Remi LACOSTE
Guillaume RIGAUD
MS DS

23/01/2020

Bayesian Statistics Project

Subject : Prediction of fever and respiratory symptoms in Jamaica

Introduction

The database provides the number of cases of fever and respiratory symptoms in Jamaica recorded per week and year, classified by patient age category. We wish to use the Bayesian framework to predict the number of cases. To do this, we divided the data into a learning sample (train set) and a test sample (test set). Weekly data is used and seems to show some seasonal trends.

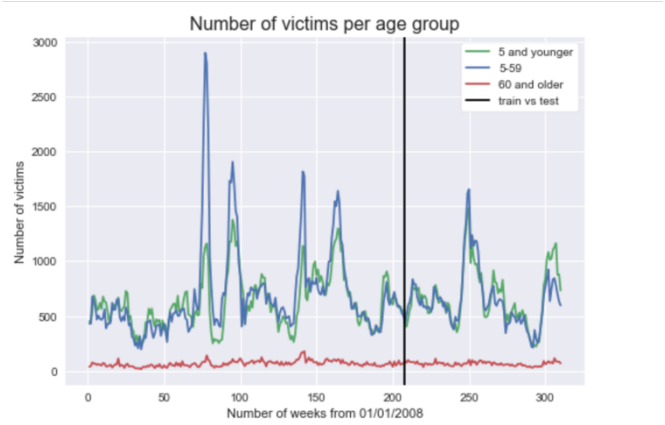

Figure 1 - Graphique du nombre de victimes en fonction du temps

Our data has a temporal aspect: the Bayesian framework appears naturally to take advantage of knowledge acquired over a learning period and then make predictions over a test period. Indeed, the Bayes formula allows to update the prior law of the parameters thanks to the information contained in the data (see appendix A). The estimation of the posterior law will allow us to add a notion of uncertainty on the parameters of the model.

A traditional approach would be the use of SARIMA (Seasonal Auto Regressive Integrated Moving Average) models. In this project, we propose an alternative method divided into two steps:
1. The estimation of a counting process to predict the number of victims. The sum of the achievements of this process gives the total number of cases over a given period of time (one year or more).
2. Assuming the known number of victims, we tried to estimate the distribution of cases of fever and respiratory symptoms per week: this means 'breaking down' the annual number of victims into a weekly number of victims.

We combined the models obtained in 1 and 2 to predict the number of victims over the entire test sample. Finally, we compared the results of our approach to the results of a SARIMA model. Our approach could be adapted to other issues such as inventory management within a company: once its annual sales are estimated, the company could look at the distribution of its sales over the year to optimize its production and transport processes.

1 Counting process

1.1 Poisson Gamma model

We observed that the empirical variance of the counting process associated with our data was greater than its empirical expectation. Based on a frequentist approach, the maximum likelihood estimate of a negative binomial distribution to model the number of weekly victims would therefore be appropriate. From a Bayesian point of view, we can use a Poisson distribution sampling model whose parameter follows a Gamma prior: indeed, a compound Poisson Gamma distribution is nothing else than a negative binomial distribution. Let y be the number of weekly victims:

$y \mid \lambda \sim Poisson(\lambda)$

$\lambda \sim Gamma\ (a, b)$.

With Bayes' formula we get :

$\pi(\lambda|y) \propto f(y|\lambda)\ \pi(\lambda)$

$\pi(\lambda|y) \propto \prod_{i=1}^{n} \frac{\lambda^{y_i}\ e^{-\lambda}}{y_i!}\ \frac{1}{\Gamma(a)b^a}\ \lambda^{a-1}e^{-\frac{\lambda}{b}}$

$\pi(\lambda|y) \propto \lambda^{\sum y_i + a - 1} e^{-\left(n+\frac{1}{b}\right)\lambda}$

$\lambda|y \sim Gamma\ (\sum_{i=1}^{n} y_i + a, \left(n + \frac{1}{b}\right)^{-1})$

Therefore $a_{posterior} = \sum_{i=1}^{n} y_i + a$ and $b_{posterior} = (n + \frac{1}{b})^{-1}$.

