A Bayesian approach to understanding the Homicide Rate in the City of Rio de Janeiro by administrative regions through their Social Progress Index indicators

Eduardo Adame Salles
School of Applied Mathematics
Getulio Vargas Foundation
eduardo.salles@fgv.br

Abstract

This study aims to investigate the relationship between the homicide rate in the city of Rio de Janeiro and the indicators of the Social Progress Index. Our approach involves employing Bayesian methodology to estimate the parameters of three multilevel models and subsequently comparing their performance. The Social Progress Index serves as a measure of the overall quality of life and social well-being of the population, and it has been regularly published by the Pereira Passos Institute for the City of Rio de Janeiro biennially since 2016. Given the well-known issue of violence in Rio de Janeiro, the homicide rate serves as a pertinent indicator of this problem. The city is divided into 33 administrative regions, and we utilize the corresponding data throughout this research.

1 Introduction

The City of Rio de Janeiro, as well as its whole state, has been facing difficulties in fulfilling their duties towards the population, especially when it comes to healthcare, basic sanitation, education, and public security. Understanding the situation of a certain region in a specific period of time is not an easy task. However, the use of indicators can be quite useful when seeking to simplify the interpretation of the presence or absence of certain policies.

The Social Progress Index (SPI) is calculated based on indicators of basic human needs, the foundations of well-being, and the opportunities in a certain region. Since 2016, the Pereira Passos Institute - through their platform [1] - has been publishing both the indicators used in its calculations and the SPI itself for the administrative regions (ARs) of the city of Rio de Janeiro.

When observing the indicators of an AR, the Homicide Rate stands out, as it directly affects all social and economic activities in a region. Therefore, it would be beneficial to understand how the Homicide Rate is influenced by other available indicators, such as Literacy Rate, Access to Sanitation, Child Labor, Family Vulnerability, or Waste Selective Collection, Cultural Access Index, and Urban Mobility - which can be improved with direct interventions from the municipal government.

The main goal of this analysis is to understand how these many indicators interact with the Homicide Rate employing Statistical Modelling technics towards the data.

1.1 Administrative Regions

As briefly discussed, the City of Rio de Janeiro is divided into 33 groups of neighborhoods - called Administrative Regions [2]. Each one is contained in one of the 8 Zones, which are administrated by sub-mayors named by the current mayor during his term.

The land sizes and populations are really different among the ARs, which turns comparisons harder. For instance, the Region of Jacarepaguá has nearly 650 thousand inhabitants while Santa Teresa's has

Preprint. Under review.

only about 40 thousand. Jacarepaguá is also comprised of 10 neighborhoods, comparing to only one in Santa Teresa. Figure 1 gives a good idea of the AR's area distributions.

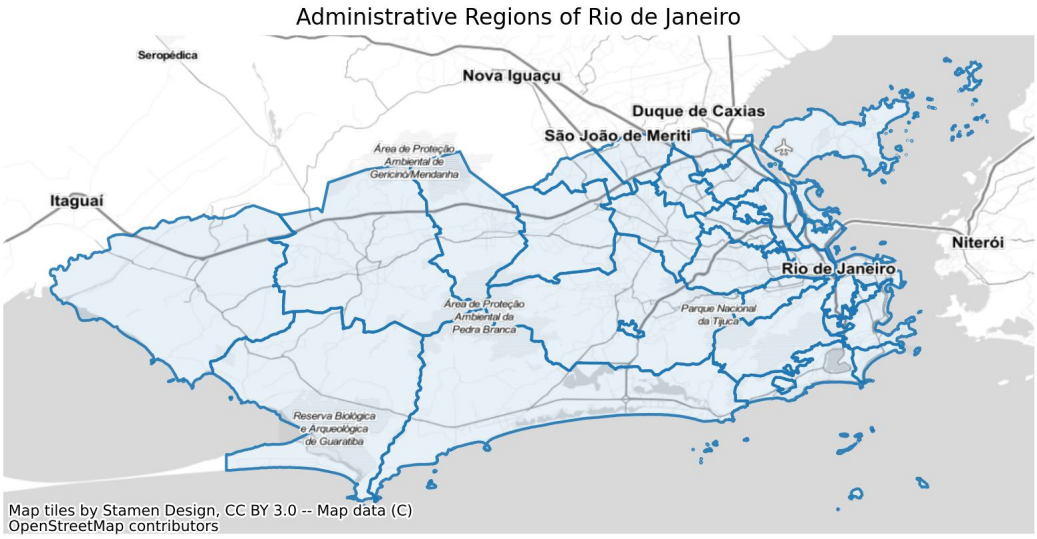


Figure 1: A simple map of the Administrative Regions of the City of Rio de Janeiro

As mentioned, it is possible to aggregate the data by Zone (South, North, etc.) or leave it separated by neighborhood. However, the SPI data is only available aggregated by ARs or by neighborhoods. This study will be based on the ARs data; later, in subsection 4.2, we will discuss the consequences of this choice.

