POLITECNICO
MILANO 1863

Teenage pregnancies across the United States

Bayesian Statistics course projects

Lu, Radišić, Santamaria

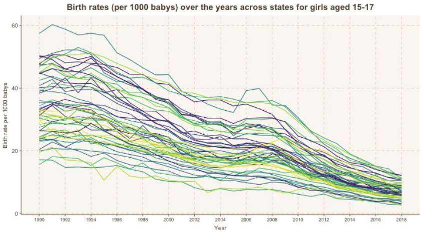
17^{th} February 2021

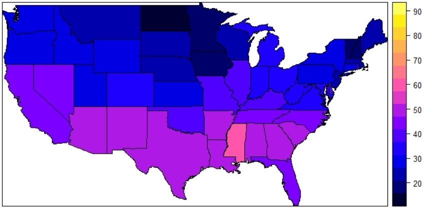
Short review for our dataset

The dataset contains the pregnancies carried out by teenage mothers over **51 states** in the US, for a period of time between **1990-2018**.

Data is present for **two age groups** of teenage girls:

- 15-17 years old (highschool)
- 18-19 years old (post-highschool)





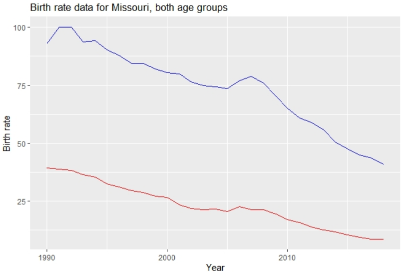
The **birth rates** are an average counting of how many girls from a sample of 1000 people in that group had a baby for a fixed year in a fixed State.

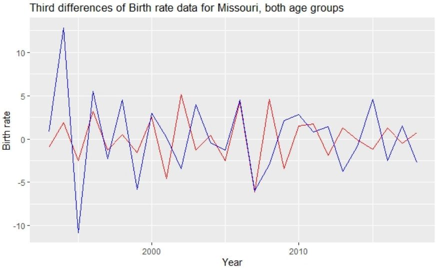
SUMMARY

1. **Previous presentation recap**
2. Spatio-Temporal ST.CARar model
3. Conclusion

1. Single state analysis: Missouri

We extracted from a discrete uniform distribution with support $S = 1, ..., 51$ the State of our analysis and we obtained Missouri.





In order to work with ARMA(p,q) models the process must be stationary and after applying 3^{rd} order differences we were able to reject ADF test's null hypotesis.

1.1 Models tested on Missouri

Model 1:
ARIMA(2,3,3) for two univariate time series with common parameter μ_0.

$$Y_{t,j}|\mu_0, \phi_j^{(1)}, \phi_j^{(2)}, \beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)}, \sigma^2_{Y_j} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_0 + \phi_j^{(1)} Y_{t-1,j} + \phi_j^{(2)} Y_{t-2,j} + \beta_j^{(1)} \epsilon_{t-1,j} + \beta_j^{(2)} \epsilon_{t-2,j} + \beta_j^{(3)} \epsilon_{t-3,j}, \sigma^2_{Y_j})$$

$$\mu_0 \sim \mathcal{N}(0, \sigma^2_{\mu_0})$$
$$\phi_j^{(i)} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2_{\phi_j^{(i)}}),\ i = 1, 2$$
$$\beta_j^{(i)} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2_{\beta_j^{(i)}}),\ i = 1, 2, 3$$
$$\sigma^2_{Y_j} \overset{\text{iid}}{\sim} InvGamma(a_Y, b_Y)$$

$j = 1, 2$ age groups, $t = 1, .., 23$

DISCARDED!

Bad performance about posterior inference.

Model 2:
ARIMA(1,3,0) for two univariate time series with information sharing over the parameters μ_{0j} and ϕ_j.

$$Y_{t,j}|\mu_{0j}, \phi_j, \sigma^2_{Y_j} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{0j} + \phi_j Y_{t-1,j}, \sigma^2_{Y_j})$$

$$\sigma^2_{Y_j} \overset{\text{iid}}{\sim} InvGamma(a_Y, b_Y)$$
$$\mu_{0j}|\mu, \sigma^2_{\mu_0} \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2_{\mu_0})$$
$$\phi_j|\phi, \sigma^2_{\phi} \overset{\text{iid}}{\sim} \mathcal{N}(\phi, \sigma^2_{\phi})$$
$$\mu \sim \mathcal{N}(0, \sigma^2_0)$$
$$\sigma^2_{\mu_0} \sim InvGamma(a, b)$$
$$\phi \sim \mathcal{N}(0, \sigma^2_0)$$
$$\sigma^2_{\phi} \sim InvGamma(a, b)$$

$j = 1, 2$ age groups, $t = 1, .., 23$

DISCARDED!