Numerical Application :

From the train sample, the hyper-parameters a and b on the age-split data are defined as follows. For each age group i = 1,2,3 (1 is Young, 2 is Middle, 3 is Old), the empirical mean $m_{emp\ i}$ and the empirical variance $\sigma_{emp\ i}$ are computed. Then, we set $a_i = \frac{{m_{emp\ i}}^2}{\sigma_{emp\ i}}$ et $b_i = \frac{\sigma_{emp\ i}}{m_{emp\ i}}$. Thus, we set the hyper-parameters as if the observations were directly the expectation of our counting process, i.e. λ.

We obtain the following table for the value of the parameters, with 261 observations :

Age Group	m_{emp}	σ_{emp}	a	b	$a_{posterior}$	$b_{posterior}$	$\hat{\lambda}$
Young	0.62	0.0594	6.61	0.095	136	0.0046	626.8
Middle	0.69	0.1746	2.74	0.252	146	0.0047	692.6
Old	0.06	0.0007	6.81	0.009	20	0.0032	67.8

Table 1-Parameters for conjugated Poisson Gamma model

We also compared theoretical results with results obtained computationally with the Metropolis Hastings algorithm. The conclusions are available in the appendix.

1.2 Relevance of a model by age groups

Graphically (Figure 1) it is clear that the 'Old' class differs from the two others. But is it legitimate to use a separate model for 'Young' and 'Middle'? Let's compare a general model containing 'Young' and 'Middle' with the two models obtained separately.

Let's note M_0 the 'Middle and Young separated' model and M_1 the 'Middle et Young combined' model. To compare these two models we used the Bayes factor, denoted $B_{01}(y)$. We took a non-informative prior : $\pi(M_0) = \pi(M_1) = \frac{1}{2}$. The Bayes factor expression is:

$B_{01}(y) = \frac{\pi(M_0|y)}{\pi(M_1|y)} / \frac{\pi(M_0)}{\pi(M_1)} = \frac{m_0(y)}{m_1(y)}$

$m_i(y) = \int_{\lambda_i} f(y|\ \lambda_i)\pi(\lambda_i)d\lambda_i$ the marginal density of y under the model i, also called *evidence*.

Since our Poisson Gamma model is conjugated, it is possible to compute $B_{01}(y)$ directly without requiring a marginal likelihood approximation method such as Laplace's method. We obtain (details of the calculation in appendix C) :

$$\frac{m_0(y)}{m_1(y)} = \frac{\Gamma(a_Y + \sum_{k=1}^{n} y_{Y,k})\left(n + \frac{1}{b_Y}\right)^{-(a_Y + \sum_{k=1}^{n} y_{Y,k})} \Gamma(a_M + \sum_{k=1}^{n} y_{M,k})(n + \frac{1}{b_M})^{-(a_M + \sum_{k=1}^{n} y_{M,k})}\Gamma(a_M){b_M}^{a_M}}{\Gamma(a_Y){b_Y}^{a_Y}\Gamma(a_M){b_M}^{a_M}\Gamma(a + \sum_{k=1}^{2n} y_k)\left(2n + \frac{1}{b}\right)^{-(a + \sum_{k=1}^{n} y_k)}}$$

For computational reasons, we couldn't compute the Bayes factor: the terms in Euler's gamma functions are not integers, so we cannot simplify them, and their value explodes towards infinity. We therefore used Laplace's approximation of the Bayes factor:

$2 \log B_{01}(y) \cong 2 \log \frac{f(y\ |\hat{\lambda}_0)}{f(y\ |\hat{\lambda}_1)} - \big(\pi(M_0) - \pi(M_1)\big) \log n$, on a supposé $\pi(M_0) = \pi(M_1)$, d'où :

$B_{01}(y) \cong \frac{f(y\ |\hat{\lambda}_0)}{f(y\ |\hat{\lambda}_1)}$, which is none other than the likelihood ratio test.

Indeed, $f(y\ |\hat{\lambda}_i)$ is the likelihood of the observations y under the model i, when $\hat{\lambda}_i$ is the maximum likelihood under the hypothesis of the model i.
We get $B_{01}(y) \cong 1.4$, which is in favour of model 0 in which a distinction is made between the 'Young' and 'Middle' classes (It is 1.4 times more likely to be in model 0 than in model 1).

