, number of commercial properties, percent of the population aged 18-24, life expectancy, number of vacant properties, percent of the workforce with less than a high school education, etc. As an aside, the shape files used to visualize the model predictions are taken from the Open Baltimore project.

Visualizing Violent Crime in 2015

Prior to enumerating the CAR/SAR models, we first visualize the incidence of violent crime in Baltimore City in 2015 by type of incident. To do so, we refer to the BPD Part 1 Victim Based Crime data to enumerate all incidents of violent crime in 2015 and create kernel density estimates by crime type using the spatial locations of each incident provided as part of the BPD data.

From these kernel density plots, we see that the incidence of violent crime is concentrated in the same geographical areas irrespective of the type of crime save for some minor distinctions. Most notably, incidents of aggravated assualt, homicide, and shootings are concentrated in the east and west portions of the inner city, while robbery is concentrated in the downtown area, where, unsurprisingly, there is the largest number of commerical properties in the city. Given these visualizations, it is indeed plausible that there is spatial dependence in the levels of violent crime in Baltimore City. However, a more rigorous investigation is needed to determine if such a dependence exists.

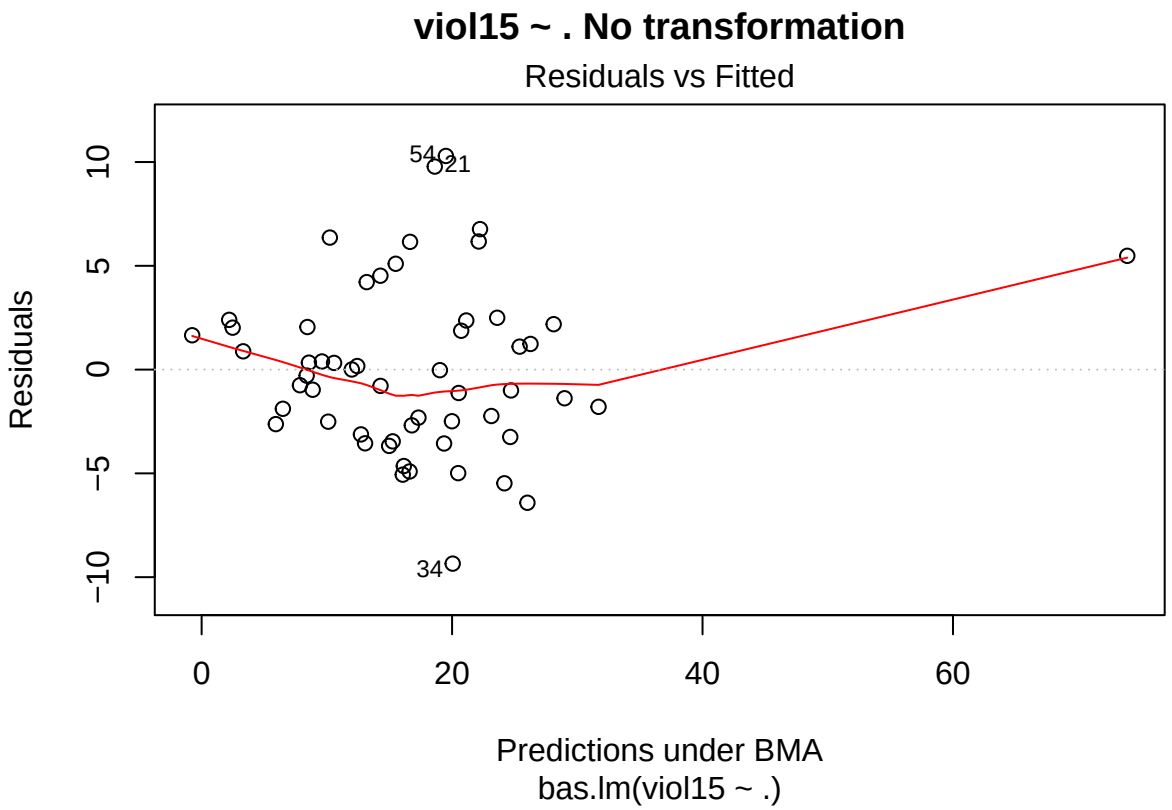
EDA and Preliminaries

Prior to enumerating the CAR and SAR models, we first conduct exploratory data analysis so as to determine if the various aforementioned predictors are linearly related to the response variable: viol15 - the number of violent crimes in a given CSA per 1,000 residents. We note that this predictor is non-negative and continuous so a standard linear model for the mean structure is appropriate. The correlation between violent crime and the list of predictors selected a priori is given below:

	Correlation between Violent Crime and predictors
viol15	1.0000000
numbus15	0.6965551
comprop15	0.6577661
vacant15	0.4865002
tanf15	0.4807282
lesshs15	0.4169799
nilf15	0.3871734
hhpov15	0.3698416
age24_15	0.2992797
mort1_15	0.2986198
teenbir15	0.2526777
unempr15	0.2419021
hcvhouseXX	0.1051269
age18_15	-0.0242679
lifexp15	-0.6284010

We see that the number of businesses and the number of commercial properties have the strongest positive correlation with the level of violent crime liekly due to the relatively high incidence of robbery compared to other types of violent crime, while life expectancy has the strongest negative correlation with violent crime.

To further investigate the relationship between the predictors and the response, we note that the residuals from a naive Bayesian linear model without transforming the predictors or the response are heteroskedastic and indicate the presence of a heavily influential point. This point corresponds to Downtown/Seton Hill, where again there is the largest number of incidents of violent crime, namely robbery, due to the disproportionate number of commercial properties and businesses in the area.



