# Teenage pregnancies across the United States

**Lu Jiajie[1], Radišić Katarina[1], Santamaria Andrea[1]**

[1] *Politecnico di Milano*

$17^{th}$ February 2021

The study of teenage pregnancies and the ways of preventing it are a subject of much study and interest in the social sciences. In the USA, the number of teenage mothers has been constantly decreasing since the 1990.

In this work we discuss and develop two different models for assessing the number of pregnancies carried out by teenage mothers happened between 1990 and 2018 in the United States of America.

The first model proposed, aims to predict the future birth rates in a given state, for which the birth rates of previous years are observed and collected. It is, intrinsically an ARIMA(1,3,0) model which shares information across two different teenage age groups in one State.

The second model we propose, aims to model the overall behaviour in the United Stated of America, and thus comes from the family of Spatio-Temporal models, more specifically it is an ST.CARar model. In this model the focus is on one single age group.

Lastly, the two models are compared, and we conclude that both of them are useful, but give answers to different questions. The ARIMA(1,3,0) model seems to do a better job in forecasting, whereas the ST.CARar model allows us to compare a state with it's neighbourhood or even sample unobserved values with high accuracy.

## Contents

**1 Dataset** **1**

**2 Single State analysis** **2**
2.1 ARIMA(2,3,3) . . . . . 3
2.2 ARIMA(1,3,0): shared information among parameters . . . . . 4
2.3 ARIMA(1,3,0): common temporal dependence parameter $\phi$ . . . . . 5
2.4 ARIMA(1,3,0): common baseline parameter $\mu_0$ . . . . . 7
2.5 Accuracy on other States . . . . . 8

**3 Spatio-Temporal model** **9**
3.1 Moran's I . . . . . 10
3.2 ST.CARar model with gaussian likelihood 11
3.3 Posterior inference . . . . . 11
3.4 Predicting birth rates in a State given its neighbours . . . . . 12
3.5 ST.CARar model with poisson likelihood 13
3.6 Comparisons . . . . . 13

**4 Conclusions** **14**

**5 Bibliography** **14**

**A Appendix** **15**

## 1 Dataset

Our dataset is made out of observations across 51 American States, over 29 years (1990-2018) for 2 separate

age groups:

- girls aged 15-17, also called high-school group
- girls aged 18-19, also called post high-school group.

This counts for a total of 51x29x2 = 2958 observations of Birth rates.

The number of pregnant girls for a given year, in a given state, is expressed through a **Birth rate**. This means that the number we will be dealing with are not the brute number of girls that gave birth, but rather a normalized number obtained by dividing the total number of girls giving birth with the total number of girls present in the specific state, for the specific year. The Birth rates are thus real and positive values.
In *section 2* we will exploit the existence of two different age groups, as we will introduce random effects on each on them. In *section 3* however, we will study each age group separately.

## 2 Single State analysis

In this section, we decided to focus on a single State. We build a model which will explain the two time series (one for each age group) and make good forecasts for the upcoming years. The state Missouri (Figure 1) was chosen at random and in this section we will focus on fitting a model for it.

**Figure 1:** *Missouri State*

As it can be seen from Figure 2, both time series have a decreasing trend between 1990 and 2018, period which considered for the analysis. This is a problem, since a decreasing, or any kind of trend, is a clear sign of non-stationarity and has to be solved since the model we want to adopt belongs to the family of ARMA models.
Our first step will be to find a transformation which fulfills the hypothesis of stationarity of the data.

| ADF p-value | $1^{st}$ diff. | $2^{nd}$ diff. | $3^{rd}$ diff. |
|---|---|---|---|
| highschool | 0.284 | 0.201 | $< 0.01$ |
| post-highschool | 0.2237 | 0.09236 | 0.01116 |

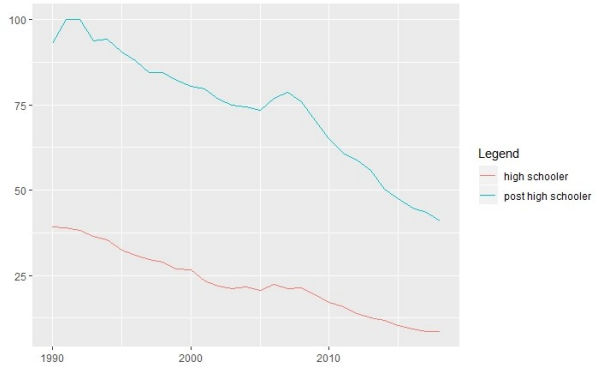


**Figure 2:** *Times series evolution from 1990 to 2018. The red line corresponds to the Birth rates of age group 15-17, while the blue line corresponds to the age group 18-19.*

As a solution we transformed the data using the d order differences and verified the stationarity assumption through the Augmented-Dickey Fuller (ADF) test. Basically, we want to find the smallest d such that it is possible to reject the null hypothesis (of non-stationnarity, or more precisely, the presence of an unit root) of the ADF test, where an order difference is:
$1^{st}$ differences: $\Delta^1 y_t = y_t - y_{t-1}$
$2^{nd}$ differences: $\Delta^2 y_t = \Delta^1 y_t - \Delta^1 y_{t-1}$
$3^{rd}$ differences: $\Delta^3 y_t = \Delta^2 y_t - \Delta^2 y_{t-1}$
and so on ...
Since the p-values are less than 0.05 when $d = 3$ for both age groups, we can assume with a 95% confidence level that our times series are now stationary. Using $d^{th}$ order differences means that we are basically adopt the ARIMA(p,d,q) model, but we still need to find the correct parameters p and q.

It's clear from Figure 3 that now the times series are stationary, both of them evolve around zero and there is no trend, but we have to pay a price: we started with 29 observations and after transformation we lost 3 obs.

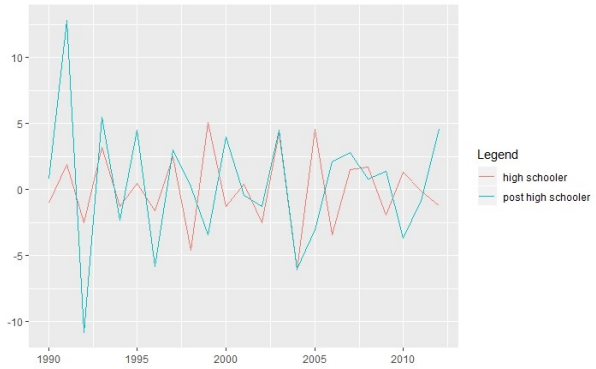


**Figure 3:** *Transformed data*

Before any kind of tuning in order to find parameters p and q we decided to split the data into a training set (23 obs) and a test set (3 obs). The former will be used to build the model and the latter will be used to test model's performances in terms of accuracy on prediction. Furthermore, one of the possible ways to find p and q is to observe the plots of ACF and PACF using our data.

Figure 4 shows a very clear pattern for the high-school group that suggests to use an ARMA(2,3) model, but this is not the case for the other group, since its ACF and PACF seem to recall an AR(1) or AR(2) model as Figure 5 shows.

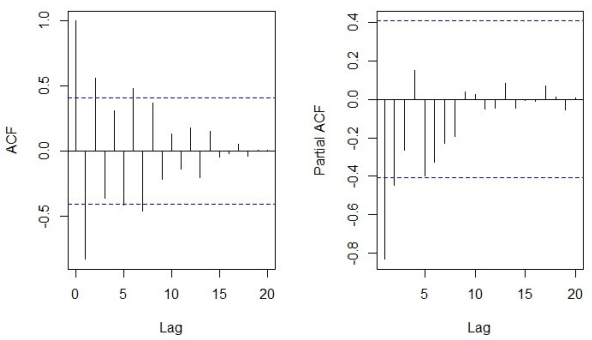


**Figure 4:** *ACF and PACF for $Hischooler$, age 15-17*

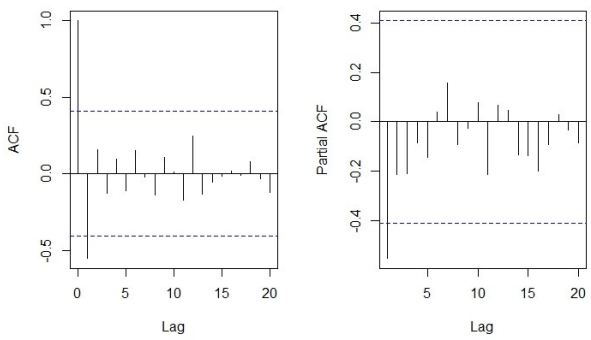


**Figure 5:** *ACF and PACF for $Post-hischooler$, age 18-19*

There is not a correct choice to make, so we decided to implement both **ARIMA(2,3,3)** and **ARIMA(1,3,0)** models in STAN programming language and then compare their performances in terms of posterior inference and overall prediction and forecasting.