Violating stationarity assumption of AR model.

1.1 Models tested on Missouri

Model 3:
ARIMA(1,3,0) for two univariate time series with common ϕ, **random effects on** $\boldsymbol{\mu_{0j}}$.

$$Y_{t,j}|\mu_{0j},\phi,\sigma^2_{Y_j} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{0j}+\phi Y_{t-1,j},\sigma^2_{Y_j})$$

$$\sigma^2_{Y_j} \overset{\text{iid}}{\sim} InvGamma(a_Y,b_Y)$$
$$\mu_{0j} \overset{\text{ind}}{\sim} \mathcal{N}(0,\sigma^2_{\mu_{0j}})$$
$$\phi \sim \mathcal{N}(0,\sigma^2_{\phi})$$

$j=1,2$ age groups, $t=1,..,23$

VS

Model 4:
ARIMA(1,3,0) for two univariate time series with common μ_0, **random effects on** $\boldsymbol{\phi_j}$.

$$Y_{t,j}|\mu_0,\phi_j,\sigma^2_{Y_j} \overset{\text{ind}}{\sim} N(\mu_0+\phi_j Y_{t-1,j},\sigma^2_{Y_j})$$

$$\sigma^2_{Y_j} \overset{\text{iid}}{\sim} InvGamma(a_Y,b_Y)$$
$$\mu_0 \sim N(0,\sigma^2_{\mu_0})$$
$$\phi_j \overset{\text{ind}}{\sim} N(0,\sigma^2_{\phi_j})$$

$j=1,2$ age groups, $t=1,..,23$

1.1 Models comparisons

Model M_j	WAIC	MSE
ARIMA(2,3,3)		
Highschool	73.48	0.128
Post-highschool	104.62	2.798
ARIMA(1,3,0), both μ_{0j}, ϕ_j rand. eff.		
Highschool	91.59	1.99
Post-highschool	136.19	5.68
ARIMA(1,3,0), μ_{0j} rand. eff., common ϕ		
Highschool	91.20	2.78
Post-highschool	135.94	2.121
ARIMA(1,3,0), ϕ_j rand. eff., common μ_0		
Highschool	91.25	1.824
Post-highschool	134.13	1.529

$$Y_{t,j}|\mu_0, \phi_j, \sigma^2_{Y_j} \overset{\text{ind}}{\sim} \mathcal{N}(\mu_0 + \phi_j Y_{t-1,j}, \sigma^2_{Y_j})$$

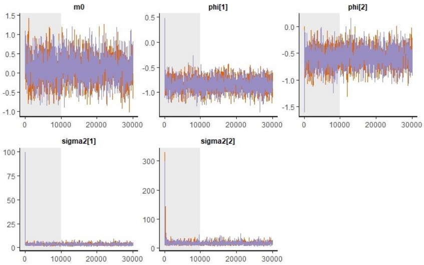
$$\sigma^2_{Y_j} \overset{\text{iid}}{\sim} InvGamma(a_Y, b_Y)$$
$$\mu_0 \sim \mathcal{N}(0, \sigma^2_{\mu_0})$$
$$\phi_j \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2_{\phi_j})$$

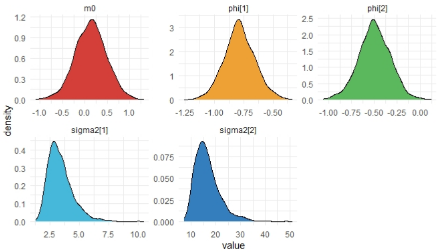
$$j = 1, 2 \text{ age groups}, \ t = 1, .., 23$$

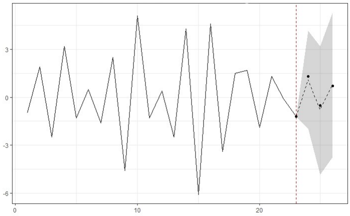
Parameters	μ_0	ϕ_1	$\sigma^2_{Y_1}$	ϕ_2	$\sigma^2_{Y_2}$
Posterior mean	0.12	-0.84	3.38	-0.57	16.67