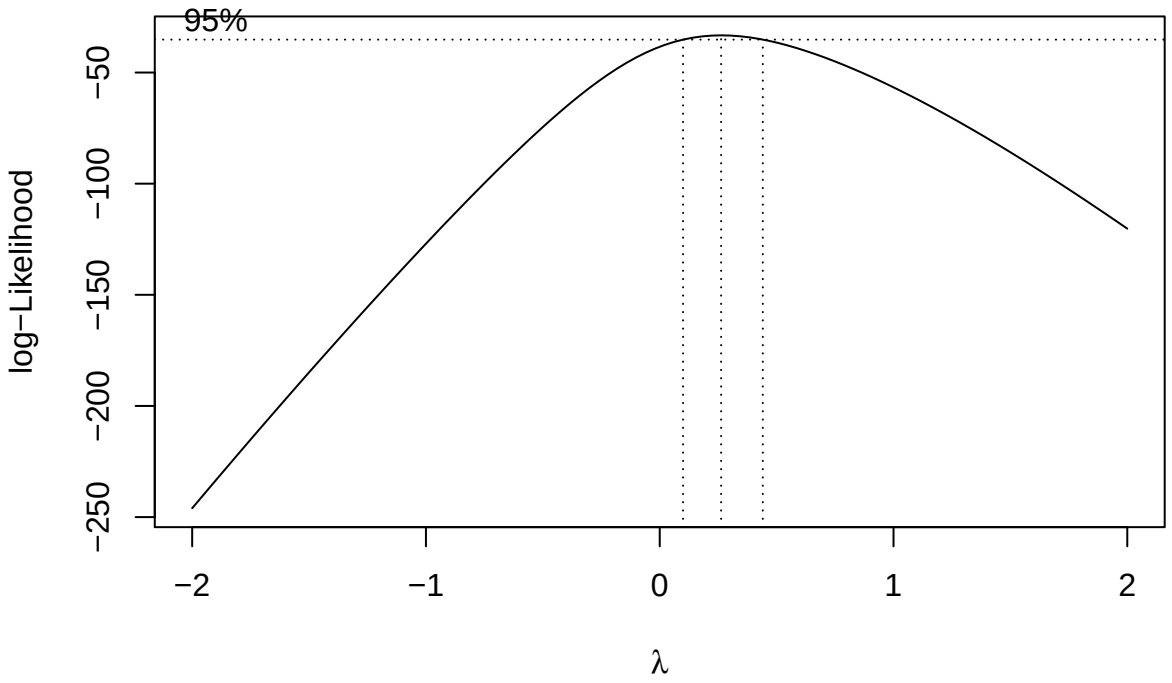
Thus, we suspect that a transformation of the response variable - perhaps a log transformation - is needed to account for the skewness of the response and to possibly stabilize the variance. Furthermore, we explore if transformations of the predictors are needed to ensure the assumption of normality is satisfied via Box-Cox transformations, the results of which are given below:

```
## [1] TRUE
```

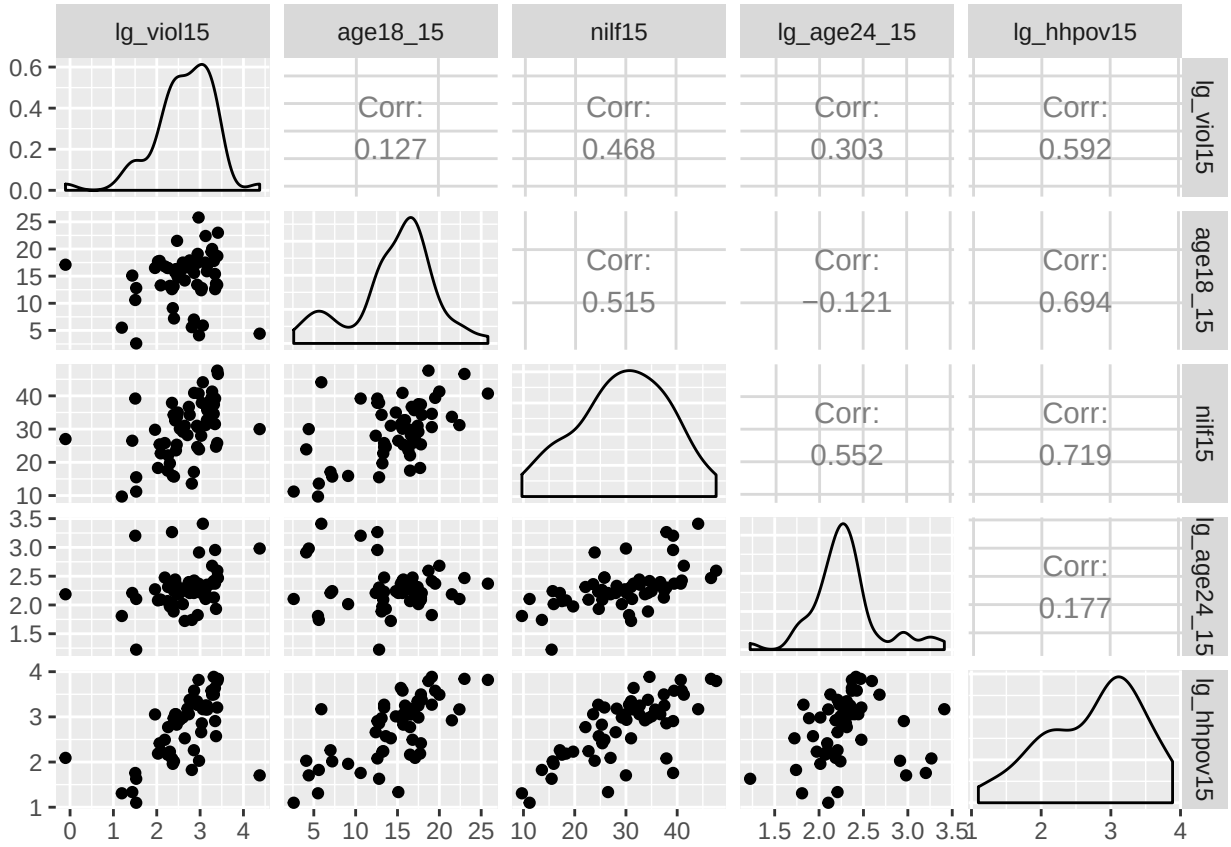
```
## bcPower Transformations to Multinormality
##             Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## age18_15      0.8064        1.00       0.4303       1.1825
## age24_15      0.0097        0.00      -0.3821       0.4014
## hhpov15       0.0249        0.00      -0.2959       0.3456
## nilf15        1.1278        1.00       0.5730       1.6827
## unempr15      0.4054        0.00      -0.0388       0.8496
## lesshs15      0.4203        0.50       0.2276       0.6130
## comprop15     0.2807        0.33       0.0898       0.4715
## numbus15      0.0283        0.00      -0.1382       0.1948
## teenbir15     0.4311        0.50       0.3318       0.5303
## tanf15        0.3394        0.33       0.2263       0.4525
## lifexp15     -2.2318        0.00      -4.8787       0.4151
## mort1_15      0.6622        0.50       0.4221       0.9023
## vacant15      0.0603        0.00      -0.0585       0.1791
## hcvhouseXX    0.3721        0.50       0.1289       0.6154
##
## Likelihood ratio tests about transformation parameters
##                                                        LRT df pval
## LR test, lambda = (0 0 0 0 0 0 0 0 0 0 0 0 0 0) 308.0247 14    0
## LR test, lambda = (1 1 1 1 1 1 1 1 1 1 1 1 1 1) 453.1140 14    0
```