It is clear that the ARIMA(1,3,0) model has less parameters to be fitted and thus is simpler.

## 2.1 ARIMA(2,3,3)

This complex model treats an observation at time $t$ as a linear combination of a baseline $\mu_0$, the previous observations at time $t-1$ and $t-2$ with coefficients $\phi_j^{(1)}$ and $\phi_j^{(2)}$ and a shock term $\epsilon_t$ evaluated at time $t$, $t-1$ and $t-2$ with coefficients, respectively, $\beta_j^{(1)}$, $\beta_j^{(2)}$ and $\beta_j^{(3)}$ for each group age j=1,2.

Our aim is trying to model the stationary time series through its past and a noisy term, assuming that the parameters in the model come from a Gaussian and an Inverse-Gamma distributions.
Here is the model we adopt:

$$
\begin{gathered}
Y_{t,j}|\mu_0, \phi_j^{(1)}, \phi_j^{(2)}, \beta_j^{(1)}, \beta_j^{(2)}, \beta_j^{(3)}, \sigma^2_{Y_j} \overset{\text{ind}}{\sim} \\
N(\mu_0 + \phi_j^{(1)} Y_{t-1,j} + \phi_j^{(2)} Y_{t-2,j} + \beta_j^{(1)} \epsilon_{t-1,j} + \\
\beta_j^{(2)} \epsilon_{t-2,j} + \beta_j^{(3)} \epsilon_{t-3,j}, \sigma^2_{Y_j}) \\
\mu_0 \sim N(0, \sigma^2_{\mu_0}) \\
\phi_j^{(i)} \overset{\text{ind}}{\sim} N(0, \sigma^2_{\phi_j^{(i)}}),\ i = 1,2 \\
\beta_j^{(i)} \overset{\text{ind}}{\sim} N(0, \sigma^2_{\beta_j^{(i)}}),\ i = 1,2,3 \\
\sigma^2_{Y_j} \overset{\text{iid}}{\sim} InvGamma(a_Y, b_Y) \\
j = 1,2 \text{ age groups},\ t = 1,..,23
\end{gathered}
$$

Note: $\epsilon_{t-1,j} = Y_{t-1,j} - \hat{Y}_{t-1,j}$, where $\hat{Y}$ is the fitted value for a certain year.

Unfortunately for what concern the diagnostics of the model we obtain terrible results. The chain from MCMC are too correlated, as shown in Figure 6, the traceplots in Figure 7 seem to have not reached convergence and last but not least the posterior distributions of the parameters have an ugly shape (Figure 8).

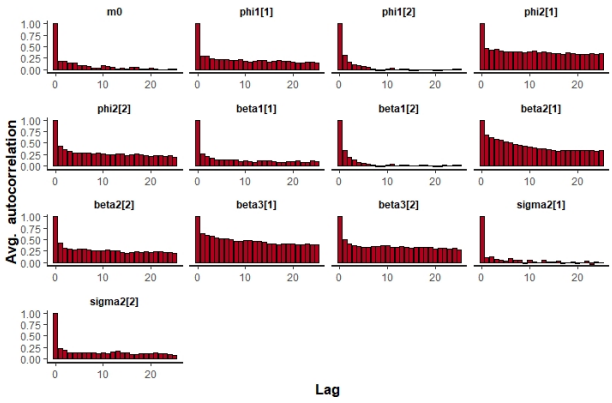


**Figure 6:** *Autocorrelation*

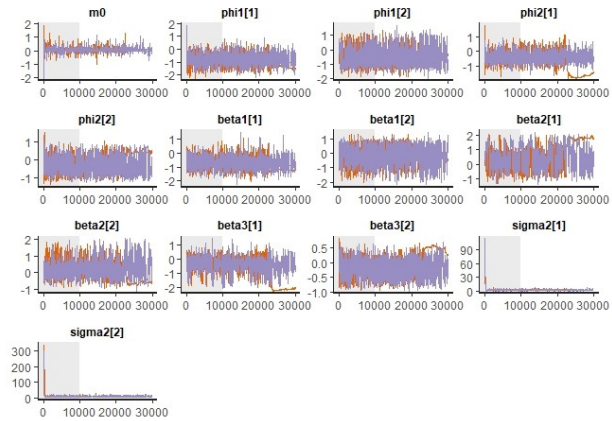


**Figure 7:** *Traceplots*

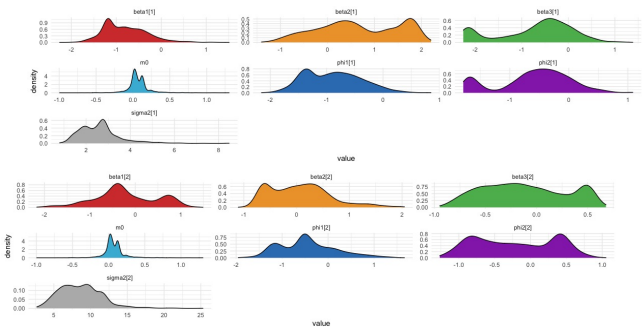


**Figure 8:** *Posterior distributions*

It seems to be the case in which having a complex model does not provide good results from a Bayesian point of view.

So we decide to simplify the model and consider the autoregressive model of order $p = 1$ perform.

However, we want to study different ways in which the two age groups could interact with each other and share information. Thus, we propose three ARIMA(1,3,0) models, each with a specific characteristic that makes it different from the others. Notice that none of them assumes complete independence between the groups because it does not seem to be an interesting case of analysis.

To put it in another way, the particularity of each of the three models proposed in the next section is the way in which the random effects are chosen.

## 2.2 ARIMA(1,3,0): shared information among parameters

In this model we assume conditional independence between data with respect to few parameters $\mu_{0j}$, $\phi_j$ and $\sigma_j^2$ for each age group $j = 1, 2$ coming from a Gaussian distribution. It is possible to consider $\mu_{0j}$ as a baseline value for group $j$ around which data fluctuate and $\phi_j$ the temporal dependence parameter that links an observation at time t with another at time $t-1$, while $\sigma_j^2$ is just the variance of the distribution for each age group $j = 1, 2$.

Moreover, we want information to be shared among parameters $\mu_{01}$, $\mu_{02}$ so that we assume them coming from the same Gaussian distribution, and the same considerations for the parameter $\phi_j$.

$$
\begin{gathered}
Y_{t,j}|\mu_{0j}, \phi_j, \sigma_j^2 \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{0j} + \phi_j Y_{t-1,j}, \sigma_j^2) \\
\sigma_j^2 \overset{\text{iid}}{\sim} InvGamma(a_Y, b_Y) \\
\mu_{0j}|\mu, \sigma_{\mu_0}^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma_{\mu_0}^2) \\
\phi_j|\phi, \sigma_\phi^2 \overset{\text{iid}}{\sim} \mathcal{N}(\phi, \sigma_\phi^2) \\
\mu \sim \mathcal{N}(0, \sigma_0^2) \\
\sigma_{\mu_0}^2 \sim InvGamma(a, b) \\
\phi \sim \mathcal{N}(0, \sigma_0^2) \\
\sigma_\phi^2 \sim InvGamma(a, b)
\end{gathered}
$$

Traceplots are shown in Figure 9 and we used two chains starting from different initial points with burn-in=$10^4$ and thin=10. They are very good traceplots since there is no trend, the support is explored enough and convergence for each parameters is achieved.

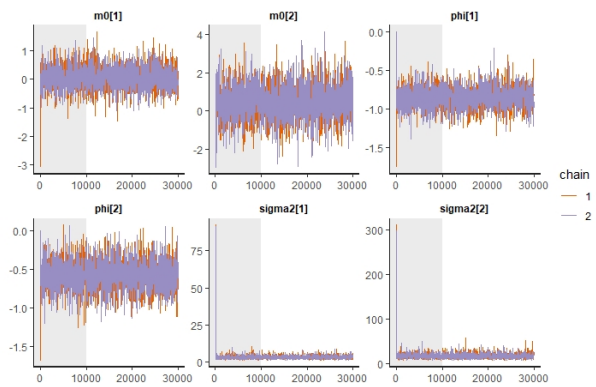


**Figure 9:** *Traceplots*

Moreover we analyzed the autocorrelation of the chain and as showed in Figure 10 seems that we can

| 95% CI for $\phi_j$ | 2.5% | 50% | 97.5% |
|---|---|---|---|
| $\phi_1$ | $-1.1257$ | $-0.83736$ | $-0.5833$ |
| $\phi_2$ | $-0.9610$ | $-0.57312$ | $-0.1937$ |

assume independence between the state of the chain.

