Project

Exploratory Project for Estimating the Reproduction Rate of Covid-19

Lucas Tramonte
Emma Greverie
Jia Yi Ang
Romain Dalcant

October 2023 - January 2024

Table des matières

1 Project Summary

Initially, our project was centered on **the analysis of the basic reproduction number R0 within the Île-de-France region**. Our initial models, however, treated Île-de-France as **an isolated region**, not accounting for the dynamic interactions with its neighboring areas. This limitation was particularly poignant **given the region's interconnected nature**, where residents frequently travel for work, leisure, and other activities.

Acknowledging this gap, **we pivoted our focus this semester to include the neighboring departments around Île-de-France**. This expansion was not just a mere enlargement of our geographical scope but a necessary step to capture the true dynamics of disease transmission. **The significant movement of people between these regions necessitated a more comprehensive approach, integrating additional data to accurately model these interactions**. As a result, our modified models now provide a more realistic and precise estimation of the R0, **reflecting the complex interplay of factors influencing the pandemic's spread**.

2 Introduction

The Covid-19 pandemic has deeply marked our youth, a time when we were primarily spectators behind our screens, following the disease's developments through terms like R0, plateau, peaks, and waves. This project transformed us into actors, allowing us to develop models and work methods based on Covid-19 data, tools useful for future situations.

The R0, or basic reproduction rate, is a key indicator in studying the spread of infectious diseases, such as Covid-19. It indicates the average number of people that an infected person can contaminate. This rate is essential for understanding the spread of Covid-19 in a population and for predicting the epidemic's evolution. These insights are crucial for health authorities to plan and implement appropriate control measures, such as vaccination campaigns, lockdowns, or social distancing measures.

During the previous semester, our project focused on utilizing various analytical models to estimate the basic reproduction number, or R0, of the Île-de-France region. This estimation of R0 was crucial for understanding the dynamics of infectious disease spread in the area. However, our initial approach had **a significant limitation : it treated Île-de-France as an isolated entity, neglecting interactions with surrounding regions.** In reality, such isolation is rarely relevant, especially in an area as dynamic and interconnected as Île-de-France.

To address this gap, this semester, **we have expanded our study to include the departments neighboring Île-de-France**. This expansion was necessary to obtain a more accurate and realistic picture of disease spread. **Indeed, the significant people flow between Île-de-France and its adjacent departments can play a vital role in the dynamics of disease transmission**.

To achieve this, we had to make substantial modifications to our existing models. These adjustments mainly involved integrating additional data. **The goal was to more accurately model the complex interactions between Île-de-France and the adjacent regions, thereby providing a more precise estimation of R0**.

3 State of the art

[1] is a concise introduction for applied mathematicians and computer scientists to basic models, analytical tools and mathematical and algorithmic results. The mathematical tools introduced include coupling methods, Poisson approximation (the Stein-Chen method), concentration inequalities (Chernoff bounds and the Azuma-Hoeffding inequality) and branching processes. The authors examine small-world phenomena, preferential attachment and classical epidemics.

[2] presents a ready-to-use tool for estimating R from incidence time series, which is implemented in popular software such as Microsoft Excel (Microsoft Corporation, Redmond, Washington). This tool produces novel and statistically robust analytical estimates of R, and incorporates uncertainty in the distribution of the serial interval (the time between the onset of symptoms in a primary case and the onset of symptoms in secondary cases). However, access to this tool is not free.

[3] has developed a function for jointly estimating the number of replicates and the outliers defined to model low-quality data.This functional also guarantees epidemiologically dictated regularity properties for reproductive number estimates, while preserving convexity, enabling the design of efficient minimization algorithms based on analytically derived proximity operators.

[4] investigated autoregressive R0 models that took into account both temporal variations at the scale of the day and spatial variations at the scale of different US states. Unlike us, they were based on the number of deaths per day and per state, and not on the number of hospitalizations.

[5] is working on a spatial analysis of the influence of socio-economic factors on the prevalence and consequences of the epidemic in French departments. Their work takes as its starting point the linear regression model of ordinary least squares, then in a second step, they correct the biases of the estimators derived from the least squares method.

[6] details the comparative and consistency analysis between statistical and compartmental models on the prediction of COVID-19 infections and deaths. He used three statistical models (ARIMA, SARIMA and Prophet) and two compartmental models (SIRD, SIRF).

4 Overview of the data

To address this gap, this semester, **we have expanded our study to include the departments neighboring Île-de-France**. This expansion was necessary to obtain a more accurate and realistic picture of disease spread. We have chosen to focus on the contributions of the following colored groups of departments.

Yellow Region : Aisne (02) + Oise (60) + Somme (80)

Pink Region : Côte-d'Or (21) + Nièvre (58) + Saône-et-Loire (71) + Yonne (89)

Blue Region : Cher (18) + Eure-et-Loir (28) + Indre (36)
+ Indre-et-Loire (37) + Loir-et-Cher (41) + Loiret (45)

Green Region : Ardennes (08) + Aube (10) + Marne (51) + Haute-Marne (52)

Violet Region : Eure (27) + Seine-Maritime (76)

Red Region : Paris (75) + Seine-et-Marne (77) + Yvelines (78) + Essonne (91)
+ Hauts-de-Seine (92) + Seine-Saint-Denis (93)

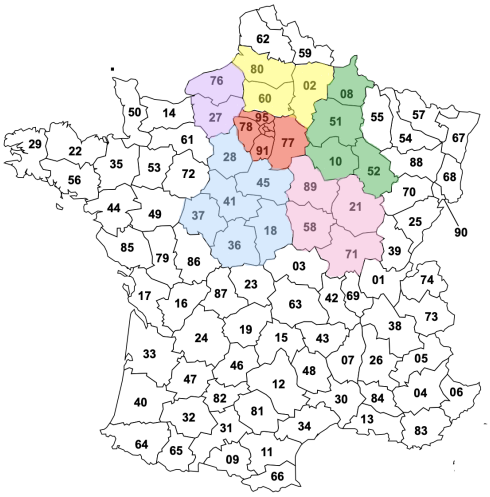


FIGURE 1 – Map of the French departments

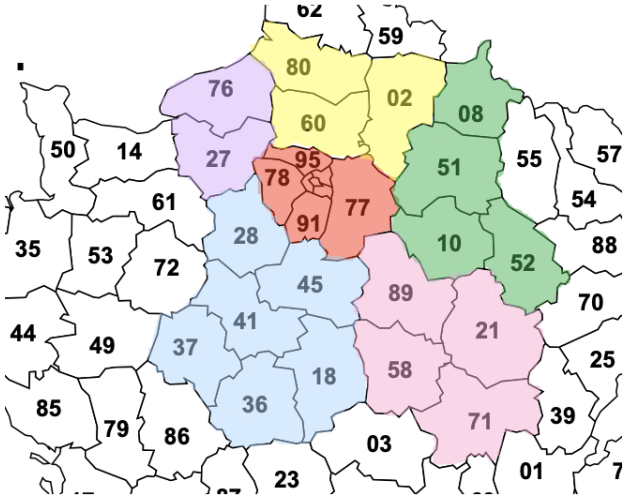


FIGURE 2 – Zoom around the Île-de-France region

We used the moving average method (order $m = 7$) to smooth out the number of cases per day for all departments - FIGURE 3 - and for each department - FIGURES 4 & 5.

