The effect of extreme rainfall on COVID-19 surveillance, the case of New York State

Master's Degree in Physics of Complex Systems

Caruso Susanna Carmen

Supervisor: **Dr. Valdano Eugenio**
Prof. Gamba Andrea Antonio

Politecnico di Torino, Sorbonne Université
July 2024

For my beloved cousin Francesco,
may this work make you proud of me as You always have been,
may the memory of your kindness give me the strength to follow always the light,
may the anger for Your loss push me to be better everyday,
may this work keep our names together forever.

Abstract

This thesis examines the impact of extreme weather on COVID-19 testing rates in New York State. Weather anomalies in temperature or precipitation were identified and their effects on daily testing analyzed. Initial findings showed a reduction in tests on anomalous weather days, which were challenging to quantify. To quantify this impact, regression models were used, considering as main ingredients the weather conditions, day of the week, and the underlying trend of tests.

The findings highlighted the effect of precipitation on testing rates. A Generalized Linear Mixed Model (GLMM) revealed heterogeneous county responses to heavy precipitation, with test variations ranging from 1.8% to -22.6% with respect to the testing trend. Additionally, counties with higher prevalence of diabetes and obesity in general population correlated with greater reductions.

This study underscores the sensitivity of disease surveillance to extreme weather, providing insights for public health improvements. Future research should explore other countries and understand all the factors that explain this sensitivity to enhance global public health strategies.

Contents

1 Introduction **1**

2 Data Collection **4**
2.1 COVID-19 Data 4
2.2 Health-Demographic Data 5
2.3 Weather Data 6

3 Methods and Algorithms **7**
3.1 Prophet 7
3.2 Isolation Forest 8
3.3 Regression Models 11
3.3.1 Poisson Regression 11
3.3.2 Generalized Linear Mixed Model regression 12

4 Results and Discussion **14**
4.1 Anomaly Detection 14
4.2 Statistical Regressions 18
4.2.1 The Number of Tests of the Day Before 20
4.2.2 Trends of tests 23
4.2.3 The county effect 25

5 Conclusions **29**

A Trend function in Prophet **31**

B An extension: France **33**

Chapter 1

Introduction

At the end of 2019, the World Health Organization (WHO) was informed about cases of atypical pneumonia with unknown etiology in the city of Wuhan, China [14]. Within less than two months, the disease was recognised as epidemic and named COVID-19 (Coronavirus Disease 2019), shortly after COVID-19 escalated from a localized outbreak to a full-blown pandemic, spreading rapidly across continents.

COVID-19 is an infectious disease caused by the beta-coronavirus SARS-CoV-2, which affects mostly the lower respiratory system presenting symptoms such as dry cough, fever and fatigue. The severity of these symptoms varies among individuals, with some experiencing mild to moderate illness while others face life-threatening complications, particularly those with underlying health conditions such as cardiovascular disease, diabetes, or cancer [17].

Since the beginning of the pandemic several studies were conducted in order to predict future outbreaks and to understand the factors influencing the disease transmission. Part of the research focused on the relationship between weather variables, climate zones and the spread of the disease.

This build upon former research, which had studied how seasonal influenza and viral respiratory illness are related to climate variability [24]. Of particular interest is the potential influence of climate variables on host susceptibility, behavior, and virus survival in the environment.

Many studies have been undertaken to explore the potential correlation between weather variables and COVID-19. While many of these studies assert the existence of an effective correlation, between new cases of COVID-19 and meteorological variables like temperature and humidity, they don't agree on the direction of this correlation, i.e. positive or negative [14].

The aim of the work here presented is to investigate further this relationship,

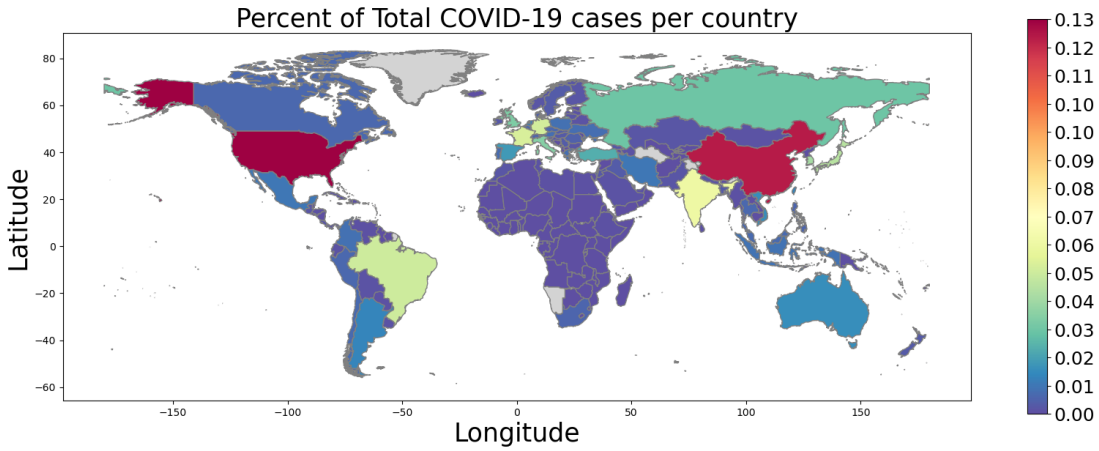

Figure 1.1: Color map according to the percentage of COVID-19 total cases that highlights the countries that contribute the most to the total number of COVID-19 cases worldwide.

with a particular focus on how the tests performed for COVID-19 may be linked to the weather conditions as well. This choice arises from the fact that the weather conditions, especially if they are extreme, may influence human behaviour and so the COVID-19 testing patterns. If these patterns are actually influenced by the weather conditions, then looking for direct link between incidence rate and climate becomes less significant. Then the human behaviour may play a decisive role.