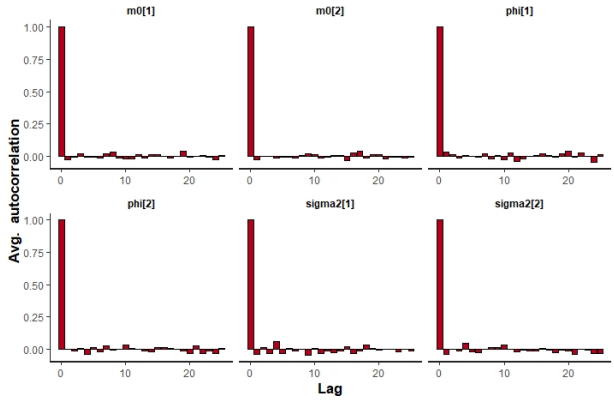


**Figure 10:** *Autocorrelation*

For what concerns the posterior distribution we have clear unimodal shapes for each parameters (Figure 11), which was not the case for the previous model.

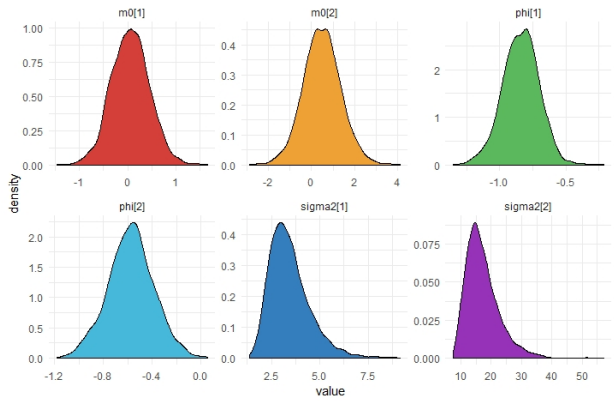


**Figure 11:** *Posterior distribution*

It's important to consider how much mass is put inside the interval $(-1, 1)$ for the parameters $\phi_j$ since stationarity was assumed at the beginning and it means that $|\phi_j| < 1$ strictly for $j = 1, 2$. From the $95\%$ credible intervals we have that for one group this condition is not respected, but this problem is solved when we consider the posterior mean, or the posterior mode, as the estimator for these parameters $\phi_j$.

As far as forecasting is concerned, we predicted the observations for the next three years (2016-2018) using the posterior mean for each parameter and the obtained values for both groups are very close to the real data coming from the test set.

In Figures 12 and 13 there are the 3rd order differences data for respectively high-school and post high-school group, while see Figures 41 and 42 for the prediction on the raw data using the inverse

| Posterior means | |
|---|---|
| $\mu_{01}$ | 0.0542 |
| $\mu_{02}$ | 0.4846 |
| $\phi_1$ | -0.8409 |
| $\phi_2$ | -0.5752 |
| $\sigma_1^2$ | 3.4437 |
| $\sigma_2^2$ | 17.1680 |

transformation in the Appendix.

The model makes very nice forecasts, the credible interval augments over the course of the years, which is to be expected, and the means follows closely the true observed Birth rates for both age groups.

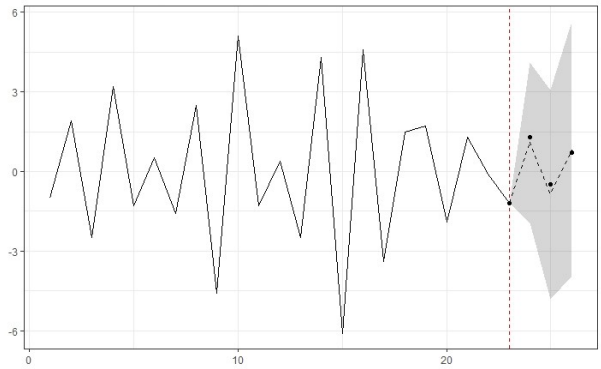

**Figure 12:** *Forecasting on transformed High-school group data*

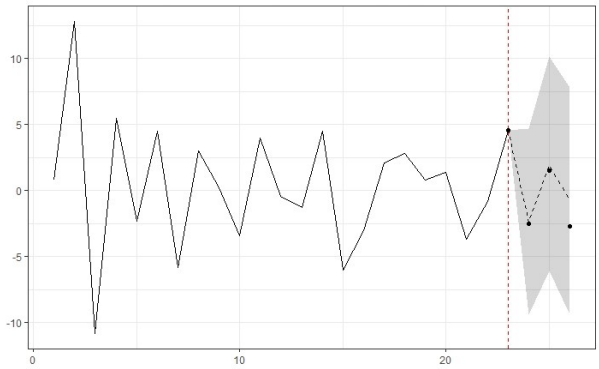

**Figure 13:** *Forecasting on transformed Post-High-school group data*

## 2.3 ARIMA(1,3,0): common temporal dependence parameter $\phi$

We thought that could be an option to implement a model that uses the same $\phi$ for both groups hoping that this simplification will solve the problem of the 95% credible interval regarding the stationarity assumption.

As previously done, we assume conditional independence between data with respect to $\mu_{0j}$, $\phi_j$ and $\sigma_j^2$ for each group j=1,2 coming from a Gaussian distribution. This model can be interpreted as an evolution for two stationary time series that have same temporal dependence but could differ in mean, since it depends on the $\mu_{0j}$.

$$Y_{t,j}|\mu_{0j}, \phi, \sigma_j^2 \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{0j} + \phi Y_{t-1,j}, \sigma_j^2)$$
$$\sigma_j^2 \overset{\text{iid}}{\sim} InvGamma(a, b)$$
$$\mu_{0j} \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2_{\mu_{0j}})$$
$$\phi \sim \mathcal{N}(0, \sigma_\phi^2)$$

Traceplots are shown in Figure 14 and two chains were used, starting from different initial points with burn-in=$10^4$ and thin=10. Again, traceplots are very nice since there is no trend, the support is explored enough and convergence for each parameters is achieved.

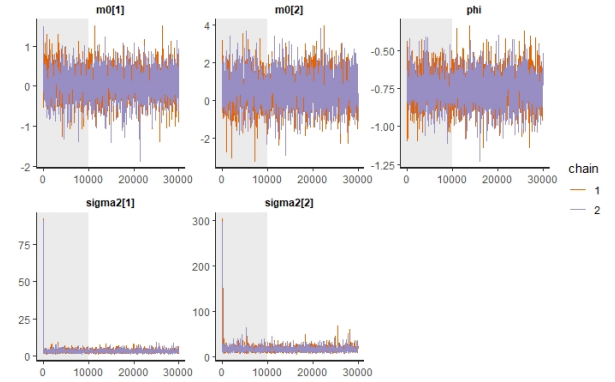


**Figure 14:** *Traceplots*

For what concern the autocorrelation (Figure 15) we can assume independence between the states of the chain as before.

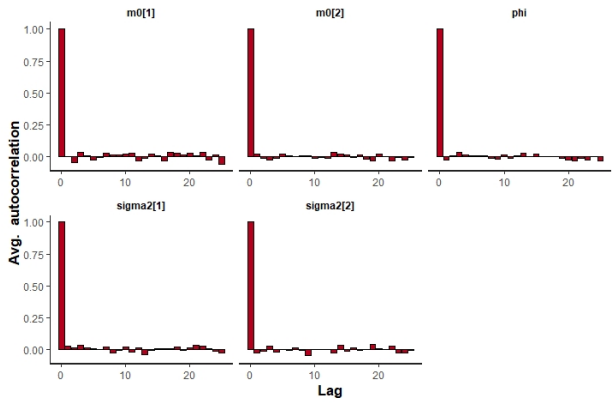


**Figure 15:** *Autocorrelation*

An improved behaviour is observed this time when it comes to the posterior inference, not only for the posterior distributions themselves (Figure 16), but also because the 95% credible interval for the temporal dependence parameter $\phi$ is inside the critical interval (-1,1), so that we have no doubts regarding the choice

| **95% CI for $\phi$** | **2.5%** | **50%** | **97.5%** |
|---|---|---|---|
| $\phi$ | $-0.9678$ | $-0.74715$ | $-0.5322$ |

| **Posterior means** | |
|---|---|
| $\mu_{01}$ | 0.04698 |
| $\mu_{02}$ | 0.57434 |
| $\phi$ | -0.74758 |
| $\sigma_1^2$ | 3.40321 |
| $\sigma_2^2$ | 17.21190 |

for the estimate of $\phi$, as it ensures the stationarity of the time series.