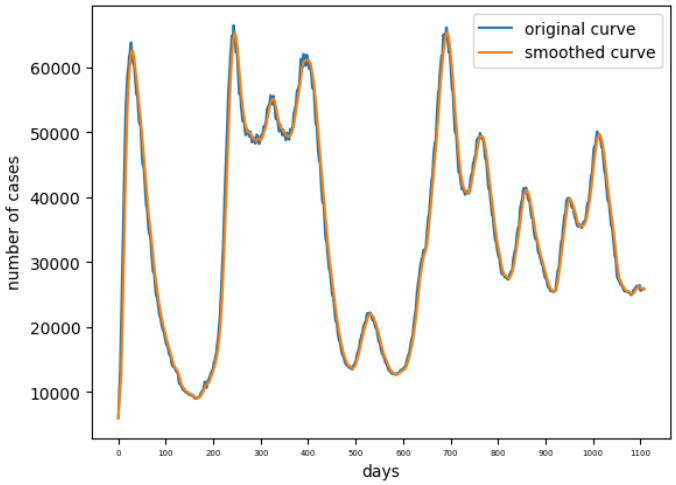


FIGURE 3 – Smoothing for all departments

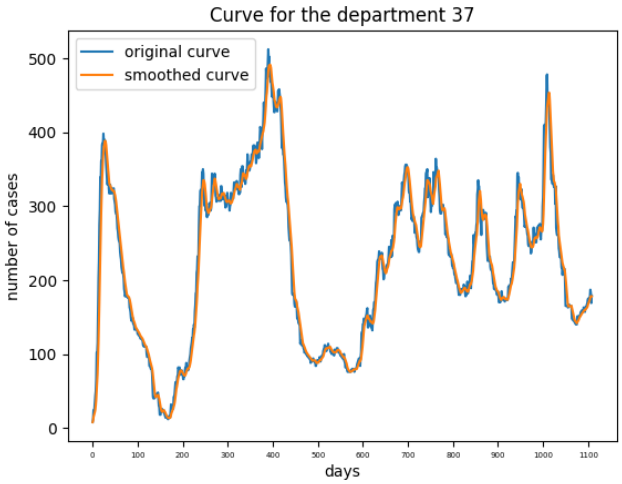


FIGURE 4 – Smoothing for department 37

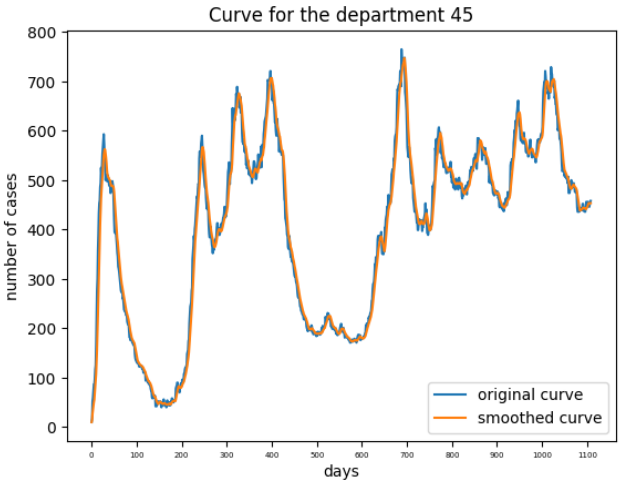


FIGURE 5 – Smoothing for department 45

5 SIR Model

A mathematical model used in epidemiology to comprehend the dynamics of infectious illnesses within a community is called the Susceptible-Infectious-Recovered (SIR) model. In this context, the parameters alpha (α) and beta (β) are used to describe the rates of infection and recovery, respectively. It is important to note that this model assumes that the total population studied is constant. Thus, we have the following equations for the SIR model :

$$\begin{aligned} \frac{dS}{dt} &= -\alpha \cdot S \cdot I \\ \frac{dI}{dt} &= \alpha \cdot S \cdot I - \beta \cdot I \\ \frac{dR}{dt} &= \beta \cdot I \\ \frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} &= -\alpha SI + \alpha SI - \beta I + \beta I = 0 \\ R_0 &= \frac{S\alpha}{\beta} \end{aligned}$$

Where S represents the number of susceptible individual, I the number of infectious individuals and R for the number of recovered individuals.

First, we plot the SIR curve for all the departments in the dataset, and a comparison was made with the real data, which can be seen in the figures 6 and 7.

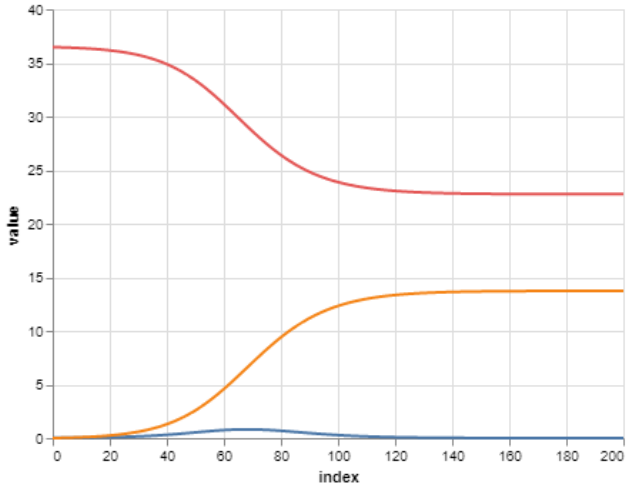


FIGURE 6 – SIR model for all departements

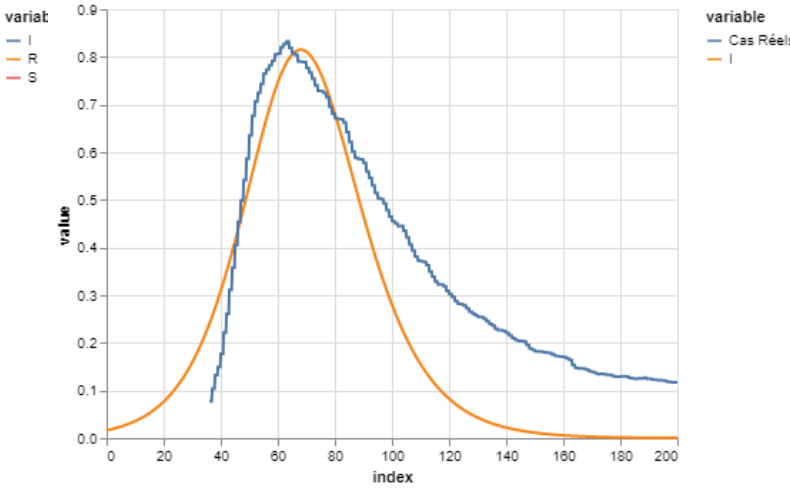


FIGURE 7 – Comparison of actual cases and I

The SIR model has many assumptions that are difficult to meet in reality. The first of these is the homogeneity of the population. The model assumes that each individual has the same probability of catching the disease, which is not true and depends on other factors such as respiratory diseases, genetic factors, among others. In addition, the SIR model assumes that individuals have permanent immunity to the disease, but in reality immunity can decrease over time and individuals can become

re-infected. It is because of these and other factors that the SIR model ends up not being such a good predictor of real data

In this context, the $R_0 = 1.25$ coefficient was calculated for the start of the pandemic, and its evolution can be seen in figure 8 for all departments, and in figures 9 and 10 for two specific departments. It is important to note that R_0 above 1 means that transmission of the disease is increasing.

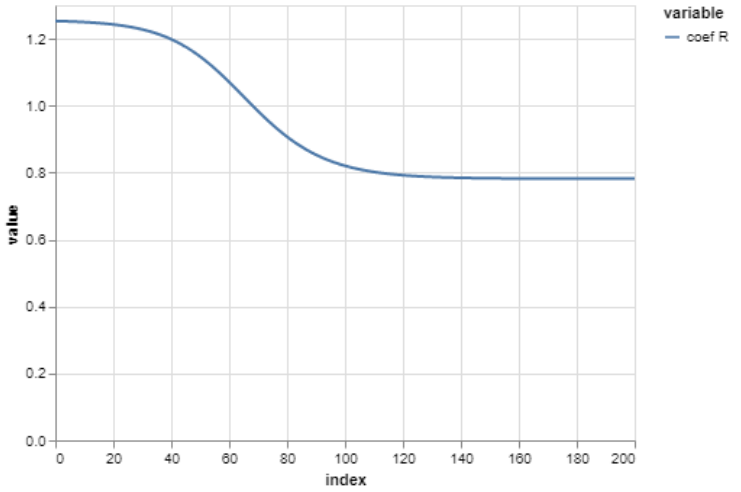


FIGURE 8 – Evolution of the reproduction coefficient R

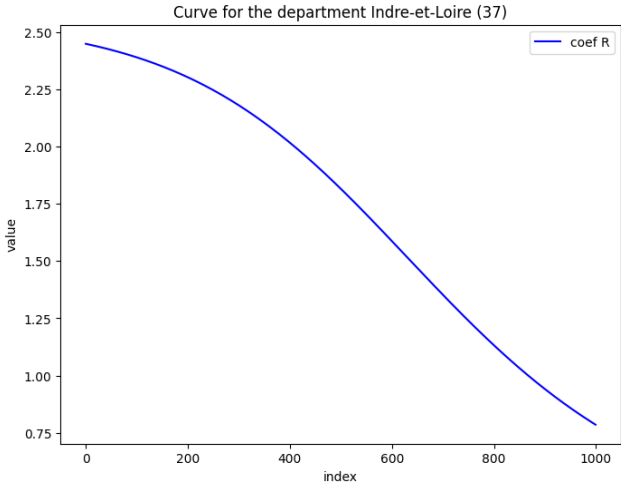


FIGURE 9 – Coefficient R for the department 37

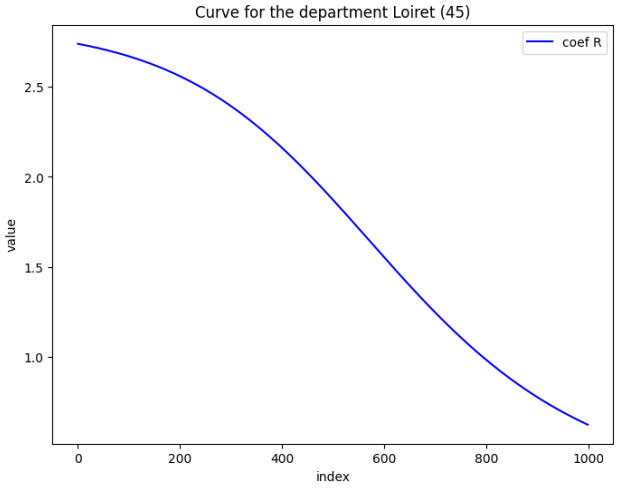


FIGURE 10 – Coefficient R for the department 45

It is interesting to note that for departments 37 and 45 the coefficient R is considerably higher than for the overall R, which is also linked to a different choice of alpha and beta for the curves to better fit the real data.

