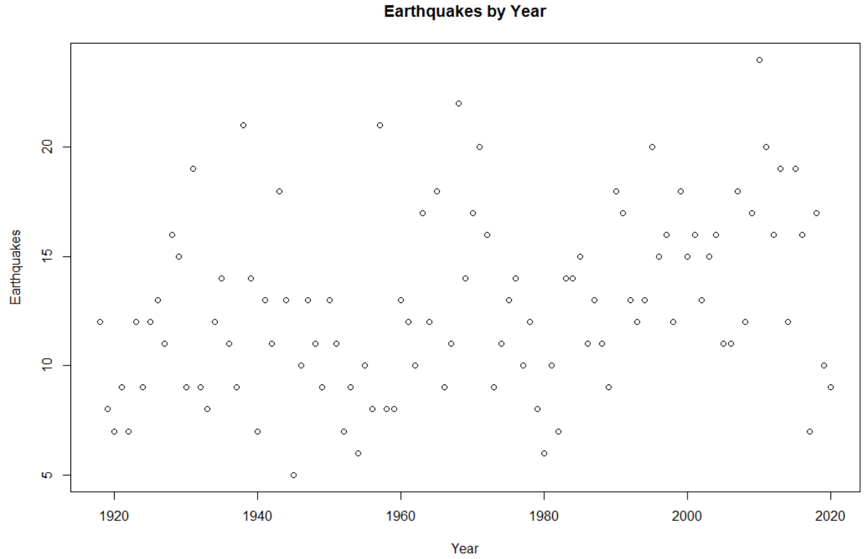
1 Problem

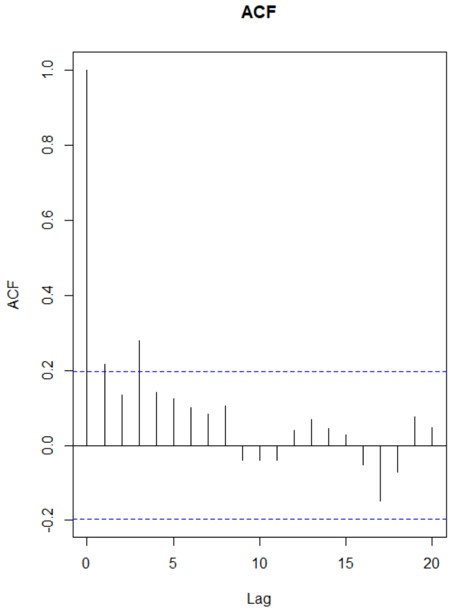
Annual earthquakes with a magnitude greater than 7 on the Richter scale will be analyzed from 1918 to 2020. The goal is to determine whether a mixture of or single autoregression (AR) model best represents the observed data. The first 100 years will be used to fit the model, while the last 3 years from 2018 – 2020 will be used to validate and compare model predictions.

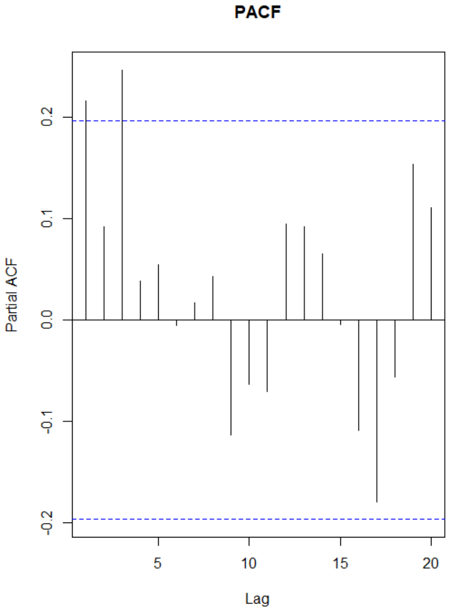
2 Data Exploration

Visual inspection of a plot of earthquakes by year displays high variance and a possible increasing trend over time. There is a 33% correlation between earthquakes and year indicating a risk to the assumption of stationarity. Given the dataset has been pre-selected, I will proceed without any modifications.