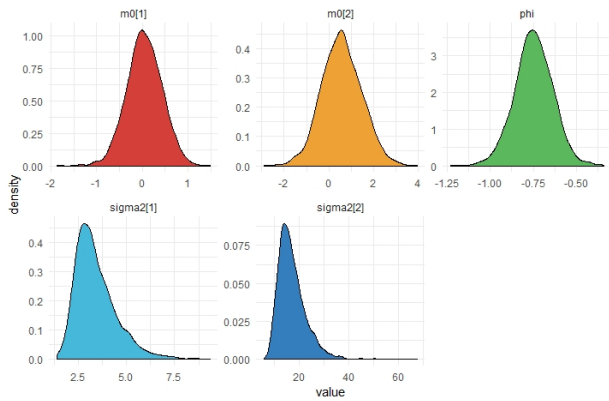


**Figure 16:** *Posterior distributions*

Once again, we want to study the forecasting properties of this model. We use the posterior mean as estimators for the parameters and in Figure 17 and 18 there are the predicted values respectively for high-school and post high-school age group data. The predictions on raw data (using the inverse transformation) are reported in Figures 43 and 44 in the Appendix.

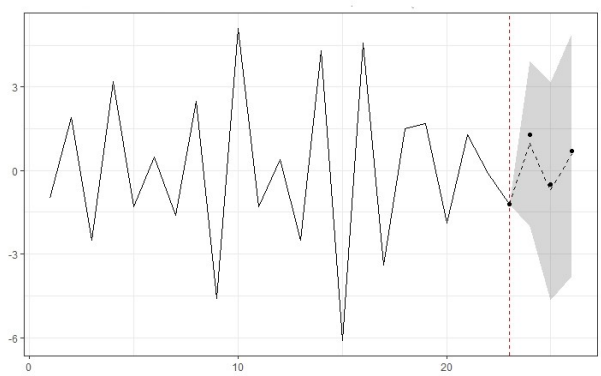

**Figure 17:** *Forecasting on transformed High-school group data*

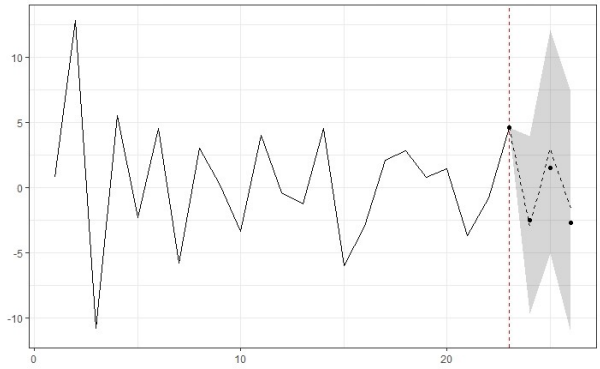

**Figure 18:** *Forecasting on transformed Post-High-school group data*

## 2.4 ARIMA(1,3,0): common baseline parameter $\mu_0$

The last model we proposed is basically the vice-versa of the previous one. Due to Figure **??** seems reasonable to assume a common baseline $\mu_0$ and a different temporal dependence parameter $\phi_j$ which could explain for each group how an observation at time t depends on one at t-1. The model we adopted is reported below:

$$
\begin{aligned}
Y_{t,j}|\mu_0, \phi_j, \sigma_j^2 &\overset{\text{ind}}{\sim} \mathcal{N}(\mu_0 + \phi_j Y_{t-1,j}, \sigma_j^2) \\
\sigma_j^2 &\overset{\text{iid}}{\sim} InvGamma(a, b) \\
\mu_0 &\sim \mathcal{N}(0, \sigma_{\mu_0}^2) \\
\phi_j &\overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma_{\phi_j}^2)
\end{aligned}
$$

Traceplots are shown in Figure 19 and two chains were used, starting from different initial points with burn-in=$10^4$ and thin=10. Again, traceplots are very nice since there is no trend, the support is explored enough and convergence for each parameters is achieved.

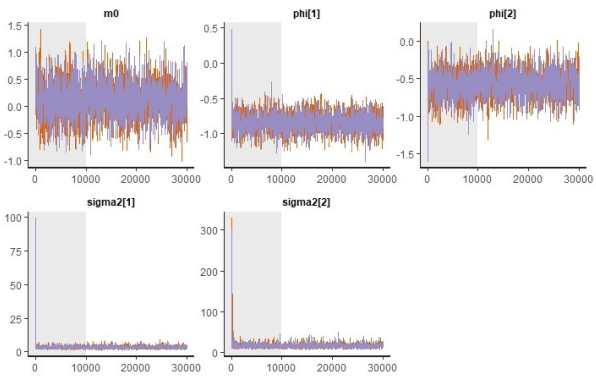


**Figure 19:** *Traceplots*

For what concerns the autocorrelation (Figure 20) we can assume independence between the states of the chain as before.

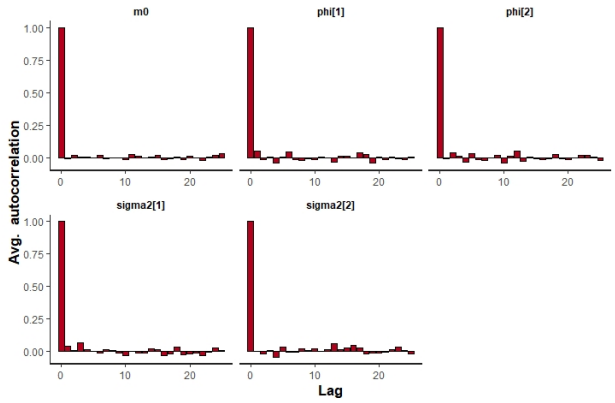


**Figure 20:** *Autocorrelation*

Unfortunately, the posterior inference has again the 95% CI problem with respect to $\phi_j$ parameters. In Figure 21 there are the posterior distribution for each parameter.

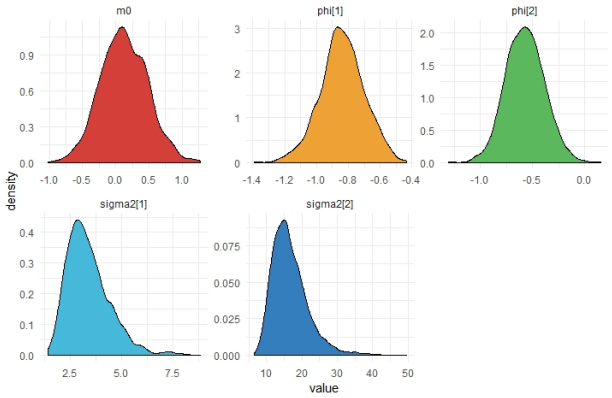


**Figure 21:** *Posterior distributions*

To conclude with this model we used the posterior means as estimators for the parameters just to solve the problem on the temporal dependence parameter $\phi_j$. In Figure 22 and 23 there are the predicted values respectively for high-school and post high-school data, while see Figures 45 and 46 for the prediction on the raw data using the inverse transformation in the Appendix.

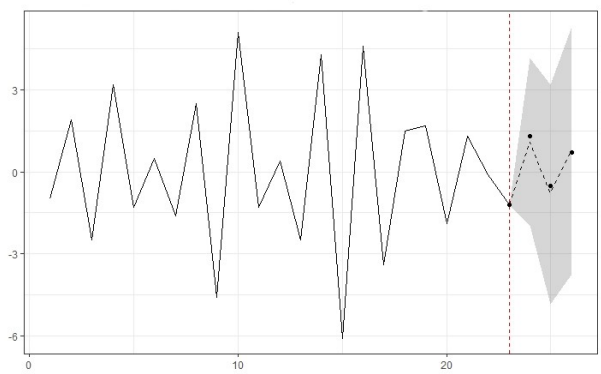

**Figure 22:** *Forecasting on transformed High-school group data*

| **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** $\mu_{0j}$, $\phi_j$ **rand. eff.** | | |
| Highschool | 91.59 | 1.99 |
| Post-highschool | 136.19 | 5.68 |
| **ARIMA(1,3,0),** $\mu_{0j}$ **rand. eff., common** $\phi$ | | |
| Highschool | 91.20 | 2.78 |
| Post-highschool | 135.94 | 2.121 |
| **ARIMA(1,3,0),** $\phi_j$ **rand. eff., common** $\mu_0$ | | |
| Highschool | 91.25 | 1.824 |
| Post-highschool | 134.13 | 1.529 |

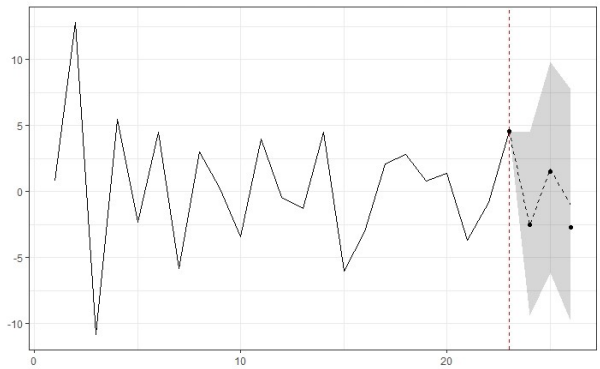

**Figure 23:** *Forecasting on transformed Post-Highschool group data*

## 2.5 Accuracy on other States

What about the other states? We would like to find out just how accurate would this model be if we tried to forecast the other states' time series, starting from the first observation.