We then implemented the same model for each department separately, and a comparison was made with the real data,. Some examples can be seen in the figures 11, 12, 13 and 14.

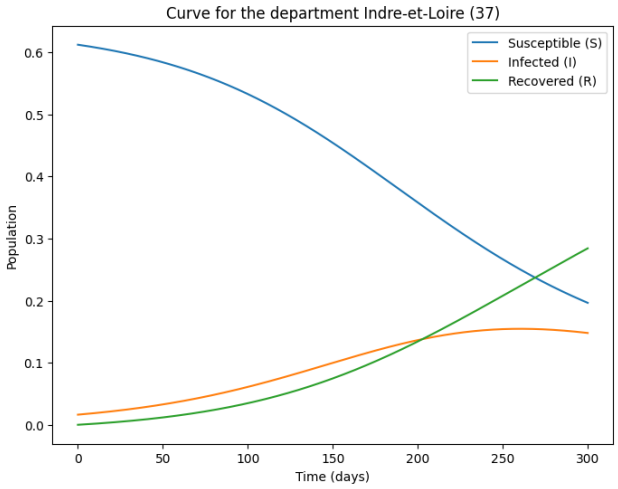


FIGURE 11 – SIR model for department 37

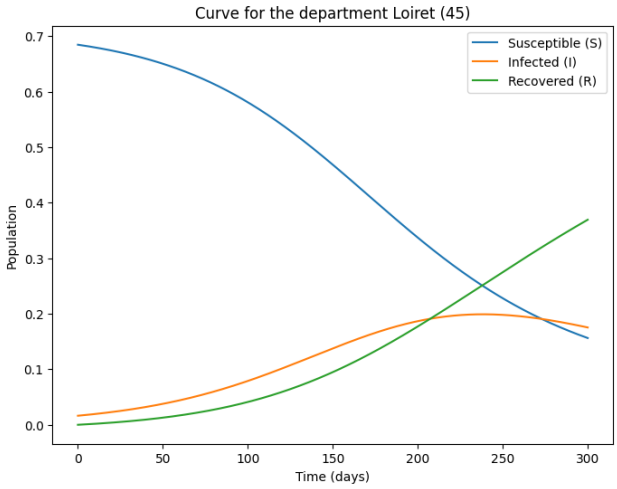


FIGURE 12 – SIR model for department 45

Thus, we compared the actual data with the theoretical number of infected people for each department.

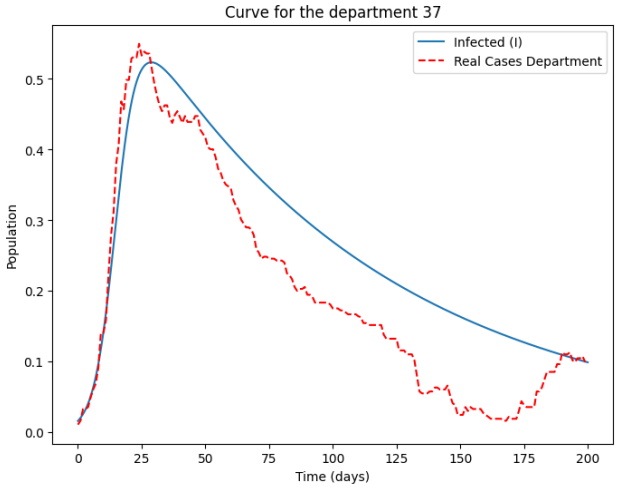


FIGURE 13 – Comparison for the department 37

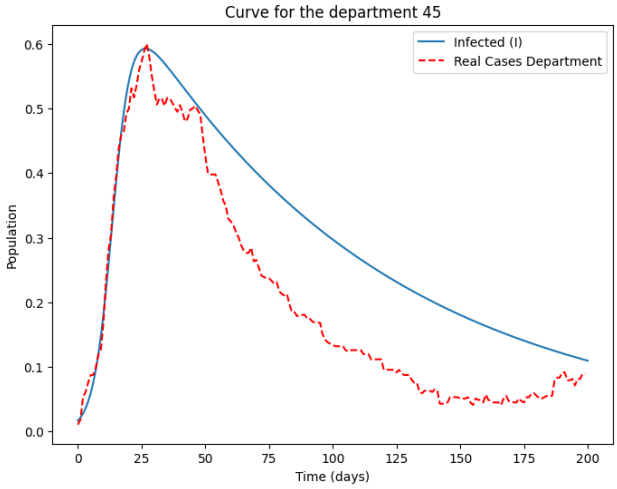


FIGURE 14 – Comparison for the department 45

6 Autoregressive model used

Autoregressive (AR) models are a type of time series model used for understanding and forecasting time-dependent data. In an AR model, the future value of a variable is assumed to be a linear combination of its past values plus an error term. The order of the model, denoted as AR(p), indicates how many past values are used for prediction.

6.1 Number of cases for a single region

The project team of last year based its project on the following expression :

$$Z(t) = \sum_{k=1}^{L} \alpha(k) Z(t-k)$$

It gives the number of cases in a chosen French department or region, on the day t, based on data from the previous L days in this department or region.

This temporal autoregressive model can be compute thanks to minimisation :

$$\hat{\alpha} = \text{argmin}_{\alpha} ||Z(t) - \sum_{k=1}^{L} \alpha(k) Z(t-k)||_2^2$$

where $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_L \end{bmatrix}$

This model allows to predicted the evolution of Covid19 in each department completely independent of each other.

6.2 Spaciorality of the new model

Now, we consider the covid trend across all departments to predict a department's trend.

The following expression gives the number of cases for one single region called c, on the day t, based on data from the previous L days as if it were isolated.

$$Z(t,c) = \sum_{k=1}^{L} \alpha(k,c) Z(t-k,c)$$

Now, it is necessary to introduce a regularization term to smooth the values between each department and introduce the spacial dimension to the model :

$$\hat{\alpha} = \text{argmin}_{\alpha} \sum_{c=1}^{6} ||Z(t,c) - \sum_{k=1}^{L} \alpha(k,c) Z(t-k,c)||_2^2 + \lambda \sum_{c,c'} \omega(c,c') ||\alpha(c) - \alpha(c')||_2^2$$

where c, and c' stand for different regions that go from 1 to 6 (the pink one, the red one, the blue one, the green one, the violet one and the blue one), where λ is the penalty parameter penalizing the difference in α between regions c and c'. ω(c,c') must be a decreasing function of distance between region. The ω(c,c') define an adjacency matrix that weights relationships between regions. We decided the ω(c,c') have to take into account populations c and c' and would be an increasing function of the population because bigger is the population, bigger is the influence on the neighbors.

Thus, we chose :

$$\omega(c,c') = \frac{pop(c) * pop(c')}{2 * d(c,c')}$$

where

$$\alpha = \begin{pmatrix} \alpha_{(1,1)} & \alpha_{(1,2)} & \alpha_{(1,3)} & \alpha_{(1,4)} & \alpha_{(1,5)} & \alpha_{(1,6)} \\ \alpha_{(2,1)} & \alpha_{(2,2)} & \alpha_{(2,3)} & \alpha_{(2,4)} & \alpha_{(2,5)} & \alpha_{(2,6)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_{(L,1)} & \alpha_{(L,2)} & \alpha_{(L,3)} & \alpha_{(L,4)} & \alpha_{(L,5)} & \alpha_{(L,6)} \end{pmatrix}$$

The actual adjency matrix is :

	0	1	2	3	4	5
0	0.000000	0.010624	0.005826	0.004468	0.040930	0.004339
1	0.010624	0.000000	0.006566	0.011398	0.097248	0.008107
2	0.005826	0.006566	0.000000	0.004680	0.047110	0.006513
3	0.004468	0.011398	0.004680	0.000000	0.094213	0.011542
4	0.040930	0.097248	0.047110	0.094213	0.000000	0.091469
5	0.004339	0.008107	0.006513	0.011542	0.091469	0.000000

FIGURE 15 – Actual adjency matrix

Where 0,1,2,3,4,5 corresponds respectively to Bourgogne, Centre, Champagne, Haute Normandie, Ile de France, Picardie.