The geographical area of research is New York state. Since, up to March 2023, USA is the country that registered most of the worldwide COVID-19 cases with approximately 13.3% of the total followed by China ($\approx$ 12.8%), India ($\approx$ 5.7%), France ($\approx$ 5.0%), and Germany ($\approx$ 4.9%) [2]. See *Figure* 1.1.

The initial recorded cases of COVID-19 in the United States dated back to January 2020, in Washington state. At the beginning the disease kept a relatively slow rate of transmission until March of the same year, hitting its first peak in early April 2020 [25]. By the end of March 2020, New York State had recorded 75795 cases of COVID-19 and 1550 related deaths [19] with the majority of cases concentrated in the metropolitan area of New York City, establishing itself as the epicenter of the pandemic in the United States [13]. The choice of the geographic area is favored also by the large presence of data both for COVID-19 testing, social-demographic parameters and weather variables.

After the data collection, a preliminary analysis was conducted using Machine Learning algorithms in order to verify the actual existence of a link between extreme

weather events and COVID-19 tests. Once shown this occurrence, some regression models were trained to quantify the impact of extreme conditions on the number of performed daily tests. Considering that the human response to an extreme event can be related to social and demographic factors, mixed regression models were also used in order to account for differences of these factors within the geographical regions studied. Eventually, this work therefore aims to identify the weather factors that can undermine the epidemiological surveillance of circulating respiratory pathogens, providing valuable insights for public health.

Chapter 2

Data Collection

2.1 COVID-19 Data

The data related to COVID-19 are taken from two different sources. In particular, data with high spatial resolution (ZCTA [1] resolution) were found in Badr et al. [2], this dataset contains the total tests and the positive tests[2] by ZCTA for the Metropolitan Area of New York City (NYC), carried out during the period from 2020-03-31 to 2023-03-30. In the dataset there are 179 ZCTA which are distributed across NYC's counties as follows: 26 for Bronx, 37 for Kings, 45 for New York, 59 for Queens, 12 for Richmond. For the former dataset some cleaning actions were made. Specifically the daily new tests were modified when their value was equivalent to the cumulative one, namely the new tests values were substituted by the difference between the corresponding and the previous cumulative data. Furthermore the rows containing negative values of new cases have been removed. Visualizing the test data one can observe some spurious outliers. In more details: the sharp peak on the 2021-08-01 has been deleted (both for daily new tests and positive tests), given the fact that also the cumulative tests showed the same jump; the sharp peak on the 2021-06-10 has been deleted, because it had no correspondence with the positive tests carried out on the same day; the relative maximum on the 2022-06-28 has been substituted by the difference between the corresponding cumulative data and the former one.

The data regarding tests at the county resolution were sourced from Chief Data Officer [10], for which the dates were moved to the day earlier, as they referred to the update.

[1] ZIP Code Tabulation Areas

[2] Positive tests, including hospitalised cases and home confinement

At the end of the data collection and after some preliminary analysis, the data with ZCTA resolution were discarded. The primary issue with this data was the frequent missing entries for several days, which resulted in an unrealistic weekly trend in the number of tests, with a unexpected peak observed on Sundays. Indeed, as one can see from the following Figure 2.1, the tests with the higher reporting delay in terms of days were registered on Sundays.

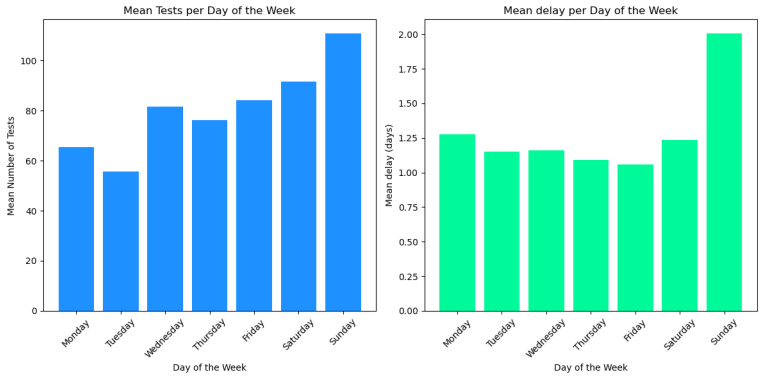

Figure 2.1: The left panel represents the mean number of tests conducted per day of the week in the metropolitan area of NYC by ZCTA. It shows an unusual increase in the number of tests as the weekend approaches. This trend can be attributed to the fact that many tests recorded on Sunday show an higher reporting delay, referring to the right panel. In detail, there are missing entries on the previous days, and it is likely that cumulative data for these missing entries were logged on Sunday, leading to the observed spike.