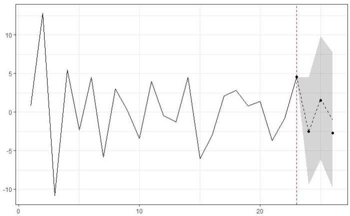
1.2 Best model for Missouri State

1.2.1 Diagnostics and forecasting









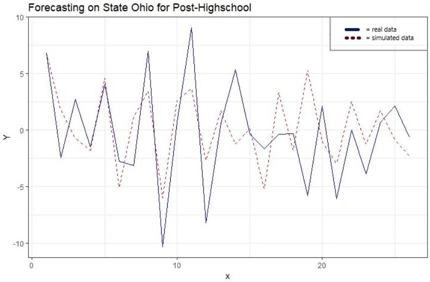
1.3 Accuracy of the model: MSE

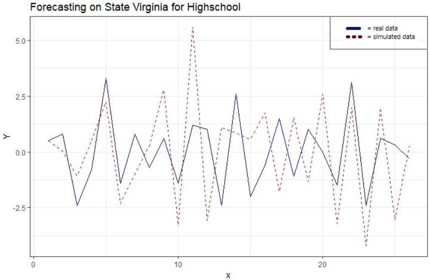
1.3.1 first half of States

	High.	Post-high.		High.	Post-high.
Alabama	22.312	69.446	**Illinois**	19.329	49.741
Alaska	26.947	195.22	**Indiana**	10.359	23.456
Arizona	33.595	55.592	**Iowa**	7.340	42.475
Arkansas	21.922	149.15	**Kansas**	19.063	64.509
California	8.912	30.393	**Kentucky**	20.171	49.478
Colorado	17.583	83.072	**Louisiana**	9.630	193.23
Connecticut	21.565	20.551	**Maine**	14.846	44.057
Delaware	47.055	95.913	**Maryland**	5.960	32.591
Dist. of Columbia	476.31	87.026	**Massachusetts**	11.594	42.709
Florida	23.882	46.929	**Michigan**	18.613	30.219
Georgia	22.319	67.183	**Minnesota**	6.510	43.305
Hawaii	73.418	83.770	**Mississippi**	32.854	67.324
Idaho	25.894	135.13	**Montana**	23.327	140.24

	High.	Post-high.		High.	Post-high.
Nebraska	21.038	176.89	**Rhode Island**	14.326	67.350
Nevada	22.930	94.604	**South Carolina**	32.776	53.714
New Jersey	18.277	56.274	**South Dakota**	29.063	248.66
New Hampshire	13.538	44.638	**Tennessee**	24.034	95.425
New Mexico	19.899	117.45	**Texas**	28.995	53.297
New York	21.953	19.881	**Utah**	22.352	100.06
North Carolina	17.187	70.141	**Vermont**	84.810	64.681
North Dakota	46.369	95.133	**Virginia**	5.081	37.197
Ohio	9.060	14.974	**Washington**	26.874	57.546
Oklahoma	19.823	68.813	**West Virginia**	6.479	61.357
Oregon	7.778	95.090	**Wisconsin**	13.539	34.520
Pennsylvania	8.741	30.455	**Wyoming**	23.116	409.58

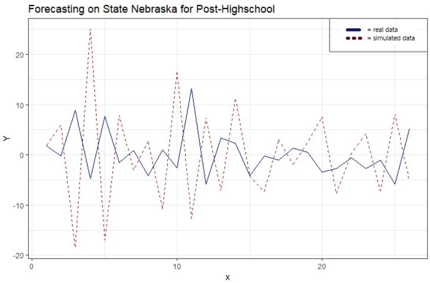
1.4 Forecasting on other States
1.4.1 smallest MSE