Here is the minimisation function and the alpha evaluation are implemented the following way :

```python
def minimization_function_bis(alpha,n, listes, lambda_penalty, adjacency_matrix):
    total_error = 0
    for i in range(n_reg):
        # On calcule la prediction avec alpha
        predicted_cases = np.zeros(len(listes[i])-n)
        for j in range (len(predicted_cases)):
            sumu = 0
            for k in range(n):
                sumu += (listes[i][j+k])*(alpha[n*i+n-k-1])
            predicted_cases[j] = sumu

        total_error += np.sum((listes[i][n:] - predicted_cases) ** 2)

    # Terme de lissage pour R0
    smoothing_penalty = 0
    for i in range(n_reg):
        for j in range(n_reg):
            c = adjacency_matrix[i][j]
            for k in range(n):

                smoothing_penalty += c*((alpha[n*i+k] - alpha[n*j+k])**2)

    # On determine l'erreur totale
    return total_error + lambda_penalty * smoothing_penalty

def alpha_calculation(liste, lambda_penalty, adjacency_matrix, n):

    # On cree notre matrice alpha initiale qu'on exprime sous la forme d'une liste pour utiliser minimize de spicy
    matrix_initial_alpha_guesses =np.array([[1]+[0]*(n-1) for _ in range(n_reg)])
    initial_alpha_guesses = matrix_initial_alpha_guesses.flatten()

    # On utilise la fonction minimize du module spicy
    result = minimize(minimization_function_bis, initial_alpha_guesses, args=(n, liste, lambda_penalty, adjacency_matrix))

    alpha_optimized = result.x  #Valeurs optimisees d'alpha pour chaque departement

    # On remet alpha sous forme d'une matrice
    alpha_final = alpha_optimized.reshape((n_reg,n))

    return alpha_final
```

Listing 1 – Minimization function

The minimisation function is able to define the alpha matrix representing the alpha coefficients for each rank and each region.

6.3 AIC model

The Akaike Information Criterion (AIC) is a measure of the quality of a statistical model. This is the criterion we have chosen to determine the processus used to model our data. The most appropriate model according to this criterion is the one that minimizes AIC. This criterion is interesting because it takes into account not only the likelihood of the model, but also the number of parameters, in order to penalize overly complex models that would lead to overfitting.

To determine the optimal rank of n we use this formula because of the complexity of the value of the real likelihood :

$$\text{AIC} = 2 \times k \times \text{len(liste)} - 2 \times \text{log\_likelihood}$$

where :

$$\text{log\_likelihood} = \Big(-\frac{n}{2}\Big)\Big(\log(2\pi) + \log\Big(\frac{\text{sse}}{n}\Big) + 1\Big)$$

```
('n', 'AIC')
(1, 10565.662802897068)
(2, 9638.818594614253)
(3, 9632.17814220558)
(4, 9630.828544399663)
(5, 9625.045440634243)
(6, 9583.684836651813)
(7, 9551.423298893184)
(8, 9538.64667703783)
(9, 9521.38039091576)
(10, 9502.962171310786)
(11, 9495.260161325694)
(12, 9505.820054841175)
(13, 9515.217273937365)
(14, 9490.550292508973)
```

FIGURE 16 – AIC of various ARMA models

The best rank n to choose is n=11 : this is the one which minimize the AIC. n = 11 will be used for the analysis.

7 Results & Analysis

In this section, we explore the analysis and validation of our autoregressive model. Validation is achieved through out-of-sample testing and error analysis. These steps ensure our model's accuracy and reliability in forecasting.

7.1 Prediction

We compare the values predicted by the model between day 751 and day 1109 and with our fixed coefficients. There is a good match between the 2 curves that illustrates the precision of the model.

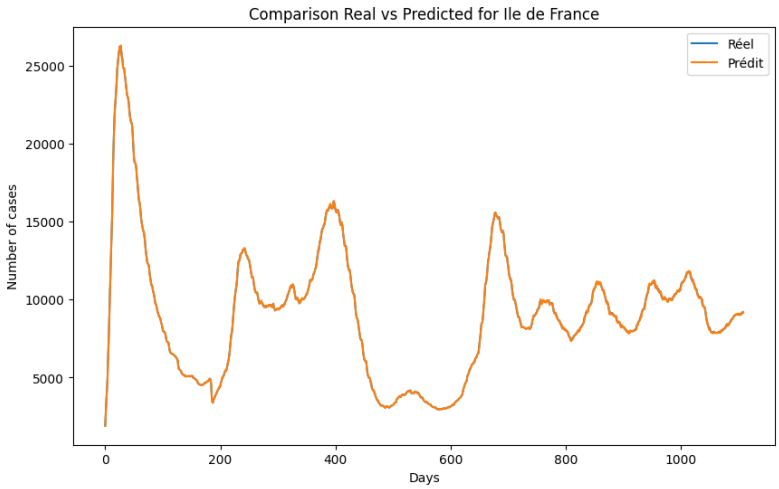


FIGURE 17 – IDF Predicted curves

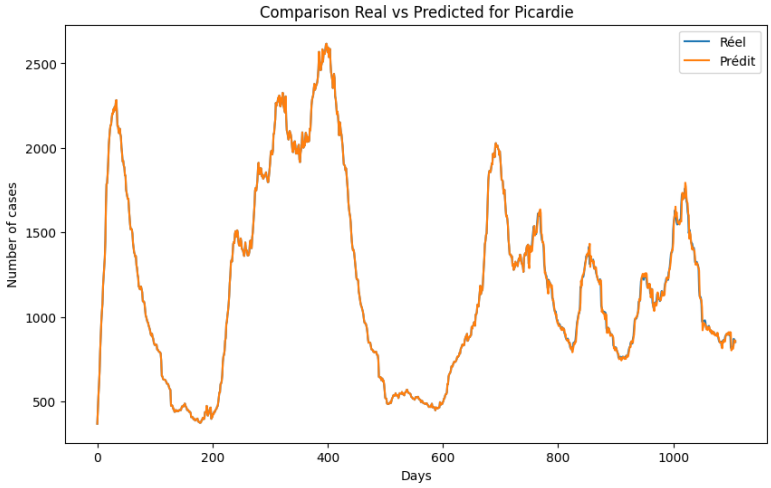


FIGURE 18 – Picardie Predicted curves

For example, for the department of Ile de France, the maximum error is 229.49 and the mean error is 63.73 which is insignificant compared to the number of cases, which is in the thousands

We also observe the relative deviation for the different department which is defined by $(Z_{real} - Z_{predicted})/(Z_{real})$.

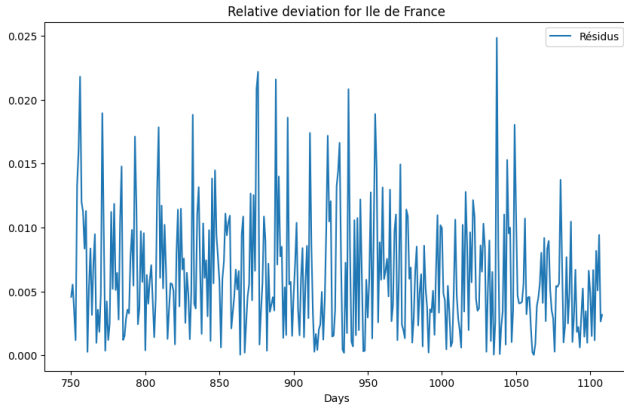


FIGURE 19 – IDF Relative deviations

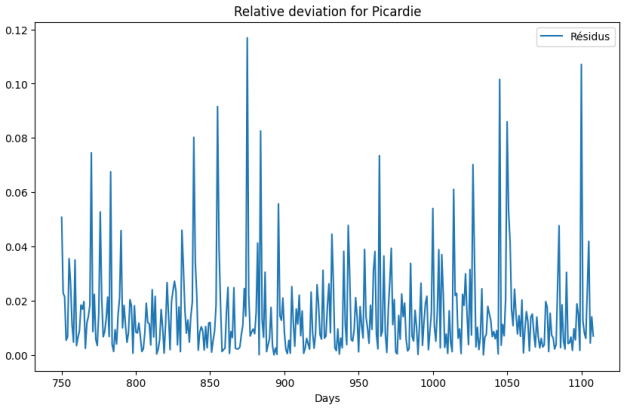


FIGURE 20 – Picardie Relative differences

We notice that the biggest deviation is around 2% for Ile de France and 10% for Picardie. This difference is due to the high population of Ile de France and the high number of cases. The spaciotemporal model is close to the Ile de France model because of the definition of the adjacency matrix which takes into account the Ile de France population. In addition, the high number of cases gives a smoother time series than for other departments with fewer cases.