2.2 Health-Demographic Data

The data for the demographic parameters are taken from the United States Census Bureau [8],[4],[7],[6],[9],[5]. For each county and ZCTA were selected the data about: Total Population; Median Income; Percent of population below the poverty level [3]; Racial composition; Percent of population under 5 years of age; Percent of population

[3]The Census Bureau determines poverty status based on income thresholds, which takes into account the family size and the composition. If a family's total income falls below this threshold, the family and all its members are considered in poverty. These thresholds are adjusted with respect the inflation but they are the same independent of the geographical regions. Poverty status is assessed based on money income before taxes and excludes capital gains or non-cash benefits.

over 60 years of age; Median age; Sex ratio for 100 females. For some geographic area the median income reported was *'250000+$'*, this was substituted by *300000*$.

While the health data are drawn from Disease Control and Prevention [11] and (CHIRS) [1], which includes: prevalence of diabetes in the general population; prevalence of high blood pressure [4]; percent of people obese (BMI>30).

2.3 Weather Data

The Weather data utilized in this study were obtained from the nClimGrid-Daily product provided by NOAA's National Centers for Environmental Information (NCEI)[16]. The data about the average temperature, the minimum temperature, the maximum temperature and precipitation for a time span from January 1, 2026, to September 30, 2023, were drawn from the archive.

[4]In particular the prevalence of high blood pressure is defined as the age-adjusted percentage of respondents who reported they had been told by a health professional they had high blood pressure.

Chapter 3

Methods and Algorithms

3.1 Prophet

For the analysis, the data were pre-processed using a seasonal-trend decomposition algorithm. The *Prophet* algorithm, developed by *Facebook* [23], was used for this study. *Prophet* is a forecasting tool able to handle time series data through an additive model where non-linear trends are fitted with yearly, monthly, weekly, daily, and even hourly seasonality. Further it can also treat holiday effects and special events. The additive models fit very quickly. The algorithm uses Stan's L-BFGS [1] to find a maximum a posteriori estimate for the model fitting. Moreover it is robust to missing data and shifts in the trends, and it effectively handles outliers. *Prophet* decomposes the time series data into various components:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

Where:

- $g(t)$ is the trend, which is the non periodic part of the time series. The trend can include change-points, i.e. the points where the slope of the trend changes.
- $s(t)$ represents the seasonal part, which captures the periodic effects at different frequencies.
- $h(t)$ denotes the holiday events that might impact the time series data.

[1] Limited-memory BFGS (L-BFGS) is an optimization algorithm that approximates the inverse Hessian matrix using a small, fixed number of past updates. It efficiently updates this approximation iteratively to guide the search direction for minimizing a function. L-BFGS is commonly used for large-scale optimization problems where storing the full Hessian matrix is impractical.

- ϵ_t represents the residuals, or error terms.

The *Prophet* algorithm was first applied to weather variables such as maximum temperature (T_{max}), minimum temperature (T_{min}), and precipitation ($Prec$) in order to separate seasonal patterns from extreme or uncommon weather events. For the weather variables, the most interesting parts after applying the *Prophet* algorithm are the residuals. Larger residual values capture events that are uncommon for that specific period or season. These residuals were used both to associate each data point to a categorical variables and to implement the *Isolation forest* algorithm to detect anomalies. The classification was done in the following way. A weather variable w is classified as:

- A if $\text{residual}(w) > 0.75 \times \max(\text{residual}(w))$,
- B if $\text{residual}(w) \in (0.25 \times \max(\text{residual}(w)), 0.75 \times \max(\text{residual}(w)))$,
- and C otherwise.

Then these results were utilized to train the regression models. The *Prophet* algorithm was also used to extract the trend from the COVID-19 tests data. Indeed, considering that the number of tests daily performed could be influenced by several factors, but first of all from the epidemic situation itself, one has to take into account the epidemic driver. Hence, the algorithm was used to extract three kinds of trend, obtained thanks to the tuning of the *changepoint prior scale* parameter, which controls the flexibility of the trend $g(t)$. See *Appendix A* for a more detailed explanation. For the statistical analysis the following trends were extracted:

- Less sensitive T_1
- Mid sensitive T_2
- Very sensitive T_3

The *Figure 3.1* displays the three trends selected for the 9 most populated counties.