We collected in a table the MSE for each state in the US for both age groups. What we would like to highlight are of course the states with the smallest MSE. In this case those are the states of Ohio for the Post-High-school group and Virginia for the High-school group.

As it can be seen from the plots, the real time series of the state after transformation is reported in blue, whereas in red we have the simulated time series obtained by the model trained on Missouri.

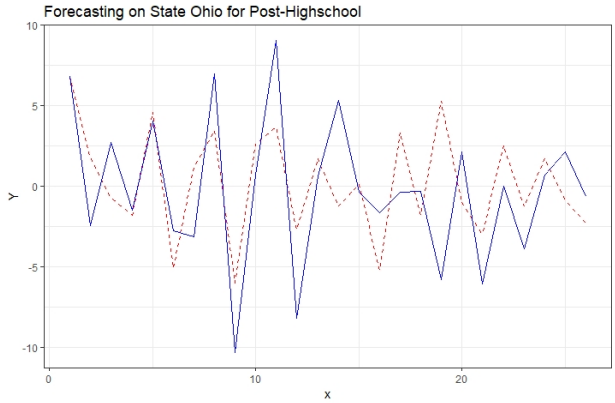


**Figure 24:** *Forecasting on Ohio State*

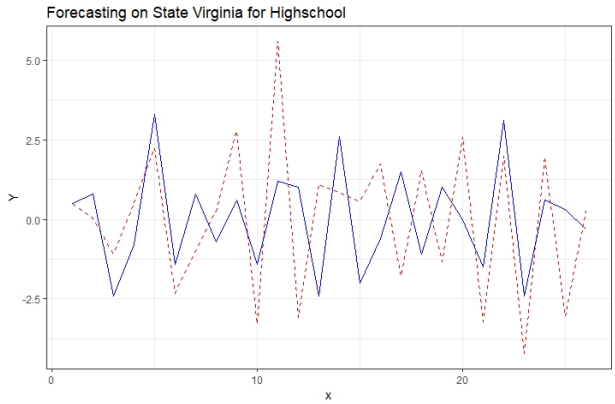


**Figure 25:** *Forecasting on Virginia State*

Both of these states are quite close to Missouri. But is this a coincidence? Does it mean that if a state is closer to Missouri than its MSE will be small? No. And we have a counter example.

Nebraska, in spite of being a first order neighbour of Missouri, has a very high MSE. Moreover, the simulated time series in red has high peaks and opposite slope at the beginning and at the end, which is quite ugly for a prediction estimation, it's not correct.

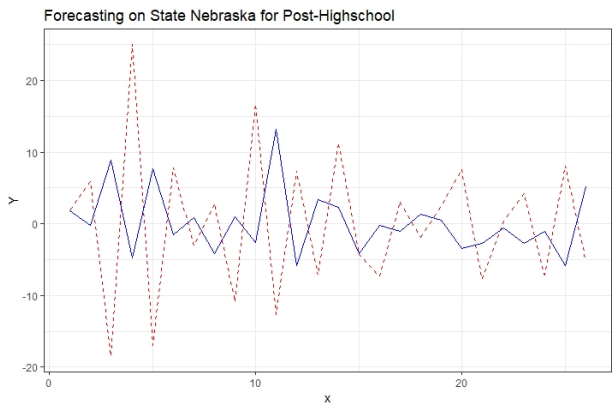


**Figure 26:** *Forecasting on Nebraska State*

Another interesting example happens with the state of Oregon, which regardless of being far away from

Missouri, has decent MSE performances. In fact, the simulated time series is quite similar to the real data.

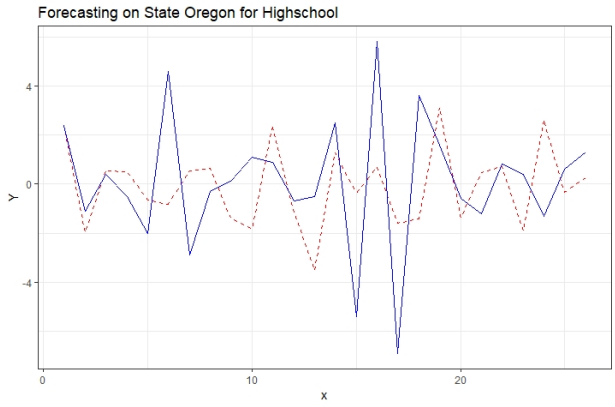


**Figure 27:** *Forecasting on Oregon State*

Why does this happen? What are the reasons behind such different behaviours of the model?

We list 3 different explanations:

1. the model is trained only on Missouri's data.
2. Moreover, this version of the model does not take into account any spatial parameters, so the information regarding the closeness of a State to Missouri is useless.
3. Last but not least we had to transform the data due to the stationarity assumption, and this is a crucial step.

The correlation between the States before the transformation is extremely high as shown in Figure 28 where the smallest value is as hgh as 84 percent!

The situation changes after data stationarisation. As it can be seen in (Figure 29) it's a completely different scenario. The correlation not only drops to lower values, but also becomes negative for some cases!

And this is the key to understand the weird performances of our model. If the test set (which in this case are the other states' time series) is similar to the training set (Missouri data) then it is reasonable to expect high accuracy, whereas if the test set is a mix of similar and dissimilar data, then, there is nothing to do. We were lucky that the model performed very good in a few states, and this was definitely not related to the geographical relationship between a state and Missouri.

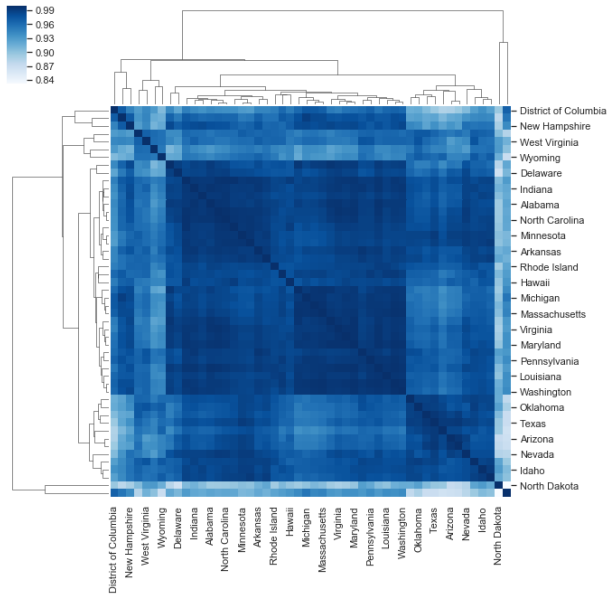


**Figure 28:** *Correlation matrix on High-school group using raw data*

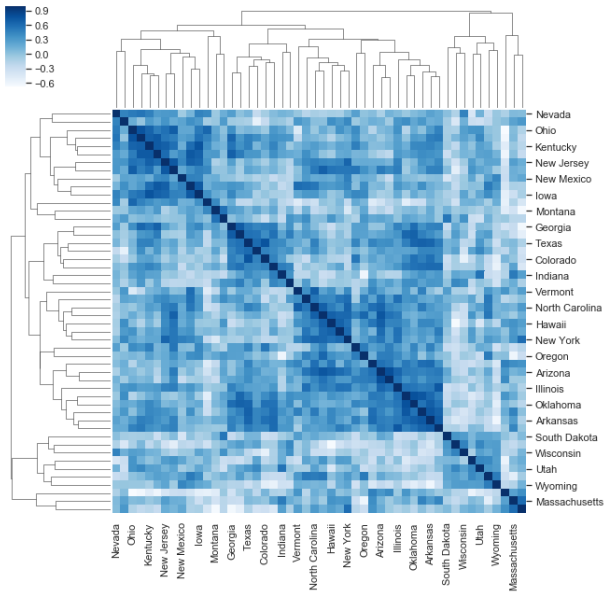


**Figure 29:** *Correlation matrix on High-school group using transformed data*

# 3 Spatio-Temporal model

In the previous section we have presented the ARIMA(1,3,0) and we could see that it does a good job in predicting future values in a given state. However, if the goal of the study conducted on Birth rates was not the mere prediction of future values in a given state, but rather the comparison between different states in the USA - a different model should be developed.

In this framework, we decide to extend our study to all states of the USA. We saw earlier that there is strong correlation in time. But this is not the only

correlation present in our data.

By visualizing the Birth rate for a fixed year on a heatmap of the USA, it is clear to see that the States which are closer share similarities in their Birth rates. There is Spatial correlation too - the areal units that are close to each other share similar behaviours.