We also performed a correlation of the residuals, the curve of which is as follows :

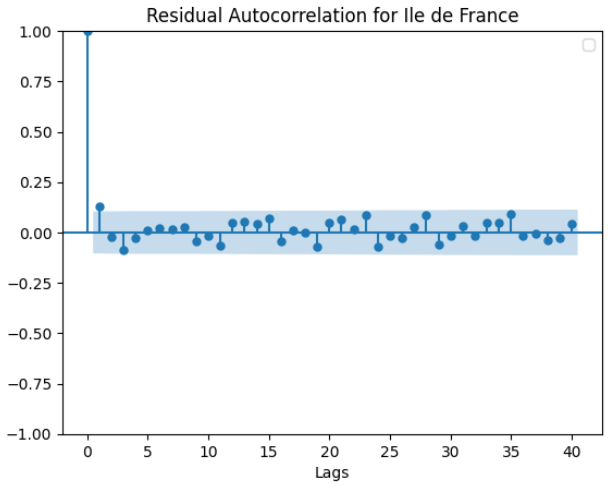


FIGURE 21 – IDF Écarts Relatifs

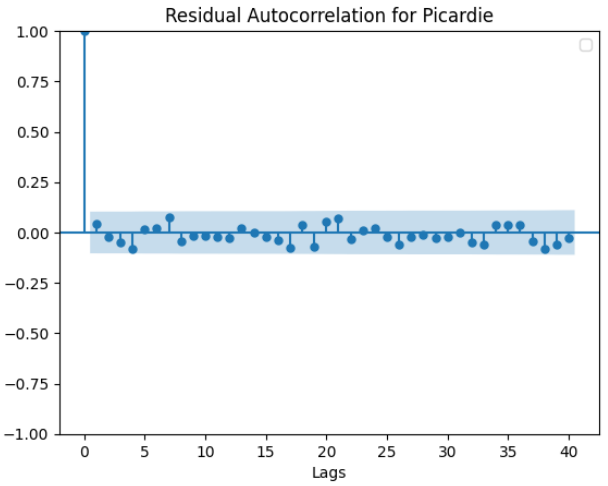


FIGURE 22 – Picardie Écarts Relatifs

In the residual autocorrelation graph, we notice that most of the bars are in the blue zone (the confidence interval). There is no residual autocorrelation which means that the model has fully captured the information in the data.

7.2 Model accuracy measurement

7.2.1 MSE (Mean Squared Error)

It's the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

$$\text{MSE} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Region	**MSE Value**
Bourgogne	359.26
Centre	308.62
Champagne	187.89
Haute Normandie	301.36
Ile de France	1785.66
Picardie	238.82

TABLE 1 – MSE Values for $\lambda = 500000$ and $n = 11$

7.2.2 MAE (Mean Absolute Error)

This is the average of the absolute differences between the predicted values and actual values. It measures the average magnitude of the errors in a set of predictions, without considering their direction.

$$\text{MAE} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}|Y_i - \hat{Y}_i|$$

Region	**MAE Value**
Bourgogne	7.39
Centre	7.00
Champagne	5.25
Haute Normandie	5.92
Ile de France	19.10
Picardie	5.72

TABLE 2 – MAE Values for $\lambda = 500000$ and $n = 11$

7.2.3 RMSE (Root Mean Squared Error)

RMSE is the square root of the average of squared differences between prediction and actual observation. It measures the standard deviation of the prediction errors or residuals.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

Region	**RMSE Value**
Bourgogne	18.95
Centre	17.56
Champagne	13.70
Haute Normandie	17.35
Ile de France	42.25
Picardie	15.45

TABLE 3 – RMSE Values for $\lambda = 500000$ and $n = 11$

7.3 Smoothing & λ Penalty

The lambda value has an influence on the model. The higher lambda is, the more the difference in alphas between regions is penalized, and the R0 values are smoothed out. A good choice of lambda is therefore essential for an accurate model. If lambda is too small, the spacial dimension of the model is useless and there is no smoothing of R0 at the scale of the different regions. Otherwise, the bias introduced by lambda is too great, and the model loses predictive accuracy.

We therefore studied the RMSE of the model as a function of lambda values (order of magnitude 100,000).

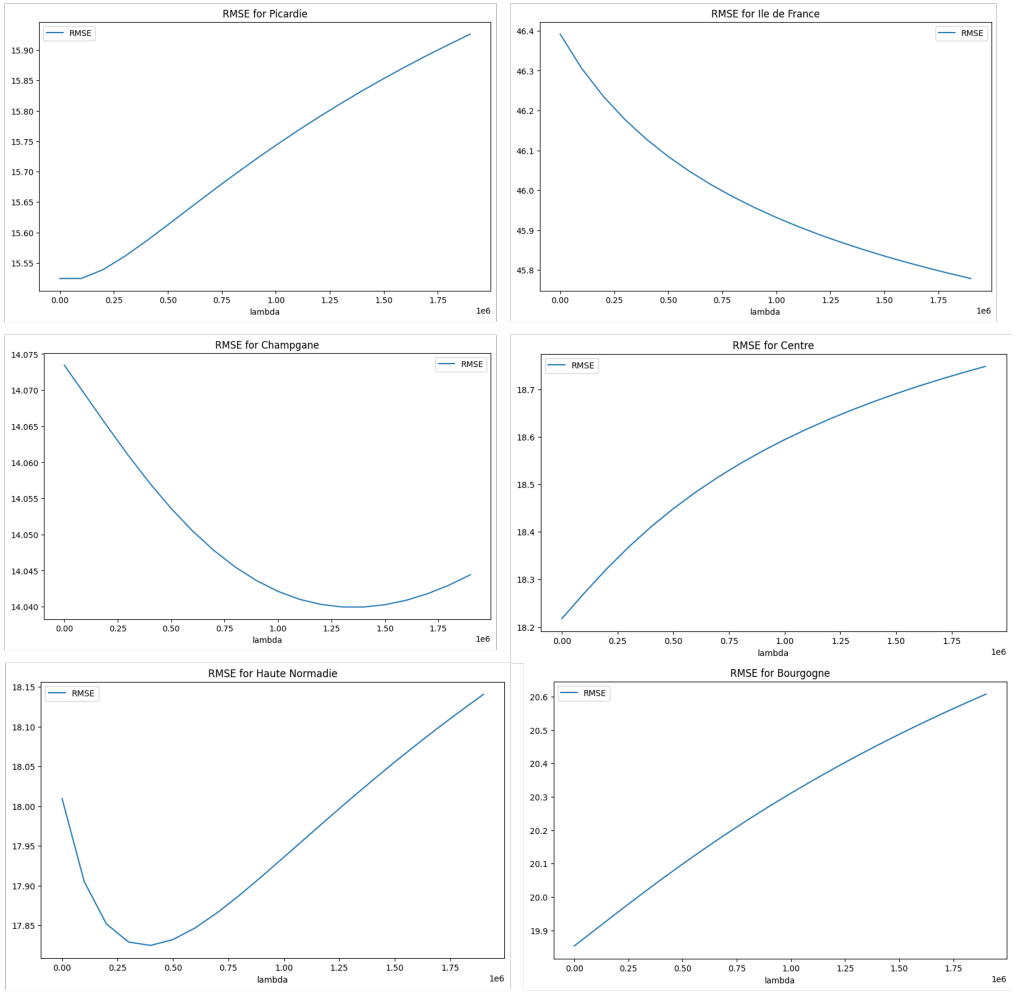


FIGURE 23 – RMSE for the 6 regions

Interestingly, for Picardy, Burgundy and Centre, RMSE is an increasing function of lambda. However, this is not the case for the other regions, and a minimum RMSE can be observed over the chosen lambda range. A minimum of the total RMSE can also be seen for lambda = 400000. This means that the model is the most accurate for this lambda value.

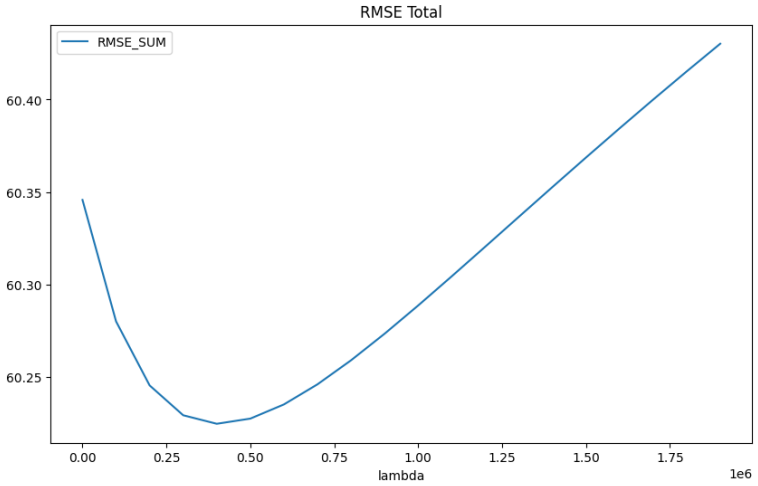


FIGURE 24 – RMSE total

The most important result is that for this lambda value, the model is more accurate than the lambda = 0 case, which corresponds to total independence of the regions. That's the whole point of creating our spaciotemporal model. It enables us to link the different regions and consider the interdependence of their evolution. As it turns out, this model is more accurate and more precise than a model with total independence.