3.2 Isolation Forest

Isolation forest is an algorithm for data anomaly detection. It uses binary trees to separate the outliers from the rest. The core of the algorithm is the fact that the

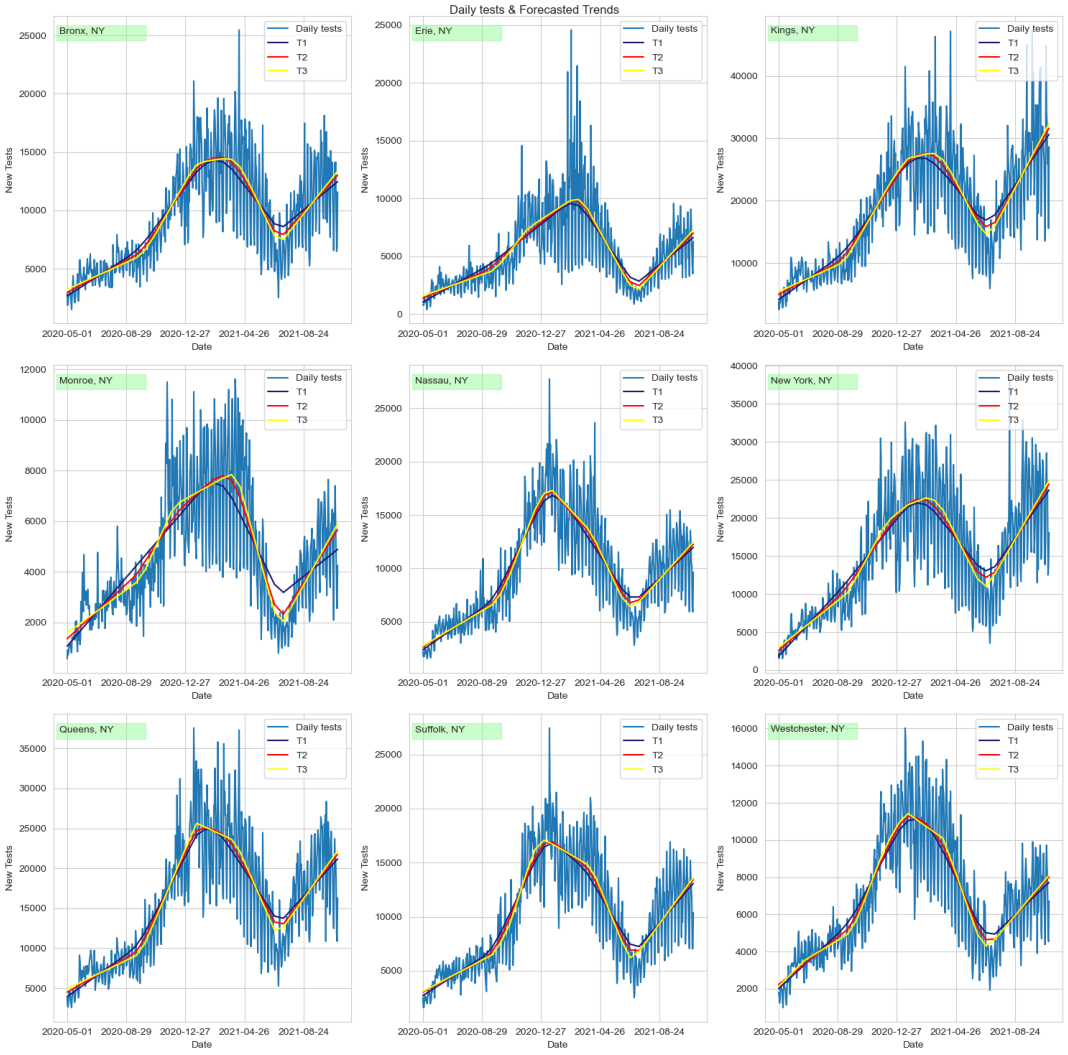

Figure 3.1: The nine panels represent daily tests performed in the period between 2020-05-01 and 2021-11-01 of the nine most populated counties of New York. Each panel also displays the different trends chosen for the analysis T_1, T_2, T_3.

anomalous points are easy to isolate. It works by creating recursively partitions by randomly selecting a feature and a split value between the minimum and maximum values allowed for that feature. The recursive partitioning is represented by a tree structure, this is the reason of the name *Isolation Forest*. The number of partitions required to isolate a point is interpreted as the length between the leaf, which marks the end of the isolation process, and the root.

Given the dataset $X = x_1, x_2, ..., x_n$ where each x_i has d-dimensions with $i \in \{1, 2, ..., n\}$, the Isolation tree (iTree) is defined with the following properties:

- A node T in the tree can be:
  - an external node with no children
  - an internal node with two child nodes T_l and T_r.
- for each node T is chosen:
  - an attribute q
  - a split value p between the maximum and the minimum of the attribute q.

When the iTree is fully grown, each data point of X can be found at one of the external nodes. Hence for each node, the path length $h(x_i)$ can be defined. The anomalous data points are the ones with smaller $h(x_i)$, so again the ones that are easier to recognize. Thanks its low memory requirements and the linear time complexity the algorithm has a simple implementation. The algorithm utilized is provided by Pedregosa et al. [18], it allows to specify parameters such as the number of estimators, the number of samples to draw from X to train each base estimator, and the contamination level, which represents the user's perception of the proportion of outliers within the dataset.

The *Isolation Forest* algorithm assigns a binary variable to each data point, $+1$ to regular points and -1 to outliers or anomalies.