1.2 Social Progress Index

The Pereira Passos Institute (Instituto Pereira Passos, in Portuguese) is the official institute of statistics and cartography of the City of Rio de Janeiro. It is responsible for the production of data and information about the city, as well as the development of studies that support the planning and management of public policies.

The Social Progress Index (SPI) is a composite index of 36 social and environmental indicators that capture three dimensions of social progress: Basic Human Needs, Foundations of Wellbeing, and Opportunity. The SPI is a comprehensive measure of real quality of life for all citizens that can be used to complement traditional measures of economic progress. Its full methodology can be found in [3].

The SPI is calculated for the administrative regions of the city of Rio de Janeiro since 2016 every 2 years. While they have published the data for 2022, it is not currently available for public download. Therefore, this study will be based on the data between 2016 and 2020 [4]. We have a lot of indicators available, and the choice of which ones to use will be discussed in subsection 3.1.

1.3 Homicide Rate

We refer to an Homicide Rate (HR) as the number of homicides per 100 thousand inhabitants in a specific region. It is calculated by the Rio de Janeiro State Institute of Public Security (ISP) and used by the Pereira Passos Institute in their SPI calculations. The SPI database does not provide any information about the population nor the number of homicides in each AR, though. Therefore, there will be some choices on how to model the HR in this study, which will be discussed in the next section.

2 Methodology

We are looking forward to modelling the Homicide Rate of the Administrative Regions of the City of Rio de Janeiro. In our case, there is no need to calculate the HR, as it is already available in the SPI database. However, it is important to understand what it expresses, as it will be used in the modelling process.

2.1 Models

In a first moment, a counting model - such as Poisson or Negative Binomial - would be a good choice to model the HR. However, the SPI database does not provide any information about the population nor the number of homicides in each AR. Therefore, we will have to use the HR as a continuous variable, which means that we cannot use counting models.

Of course, there is still many possibilities to model the HR - specially in terms of Generalized Linear Models (GLMs) [5, 6], which are usually a good choice when we have a response variable that is not normally distributed and we want to model it with linear predictors, which is our case.

For our purposes, we can define a GLM as follows:

Definition 1 (Generalized Linear Model (GLM))**.** Given an independent sample of the response variable $\boldsymbol{Y} \in \mathbb{R}^N$ and a set of p covariates $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p \in \mathbb{R}^N$, a GLM is a model which assumes that, for all $i \in \{1, \ldots, N\}$, Y_i follows a distribution F in the exponential family, with mean μ_i and variance σ^2, and that μ_i is related to the covariates $X_{i1}, X_{i2}, \ldots, X_{ip}$ through a function g of a linear predictor η_i.

$$\begin{aligned} Y_i \mid \boldsymbol{X}_i &\sim F(\mu_i, \sigma^2), \\ \eta_i &= X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots + X_{ip}\beta_p, \\ \mu_i &= g^{-1}(\eta_i). \end{aligned} \tag{2.1}$$

We call g the link function.

If we had both the population and the number of homicides, we could use a Poisson or Negative Binomial model, and the rate could be accommodated as an offset. However, as mentioned, we do not have this information. Therefore, we will have to model the HR directly.

In this case, employing a Log-Normal regression model would be a good choice. In fact, we will use some of its variations, and compare the results. In all cases, we will use a logarithmic function as link function. This is robust to outliers and non-additive effects. Another approach would be to employ a Gamma regression model [7], but we do not consider it here because it makes the model more complex and it is not necessary for our purposes.

So, our first model is defined as follows:

Definition 2 (Model M1: Log-Normal regression)**.** Given an independent sample $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$ and covariates $\boldsymbol{X} \in \mathbb{R}^{N \times T \times P}$, we define the model as follows:

$$\begin{aligned} Y_{it} \mid \boldsymbol{X}_{it} &\sim \mathcal{LN}(\mu_{it}, \sigma^2), \\ \eta_{it} &= X_{it1}\beta_1 + X_{it2}\beta_2 + \cdots + X_{itP}\beta_P, \\ \mu_{it} &= \exp(\eta_{it}). \end{aligned} \tag{2.2}$$

This is only a starting point. We can try to build more appropriate models. For instance, nextly we discuss the evolution of the HR in AR i and year t, represented as Y_{it}.

The approach for this problem is to consider that the HR in a certain year has its own baseline. Considering that the HR in a certain AR and year is Y_{it}, we can define the baseline as $\beta_0^{(t)}$. This kind of approach is often referred to as *random intercept* [8].