1.3 Predicted annual number of victims

The model fitted on the train sample generates a weekly number of victims. To obtain an annual number, we sum up 52 simulations of the Poisson Gamma process. By repeating the procedure 1000 times, we obtain the following distribution for the total number of victims over the test period :

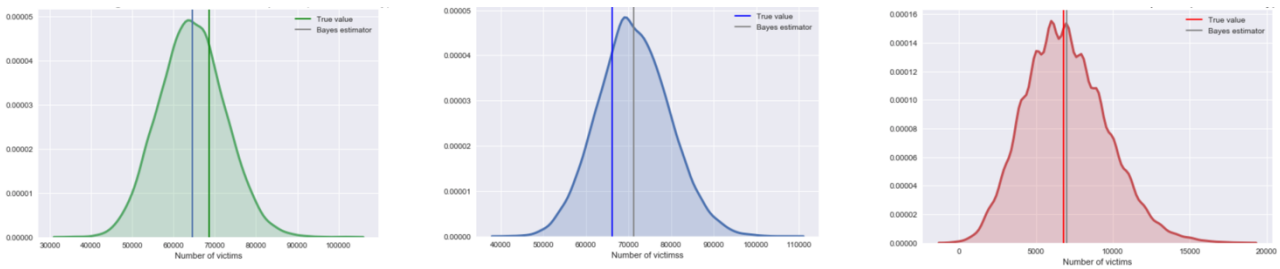

Figure 2 – Predicted total number of victims over the test period, classes Young, Middle, Old

The results are satisfying since the error is maximum 7.75% of the correct value. Prediction results are available in Appendix D.

The conclusions we are making here concern the number of victims. When we cross-reference the number of victims with the demographics of Jamaica, we see that by far the most vulnerable class is the 'Young' class.

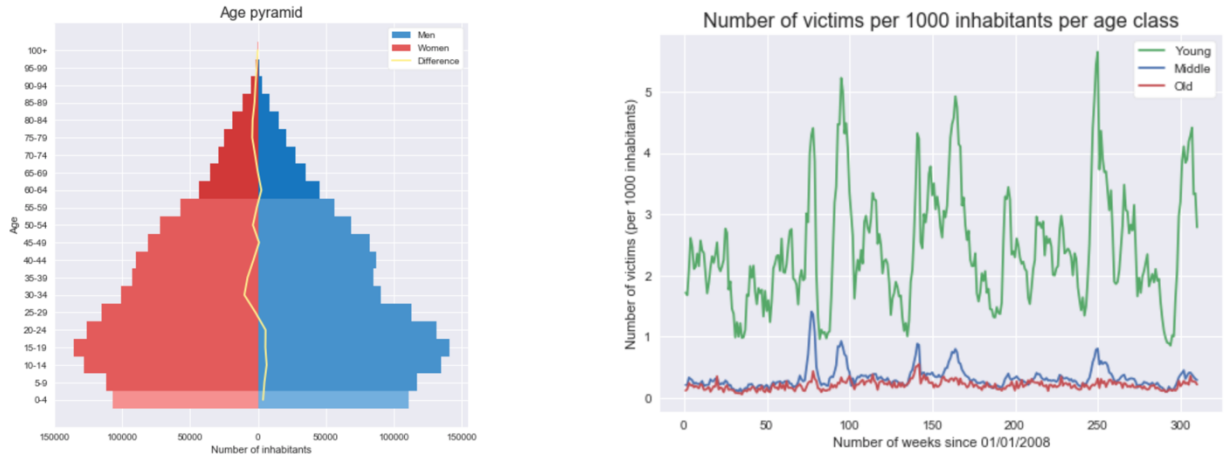

Figure 3 - Vulnerability within each age group

The population pyramid shows that the 'Young' class represents a small proportion of the population. To make the graph on the right, for each category the weekly number of victims was divided by the number of individuals in the class and multiplied by 100 to obtain a percentage. Thus, it is clear that for every 1000 individuals in a class, the Young class is the most affected by the epidemic: it has up to 5 victims per week while the other two classes have a maximum of one case per week.

2 Seasonal trend: Multinomial sampling a Dirichlet

It is clear that the Poisson Gamma model adjusted on the train period does not capture a particular seasonality. To capture seasonality we used a multinomial sampling model coupled with a non-informative Dirichlet prior law.

Our idea is the following: knowing the total number of victims over a year, what is the probability for a victim to be sick in a given week?