In order to investigate the impact of extreme weather events on daily number of tests for COVID-19, one should analyze how the distribution of daily performed tests varies during a weather anomaly. To achieve this, the relative change of number of tests, denoted as $\rho(t)$, is defined as:

$$\rho(t) = \frac{\tau(t) - \tau(t-1)}{\tau(t-1)} \tag{3.1}$$

where $\tau(t)$ is the number of tests performed at day t. In this work the *Isolation Forest* was applied to the residuals obtained from the *Prophet* algorithm on the

weather variables. This approach was used to classify these residuals in terms of anomalies.

3.3 Regression Models

3.3.1 Poisson Regression

Poisson regression is a statistical model belonging to the family of generalized linear models (GLMs). It is designed to model count data, i.e. the number of occurrences of an event. This regression model assumes that the expected value of the response function $\boldsymbol{Y}$ follows a Poisson distribution. Considering a dataset $\boldsymbol{X}$ with n observations and m features, $\boldsymbol{X} \in \mathbb{R}^{n,m}$, where each row represents an independent variables $\boldsymbol{X_i} \in \mathbb{R}^m$ with $i \in \{1, 2, 3, ...n\}$, and m represents the number of features of each variable. As customary, an exponential link it is used for the Poisson regression, which assumes that the expected value of the dependent variable $\boldsymbol{Y} \in \mathbb{R}^n$ is the exponential of a linear combination of the regressors:

$$\log(\mathbb{E}(Y_i|\boldsymbol{X_i})) = \alpha + \boldsymbol{\beta x_i}$$

where $Y_i \in \boldsymbol{Y}$ and $i \in \{1, 2, ..., n\}$, $\boldsymbol{\beta} \in \mathbb{R}^{1,m}$ is the vector of coefficients and $\alpha \in \mathbb{R}$ the intercept term. To understand how the features affect the dependent variables Y_i one has to find the vector $\boldsymbol{\beta}$ and make reasoning on the sign and on the order of magnitude of each coefficients. The expected value of the i^{th} sample is given by:

$$\mathbb{E}(Y_i|\boldsymbol{X_i}) = \lambda_i = e^{\alpha + \boldsymbol{\beta x_i}}$$

Which has a Poisson distribution:

$$P(Y_i|\boldsymbol{X_i}) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

To find $\boldsymbol{\beta}$ one has to maximise the Likelihood $\mathbf{L}(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X})$:

$$\mathbf{L}(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{\boldsymbol{\beta x_i} + \alpha}} e^{y_i(\alpha + \boldsymbol{\beta x_i})}}{y_i!}$$

Or, thanks to the convexity of the logarithmic function, one can maximize the Log-likelihood $\log \mathbf{L}(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X})$:

$$\log \mathbf{L}(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} (-e^{\alpha + \boldsymbol{\beta x_i}} + y_i(\alpha + \boldsymbol{\beta x_i}) - \log y_i!)$$

Hence,

$$\frac{\partial}{\partial \beta} \log \mathbf{L}(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} (-x_i e^{\alpha+\boldsymbol{\beta}\boldsymbol{x_i}} + y_i \boldsymbol{x_i}) = 0$$

By solving this equation for the regression coefficients $\boldsymbol{\beta}$, one obtains the Maximum Likelihood Estimate (MLE) for $\boldsymbol{\beta}$. Rather than by hand computation, an iterative method like Iteratively Reweighted Least Squares (IRLS), implemented in the Python package *statsmodels* and in the R package *glm* was used. The Poisson regression algorithm was trained with different models listed in the following subsections [22].

3.3.2 Generalized Linear Mixed Model regression

The generalized linear mixed model (GLMM) is an extension of the generalized linear models, such as the Poisson model, where both fixed and random effects are taken into account.

The GLMMs are useful to handle multilevel or grouped data. The use of random effects allows to estimate variability at different levels of the hierarchy. Similar to generalized linear models (GLMs), the expected value of the response function is linked to the linear predictor through a link function g. Consider m groups and n_i observations for each group. In general it is assumed that the random effects are independently normally distributed with zero mean and some covariance matrix $\boldsymbol{\Sigma}$. While the response Y_{ij}, conditioned on the random effect, is assumed to be distributed according to the exponential family f with $1 \leq i \leq m$ and $1 \leq j \leq n_i$. Hence the generic formulation for the GLMM is:

$$Y_{ij}|\boldsymbol{U_i} \backsim f_{Y_{ij}|\boldsymbol{U_i}}(y_{ij}|\boldsymbol{u_i}),$$

$$\boldsymbol{U_i} \backsim N(0, \boldsymbol{\Sigma})$$

where $\boldsymbol{U_i} \in \mathbb{R}^{1,q}$ with q the number of random effects and with a natural parameter η_{ij} for the exponential family distribution $f_{Y_{ij}|\boldsymbol{U_i}}$ defined as:

$$\eta_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T \boldsymbol{U_i}$$

Here $\boldsymbol{X}_{ij} \in \mathbb{R}^p$, where p is the number of predictors with coefficients $\boldsymbol{\beta}$ (fixed effects) and $\boldsymbol{Z}_{ij} \in \mathbb{R}^q$ with coefficient U_i (random effects). Then the expected value of the response function Y_{ij} is given by:

$$g(\mathbb{E}(Y_{ij}|\boldsymbol{U_i})) = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \boldsymbol{Z}_{ij}^T \boldsymbol{U_i}$$

where for g, in the case of the Poisson regression, the exponential link is normally chosen. For the work here presented, it was assumed that the expected value of response function $\lambda_{ij} = \mathbb{E}(Y_{ij}|\boldsymbol{U_i})$ is distributed according to a Poisson distribution. Hence,

$$P(Y_{ij}|\boldsymbol{U_i}) = \frac{e^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}}{y_{ij}!}$$