Then, we can model the HR in a certain year as a function of the baseline and a set of covariates. In this case, we can define the HR in a year t as follows:

$$\mathbb{E}[Y_{it} \mid X_{it}] = \mu_{it} = \exp\Big(\eta_{it} + \beta_0^{(t)}\Big), \tag{2.3}$$

where $\beta_0^{(t)} \sim \mathcal{N}(0, s^2)$ and η_{it} is a linear predictor.

Therefore, our next model is the following:

Definition 3 (Model M2: Time-Multilevel Log-Normal regression)**.** Given an independent sample $\boldsymbol{Y} \in \mathbb{R}^{N \times T}$ and covariates $\boldsymbol{X} \in \mathbb{R}^{N \times T \times P}$, we define the model as follows:

$$
\begin{aligned}
Y_{it} \mid \boldsymbol{X}_{it} &\sim \mathcal{LN}(\mu_{it}, \sigma^2), \\
\eta_{it} &= X_{it1}\beta_1 + X_{it2}\beta_2 + \cdots + X_{itP}\beta_P, \\
\mu_{it} &= \exp\Big(\eta_{it} + \beta_0^{(t)}\Big), \\
\beta_0^{(t)} &\sim \mathcal{N}(0, s^2).
\end{aligned}
\tag{2.4}
$$

In an analogue way, we can define a model that considers the HR in a certain AR as a function of the baseline and a set of covariates.

Definition 4 (Model M3: Spatial-Multilevel Log-Normal regression)**.** Given an independent sample $\boldsymbol{Y} \in \mathbb{R}^{N\times T}$ and covariates $\boldsymbol{X} \in \mathbb{R}^{N\times T\times P}$, we have:

$$
\begin{aligned}
Y_{it} \mid \boldsymbol{X}_{it} &\sim \mathcal{LN}(\mu_{it}, \sigma^2), \\
\eta_{it} &= X_{it1}\beta_1 + X_{it2}\beta_2 + \cdots + X_{itP}\beta_P, \\
\mu_{it} &= \exp\Big(\eta_{it} + \beta_0^{(i)}\Big), \\
\beta_0^{(i)} &\sim \mathcal{N}(0, s^2).
\end{aligned}
\tag{2.5}
$$

It is relevant to mention that the models M2 and M3 do not enlarge the number of parameters, since we are modelling with respect to just one more parameter than the model M1 (s).

2.2 Fitting

Although it is possible to fit the models using a frequentist approach, we will use a Bayesian approach. In this case, we will use the `Stan` software [9] through the `rstanarm` package [10] for `R` [11].

In this case, we can use the posterior distribution of the parameters to make inference about the HR in a certain AR and year, and also to make inference about the parameters themselves.

In Mathematical terms, we can define the posterior distribution of the parameters as follows:

$$
\xi(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{X}) \propto f_N(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\xi(\boldsymbol{\theta}),
\tag{2.6}
$$

where $\xi(\boldsymbol{\theta})$ is the prior distribution of the parameters and $f_N(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})$ is the likelihood function.

Calculating this posterior distribution is not a trivial task. In fact, it is often impossible to calculate it analytically. Therefore, we need to use numerical methods to approximate it - and this is one of the main reasons why we use `Stan`, which employs the Hamiltonian Monte Carlo (HMC) algorithm [12] and the no-U-turn sampler (NUTS) [13].

2.2.1 Priors

Fitting a model in a Bayesian framework requires the definition of a prior distribution for the parameters. In this case, we will use a independent weakly informative priors as follows:

$$
\begin{aligned}
\beta_j &\sim \mathcal{N}(0, 1), \\
\beta_0^{(t)} &\sim \mathcal{N}(0, 1), \\
\beta_0^{(i)} &\sim \mathcal{N}(0, 1), \\
\sigma &\sim \exp(1)
\end{aligned}
\tag{2.7}
$$

We choose those priors to avoid imposing any heuristic information on the model. For models M2 and M3, s will be implicitly distributed by a covariance distribution. We will use the default on `rstanarm`, which is a *decov* (it is specified in the priors reference of [10]) with all the parameters set to 1.

2.3 Assessment

In order to assess the models, we will use the following metrics:

- **WAIC**: The WAIC is a measure of the out-of-sample deviance of the model. The lower the WAIC, the better the model fits the data. Specially used to compare models.
- **RMSE**: The RMSE is a measure of the error of the model. The lower the RMSE, the better the model fits the data.
- **Residuals**: The residuals are a measure of the error of the model. The lower the residuals, the better the model fits the data. It is important to look at their distribution as well.

To assess the parameters, we will use the following metrics:

- **Rhat**: The Rhat is a measure of the convergence of the chains. The closer the Rhat is to 1, the better the convergence of the chains.
- **ESS**: The ESS is a measure of the effective sample size of the chains. The higher the ESS, the better the convergence of the chains.
- **MCSE**: The MCSE is a measure of the Monte Carlo Standard Error of the chains. The lower the MCSE, the better the convergence of the chains.
- **Credibility Intervals**: The credibility intervals are a measure of the uncertainty of the parameters. The wider the credibility intervals, the higher the uncertainty of the parameters.