More formally, we have as a sampling space $X = \{1, \dots, K\}^n$ with $K = 52$ (rounded number of weeks in a year).
Let's denote $p = \{p_1, \dots, p_K\}$ such that $p_k \geq 0 \ \forall k = 1, \dots, K$ and $\sum_{k=1}^{K} p_k = 1$. p is the probability vector whose k^{th} coordinate corresponds to the probability of drawing the k^{th} week.
The density of the multinomial sampling model is therefore :

$$P\big((x_1, \dots, x_K) = (n_1, \dots, n_K)\big) = \frac{n!}{n_1!, \dots, n_K!} \prod_{k=1}^{K} {p_k}^{n_k}$$

Let p follow a Dirichlet law, $Dir(K, \rho_0)$. That is : $\rho_0 = (\rho_{0,1}, \dots, \rho_{0,K}) > 0$ and
$\pi(p_1, \dots, p_K | \rho_0) = \frac{\Gamma(\bar{\rho}) \prod_{k=1}^{K-1} {p_k}^{\rho_{0,k}}}{\prod_{k=1}^{K} \Gamma(\rho_{0,k})} \times (1 - \sum_{k=1}^{K-1} p_k) \prod_{k=1}^{K-1} 1_{\{p_k \geq 0\}} \, 1_{\{\sum_{k=1}^{K-1} p_k \leq 1\}}$, where $\bar{\rho} = \sum_{k=1}^{K} \rho_{0,k} = 1$
We set $\forall k, \rho_{0k} = 2$. By assigning the same value to each component of ρ_0 we take a non-informative a priori: we do not assume a week has higher probability.
The posteriori law is still a Dirichlet law, with parameters : $\rho_{*\,k} = \rho_{0\,k} + n_k$
In our case, $n_k = \sum sick\ indivuals\ at\ week\ k\ on\ the\ training\ set$
Since n_k is high compared to ρ_{0k} : the choice of ρ_0 has a little influence.

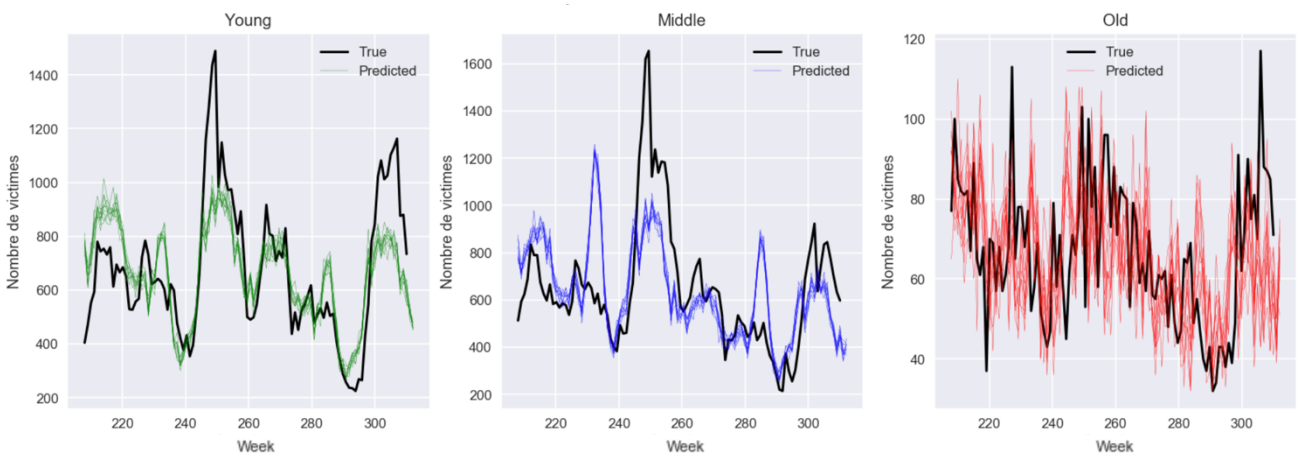

Figure 4 - Allocation of victims on a weekly basis, knowing the number of victims per year

We are quite satisfied with the results, since the trend seems to be captured even on the test sample. The graphs show several simulated trajectories for each age group. The model fails to capture large peaks that are difficult to predict, causing a root mean squared error (RMSE) higher than the mean absolute error (MAE). The results in terms of error metrics are available in Appendix E.