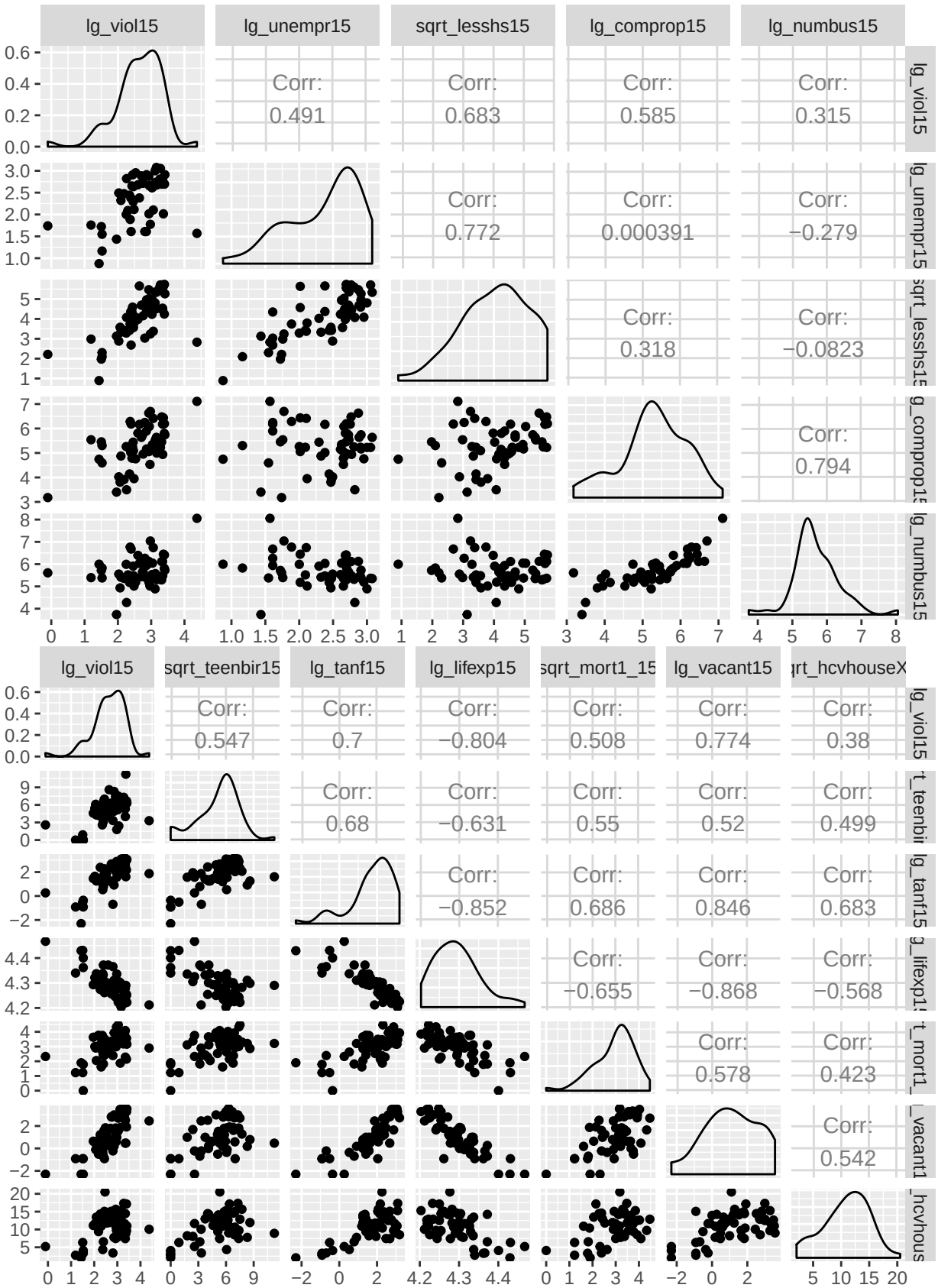
The results of the Box-Cox transformation indicates that the log transformation is indeed recommended, as

indicated in the plot below:



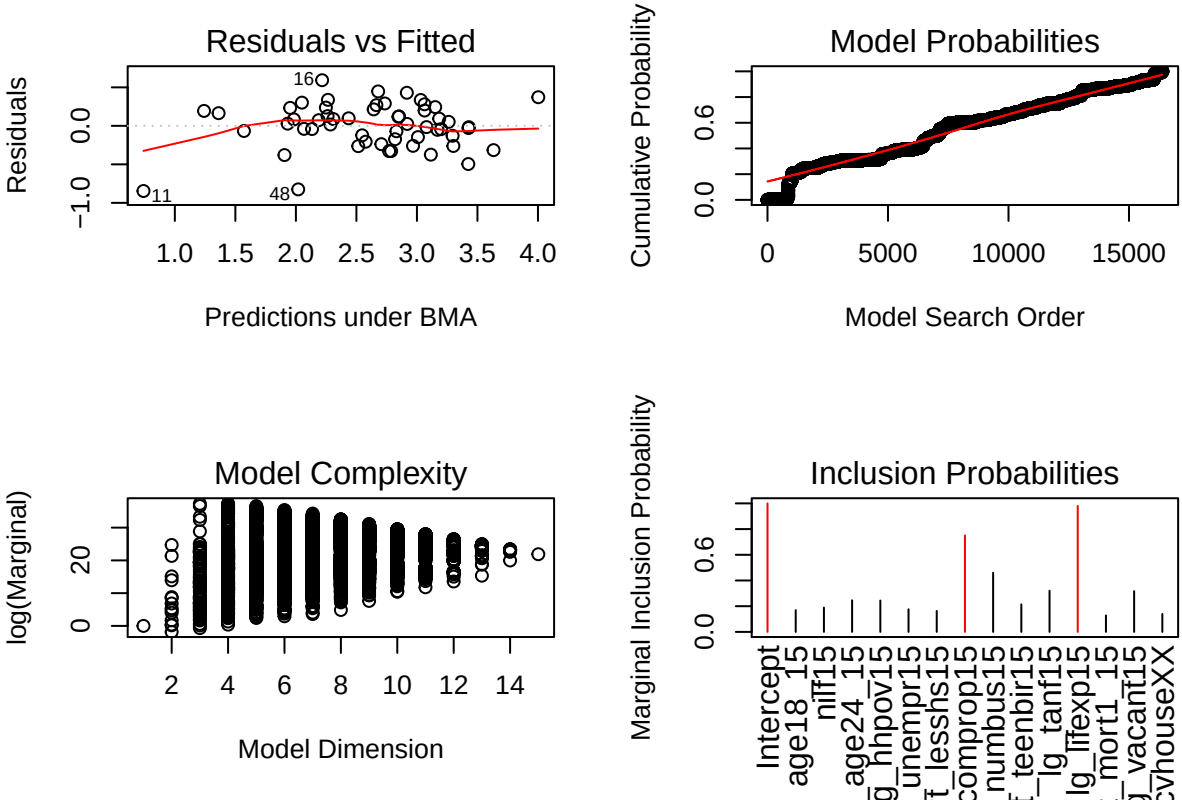
After incorporating the recommended transformations to both the response and predictors, we see that the relationships between the predictors and the response appear to be linear as shown by the following pairs plots:





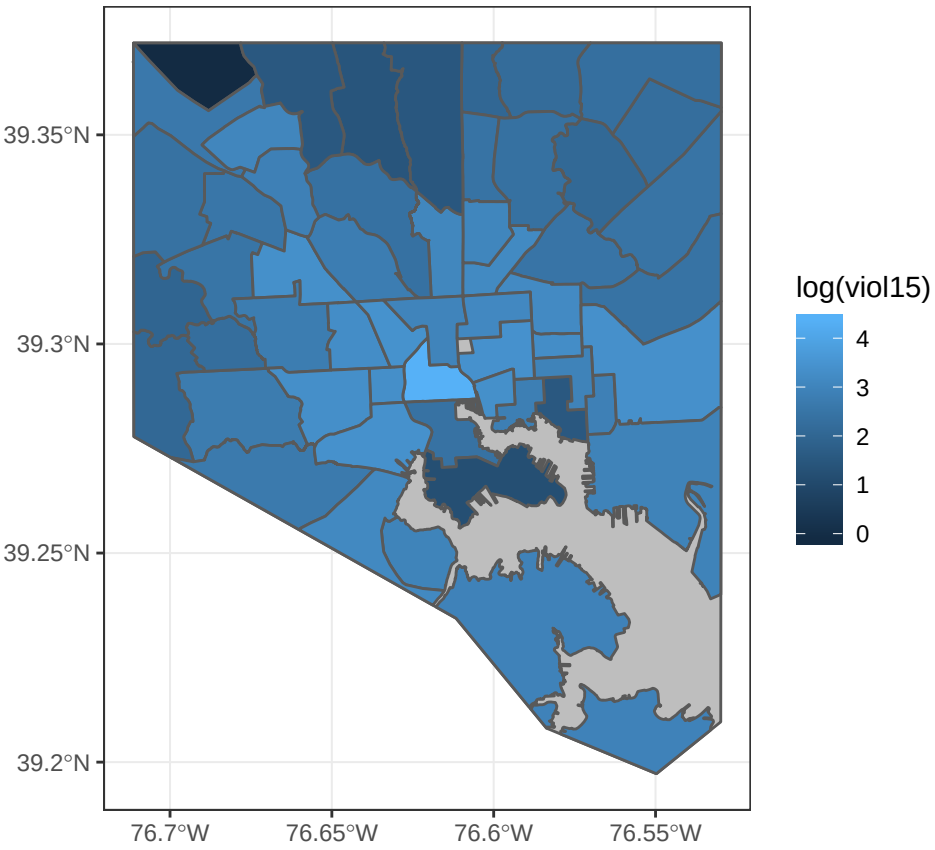
Following the above transformations, we refine the model of the mean structure by performing Bayesian

Variable selection via BMA using a mixture of g-priors. We select variables for which the posterior inclusion probability is above 0.5, which are summarized in fourth plot, below:



Using this criterion, performing Bayesian variable selection suggests that the log of the number of commercial properties and the log of life expectancy in a CSA should be used in order to model the mean of the log of violent crime per 1,000 residents in said CSA. For completeness, we also performed variable selection via stepwise methods, more specifically with backward selection. Using BIC as the selection criterion resulted in the same choice of predictors as the Bayesian variable selection, while AIC predictably selected more predictors. The overall predictive performance, however, using these additional predictors: namely the log of the percent of the population aged 18-24, log of the number of businesses, log of the number of vacant properties and the square root of the number of teen pregnacies resulted in worse predictive performance when incorporated into the mean structure of the CAR and SAR models. We exclude explicitly enumerating these variables into the mean structure for each model and simply state the result in the interest of brevity.

The map displaying the log of violent crime rates in each CSA of Baltimore City in 2015 is displayed below; at a cursory glance there does appear to be some spatial dependence in the data as neighboring areas appear to share similar levels of the response variable.



To formally test for the presence of spatial autocorrelation, we first construct an adjacency matrix. We note that while Baltimore city is a continguous shape, the function st_touches returns elements for which there are not adjacent neighborhoods, namely Harbor East and Hawkins Point. We reference the map of Baltimore City's CSA's in order to fill in these elements manually making sure to satisfy the symmetry of the adjacency matrix when doing so. After doing so, we compute Moran's I for the log of violent crime rates in each CSA and conclude that there is statistically significant positive spatial autocorrelation present.

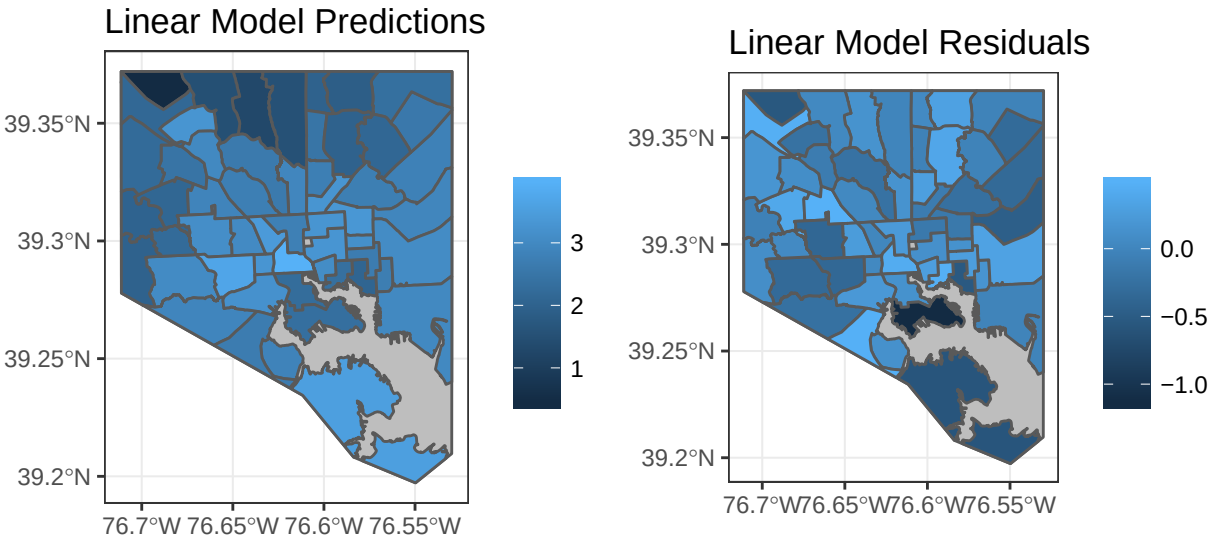
```
##
##  Moran I test under randomisation
##
## data:  log(bas_data$viol15)
## weights: listW
##
## Moran I statistic standard deviate = 4.1114, p-value = 1.967e-05
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.312956149      -0.018518519       0.006500193
```

We further test if there is significant spatial structure in the data upon controlling for the relevant covariates implied by the Bayesian variable selection by checking if the residuals from the implied linear model sans any spatial random effect exhibit any significant autocorrelation. The summary of the naive linear model is given below:

```
##
## Call:
## lm(formula = log(viol15) ~ lg_comprop15 + lg_lifexp15, data = bas_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.13468 -0.19278  0.03524  0.21688  0.48111
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.77663    3.43294  11.004 3.44e-15 ***
## lg_comprop15  0.33330    0.05204   6.404 4.40e-08 ***
## lg_lifexp15  -8.59350    0.78050 -11.010 3.37e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 52 degrees of freedom
## Multiple R-squared:  0.8027, Adjusted R-squared:  0.7951
## F-statistic: 105.8 on 2 and 52 DF,  p-value: < 2.2e-16
```