3 Results

3.1 Covariate selection

As discussed in subsection 1.2, the SPI is a composite of 36 indicators which is a lot of information to work with - considering we have only 33 ARs. Therefore, we performed an Exploratory Data Analysis (EDA) to understand the data and choose which covariates to use in the models.

The data is available in the Excel format, which is not the best for data analysis. Therefore, we used `Python`'s module `pandas` [14] to clean this data.

The first selection was made based in the variance of each covariate. If a covariate has low variance, it means that it does not change much among the ARs, which makes it less interesting to be used in the models. Therefore, we considered the 22 covariates with higher variance when normalized by the maximum value. Which keeps us with all variances greater than 0.13. This was arbitrary, but it is a good starting point.

Following, all covariates were standardized to have mean 0 and standard deviation 1.

The next step was to look further into the correlation between the covariates. And this highly correlated block was found:

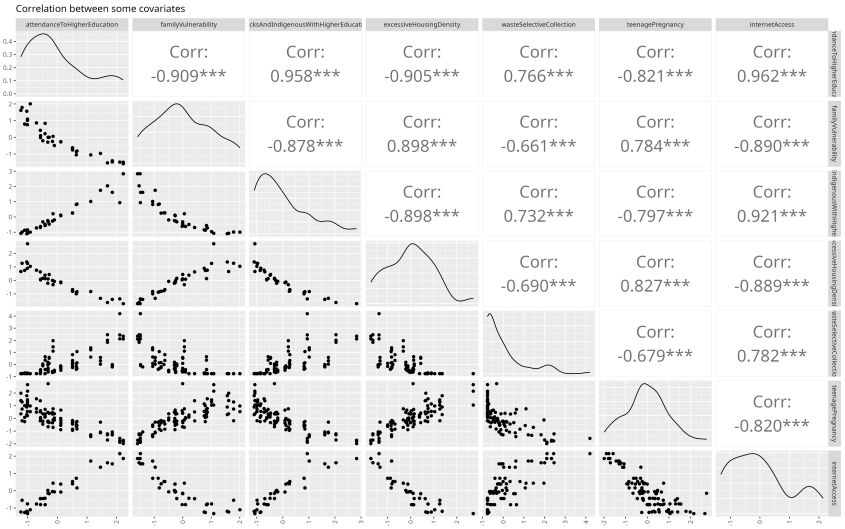


Figure 2: Correlation between some of the covariates

As can be seen, Attendance to Higher Education is the most correlated with all other covariates in this block, and they are all highly correlated with each other. Therefore, we will only consider Attendance to Higher Education out of the covariates in this group. This keeps us left with 16 covariates.

Keeping this idea, we will also remove those covariates which are still correlated (> 0.4) with Attendance to Higher Education, once it is now representing that whole block. This leaves us with 9 covariates. We can see that the covariates are not linearly correlated with one another in Figure 3.

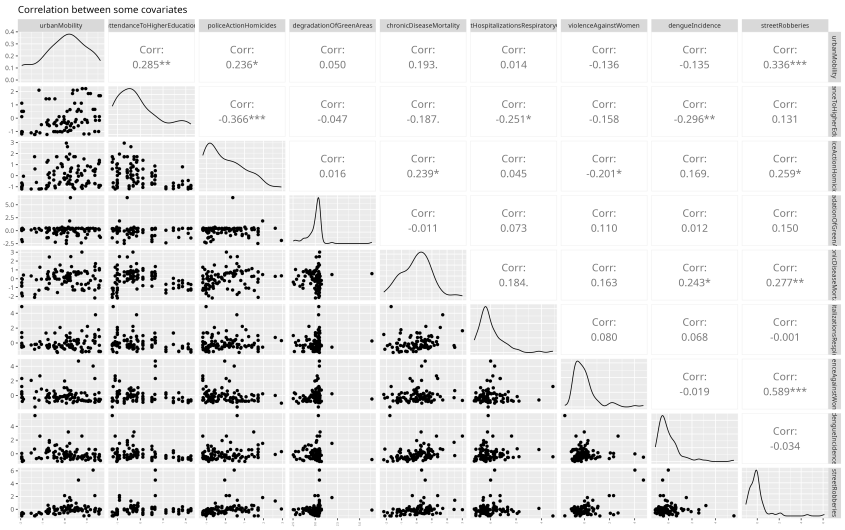
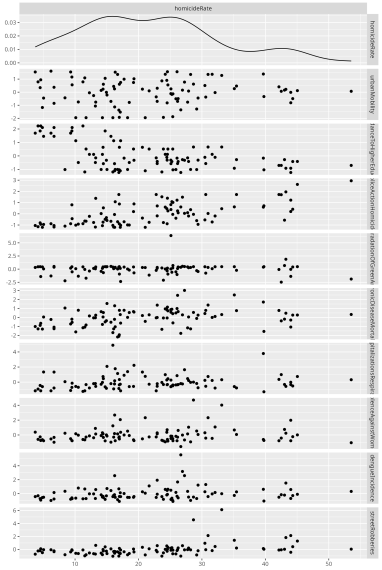