3 Predictions with our two models

Let's put our Poisson Gamma model and our Dirichlet model together :

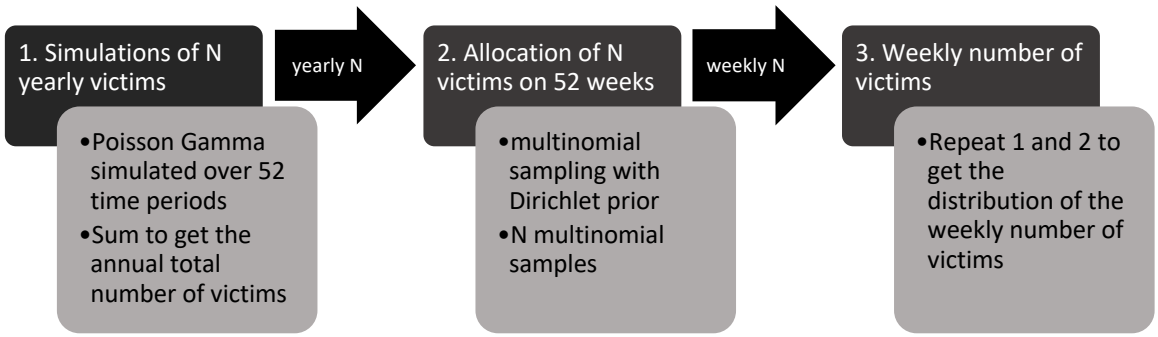

Figure 5 – Full model explanations

We fitted the Poisson Gamma model and our multinomial sampling model over the training period and then tested the model by comparing the mean trajectory (which is therefore the Bayes estimator at each point in time) to the actual trajectory. The first model gave a good estimate of the annual number of casualties, so it contributes little to the prediction error. The MAE remains stable compared to the previous stage while the RMSE has increased, indicating the presence of large errors. In some weeks the second model fails to reproduce a very high peak. We are nevertheless satisfied since the model is better than the constant prediction equal to the mean. The results of the error metrics are available in Appendix F. When we plot the results we obtain the following graph:

:

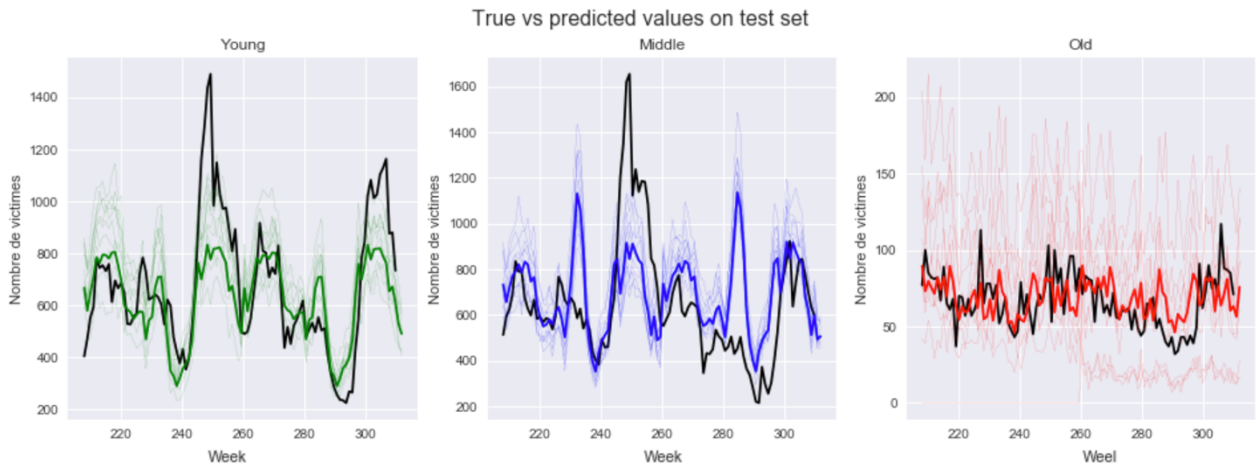

Figure 6 – True trajectory (black) vs mean predicted trajectory (bold)

4 Comparison with SARIMA

From now on, we wish to compare the results obtained using our Bayesian method with those that we would obtain using a classic approach.
To do so, after a quick analysis, we have set up for each age group an auto arima which tests different SARIMA models and chooses the best one based on the Akaike information criterion (the models are likely to be improved. The idea here is to have a frequentist baseline to compare with our Bayesian models).