7.4 Correlation

To determine if our model fits correctly and to discover how the region's distance between each other and their respective populations affects infection, we can perform some data analysis using data visualisation tools as well as statistical graphs.

7.4.1 Correlation Matrix

To see if there is a pattern between distance and population and infection, 2 correlation plots between regions are plotted. An additional heat map representing the adjacency matrix is plotted to determine trends. The first plot represents the heat map of the relation of the distance and population between regions. The second plot represents the correlation between regions using the real infection values. The third plot represents the correlation between regions using the predicted infection values. The lambda penalty used for this plot is 10000.

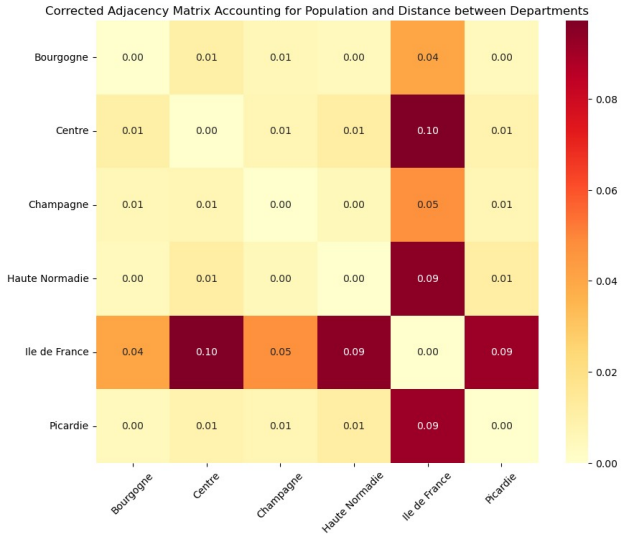
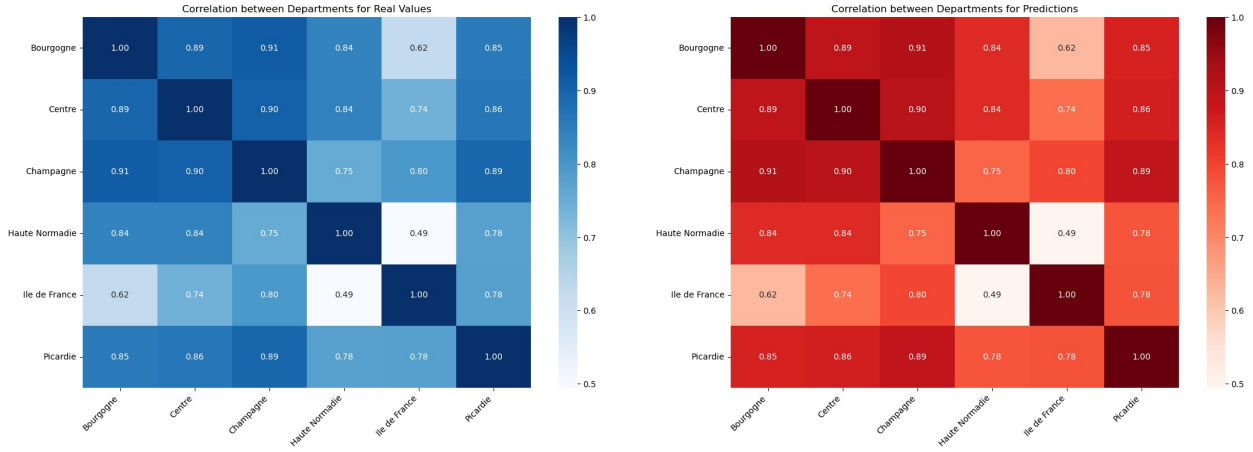


FIGURE 25 – Heatmap of adjacency matrix



(a) Correlation plot between regions' real infection values

(b) Correlation plot between regions' infection values

FIGURE 26 – Correlation plots

Based on the plots, we can see that there is a pattern between the second, the first and the third plot. In the first plot (heat map), we observe that Ile de France has a more positive relation to other regions compared to other regions to other regions. This pattern can be seen in the neighbouring plots whereby regions correlated with Ile de France produced a lower positive correlation compared to them being correlated with other regions. Based on this result, we can determine that the model has been fitted well. Moreover, as there is an observable pattern, we can plot a correlation graph between distance and population against both predicted and real infection values to determine how they are related to each other.

Given the geographical context of our study, conducting a spatial correlation analysis is a prudent next step. For this purpose, employing Moran's I statistic offers a robust method for quantifying spatial autocorrelation, providing valuable insights into the spatial dynamics at play.

7.4.2 Moran's I

Moran's I is a statistic used to measure spatial autocorrelation, which is the degree to which one object is similar to others nearby. It is particularly useful in spatial analysis to determine whether the pattern observed is clustered, dispersed, or random compared to a random distribution. Moran's I is defined as :

$$I = \frac{N}{W} \cdot \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

where :

— N is the number of spatial units indexed by i and j,

— x_i and x_j are the observations for spatial units i and j,

— $\bar{x}$ is the mean of all observations,

— w_{ij} is the spatial weight between units i and j,

— W is the sum of all spatial weights, $W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$.

Value Range : Moran's I values range from -1 to 1 :

— A value close to +1 indicates a strong positive spatial autocorrelation, where similar values cluster together in the space.

— A value close to -1 indicates a strong negative spatial autocorrelation, where dissimilar values are adjacent to each other, suggesting a dispersed pattern.

— A value close to 0 indicates a random spatial pattern, with no significant autocorrelation.

Moran's I values range from -1 (indicating perfect dispersion) to +1 (indicating perfect clustering), with a value of 0 suggesting a random spatial pattern. A positive Moran's I value indicates that similar values are clustered together, while a negative value suggests that dissimilar values are adjacent to each other.

The significance of Moran's I is usually assessed through a permutation test, comparing the observed value to a distribution of Moran's I values generated under the null hypothesis of spatial randomness.

To interpret Moran's I in the context of spatial analysis :

— A significantly positive Moran's I indicates that the variable exhibits a clustered pattern.
— A significantly negative Moran's I suggests a dispersed pattern.
— A Moran's I near zero (not significant) implies a random spatial distribution.

This measure provides valuable insights into the spatial structure of the data, guiding further spatial statistical analysis and modeling.

Apply moran's I to the last predicted value of infection, we get a value of -0.7668855733863262. This shows that the distance and population values are negatively correlated with infection values.

As different time periods (days) have differing values, we can plot moran's I over a period of time to determine how it changes over time, allowing us to observe and hence further analyse how distance and population affects infection over time.

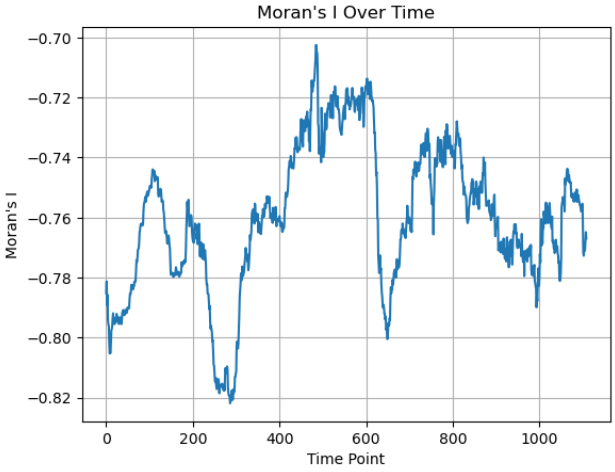


FIGURE 27 – Moran's I over time (days)

In general, the values of moran's I is negative and closer to -1 over the whole time period. We can infer that distance and population is always negatively correlated to infection. On certain days, moran's I is much lower.

For example, the day with the lowest moran's I is day 286 with a value of -0.8218451720232388. The prediction values for this day are : [2279, 1736, 1509, 1422, 9666, 1831] (each element in the list represents a region's predicted infection value) On this day, the predicted infection values were exceptionally high compared to day 484, which has the highest moran's I of -0.7025043726522623 and low prediction values of [463, 439, 194, 606, 3158, 796].

From the value and the plot, we can determine that the infection spread is more dispersed over the regions.