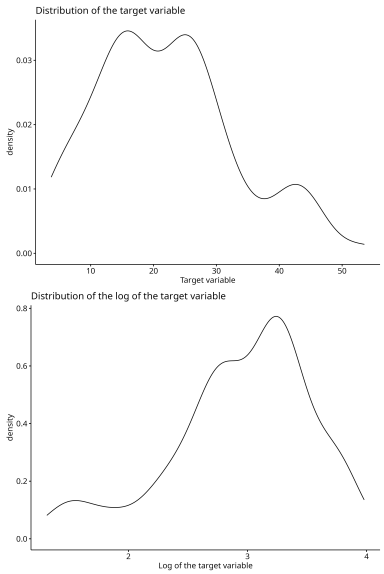
Figure 3: Correlation between the chosen covariates

3.2 Descriptive analysis

Here we will look into the data to understand the distribution of the covariates and the homicide rate, as well as the homicide rate itself. Below, we can check the distribution of the covariates and the homicide rate.



(a) Correlation between the covariates and the homicide rate



(b) Density of the homicide rate and its logarithm

Figure 4: Homicide rate descriptive analysis

Which shows that the homicide rate can be considered to be log-normally distributed. And that the covariates are not linearly correlated with the homicide rate.

It is also a good idea to have some tabular information about the homicide rate. Therefore, we will look into the mean, standard deviation, minimum, maximum, and quantiles of the homicide rate.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd.
Homicide Rate	3.72	14.39	21.37	21.97	27.79	53.54	11.09
log(Homicide Rate)	1.31	2.66	3.06	2.93	3.32	3.98	0.6

Viewing the quantiles, we can reinforce the idea that the homicide rate is log-normally distributed.

3.3 Fitting the models

Despite having some problems to make the models converge, we were able to fit the models. It was possible to get through the problems by increasing the number of iterations and the number of chains - as recommended by the documentation. All three models were fitted using 8 chains with 8000 iterations each.

Below, we can see the results of the fittings.

	mean	mcse	sd	10%	50%	90%	n_eff	Rhat
(Intercept)	2.99	0.00	0.04	2.94	2.99	3.05	24309.00	1.00
urbanMobility	-0.06	0.00	0.05	-0.12	-0.06	-0.00	23607.00	1.00
attendanceToHigherEducation	-0.28	0.00	0.06	-0.35	-0.28	-0.21	20213.00	1.00
policeActionHomicides	0.14	0.00	0.05	0.09	0.14	0.20	16116.00	1.00
degradationOfGreenAreas	-0.04	0.00	0.03	-0.09	-0.04	0.00	29604.00	1.00
chronicDiseaseMortality	0.05	0.00	0.03	0.01	0.05	0.09	29512.00	1.00
infantHospitalizationsRespiratoryCrisis	0.02	0.00	0.04	-0.02	0.02	0.07	27595.00	1.00
violenceAgainstWomen	-0.11	0.00	0.07	-0.20	-0.11	-0.02	13783.00	1.00
dengueIncidence	-0.02	0.00	0.03	-0.07	-0.02	0.02	27423.00	1.00
streetRobberies	0.22	0.00	0.06	0.15	0.22	0.30	13223.00	1.00
sigma	7.40	0.00	0.57	6.69	7.36	8.15	26009.00	1.00
mean_PPD	21.68	0.01	1.07	20.31	21.68	23.04	30691.00	1.00
log-posterior	-342.81	0.02	2.53	-346.21	-342.45	-339.89	12580.00	1.00

Table 1: Model M1 summary

For model M1, we can see that some of the predictors have small credibility intervals (that do not cross 0). The Rhat values are all 1, which implies that the chains converged. The effective sample size is also high, which suggests that the chains are not autocorrelated. The monte carlo standard error is also small, which shows that the chains are stable. The predictive posterior mean is close to the mean of the data, which expresses that the model is not biased.