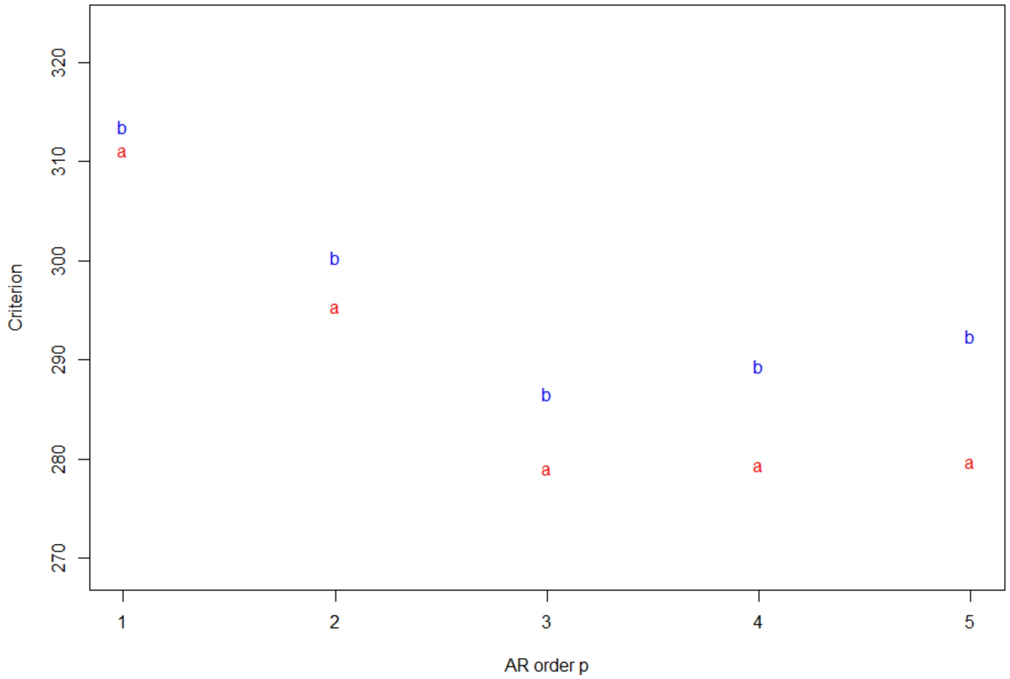
The ACF shows that the combined direct and indirect effects of earthquakes in prior years have a statistically significant correlation with the future number for up to 4 years, after which correlations fall strictly blue the critical threshold at 0.2. The PCF, however, indicates that when considering only the direct effect across time periods, looking back over 3 years is sufficient to capture the signal across periods. All correlations above the critical thresdhold in both ACF and PACF are positive, in line with intuition that more earthquakes in recent years translates to more earthquakes in the current year.