8 Windowing & R0

At this stage, we have a model that gives satisfactory results. However, one very important parameter is missing from this equation : the basic reproduction rate R0(t). This variable is the manifestation of the environmental, biological and political effects that influence the spread of the virus.

8.1 Windowing

The α coefficients are determined from a database. Considering socio-political and climatic changes, whose consequences on the number of cases can be seen, the α's change continuously. For this reason, the model with stationary coefficients is not sufficient.

We have to introduce non-stationnary coefficients. The precedent expressions of the model become :

$$Z(t) = \sum_{k=1}^{L} \alpha(k,t) Z(t-k)$$

$$\hat{\alpha} = \text{argmin}_{\alpha} ||Z(t) - \sum_{k=1}^{L} \alpha(k,t) Z(t-k)||_2^2$$

$$Z(t,c) = \sum_{k=1}^{L} \alpha(k,t,c) Z(t-k,c)$$

$$\hat{\alpha} = \text{argmin}_{\alpha} \sum_{c=1}^{6} ||Z(t,c) - \sum_{k=1}^{L} \alpha(k,t,c) Z(t-k,c)||_2^2 + \lambda \sum_{c,c'} \omega(c,c') ||\alpha(t,c) - \alpha(t,c')||_2^2$$

In reality, α's change from day to day. However, setting coefficients that depend on a single day creates a risk of overfitting. We have to work with windows of several days. Indeed, to determine $\alpha(t)$, we need to use data over the window period T ; [t - T/2 ; t + T/2].

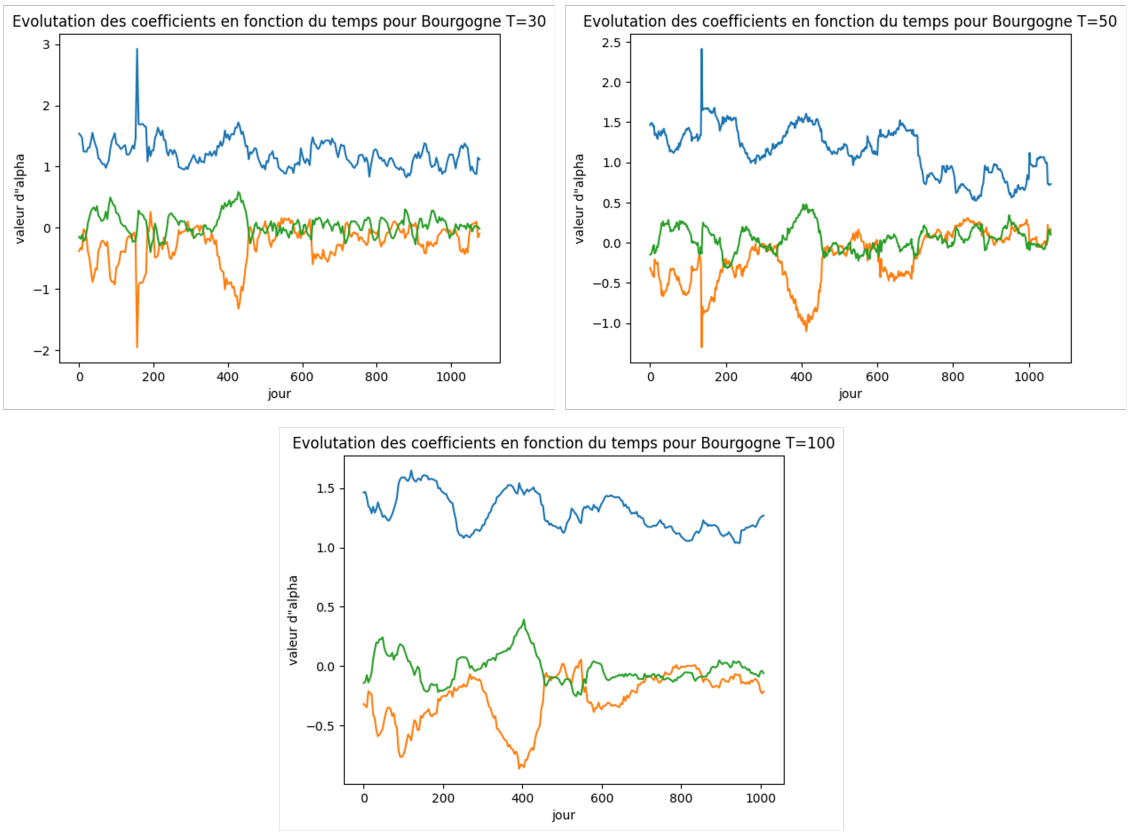


FIGURE 28 – Alpha Evolution for Bourgogne over time

This is the evolution of alpha for a window of period for perdiod T=30 ; T = 50 and T = 100. The variation in coefficients is further evidence of the non-stationarity of the time series.

As T increases, we notice a less noisy evolution of the coefficients. This is due to the fact that $\alpha(t)$ and $\alpha(t+1)$ will have close values, as they will be determined by a larger number of data. There will then be fewer abrupt changes and irregularities. In fact, noise is reduced : As the window increases, fluctuations and abrupt changes in α coefficients tend to diminish. This is due to the fact that the α coefficients are determined by a larger number of data, smoothing out temporary variations and random effects.

On the other hand, the longer the window, the more the coefficients separate and differentiate. This reveals more subtle variations and trends that are not apparent with a short window. One way of thinking about this is to consider the cyclical and seasonal effects of the data. In fact, increasing the window may allow the coefficients to capture these longer variations, which may result in larger gaps between coefficients.

T = 50 will be used for the next part.

8.2 R_0 definition

Remember that the R_0, or basic reproduction rate, indicates the average number of people that an infected person can contaminate.

Let assume that according to [3] :

$$\alpha(k,t) = R_0(t)\beta(k)$$

Then, because of the definiton of R_0, we could assume that the beta coefficient must be constrained to :

$$\sum \beta(k) = 1$$

Therefore, we can obtain directly R_0 :

$$R_0(t) = \sum \alpha(k,t)$$

Then we can create a time series with all the values of R0 based on the alpha calculation based on the real number of cases and a time series with all the values of R0 based on the alpha calculation based on the predicted number of cases. So, we obtain a kind of $R_{0_{\text{real}}}$ and $R_{0_{\text{predicted}}}$.