In order to find the fixed coefficients $\boldsymbol{\beta}$, and the variability of these coefficients between the groups described by $\boldsymbol{\Sigma}$, one must maximize the Likelihood $\mathbf{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z})$.

$$\mathbf{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}|\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}) = \prod_{i=1}^{m} \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} f_{Y_{ij}|\boldsymbol{U_i}}(y_{ij}|\boldsymbol{u_i}) f_{\boldsymbol{U}_i}(\boldsymbol{u}_i) d\boldsymbol{u}_i = \prod_{i=1}^{m} \int_{\mathbb{R}^q} \prod_{j=1}^{n_i} \frac{e^{-\lambda_{ij}}\lambda_{ij}^{y_{ij}}}{y_{ij}!} f_{\boldsymbol{U}_i}(\boldsymbol{u}_i) d\boldsymbol{u}_i$$

Here, $f_{\boldsymbol{U}_i}(\boldsymbol{u}_i)$ represents the distribution of the random effects. In general, this maximization problem does not have closed form, and integrating over the random effects is usually extremely computationally intensive. To address this issue, most statistical software programs, including R which was used in this study, employ the Laplace approximation to implement this model [3].

Chapter 4

Results and Discussion

4.1 Anomaly Detection

The *Prophet* algorithm was applied to the weather variables of T_{min},T_{max}, $Prec$ for the period from 2016-01-01 to 2023-09-30. The residuals obtained were used to implement the *Isolation Forest* for a shorter period, specifically from 2020-05-01 to 2022-05-01. This time period was chosen to avoid moments in which the surveillance on COVID-19 was weak, like at the beginning of the outbreak and far enough from the beginning. To account for extreme weather events, such as heatwaves and extreme cold, the residuals of T_{min} and T_{max} were processed as follows: negative residuals of T_{min} and positive residuals of T_{max} were retained, while all other residuals were set to zero.

Firstly the most populated county of New York, i.e. Kings, was taken in consideration. *Figure 4.1* illustrates the daily test time series for Kings County, with colored dots indicating days when a weather anomaly occurred, as described in the legend. According to (3.1), the relative change in the number of tests performed were computed for different scenarios across all days of the week:

- $\rho_0(t)$ baseline, neither on day $t-1$ nor on day t an anomaly occurred;
- $\rho_{tmin}(t)$, on day $t-1$ no anomaly occurred while on day t an anomaly of T_{min} occurred, extreme cold;
- $\rho_{tmax}(t)$, on day $t-1$ no anomaly occurred while on day t an anomaly of T_{max} occurred, extreme hot;
- $\rho_{prec}(t)$, on day $t-1$ no anomaly occurred while on day t an anomaly of $Prec$ occurred, extreme rainfall;

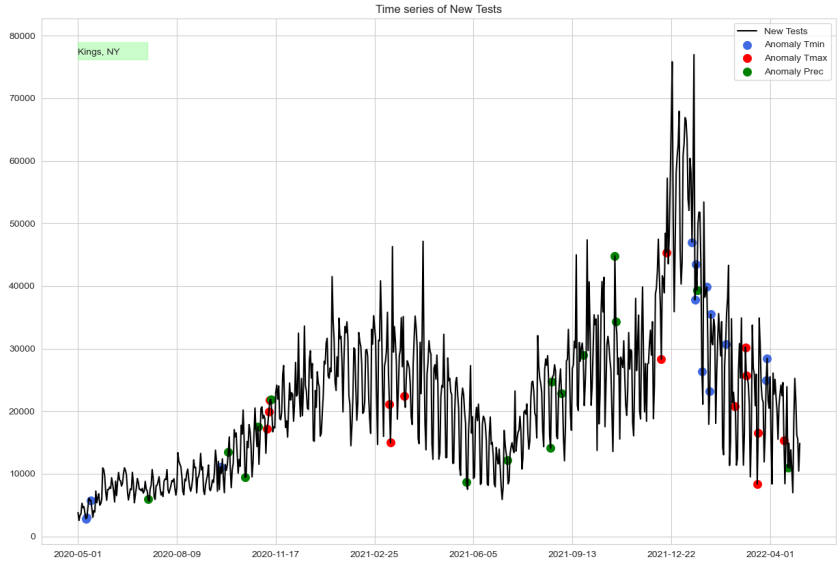

Figure 4.1: The figure illustrates the daily test time series for Kings County. The colored dots indicate when the weather anomalies occurred identified using the Isolation Forest algorithm on the residuals of weather variables after seasonal decomposition.