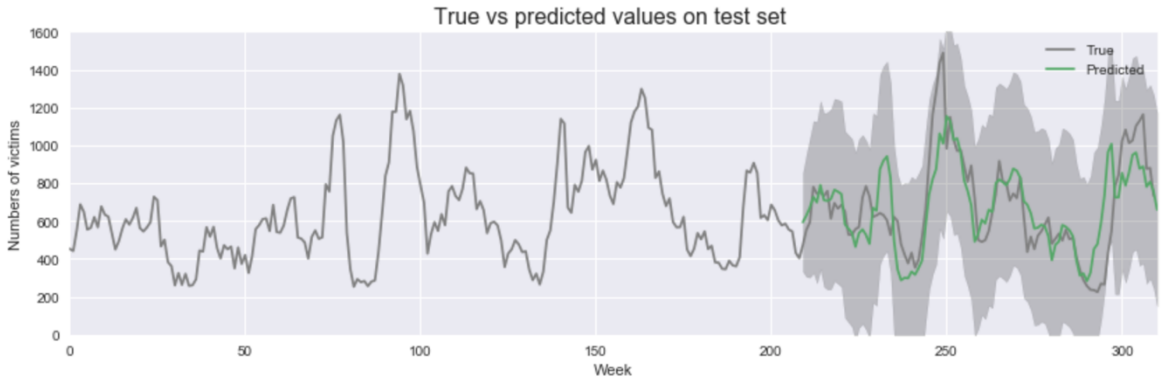

Figure 7 – True trajectory (black) vs mean predicted for Young age group

Let's compare the results between these two approaches:

Age Group	RMSE SARIMA	RMSE with our approach
Young	187	183
Middle	435	250
Old	26	18

Tableau 2 -Comparison between SARIMA and our approach

We notice that the results are similar for the Young and Old classes. On the other hand, for the Middle class, the Bayesian method is better. We understand why this difference exists when we observe the curve 'True vs predicted values on test set' (see in Python code): the SARIMA approach has learned peaks on the training sample that no longer exist on the test sample.

Conclusion and areas of improvement

Our work has allowed us to predict the number of victims on our test sample with different methods and thus to obtain conclusive results.

In order to better take into account the evolution of trends in the annual number of victims and their weekly distribution, we could have 'dragged' the learning periods and obtained the best window size by cross-validation. The annual number of victims simulated by the first model is close to reality, but the second model that reproduces seasonality is sometimes too far from reality at certain peaks: a shorter (or weighted) learning period and therefore closer in time might have allowed us to better take into account changes in seasonality.

Appendix

A. Recall on Bayes Formula

We denote $f(z|\theta)$ the likelihood of a the sampling model.
In the Bayesian framework, an uncertainty is introduced on the parameter of interest θ given by a probability law, called the *prior* (law of probability) and noted $\pi(\theta)$. The *posterior* (law of probability) is an update of the prior conditional on the observations and is given by the Bayes formula*:*

$$\pi(\theta|z) = \frac{f(z|\theta)\pi(\theta)}{m(z)}$$

$$\propto f(z\ |\theta)\pi(\theta)$$

with :
$m(Z) = \int_{\theta\in\mathbb{R}^4} f(\ z|\ \theta)\pi(\theta)d\theta$ the marginal density of Z also called *evidence*.

B. Theoretical and computational results comparison

Classe d'âge	$\hat{\lambda}$ Metropolis	$\hat{\lambda}$ Empirical	*Erreur* %
Young	429.22	626.8	31
Middle	468.5	692.6	32
Old	53.5	67.8	21

We are surprised by the difference between the values calculated using the theoretical expression (and empirical moments) and the values of the estimators obtained with Metropolis Hastings. With more time, we would have investigated this question, but computational methods are not the core of this project. A longer burn-in period coupled with a higher number of simulations could maybe solve the problem. For the remainder of this project, we used the values of $\hat{\lambda}$ Empirical.

C. BIC calculation

We have :

$$m_i(y) = \int_{\theta_i} f(\ y|\ \lambda_i)\pi(\lambda_i)d\lambda_i$$

For M_0 'Young' and 'Middle' disctinct, we therefore have two different laws on λ_0 depending on whether the victim belongs to 'Young' or 'Middle', we denote respectively $\lambda_{0,Y}$ and $\lambda_{0,M}$ these parameters. These two parameters prior are respectively a *Gamma* $(a_Y\ , b_Y)$ and a *Gamma* $(a_M\ , b_M)$. For M_1, the two populations follow the same probability law, so we have only one parameter λ_1 whose prior is a *amma* $(a\ , b)$.
Assuming the independence between the 'Young' and 'Middle' classes, we have :