	mean	mcse	sd	10%	50%	90%	n_eff	Rhat
(Intercept)	2.09	0.03	1.03	0.52	2.49	3.05	1435.00	1.01
urbanMobility	-0.02	0.00	0.05	-0.08	-0.01	0.05	12267.00	1.00
attendanceToHigherEducation	-0.30	0.00	0.06	-0.37	-0.29	-0.22	11209.00	1.00
policeActionHomicides	0.15	0.00	0.04	0.09	0.15	0.20	10352.00	1.00
degradationOfGreenAreas	-0.06	0.00	0.03	-0.10	-0.06	-0.02	14225.00	1.00
chronicDiseaseMortality	0.10	0.00	0.04	0.06	0.10	0.15	14467.00	1.00
infantHospitalizationsRespiratoryCrisis	-0.01	0.00	0.04	-0.06	-0.01	0.03	15727.00	1.00
violenceAgainstWomen	-0.13	0.00	0.08	-0.23	-0.13	-0.04	7204.00	1.00
dengueIncidence	-0.04	0.00	0.04	-0.09	-0.04	0.00	13402.00	1.00
streetRobberies	0.22	0.00	0.07	0.13	0.22	0.30	6977.00	1.00
b[(Intercept) year:2016]	1.05	0.03	1.04	0.06	0.66	2.63	1439.00	1.01
b[(Intercept) year:2018]	0.94	0.03	1.03	-0.02	0.54	2.51	1427.00	1.01
b[(Intercept) year:2020]	0.72	0.03	1.02	-0.23	0.31	2.29	1428.00	1.01
sigma	6.96	0.00	0.53	6.30	6.93	7.66	15061.00	1.00
Sigma[year:(Intercept),(Intercept)]	4.98	0.35	12.85	0.02	0.70	13.36	1323.00	1.01
mean_PPD	21.74	0.01	1.01	20.46	21.75	23.02	29877.00	1.00
log-posterior	-343.14	0.07	4.09	-348.41	-342.98	-337.86	3088.00	1.00

Table 2: Model M2 summary

Now, on the other hand, for model M2, we can see that the intercept has a high standard deviation, which means that it is not very credible. And the s estimation is really poor, with high mcse and sd. For the other predictors, and mainly for the varying intercepts, we have the same situation as M1.

	mean	mcse	sd	10%	50%	90%	n_eff	Rhat
(Intercept)	2.98	0.00	0.05	2.91	2.98	3.04	15638.00	1.00
urbanMobility	-0.05	0.00	0.06	-0.12	-0.05	0.02	17493.00	1.00
attendanceToHigherEducation	-0.33	0.00	0.08	-0.43	-0.33	-0.24	8196.00	1.00
policeActionHomicides	0.08	0.00	0.06	0.01	0.08	0.16	6491.00	1.00
degradationOfGreenAreas	-0.01	0.00	0.04	-0.06	-0.01	0.03	12148.00	1.00
chronicDiseaseMortality	0.03	0.00	0.04	-0.03	0.03	0.08	10994.00	1.00
infantHospitalizationsRespiratoryCrisis	0.03	0.00	0.04	-0.02	0.03	0.08	23560.00	1.00
violenceAgainstWomen	-0.13	0.00	0.08	-0.23	-0.13	-0.03	12224.00	1.00
dengueIncidence	-0.04	0.00	0.04	-0.09	-0.04	0.01	14637.00	1.00
streetRobberies	0.24	0.00	0.07	0.16	0.24	0.33	12827.00	1.00
sigma	6.62	0.01	0.66	5.81	6.57	7.50	6769.00	1.00
Sigma[administrativeRegion:(Intercept),(Intercept)]	0.04	0.00	0.03	0.00	0.03	0.08	4412.00	1.00
mean_PPD	21.72	0.01	0.96	20.50	21.73	22.94	34183.00	1.00
log-posterior	-382.45	0.14	8.28	-393.37	-382.22	-371.93	3521.00	1.00

Table 3: Model M3 summary

Now, for Model M3, in order to simplify the interpretation of the results, the varying intercepts were removed from the table. Now, s has a better estimation, with lower mcse and sd. The intercept is also more credible, with a lower sd. The other predictors have the same situation as M2. We can look at the distribution of the varying intercepts in the Figure 5.