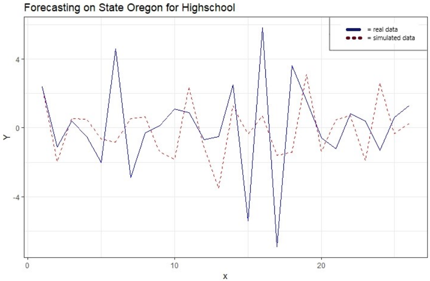




1.4 Forecasting on other States

1.4.2 counter examples



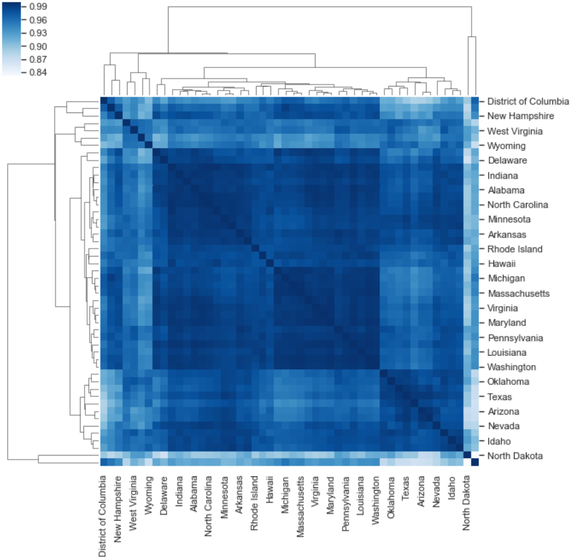


... but why?

1. model is trained on Missouri's data
2. model doesn't take into account spatial informations
3. data are transformed using 3^{rd} order differences

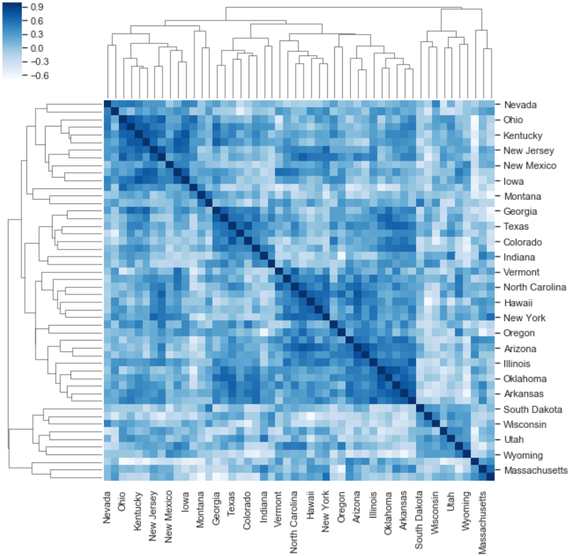
1.5 Correlation between States

1.5.1 Highschool group raw data



1.5 Correlation between States

1.5.2 Highschool group transformed data



SUMMARY

1. Previous presentation recap
2. **Spatio-Temporal ST.CARar model**
3. Comparison and Conclusion

$$I = \frac{K}{W} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}, \; w_{ij} = 1 \text{ if } i, j \text{ are neighbours, 0 otherwise}$$

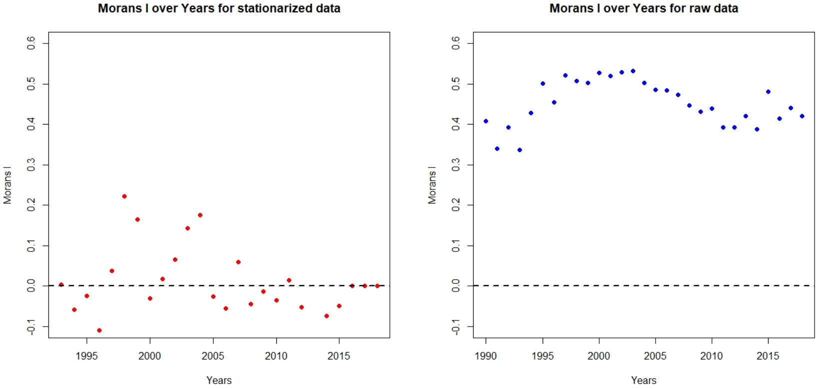
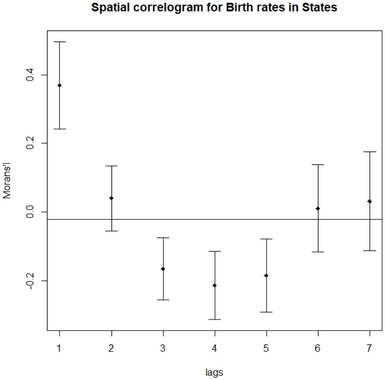
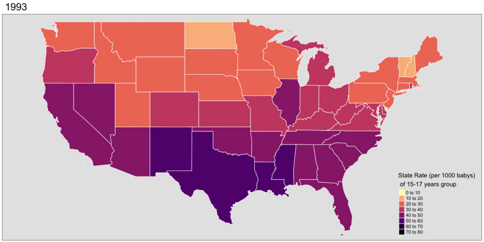


Figura: Moran's I shows evidence of **strong spatial correlation** in Birth rates.