The boxplot [1] in *Figure* 4.2 displays distribution of the different ρ's. One can observe, how in particular for ρ_{tmax} and ρ_{prec} the distribution is shifted towards smaller values of ρ's.

ρ_0	ρ_{tmin}	ρ_{tmax}	ρ_{prec}
-0.007	0.014	-0.144	-0.096

Table 4.1: The table shows the medians of the distribution of ρ's. Except for ρ_{tmin}, referring to an event of extreme cold, the medians of ρ_{tmax} and ρ_{prec} are smaller than the one of ρ_0, indicating a reduction of tests performed.

The effect of an anomaly of T_{max} and $Prec$ is a reasonable slightly reduction of the number of tests performed during the day of the anomaly. While the effect of an anomaly of T_{min} is not very clear, moreover it could be contaminated by the presence of seasonal influences during the colder months, that push people to perform

[1] The boxplot is a tool to show graphically the locality, spread of numerical data through their quartiles. The box extends from the first quartile (Q1) to the third quartile (Q3) of the data, with a line at the median. The whiskers extend from the box to the farthest data point lying within 1.5x the inter-quartile range (IQR) from the box. Flier points are those past the end of the whiskers.

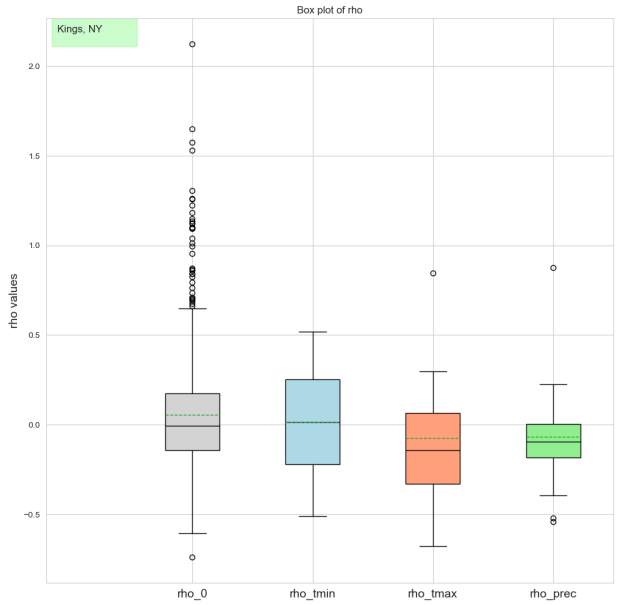

Figure 4.2: Boxplot for the distribution of ρ's for the Kings county. As expected in case of no anomalies the distribution of ρ_0 is almost centered on zero, given that all the days of the week are considered. Instead the distributions of ρ_{tmax} and ρ_{prec}, are slightly shifted to more negative values, indicating a reduction of the number of test compared to the baseline (no anomalies). The black line in each box represents the median.

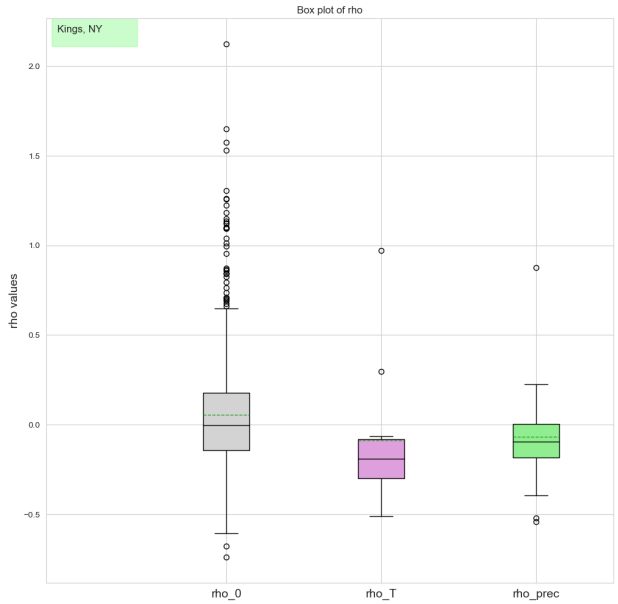

Figure 4.3: Boxplot for the distribution of ρ's for the Kings county. The figure displays a shift in the distribution towards smaller values of ρ, both in case of an unusual temperature and in case of an extreme rainfall. The black line in each box represents the median.

more tests [2]. To further investigate the response to an anomaly in temperature, the *Isolation Forest* was applied to a 2-dimensional vector containing both the residuals of T_{min} and T_{max}.

Hence, one can define :

- $\rho_T(t)$, on day $t-1$ no anomaly occurred while on day t an anomaly of T_{min} or T_{max} occurred;

The resulting boxplot, see *Figure* 4.3, shows a more clear signal about the effect of an atypical temperature.

These results account for all days of the week in Kings County.

However, it can be interesting to understand if the specific day of the week amplify or shrink the effect, and whether similar results are observed in other counties. Then, the same kind of approach was applied to the 35 most populated counties in New

[2] *Insight*: A limitation of this approach is the fact that, according to their definition, ρ_{tmin} accounts mostly for the colder seasons and ρ_{tmax} for the warmer ones. Instead ρ_0 does not have this kind of unbalance.