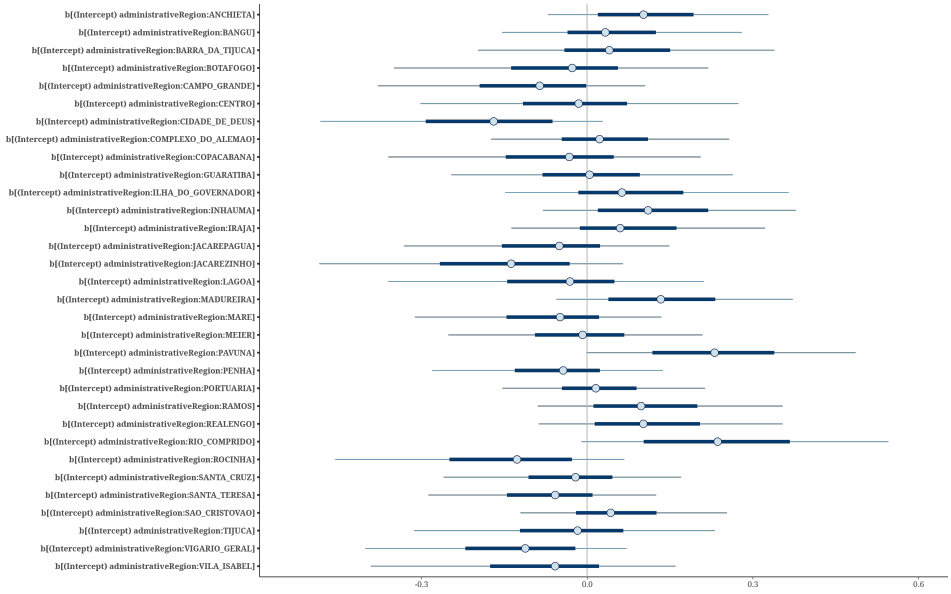


Figure 5: Varying intercepts distribution for model M3

It is possible to see that we have some regions with good credibility, but most cross the zero line, which means that they are not very credible. Although we can see employing a model with varying intercepts accommodates the data better

3.4 Comparing the models

First, it is a good idea to look at their RMSE and WAIC. We have the following results:

Model	RMSE	waic	p_waic
M1	6.93	668.0	10.4
M2	6.47	658.6	12.2
M3	5.86	663.9	24.2

Table 4: RMSE for the models

For both waic and p_waic, we got lower standard errors. But it is important to notice that a high p_waic means that the waic is less reliable. Another approach would be to use a LOO cross-validation or a k-fold cross-validation. But, for the sake of simplicity, we will use the waic and p_waic. In the expressed metrics, all models are very similar, but M3 could be preferred because it has the lowest RMSE by some considerable margin.

Now we can take a look at the residuals for each model. We have the following results:

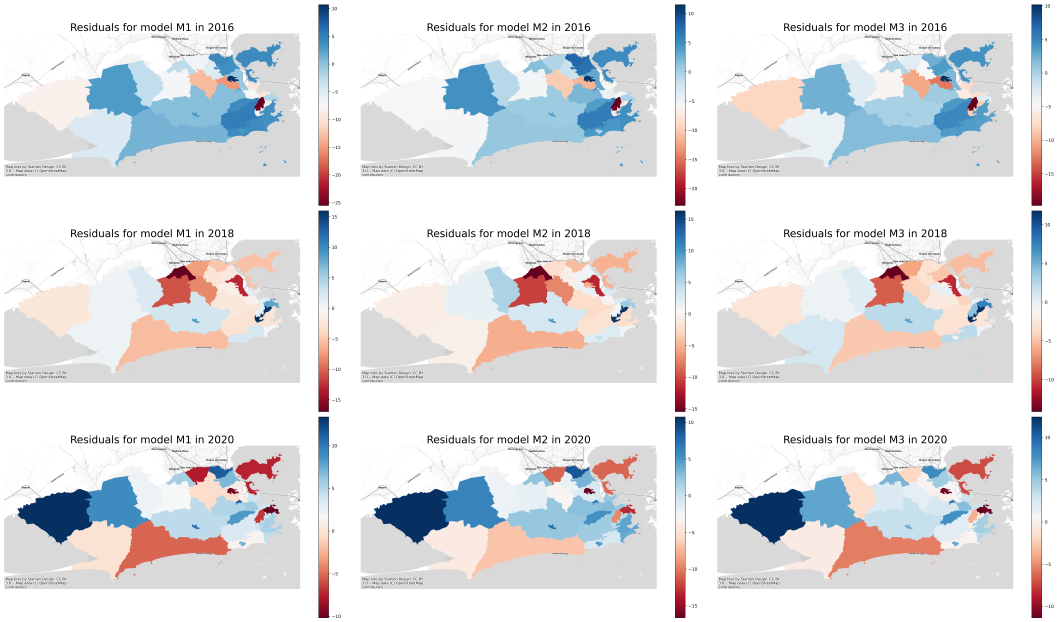


Figure 6: Residuals for each model and year

Each line is a year and each column is a model. Although there is a lot of information in this plot, it is possible to see that the residuals look alike between models in a same year. But, overall, M2 has the lowest residuals. This can conclude our choice for the best model among the three.

4 Conclusion

4.1 Summary

In this study, modeling the homicide rate (HR) directly provided a valuable opportunity to gain practical knowledge and implement the concepts of log-normal regression. While the model effectively captured the non-additive effects of the covariates, its flexibility was somewhat limited.

Although the findings are not definitively conclusive, they align with existing literature. The model successfully captured the impact of the covariates, and utilizing a multilevel model was a suitable choice for accommodating the hierarchical structure of the data.