Upon taking the number of commerical properties and the life expectancy into account, we see that there is no longer any statistically significant spatial autocorrelation with regard to violent crime, as per the results of the Moran I test. Nonetheless, we proceed with developing CAR and SAR models to see if we can more accurately predict the level of violent crime in each CSA in 2015.




```
##
##  Moran I test under randomisation
##
## data:  bas_data$lm_resid
## weights: listW
##
## Moran I statistic standard deviate = -0.27549, p-value = 0.6085
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      -0.041038939      -0.018518519       0.006682698
```

CAR and SAR models

The model under consideration is given by:

$$log(viol15_i) = \beta_1 + \beta_2 log(comprop15_i) + \beta_3 log(lifexp15_i) + \omega_i + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma_w^2)$ and the spatial random effect $\omega \sim N(0, \Sigma_j)$ and the covariance matrix takes the usual forms:

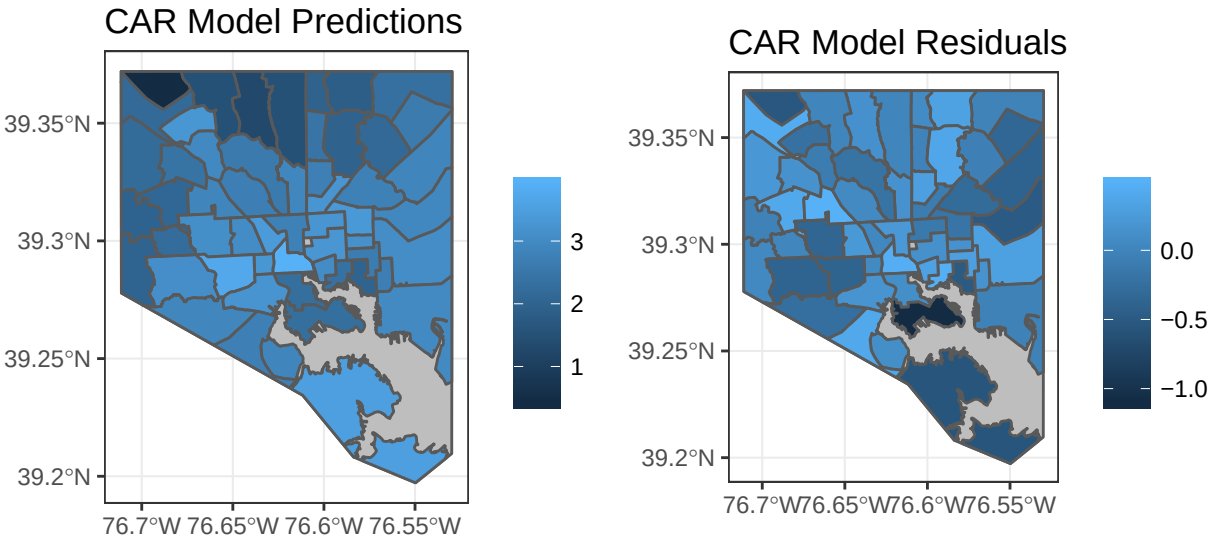
$$\Sigma_{CAR} = \sigma^2 (D - \phi W)^{-1}$$

$$\Sigma_{SAR} = (I - \phi D^{-1} W)^{-1} \sigma^2 D^{-1} \left((I - \phi D^{-1} W)^{-1} \right)^T$$

for the CAR and SAR models, respectively. We first use the spdep library to fit each model and compare the results.

The summary of the CAR model, the predictions and residuals for each CSA and the Moran's I test on the residuals are given below:

```
##
## Call: spautolm(formula = log(viol15) ~ lg_comprop15 + lg_lifexp15,
##     data = bas_data, listw = listW, family = "CAR")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.099257 -0.198009  0.051148  0.207473  0.483668
##
## Coefficients:
##               Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)  37.977201   3.214277  11.8152 < 2.2e-16
## lg_comprop15  0.330635   0.048662   6.7945 1.087e-11
## lg_lifexp15  -8.635641   0.730046 -11.8289 < 2.2e-16
##
## Lambda: -0.048132 LR test value: 0.25367 p-value: 0.6145
## Numerical Hessian standard error of lambda: 0.096096
##
## Log likelihood: -14.21453
## ML residual variance (sigma squared): 0.097665, (sigma: 0.31251)
## Number of observations: 55
## Number of parameters estimated: 5
## AIC: 38.429
```



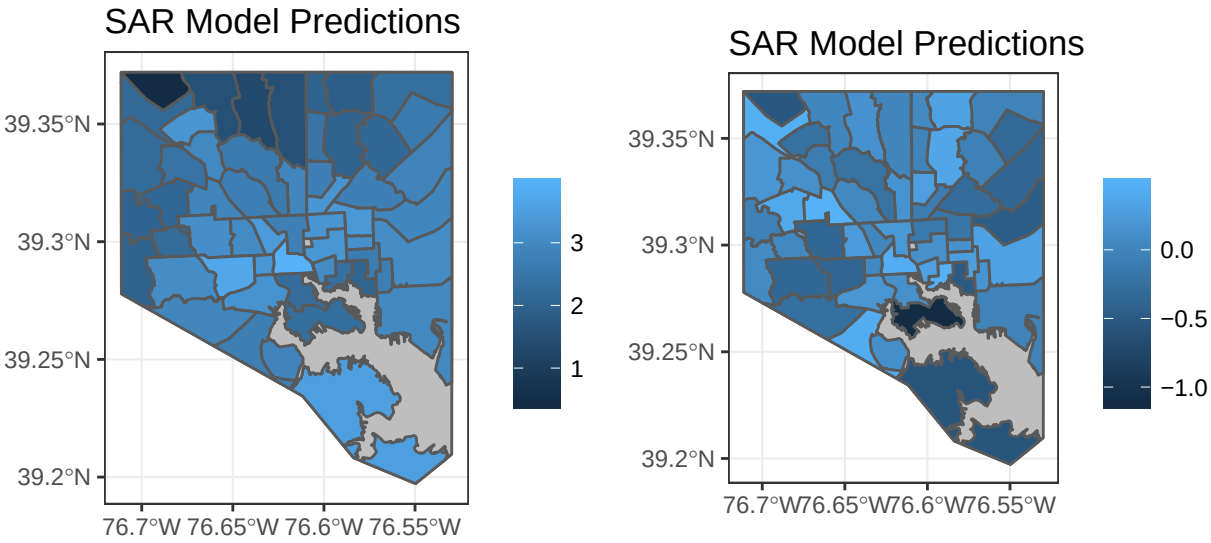

```
##
##  Moran I test under randomisation
##
## data:  bas_data$car_resid
## weights: listW
##
## Moran I statistic standard deviate = 0.59425, p-value = 0.2762
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.030221773      -0.018518519       0.006727263
```

The fitted value of $\hat{\phi} \approx -0.048132$ along with the corresponding likelihood ratio test suggests that there is no significant spatial autocorrelation present, a conclusion buttressed by the Moran's I test. This is as expected considering the previous result that there is no remaining spatial structure in the incidence in violent crime upon controlling for the relevant covariates contained in the mean structure of the model.