By a simple inspection of the plot of Birth rate values in the USA in Figure 30, it is natural to want to consider a Spatio-Temporal model.

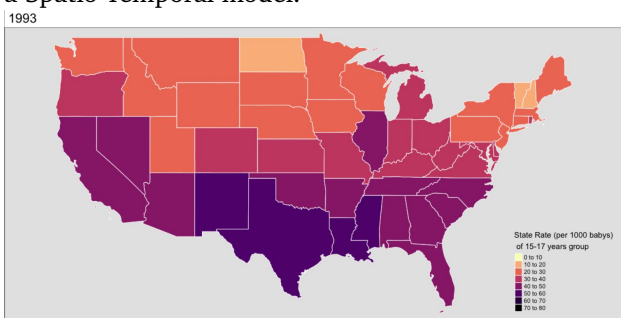


**Figure 30:** *Heatmap of Birth rate values across the USA for the Highschool age group 15-17 years old. Close areal units share similar behaviour*

This section will be developed as follows :

1. Choice of the matrix of spatial weights, order of neighbours and study of Moran's I
2. Presentation of the ST.CARar model proposed in the CARBayesST library
3. Posterior inference of our model
4. Prediction of Birth rate values in an unobserved state, given the evolution of Birth rate values in its neighbours
5. Forecasting with the ST.CARar model and comparison with ARIMA(1,3,0).

Notice: in this section we will only study the 48 contiguous US states, since Hawaii and Alaska do not have neighbours. We will add District of Columbia as another areal unit, for a total of **K=49 areal units** over **N=29 observations in time**.

## 3.1 Moran's I

We use Global Moran's I to measure whether there is significant spatial correlation among our 49 areal units. The Global Moran's I is defined as ,

$$I = \frac{K}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

where $K$ is the number of areal units indexed by $i$ and $j$; $x$ is the variable of interest (in our case the Birth rate); $\bar{x}$ is the mean of $x$; $w_{ij}$ is a matrix of spatial weights with zeroes on the diagonal and $W$ is the sum of all $w_{ij}$.

The matrix used for the spatial weights is defined as:

$$w_{ij} = 1 \text{ if areal units } i \text{ and } j \text{ are neighbours;}$$
$$w_{ij} = 0 \text{ otherwise.}$$

Considering that an areal unit is not a neighbour of itself.

Moran's I for every year is shown in Figure 31. We see that high spatial correlation is observed for each year considered in our study, which proves that there is significant spatial correlation (when using the spatial weights matrix W). Thus, considering spatial models is reasonable.

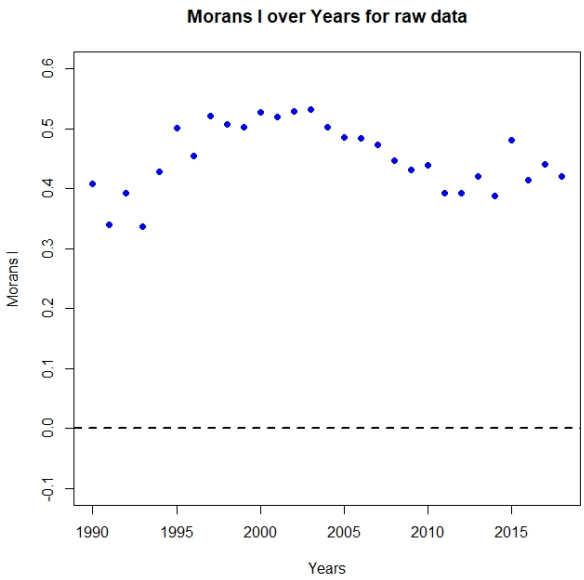


**Figure 31:** *Moran's I obtained for every year 1990-2018, we observe high spatial correlation among the countries for every year*

Notice that this should be done on the raw observations, and not on stationarized data, since spatial correlation is lost by doing the third order differences, as seen in the Figure 32.

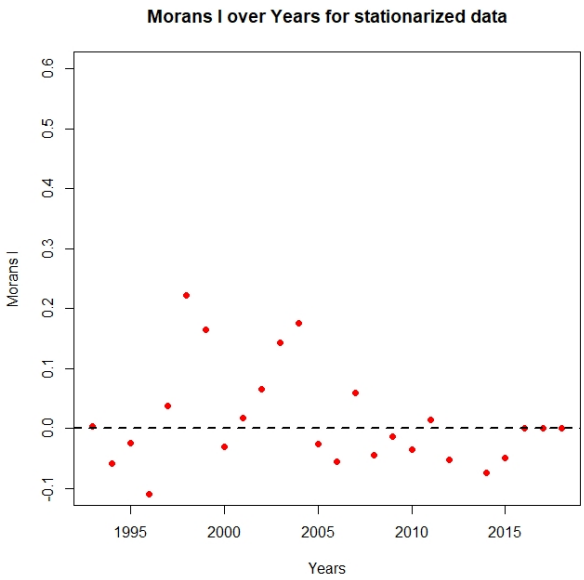

**Figure 32:** *Moran's I obtained for every year 1990-2018 for third order differences. We notice that this type of transformation makes us lose the spatial dependency*

Thus for the Spatio-Temporal models we will not be stationarizing our data as we did in the ARIMA(1,3,0) model.

## 3.2 ST.CARar model with gaussian likelihood

For the Spatio-Temporal modelling, we use the ST.CARar model from the CARBayesST package.

The ST.CARar model represents the spatio-temporal structure with a multivariate first order autoregressive process and a spatially autocorrelated precision matrix.

In this model $\phi_t$ is the 49x1 dimensional vector of random effects for time period $t$, which evolves over time via a multivariate first order autoregressive process with temporal autoregressive parameter $\rho_T$. Thus temporal autocorrelation is induced via the mean $\rho_T\ \phi_{t-1}$, while spatial autocorrelation is induced by the variance $\tau^2\mathbf{Q}(\mathbf{W},\rho_S)^{-1}$.

Flat and conjugate priors are specified for $(\rho_S, \rho_T)$ and $\tau^2$. Finally, missing (NA) values are allowed in the response data $Y$ for this model.

$$
\begin{gathered}
Y_{kt}|\mu_{kt}, \nu^2 \overset{\text{ind}}{\sim} \mathcal{N}(\mu_{kt}, \nu^2) \\
\mu_{kt} = \beta_0 + \phi_{kt} \\
\beta_0 \sim \mathcal{N}(\mu_\beta, \sigma_\beta) \\
\nu^2 \sim Inv - Gamma(a, b) \\
\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}, \rho_S, \rho_T, \tau^2 \sim \mathcal{N}_{49}(\rho_T\boldsymbol{\phi}_{t-1}, \tau^2\mathbf{Q}(\mathbf{W}, \rho_S)^{-1}) \\
\boldsymbol{\phi}_1 \sim \mathcal{N}_{49}(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W}, \rho_S)^{-1}) \\
\tau^2 \sim Inv - Gamma(a, b) \\
\rho_S, \rho_T \overset{\text{iid}}{\sim} Uniform(0, 1)
\end{gathered}
$$

The corresponding precision matrix $\mathbf{Q}(\mathbf{W}, \rho_S)$ is defined as

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

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

The precision matrix $\mathbf{Q}(\mathbf{W}, \rho_S)$ can be seen as a linear convex combination of two parts, $[diag(\mathbf{W1}) - \mathbf{W}]$ which can be considered the part that accounts for high spatial correlation, and $\mathbf{I}$ which would show no spatial correlations.

Thus, a high value (close to 1) for $\rho_S$ is an indicator of significant spatial correlation.

## 3.3 Posterior inference

We fit the ST.CARar model specified in the previous section by producing a total of 220 000 samples, of which 20 000 used for the burn in period, and thin = 100, getting finally 2000 samples of each parameter.

The posterior distributions are shown on Figures 33 - 36.