The fitting process posed some challenges, particularly in dealing with numerical and convergence issues. This experience emphasized the significance of weakly-informed priors and a comprehensive understanding of the model prior to fitting.

Handling geographical data also presented its own difficulties, given the high correlation among the covariates of the different regions. Undoubtedly, there is ample room for improvement in this aspect of the research.

4.2 Limitations

The first of all is the lack of data. The data used in this study is limited to three years. This is a very short period of time to make any conclusions. The model is not able to capture the year effect, which is a very important factor in the HR.

We are also limited by not having the number of homicides nor the population of each region. If we had, our approach would not be necessary. We could have used the number of homicides as the response variable and the population as the offset. Which are likely to be a much better approach. Obviously, it would have been possible to get this data from other source. But, the idea of this study was to use the data that is used on SPI calculation.

Modelling grouped data usually gets to the called ecological fallacy [15]. This is a very important limitation of this study. The model is not able to capture the individual effect of each covariate. It is only able to capture the effect of the covariates in the group level.

By choosing the log-normal regression, we are assuming that the HR is log-normally distributed. This is a very strong assumption. Even though it seems to be a good one, we could have used a more flexible model that would accommodate other shapes.

4.3 Future work

Getting more data is not an easy option. But, there are many other ways to improve the model. Mainly, in terms of assessments and priors.

Employing a LOO cross-validation would be a good way to assess the model. I had tried some approaches, but I was not able to make it work - the same happened with the k-fold cross-validation. Putting some effort on this could be a good following step.

Understanding the data better would also be a good way to improve the model. During our analysis, there were some variables that were not used. And the year was not used as a covariate or during the plots.

The priors used in this study were very weakly-informed. This was a good choice, given the lack of knowledge about the data. But, it would be possible to use more informative priors.

The model could also be more flexible. We could have used a more flexible model, like the Gaussian process regression. This would allow us to capture the year effect and the individual effect of each covariate.

But, after improving the modelling itself, it would be a great idea to add some political events in the analysis. This would allow us to understand the impact of the political events in the HR. Like elections and the pandemic.

Finally, we could use the model to make some counterfactual analysis. This would allow us to understand the impact of the covariates in the HR.

4.4 Acknowledgements

I would like to thank the Botafogo de Futebol e Regatas soccer team for the support during this work. It has been a pleasure to watch the games and see the team winning.

In the same way, I would like to thank the professor Luiz Max de Carvalho and the teaching assistant Isaque Pim - the staff of the Statistical Modelling course - for helping me to dive deeper into the world of statistics.

Finally, I would like to thank my family and friends for the support during this journey. Namely, my friends Caio Lins and Tulio Koneçny.

As funny as it might seem, there is no ChatGPT involved in this work. It failed at the easiest tasks. Sorry, OpenAI, statistics is not that easy.

4.5 Code

The code used in this study is available at `https://github.com/adamesalles/homicide-rate-rj`.

References

[1] Instituto Pereira Passos. Data Rio, . URL https://www.data.rio/. [Accessed 22-Jun-2023].

[2] Prefeitura do Rio de Janeiro. Regiões Administrativas do Rio de Janeiro. URL https://www.rio.rj.gov.br/web/cvl/ra. [Accessed 23-Jun-2023].

[3] Instituto Pereira Passos. Metodologia SPI, . URL https://ips-rio-pcrj.hub.arcgis.com/pages/metodologia2. [Accessed 26-Jun-2023].

[4] Instituto Pereira Passos. Base de dados do Índice de Progresso Social - IPS por Regiões Administrativas (RA) - Município do Rio de Janeiro - 2016/2018/2020 — ips-rio-pcrj.hub.arcgis.com, . URL https://ips-rio-pcrj.hub.arcgis.com/documents/918dd39478594792a9cfa7080b84c0b5/about. [Accessed 26-Jun-2023].

[5] A. Gelman, J. Hill, and A. Vehtari. *Regression and Other Stories*. Analytical Methods for Social Research. Cambridge University Press, 2020. ISBN 9781107023987.

[6] P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1989. ISBN 9780412317606.

[7] Brian L. Wiens. When log-normal and gamma models give different results: A case study. *The American Statistician*, 53(2):89–93, 1999. ISSN 00031305. URL http://www.jstor.org/stable/2685723.

[8] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2007. ISBN 9780521686891.

[9] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[10] Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2023. URL https://mc-stan.org/rstanarm/. R package version 2.21.4.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

[12] Michael Betancourt. A conceptual introduction to hamiltonian monte carlo, 2018.

[13] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014. URL http://jmlr.org/papers/v15/hoffman14a.html.

[14] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

[15] Steven Piantadosi, David P Byar, and Sylvan B Green. The ecological fallacy. *American journal of epidemiology*, 127(5):893–904, 1988.