$$m_0(y) = \int_{\lambda_{0,Y}} f(\, y_Y |\, \lambda_{0,Y})\pi(\lambda_{0,Y})d\lambda_{0,Y} \int_{\lambda_{0,M}} f(\, y_M |\, \lambda_{0,M})\pi(\lambda_{0,M})d\lambda_{0,M}$$

with y_Y the 'Young' group observations and y_M the'Middle' group observations
We have the same number n of observations in each group :

$$m_0(y) = \frac{\int_{\lambda_{0,Y}} \lambda^{\sum_{k=1}^n y_{Y,k}+a_Y-1} e^{-\left(n+\frac{1}{b_Y}\right)\lambda_{0,Y}}\, d\lambda_{0,Y} \int_{\lambda_{0,M}} \lambda^{\sum_{k=1}^n y_{M,k}+a_M-1} e^{-\left(n+\frac{1}{b_M}\right)\lambda_{0,M}}\, d\lambda_{0,M}}{\prod_{k=1}^n y_{Y,k}!\,\Gamma(a_Y) {b_Y}^{a_Y} \prod_{k=1}^n y_{M,k}!\,\Gamma(a_M) {b_M}^{a_M}}$$

The multiplicative constants allowing to have the product of densities of two gamma laws are introduced in the expression and we obtain:

$$m_0(y) = \frac{\Gamma\big(a_Y + \sum_{k=1}^n y_{Y,k}\big)\left(n + \frac{1}{b_Y}\right)^{-\left(a_Y + \sum_{k=1}^n y_{Y,k}\right)} \Gamma(a_M + \sum_{k=1}^n y_{M,k})(n + \frac{1}{b_M})^{-(a_M + \sum_{k=1}^n y_{M,k})}}{\prod_{k=1}^n y_{Y,k}!\,\Gamma(a_Y) {b_Y}^{a_Y} \prod_{k=1}^n y_{M,k}!\,\Gamma(a_M) {b_M}^{a_M}}$$

The calculation for $m_0(y)$ is the same, with only one integral since the data comes from the same model, we have 2n observations and we obtain:

$$m_1(y) = \frac{\Gamma\big(a + \sum_{k=1}^{2n} y_k\big)\big(2n + \frac{1}{b}\big)^{-(a + \sum_{k=1}^n y_k)}}{\prod_{k=1}^{2n} y_k!\,\Gamma(a) b^a}$$

Taking the quotient, $\frac{\prod_{k=1}^{2n} y_k!}{\prod_{k=1}^n y_{Y,k}! \prod_{k=1}^n y_{M,k}!}$ simplifies and we get :

$$\frac{m_0(y)}{m_1(y)} = \frac{\Gamma(a_Y + \sum_{k=1}^n y_{Y,k})\left(n + \frac{1}{b_Y}\right)^{-(a_Y + \sum_{k=1}^n y_{Y,k})} \Gamma(a_M + \sum_{k=1}^n y_{M,k})(n + \frac{1}{b_M})^{-(a_M + \sum_{k=1}^n y_{M,k})} \Gamma(a_M) {b_M}^{a_M}}{\Gamma(a_Y) {b_Y}^{a_Y} \Gamma(a_M) {b_M}^{a_M} \Gamma\big(a + \sum_{k=1}^{2n} y_k\big)\left(2n + \frac{1}{b}\right)^{-(a + \sum_{k=1}^n y_k)}}$$

D. Number of victims simulations on the test period results

Age group	True value	Prediction	Error (%)
Young	68 722	64 591	-6.01
Middle	66 216	71 345	7.75
Old	6 787	6 968	2.66

E. Seasonal trend simulations results

Age group	Mean Absolute Error	Root Mean Squared Error
Young	134	175
Middle	160	219
Old	15	19

F. Results of the two models put together

Classe d'âge	Mean Absolute Error	Root Mean Squared Error	% erreur moyenne	RMSE estimateur égal à la moyenne
Young	133	183	19	258
Middle	180	250	30	262
Old	14	18	24	17

G. SARIMA results

Classe d'âge	Mean Absolute Error	Root Mean Squared Error	% erreur moyenne	RMSE estimateur égal à la moyenne
Young	138	187	22	258
Middle	248	435	47	262
Old	19	26	36	17

Data source :

https://www.kaggle.com/campbellsinvestment/jamaican-islandwide-fever-and-respiratory-symptoms