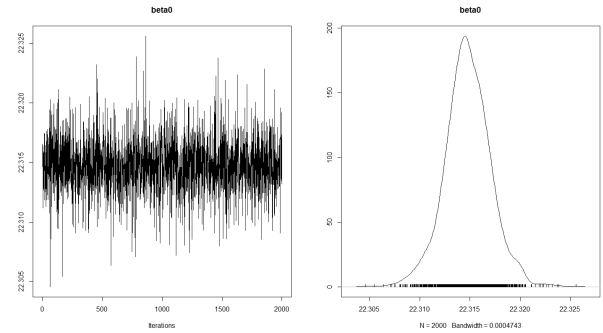


**Figure 33:** *Posterior distribution of $\beta_0$*

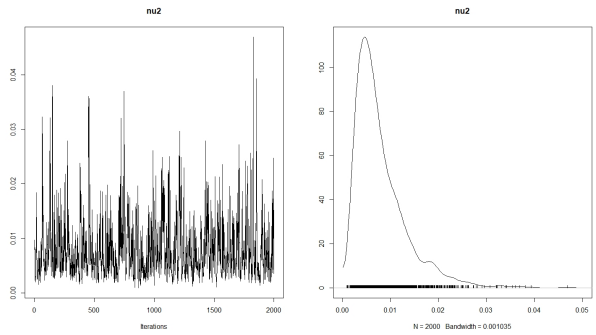


**Figure 34:** *Posterior distribution of $\nu^2$ the MCMC chains are highly correlated*

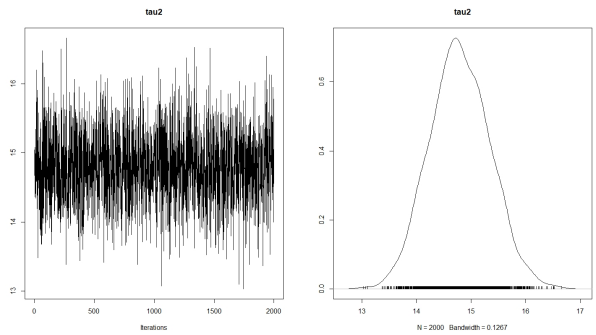


**Figure 35:** *Posterior distribution of $\tau^2$, nice MCMC chains*

| | Median | 2.5% | 97.5% | n.effect. | Geweke.diag |
|---|---|---|---|---|---|
| $\beta_0$ | 22.3147 | 22.3100 | 22.3196 | 2000.0 | 0.9 |
| $\tau^2$ | 14.7783 | 13.7520 | 15.8926 | 2000.0 | 0.2 |
| $\nu^2$ | 0.0064 | 0.0020 | 0.0225 | 346.5 | 0.2 |
| $\rho_S$ | 0.9556 | 0.9307 | 0.9734 | 2000.0 | -0.4 |
| $\rho_T$ | 0.9426 | 0.9247 | 0.9602 | 2000.0 | -0.8 |

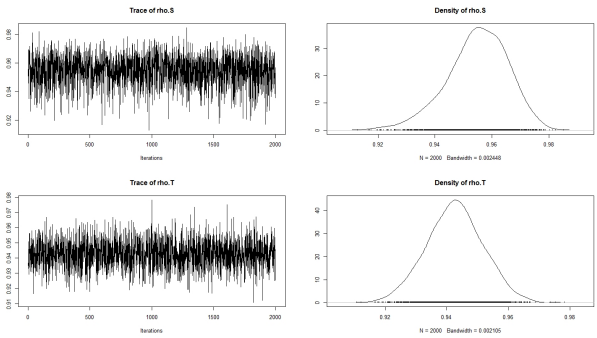


**Figure 36:** *Posterior distribution of $\rho_S$ and $\rho_T$, nice MCMC chains. Both $\rho_S$ and $\rho_T$ tend to have values close to 1, which implies strong spatial and temporal correlation*

The posterior credible intervals for $\rho_S$ and $\rho_T$ show that there is significant spatial and temporal correlation. The effective number of independent samples is equal to the number of samples for all parameters, except for $nu^2$, which could be seen already in the traceplots for $nu^2$ that show correlation.

Despite the samples of $\nu^2$not being ideal, we continue the study of this model as we consider that $nu^2$ is not a crucial parameter and we are curious to see further results.

## 3.4 Predicting birth rates in a State given its neighbours

An interesting use of the ST.CARar model proposed in Section 4.2 could be predicting the Birth rates in an unobserved State, when only the observation of Birth rates of its neighbours is available.

In order to study this scenario, we suppose now that the Birth rate of state Missouri are completely unknown (NA) for the whole time period (1990-2018).

We use the ST.CARar model to discover these values. Then, we compare the posteriors obtained for the whole period 1990-2018 for the state of Missouri, and its real (observed) values and plot it in Figure 37.

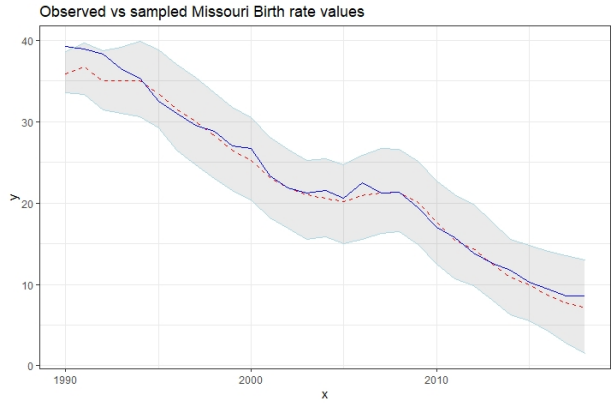


**Figure 37:** *Posterior mean and 95% credible interval for Birth rates in Missouri in years 1990-2018 in red. In blue, observed data for Missouri. The observed data falls into the credible interval obtained without knowing Missouri data*

We find that Missouri's Birth rate data belongs to its posterior credible interval. This goes to say that the estimated values are reasonable, but Missouri's observed Birth rate values are slightly lower than what we would expect them to be, given its neighbours.

When conducting the same study on Texas, again we find that the observed values belong to the 95% credible interval, as seen in Figure 38. Nonetheless, the observed values this time are higher than expected. Again, the estimated values are reasonable, but Texas's observed Birth rate values are higher than what we would expect them to be, given its neighbours.

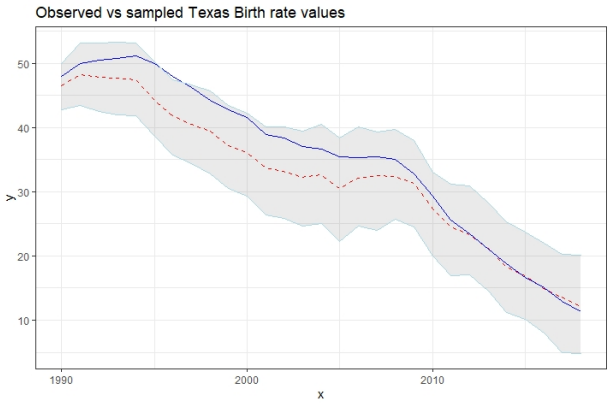


**Figure 38:** *Posterior mean and 95% credible interval for Birth rates in Texas in years 1990-2018 in red. In blue, observed data for Texas. The observed data falls into the credible interval obtained without knowing Texas data*

We have observed previously that southern states have higher Birth rates than northern states, and given that Texas has only neighbours on its north border, it only confines with states with lower Birth rates than its own. Hence, having observed values higher than the posterior mean obtained with neighbour data seems

reasonable.

A similar study is repeated for Idaho, a state with mostly southern neighbours, and the opposite behaviour is found. On Figure 39 Idaho's Observed Birth rates are slightly lower than what would have been expected by only observing its neighbours.

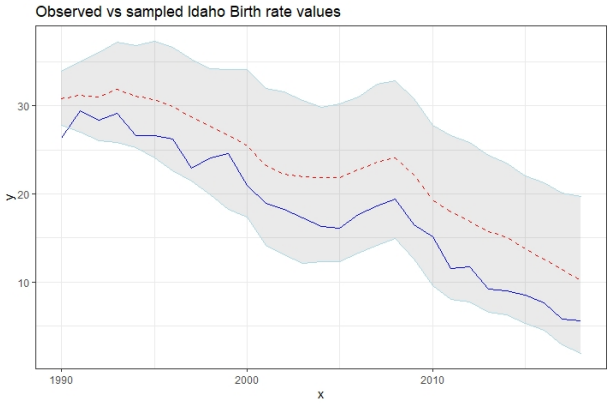


**Figure 39:** *Posterior mean and 95% credible interval for Birth rates in Idaho in years 1990-2018 in red. In blue, observed data for Idaho. The observed data falls into the credible interval obtained without knowing Idaho data*

## 3.5 ST.CARar model with poisson likelihood

One possible problem that the previous model could have is to predict a negative value for the birth rates, and this a big error because our data are an average counting, so they must remain positive. In order to solve this we provide two models with the same structure as below:

$$
\begin{gathered}
Y_{kt}|\mu_{kt} \overset{\text{ind}}{\sim} Poisson(\mu_{kt}) \\
ln(\mu_{kt}) = \beta_0 + \phi_{kt} \\
\beta_0 \sim \mathcal{N}(\mu_\beta, \sigma_\beta) \\
\nu^2 \sim InvGamma(a, b) \\
\boldsymbol{\phi}_t|\boldsymbol{\phi}_{t-1}, \rho_S, \rho_T, \tau^2 \sim \mathcal{N}_{49}(\rho_T \boldsymbol{\phi}_{t-1}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1}) \\
\boldsymbol{\phi}_1 \sim \mathcal{N}_{49}(\mathbf{0}, \tau^2 \mathbf{Q}(\mathbf{W}, \rho_S)^{-1}) \\
\tau^2 \sim InvGamma(a, b) \\
\rho_S, \rho_T \overset{\text{iid}}{\sim} Uniform(0, 1)
\end{gathered}
$$

The idea behind the first model is trying to model only the integer part of the time series, letting the fractional part coming from an uniform distribution with support in [0,1]. The second model simply multiply by 10 each observation in order to obtain integer value for each time series.