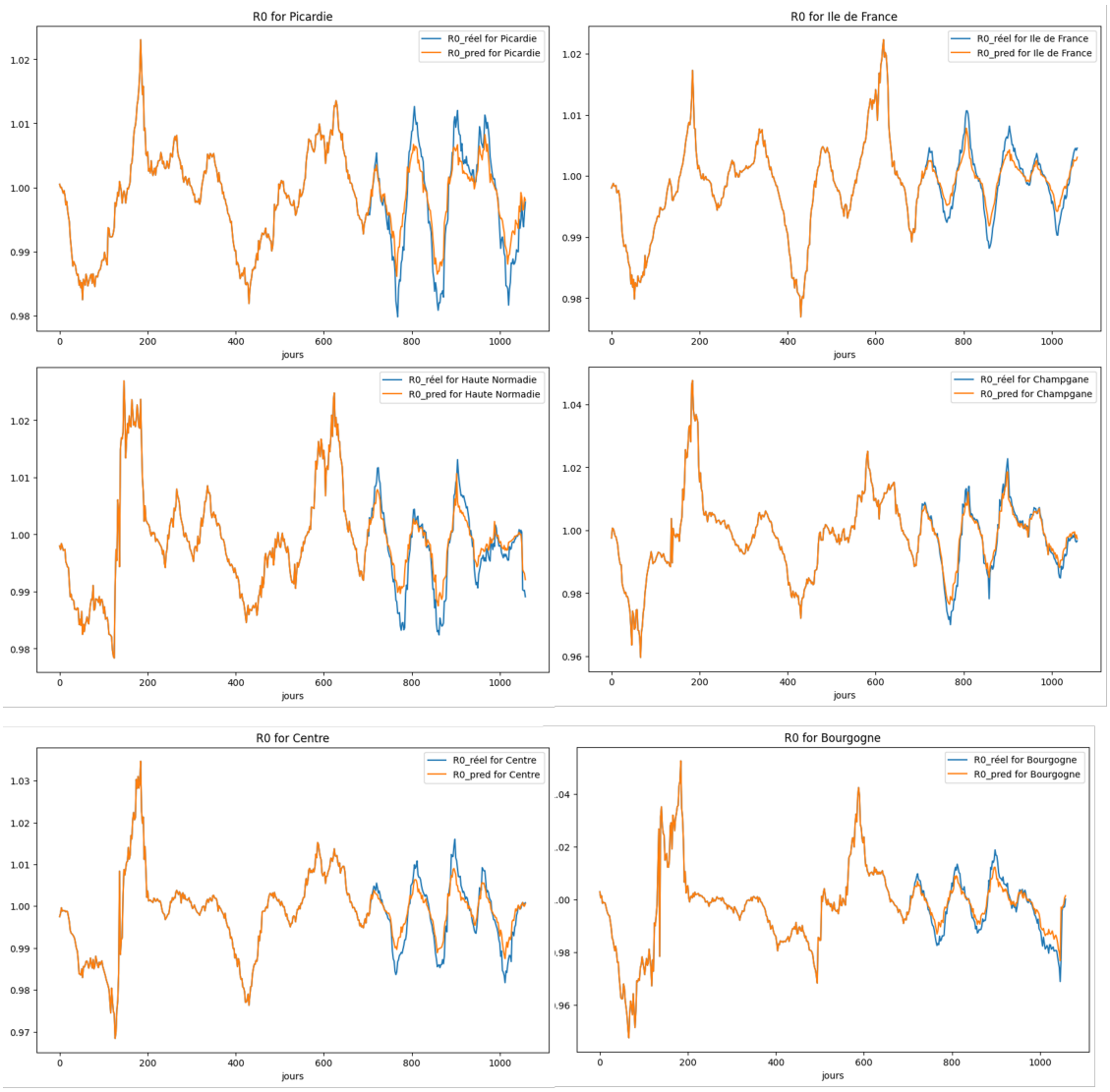


FIGURE 29 – R_0 for each region

Here are the R_0 predictions for each department. Note that the predictions have a certain lag time to follow the variations, but overall the results are satisfactory.

We can also see that R_0 values are always relatively close to (consistent), bearing in mind that if $R_0 > 1$, the epidemic is expanding, and if the opposite is true, the epidemic is slowing down.

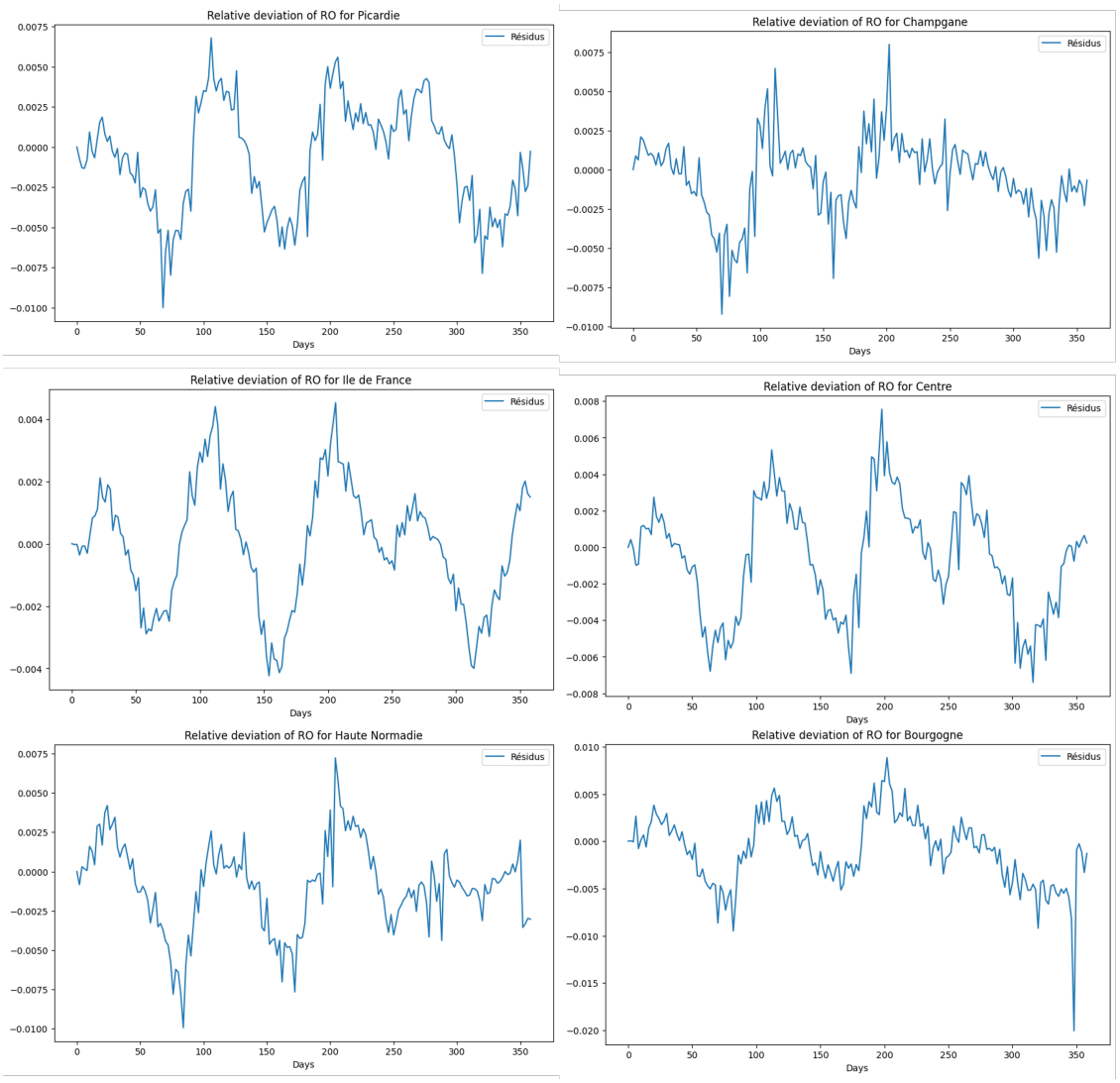


FIGURE 30 – Relative Deviation of R_0 for each region

Here are the R_0 relative deviation $(R_{0_{\text{real}}} - R_{0_{\text{predicted}}})/R_{0_{\text{real}}}$ for each department. Note that the relative deviation does not exceed 1%.

Note :$R_{0_{\text{real}}}$ is not really real, it's just an estimation based on our assumption, our model and the real number of cases of the database.

9 Conclusion

9.1 Summary

First and foremost, as we learned from the government's televised interventions throughout the Covid-19 period, determining R0 is a crucial measure for assessing the spread of the virus and making informed public health decisions. Because by understanding how the virus spreads in a population, it is possible to put in place effective preventive measures and control strategies to reduce trans- mission.

The SIR model provides an estimate of R0 by analyzing observed infection and cure rates. The value created lies in the precision of these estimates, which can help predict the scale of the epidemic and assess the effectiveness of measures taken to contain it.

In parallel, the use of autoregressive models offers a complementary approach to determining R0. These models exploit historical data on infections and transmission behavior to predict future trends. They can take into account factors such as individual mobility, social distancing policies and other relevant vari- ables to refine R0 estimates. Value creation here lies in the ability to provide accurate and reliable forecasts, enabling decision-makers to take timely preventive action.

Incorporating different models to determine Covid's R0 also offers additional advantages in terms of robustness and cross-validation of results. By comparing the estimates obtained from the SIR model and the autoregressive models, any discrepancies or inconsistencies can be detected, enabling estimates to be refined and errors minimized. This adds value to the project by reinforcing the reliability of results and providing a better understanding of virus transmission dynamics.

9.2 Review and Perspectives

We could have focused more on alpha and its predictions than on the number of cases. It would be interesting to study predictions over a longer period - our model is based on a one-day prediction only - and to determine up to how many days the model is still relevant (at 95It would also be interesting to extend the model to the whole of France, although calculation times are already very long. In fact, we could considerably improve calculation times by using the least-squares method to determine the exact alpha matrix without using a minimization function and the minimize function of the spicy module. We could also seek to better define the adjacency matrix that represents the links between each region : take into account regular flows of people such as Paris-Bordeaux, for example. Finally, it would be relevant to extend the model into a stochastic model, using fish law in the autoregression. We could also add noise to our model.

10 References

[1] Draief, M., & Massouli, L. (2010). Epidemics and rumours in complex networks. Cambridge University Press.

[2] Cori, A., Ferguson, N. M., et al. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. American Journal of Epidemiology, 178(9), 1505-1512.

[3] Pascal, B., Abry, P., et al. (2022). Nonsmooth convex optimization to estimate the Covid-19 reproduction number space-time evolution with robustness against low-quality data. IEEE Transactions on Signal Processing.

[4] Amdaoud, M., Arcuri, G., & Levratto, N. (2020). Covid-19 : analyse spatiale de l'influence des facteurs socio-économiques sur la prévalence et les conséquences de l'épidémie dans les départements français.

[5] Ives, A. R., & Bozzuto, C. (2020). State-by-State estimates of R0 at the start of COVID-19 outbreaks in the USA.

[6] Fowe, E. P. (2022). Analyse comparative et de cohérence entre les modèles statistiques et par compartiments sur la prédiction des cas d'infections et décès de la COVID-19 au Québec.