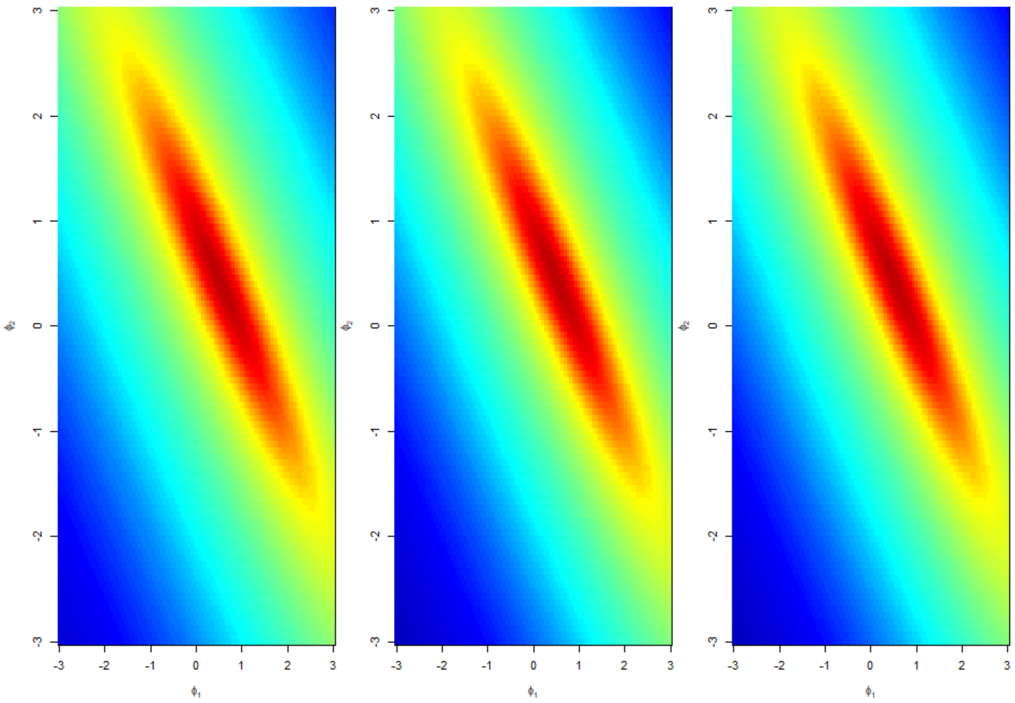
3 Single AR Model

As with PACF, both AIC and BIC indicate the selection of an order 3 AR model. As a result, an order 3 will be used for the single AR model.



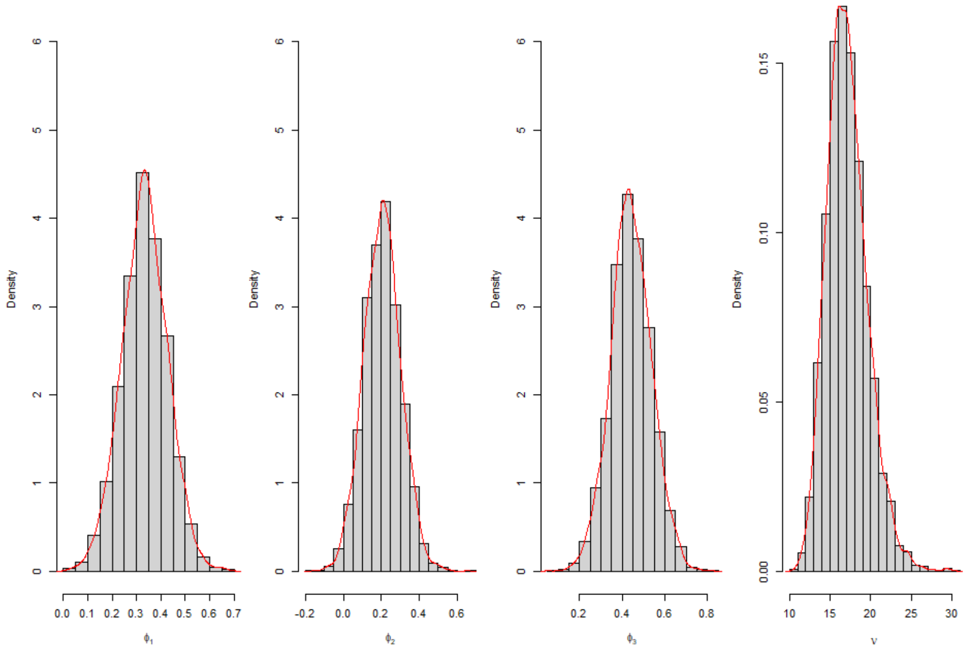
Prior to proceeding, various hyperparameters specifications are tested using an order 2 model.