The SAR model yields analogous results, with slightly different coefficient estimates and therefore predictions as well.

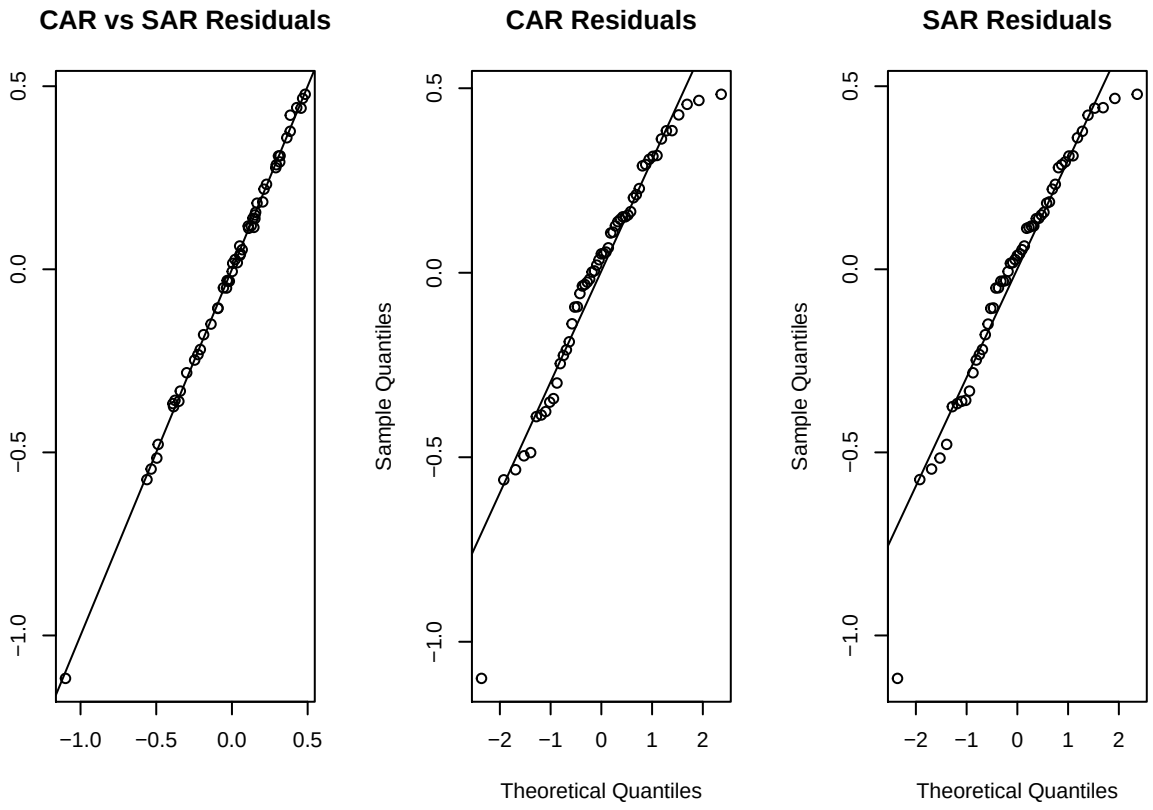
```
##
## Call: spautolm(formula = log(viol15) ~ lg_comprop15 + lg_lifexp15,
##     data = bas_data, listw = listW, family = "SAR")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.117236 -0.198795  0.037864  0.201434  0.477997
##
## Coefficients:
##               Estimate Std. Error  z value  Pr(>|z|)
## (Intercept)  37.981532   3.200625  11.8669 < 2.2e-16
## lg_comprop15  0.330431   0.048439   6.8216 9.001e-12
## lg_lifexp15  -8.636258   0.726884 -11.8812 < 2.2e-16
##
## Lambda: -0.026255 LR test value: 0.28051 p-value: 0.59637
## Numerical Hessian standard error of lambda: 0.049397
##
## Log likelihood: -14.20111
```

```
## ML residual variance (sigma squared): 0.097818, (sigma: 0.31276)
## Number of observations: 55
## Number of parameters estimated: 5
## AIC: 38.402
```




```
##
##  Moran I test under randomisation
##
## data:  bas_data$sar_resid
## weights: listW
##
## Moran I statistic standard deviate = 0.18968, p-value = 0.4248
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      -0.002983217      -0.018518519       0.006708372
```

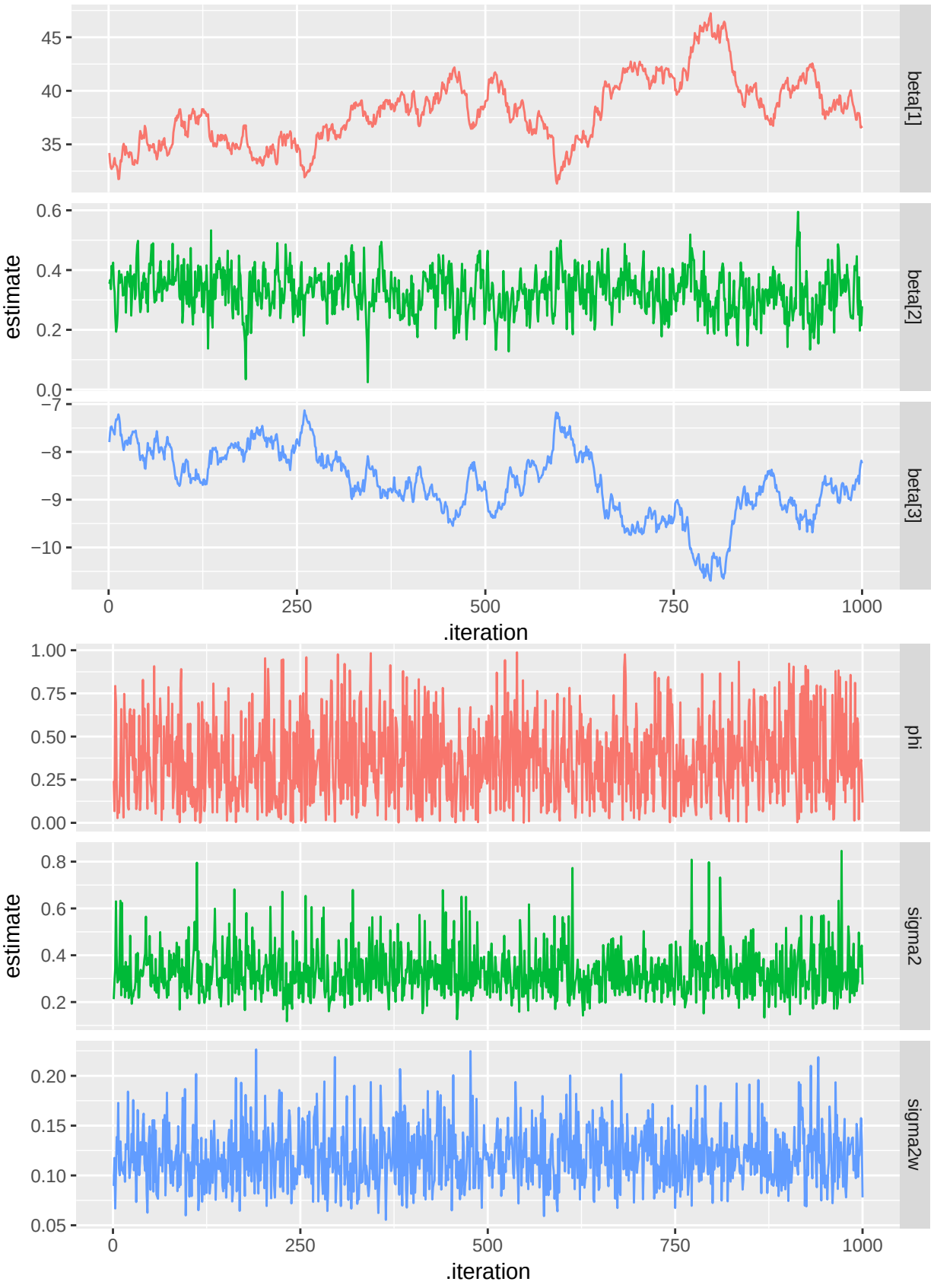
The following plots show the similarity in the residuals - and therefore the predictions - given by each model, and that the assumption regarding the normality of the errors for each model is justifiable save for the outlier of South Baltimore, for which both models overestimate the aggregate level of violent crime in 2015.