2. Spatio-Temporal ST.CARar model

2.1.2 Moran's I





Spatial correlogram for Birth rates among USA states shows that it is reasonable to consider only first-order neighbours, as Moran's I drops when considering higher order neighbours.

$$Y_{kt}|\mu_{kt}, \nu^2 \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{kt}, \nu^2)$$
$$\mu_{kt} = \beta_0 + \phi_{kt}$$
$$\beta_0 \sim \mathcal{N}(\mu_\beta, \sigma_\beta)$$
$$\nu^2 \sim Inv - Gamma(a, b)$$

$$\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}, \rho_S, \rho_T, \tau^2 \sim \mathcal{N}_{49}(\rho_T \boldsymbol{\phi}_{t-1}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1})$$
$$\boldsymbol{\phi}_1 \sim \mathcal{N}_{49}(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1})$$
$$\tau^2 \sim Inv - Gamma(a, b)$$
$$\rho_S, \rho_T \overset{\text{iid}}{\sim} Uniform(0, 1)$$

$$\mathbf{Q}(\mathbf{W}, \rho_S) = \rho_S[diag(\mathbf{W1}) - \mathbf{W}] + (1 - \rho_S)\mathbf{I}$$

Where **1** is a $49x1$ vector of ones, while **I** is the $49x49$ identity matrix. For each State $k = 1, ..., 49$, with $t = 1, ..., 29$ and $a = 1, b = 0.01$.

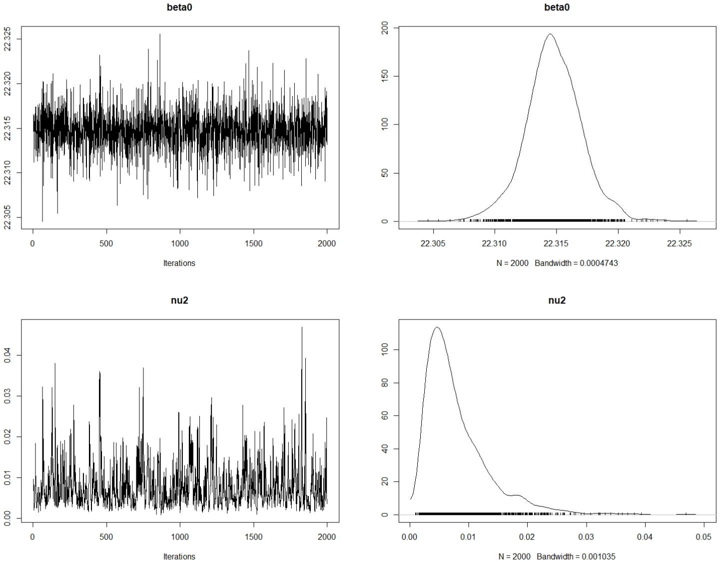
Posterior quantities for selected parameters

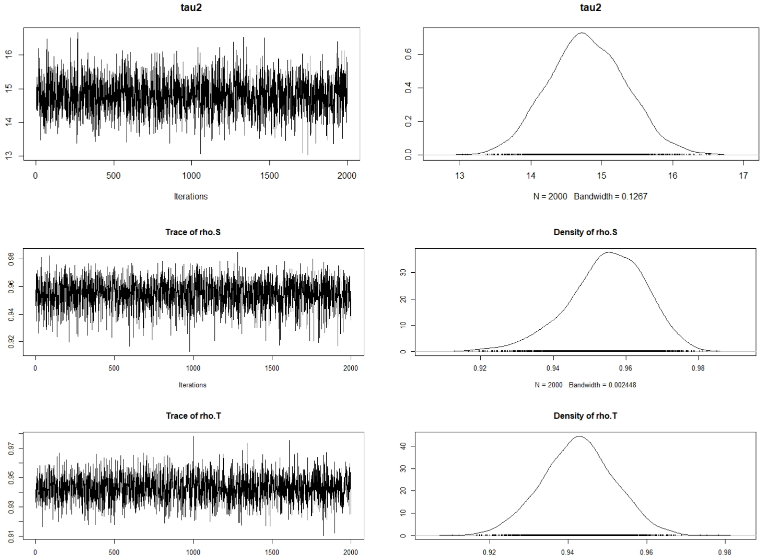
	Median	**2.5%**	**97.5%**	**n.effective**	**Geweke.diag**
(Intercept)	22.3147	22.3100	22.3196	2000.0	0.9
τ^2	14.7783	13.7520	15.8926	2000.0	0.2
ν^2	0.0064	0.0020	0.0225	346.5	0.2
ρ_S	0.9556	0.9307	0.9734	2000.0	-0.4
ρ_T	0.9426	0.9247	0.9602	2000.0	-0.8