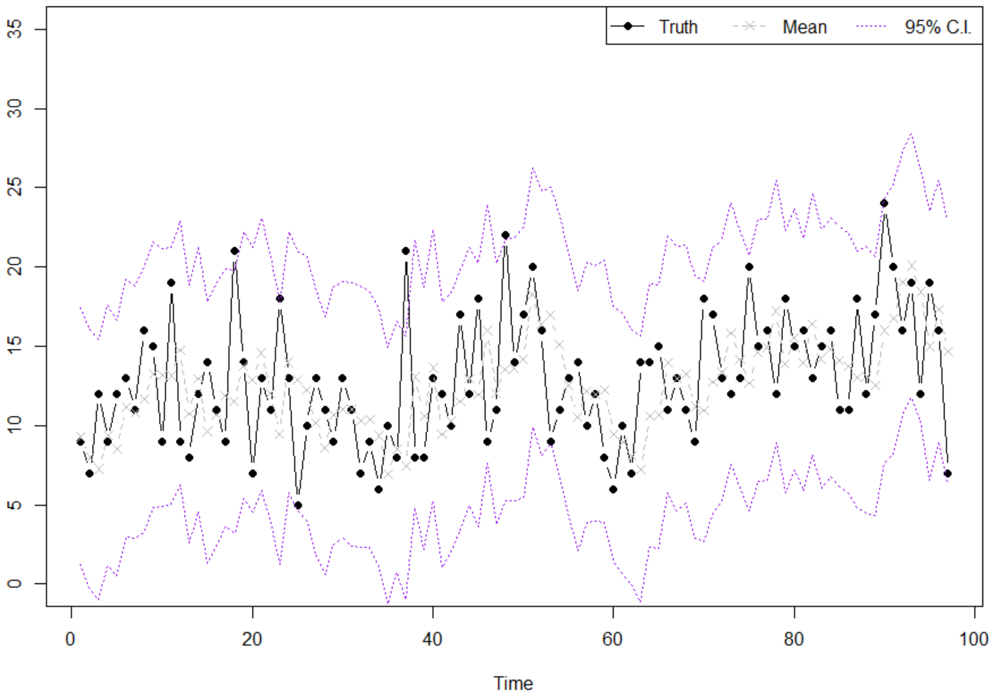
	Option 1	Option 2	Option 3
N0	2	6	6
D0	2	1	1
M0	Matrix(0, 0)	Matrix(0, 0)	Matrix(-0.5, -0.5)



As a result, option 1 will be selected for simplicity, and will now be used for an order 3 model.

A Markov Chain Monte Carlo (MCMC) for 5000 iterations led to the following posterior distributions for AR coefficients and variance.





The charts indicate a similar AR coefficient probability distribution across each lag, albeit slightly stronger at the first lag. The average sample variance is ~17. Earthquake inferences for the average sample are generally consistent with observed data minus anomalous years that fall close to the 95% CI intervals, such as the last year in the training set (2017) where only 7 qualifying earthquakes occurred. Less than 5 points in the last 100 years fell outside the 95% CI, adding clout to the model's 'goodness of fit'.

4 Mixture AR Model

For simplicity, the same order (3) will be used across two components of a mixture model. Furthermore, the same hyperparameters will be used as before with an added Dirichlet prior with constant alpha of 1 across each mixture component. The components will have separate scales and locations but consistent priors.