Bayesian CAR/SAR models

We also fit Bayesian models to see if there is any appreciable improvement in predictive accuracy over the frequentist versions fitted above. For both the Bayesian CAR and SAR models, we use uniformative normal priors for the regression coefficients: $\beta_i \overset{iid}{\sim} N(0, 100^2)$, gamma priors for the precisions $\frac{1}{\sigma^2}, \frac{1}{\sigma_w^2} \overset{iid}{\sim} Gamma(1, 1)$ and a uniform prior on $\phi \sim U[0, 1)$ where the support is such that the value of ϕ will be restricted to be less than one so that the covariance matrices are invertible. We run the chains for 50,000 iterations, thinning every 50 observations.

The chains and posterior summaries for each of the parameters for the CAR model are given below:



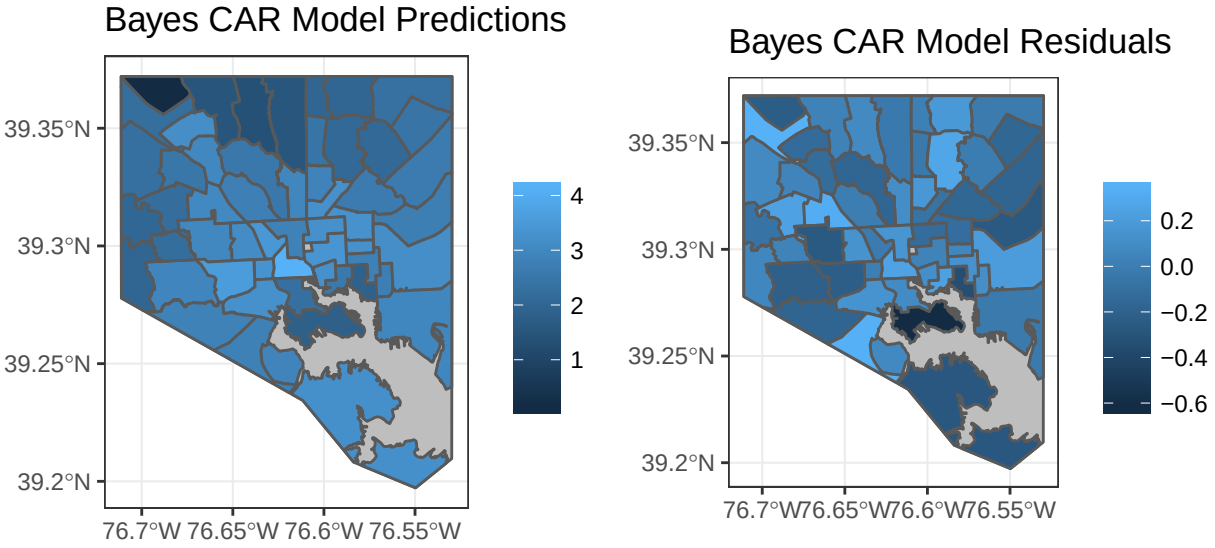
We see that despite thinning, there is significant autocorrelation in the chains for the coefficients β_1 and β_3;

the remaining chains appear to have explored the parameter space sufficiently well to conclude that they appear to have converged. The posterior summaries for each of the model parameters are given below:

Table 2: Bayesian CAR model

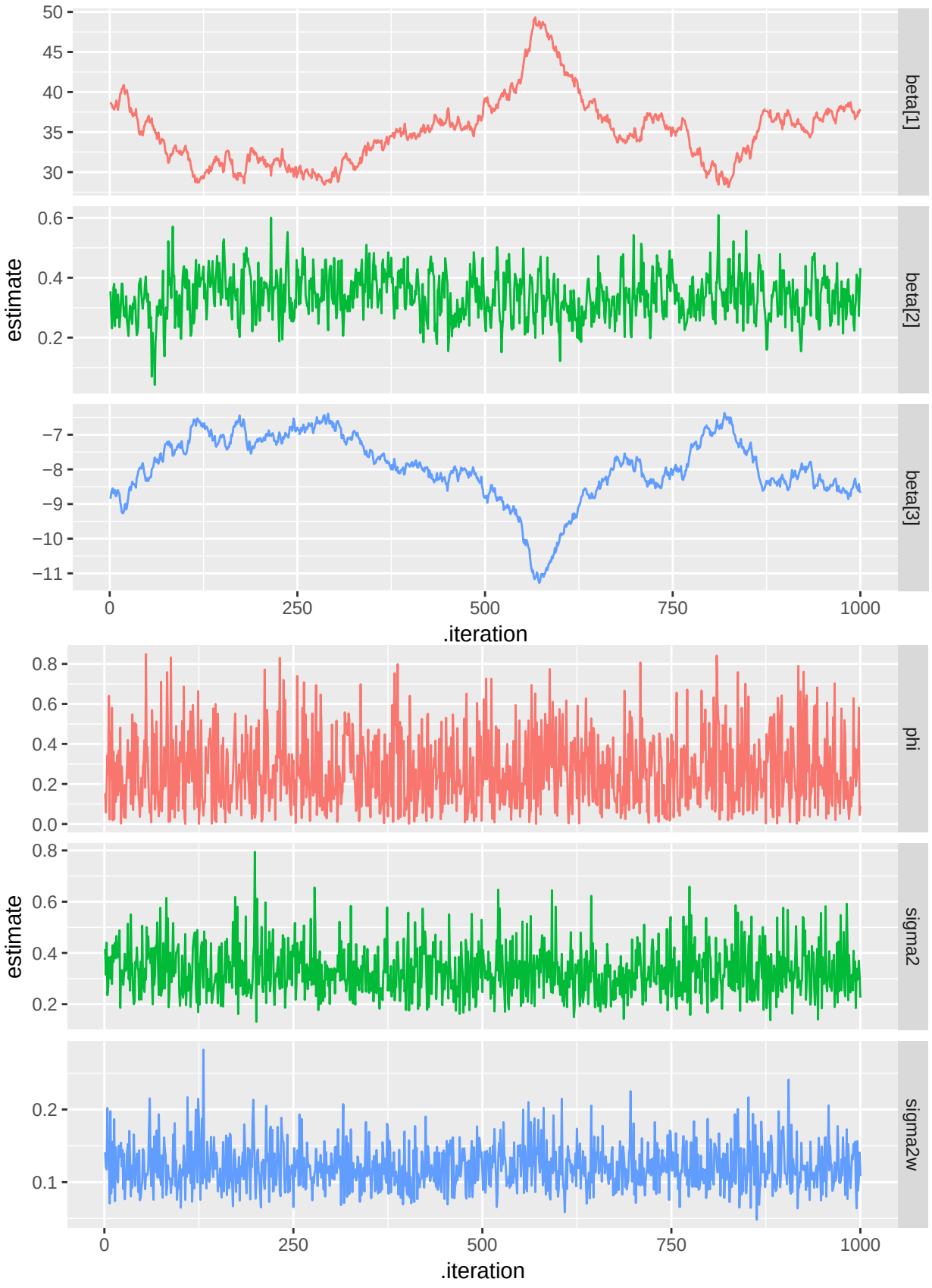
term	post_mean	post_med	post_lower	post_upper
beta[1]	38.089	38.048	32.900	45.432
beta[2]	0.327	0.328	0.188	0.473
beta[3]	-8.655	-8.663	-10.354	-7.457
phi	0.362	0.336	0.013	0.871
sigma2	0.329	0.314	0.177	0.573
sigma2w	0.119	0.116	0.072	0.187

The posterior means for the coefficient estimates for the mean structure are markedly similar to the estimates from the previously fitted frequentist CAR model. The posterior mean for $\phi \approx 0.362$ is markedly higher than the estimate produced by the previous CAR model indicating that the Bayesian model shares more information across observations than the frequentist model. As to be expected the residuals from the model display no evidence of significant spatial autocorrelation, as indicated by the Moran's I test.




```
##
##  Moran I test under randomisation
##
## data:  bas_data$bayes_car_resid
## weights: listW
##
## Moran I statistic standard deviate = -0.7796, p-value = 0.7822
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      -0.082765587      -0.018518519       0.006791508
```

The chains from the SAR model, given below, display the same results as that of the CAR model; the chains for β_1 and β_3 suffer from significant autocorrelation despite thinning while the chains for the remaining coefficients have explored the parameter space well and appear to have converged.

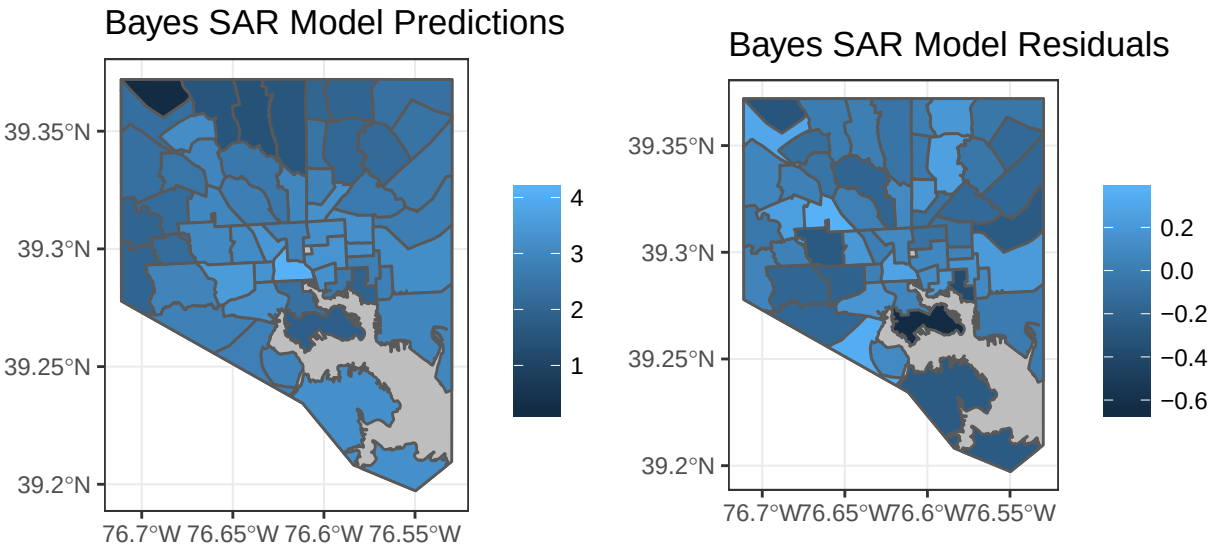

estimate
beta[1]
beta[2]
beta[3]
.iteration
phi
sigma2
sigma2w
estimate
.iteration

The posterior summaries for the parameters of the SAR model show similar behavior as those of the CAR model:

Table 3: Bayesian SAR Model

term	post_mean	post_med	post_lower	post_upper
beta[1]	35.097	35.119	28.931	46.884
beta[2]	0.339	0.338	0.199	0.481
beta[3]	-7.974	-7.956	-10.712	-6.584
phi	0.267	0.233	0.012	0.682
sigma2	0.330	0.315	0.181	0.553
sigma2w	0.121	0.117	0.073	0.192

The coefficients of the mean structure are similar to those from the frequentist SAR model while the estimate of ϕ is significantly higher indicating the Bayesian model, again, permits more sharing of information across observations. Again, the model residuals display no evidence of significant spatial autocorrelation unaccounted for by the model.




```
##
##  Moran I test under randomisation
##
## data:  bas_data$bayes_sar_resid
## weights: listW
##
## Moran I statistic standard deviate = -0.72062, p-value = 0.7644
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      -0.077619560      -0.018518519       0.006726325
```

Regarding predictive performance of each of the models, the areal data models have a lower RMSE than the simple linear model, with the Bayesian models showing an appreciable improvement over said model as well as the frequentist areal models. This is likely due to the higher estimate of ϕ allowing for more sharing of information across observations and thus improving predictive performance. As an aside, we note that the CAR model, both in the Bayesian and frequentist setting, marginally outperforms the SAR model.

	RMSE
bayes_car	0.1869894
bayes_sar	0.1884865
car	0.3122276
sar	0.3127594
lm	0.3140553