We managed to obtain some results but they were not good, on one side we use a different dataset from the gaussian model, so in principle a comparison is impossible. Moreover the posterior inference for these model was very bad.

| **Model** $M_j$ | **Gaussian** | **Poisson v.1** | **Poisson v.2** |
|---|---|---|---|
| DIC | $-1974.679$ | $7368.810$ | $11488.150$ |
| LPLM | $412.939$ | $-3598.783$ | $-5645.510$ |

## 3.6 Comparisons

The ARIMA(1,3,0) model studied in section 2 has shown good forecasting properties. A natural question is to see whether with the Spatio-Temporal ST.CARar model we have further increased our forecasting accuracy.

In Figure 40 we observe that the forecasting obtained with the ST.CARar model performs very poorly. The ARIMA(1,3,0) model is to be preferred if the goal of the study is Forecasting in a single state.

The possible reason behind this poor performances is the fact that the data used in ST.CARar model is not stationarised, thus predictions in time are not behaving as we would expect.

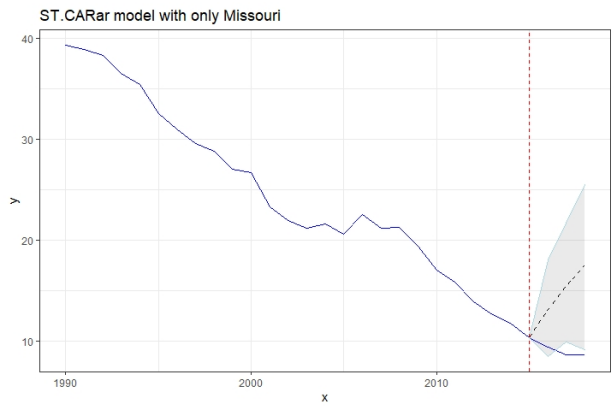


**Figure 40:** *Posterior mean and 95% credible interval for Birth rates forecasting in Missouri in years 2016-2018 (dotted black line). In blue, observed data for Missouri*

The ARIMA(1,3,0) model should be used for forecasting in a single state, but the ST.CARar model has other advantages.

As far as it goes between the comparisons among Poisson or Gaussian prior for the ST.CARar model, the two models cannot really be compared since they work on different data (Poisson needs integer numbers).

# 4 Conclusions

With this work we were able to apply some of the techniques offered by the Bayesian statistical approach to a real case study.

We were able to successfully fit two models, the ARIMA(1,3,0) and ST.CARar model.

We have implemented the ARIMA(1,3,0) in STAN language. Our model takes into account only one state, and introduces random effects on the two age groups. It can be used to forecast future Birth rates in a state for the upcoming years, for both age groups. Before fitting the model, data is stationnarized by means of third order differences.

As a second model, the ST.CARar model from CARBayesST was used, in order to work with areal data. This model works with raw (not-stationnarized) data, and takes into account only one age group.

We can imagine to use this model in order to assess whether the data observed in one state seems reasonable with respect to what would be expected regarding its neighbouring states, to obtain Birth rates for those states that hide their Birth rate data, or maybe even detect fraudulent data.

This being said, no model is better than the other. The two models are complementary and give answers to different questions.

# 5 Bibliography

[1] 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.

[2] 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.

# A Appendix

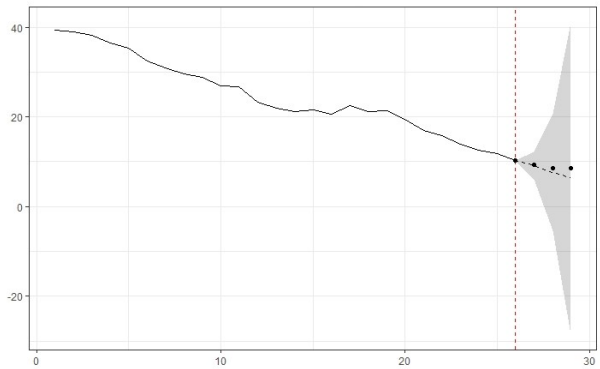


**Figure 41:** *Forecasting on real Highschool group data*

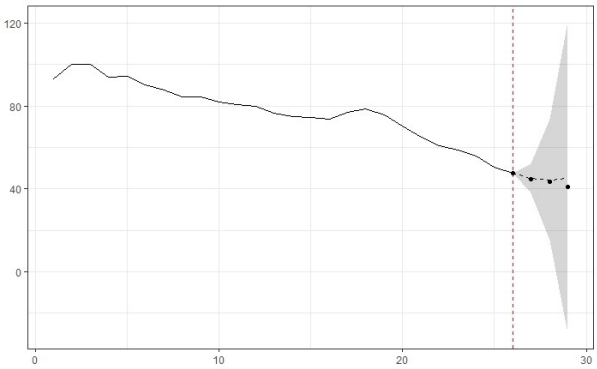


**Figure 42:** *Forecasting on real Post-Highschool group data*

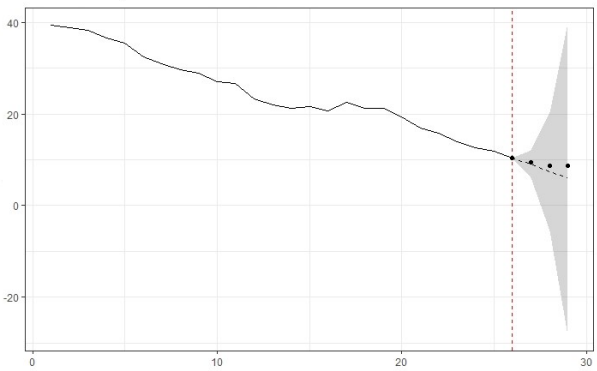


**Figure 43:** *Forecasting on real Highschool group data*

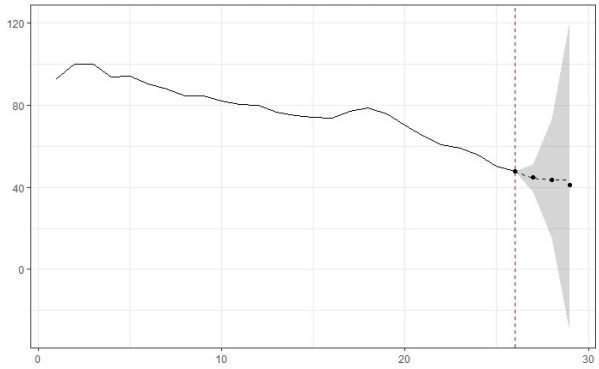


**Figure 44:** *Forecasting on real Post-Highschool group data*

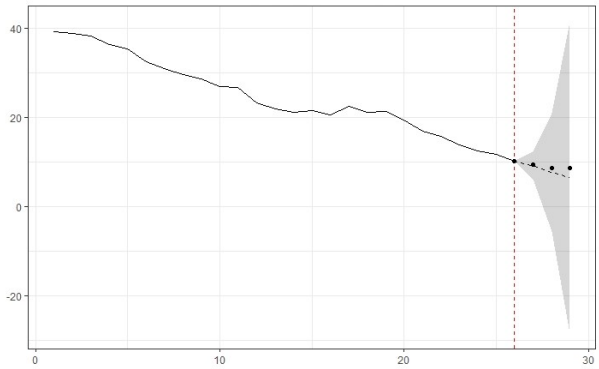


**Figure 45:** *Forecasting on real Highschool group data*

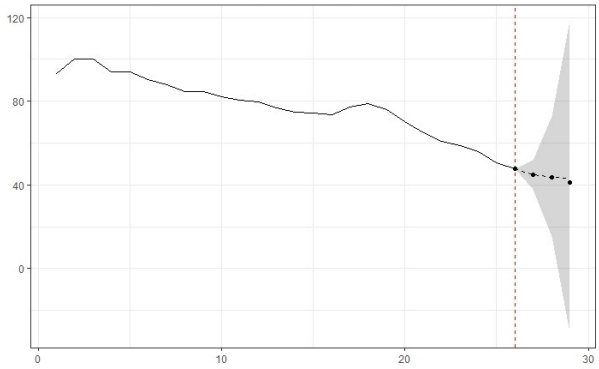


**Figure 46:** *Forecasting on real Post-Highschool group data*