$total$: 220000 $burnin$: 20000 $thin$: 100 $samples$: 2000

2. Spatio-Temporal ST.CARar model

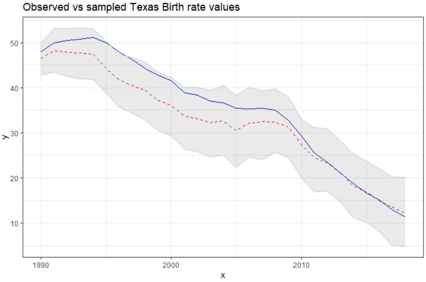
2.3.1 Posterior inference: traceplots and posterior densities

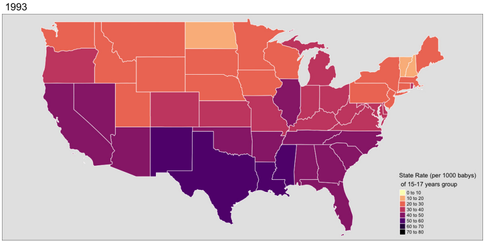



tau2
Iterations
tau2
N = 2000 Bandwidth = 0.1267
Trace of rho.S
Iterations
Density of rho.S
N = 2000 Bandwidth = 0.002448
Trace of rho.T
Density of rho.T

Montana (left) and Missouri (right) observed values of Birth rate reported in blue, simulated values in red while 95% Credible Intervals in grey.

Texas has higher Birth rate values than what would be expected given its neighbours. It seems to have had a bigger drop in Birth rates after 2010.





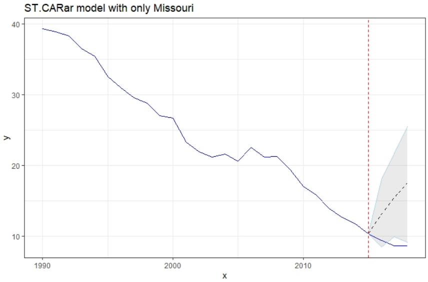
Texas' simulated values for Birth rates reported in red, observed values in blue, while 95% Credible Interval in grey.

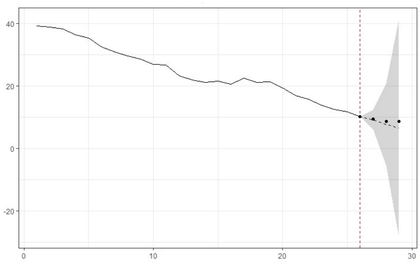
SUMMARY

1. Previous presentation recap
2. Spatio-Temporal Bayesian model
3. **Comparison and Conclusion**

3. ST.CARar vs ARIMA(1,3,0)

Predicting last 3 years of Missouri times series





Results when forecasting 2016-2018 with **ST.CARar** are much **worse** than when forecasting with **ARIMA(1,3,0)**.

3.1 Conclusion

Which model is better?

It depends on the goal of the study.

- If the goal is to predict the future values of a given state for which we have previous measures, the **ARIMA(1,3,0)** model makes better predictions.
- If the goal is to study the overall behaviour in the USA, or to predict the values of a state only by observing the measures in other states, the **ST.CARar** model should be used.

REFERENCES

- Lee, D., Rushworth, A., and Napier, G. (2018). Spatio-Temporal Areal Unit Modeling in R with Conditional Autoregressive Priors Using the CARBayesST. *Journal of Statistical Software, 84*(9), pp. 1–39.
- Centers for Disease Control and Prevention (2018) *NCHS - U.S. and State Trends on Teen Births,* [online], URL https://data.cdc.gov/NCHS/NCHS-U-S-and-State-Trends-on-Teen-Births/y268-sna3/data/.
  Accessed on October 2020.