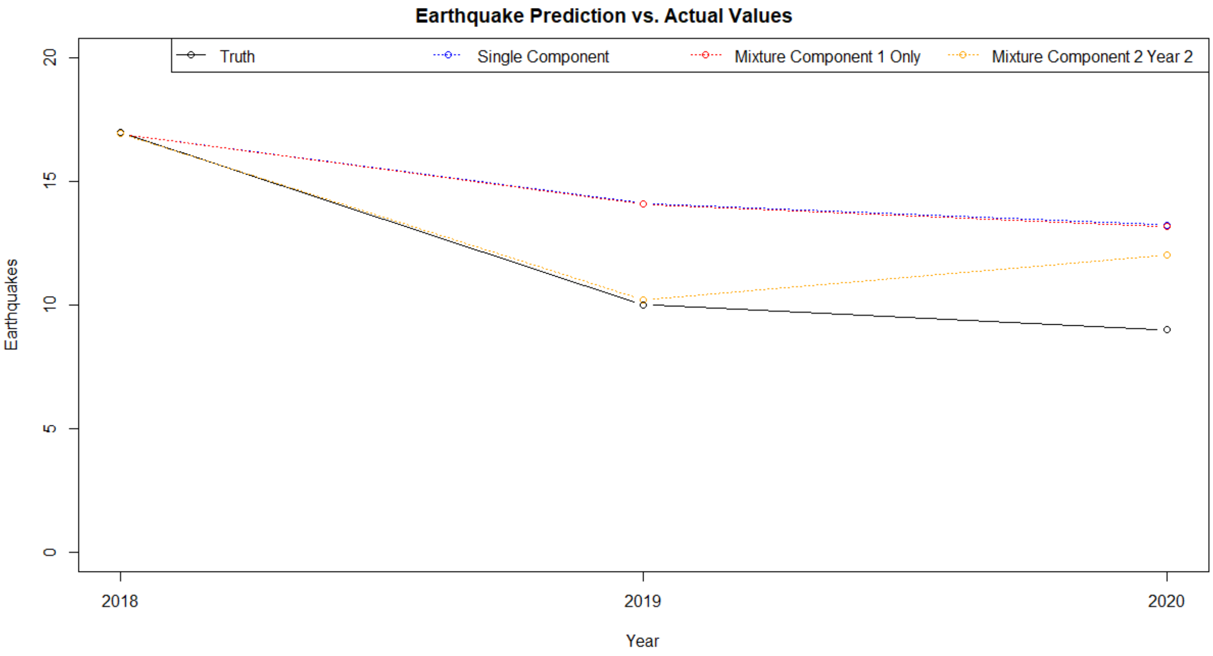
A Markov Chain Monte Carlo (MCMC) for 20000 iterations led to the following posterior means for AR coefficients and variance.

Mixture Component	Ω	Φ_1	Φ_2	Φ_3	v
1	0.97362189	0.33485395	0.19924020	0.44532631	16.153703
2	0.02637811	0.29347500	0.46304163	0.09212278	4.249794

By observing the omega values alone, we can see that the model strongly favors mixture component 1. This could be indicative of the preference for single mixture component in appropriately fitting the data especially considering a second component would ~ double the model's complexity. Furthermore, phi coefficients in mixture 2 have an average variance of ~38 in contrast to ~0.01 in mixture 1, indicating coefficient instability.

5 Model Comparison

The following charts show the average model predictions across simulated parameters on the validation set. Using mixture 1 only leads predictions consistent with the single component model, which provides a reasonably good fit to the observed data. The orange line indicates a scenario where the second mixture component is used in 2019, followed by mixture component 1 in 2020. While it provides the best fit, it's been selected specifically and is in effect an effort to overfit to the data. Average earthquake predictions are unstable and jump wildly when the 2nd mixture component occurs for sequential periods, indicative of the high variance implicit in the second mixture component coefficients. Due to the high variance of coefficients in additional components, a single component model is deemed more appropriate.



6 Conclusion

A single component AR model of order 3 provides a good fit to earthquake data from 1918 to 2020. Lack of coefficient stability and small posterior omegas in a multiple component model leads to the conclusion that a single component model may be more appropriate.

A future model may look to account for the upward trend observed in earthquakes over the course of the dataset, evidenced by a 33% correlation, to ensure the stationarity assumption holds. Furthermore, additional covariates on top of AR coefficients could be considered such as geological measurements from prior years to produce the best model.