ρ_0	ρ_T	ρ_{prec}
-0.002	-0.191	-0.096

Table 4.2: The table shows the medians of the distribution of ρ's. Both the medians of ρ_T and ρ_{prec} are smaller than the one of ρ_0, indicating a reduction of tests performed when a weather anomaly occurs.

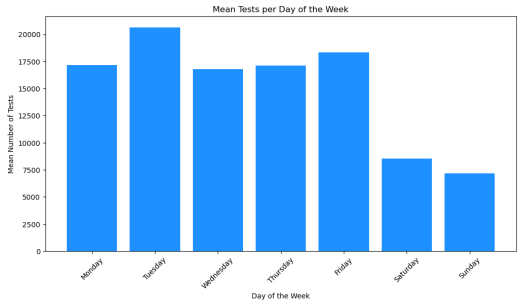

Figure 4.4: Mean number of tests performed for each day of the week, for the 35 most populated counties.

York State. By viewing the average number of tests performed for each day of the week, as shown in *Figure* 4.4 , it can be observed that the number of tests decreases sharply at the weekend, while for the weekdays the higher mean is registered on Tuesday.

Take Wednesday as the reference day. On average fewer tests are performed on Wednesdays compared to Tuesdays. By computing the ρ values, one can determine whether an anomaly on Wednesday amplify or mitigate this reduction in testing.

The overall effect of the anomaly is to increase the reduction of tests performed on Wednesday compared to Tuesday, when an atypical event occurs on Wednesday. This is clear to a greater extent for ρ_{tmin}, ρ_{tmax} and ρ_T, while the distribution of ρ_{prec} results wider, see *Figure* 4.5.

In summary, the results of this section suggest that anomalous weather conditions generally lead to a reduction in surveillance activities. However, in some cases these results are challenging to interpret and quantify.

4.2 Statistical Regressions

To overcome the challenges of the previous section and quantify the impact of weather events, the analysis was extended to include statistical regression methods. The aim

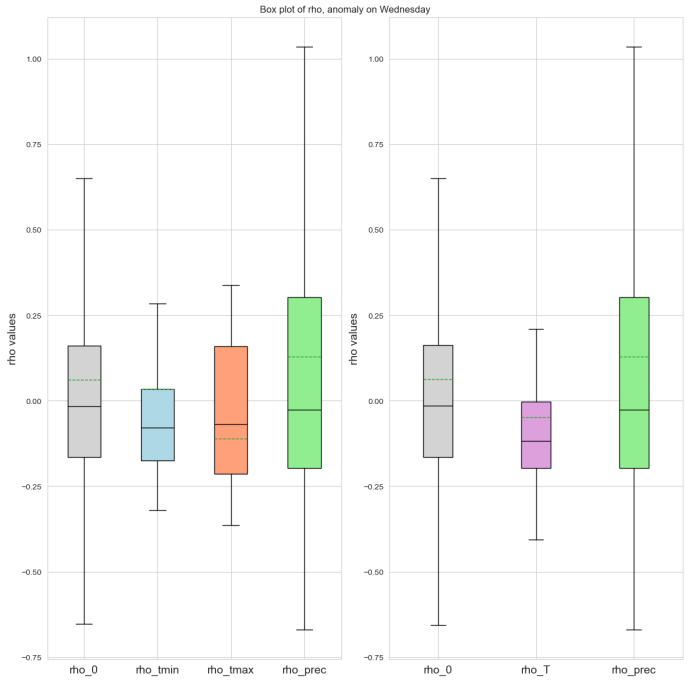

Figure 4.5: Boxplot for the distribution of ρ's for the 35 most populated counties in New York State. The left panel displays the distributions of ρ for anomalies in T_{min} (extreme cold) and T_{max} (extreme heat) separately. The right panel shows the distribution of ρ_T, which combines the temperature anomalies. The temperature anomaly seems to have an amplifying behaviour on the reduction in tests of Wednesday compared to Tuesday. While the effect of precipitations appears less clear, and the distribution of ρ_{prec} is wider. The black line in each box represents the median value.

of this approach is to estimate the relationship between the dependent variable, the number of tests daily performed, and independent variables such as weather conditions, weekly effects, and other relevant factors.

The number of tests daily performed can be interpreted as counts data, which may depend on various factors including weather conditions, day of the week and the individual's awareness of the epidemic's spread. In particular, weather anomalies such as extreme temperatures or heavy precipitation can affect people's behavior, potentially leading to changes in the number of tests performed. Additionally, the day of the week can introduce systematic variations. Finally, public awareness and concern about the epidemic, driven by factors such as media coverage, public health announcements, and the reported number of cases, can significantly impact testing behavior.

To model count data with these three main factors, Poisson regression was used. In the following models, weather variables and the week effect were considered as predictors, while public awareness was treated as an offset. Different kinds of offset were chosen in order to reflect the underlying rate of testing, which depends on the consciousness of people, probably the season (academic years, holidays) and from the epidemic itself.

4.2.1 The Number of Tests of the Day Before

The initial approach to model individual's awareness was to consider the *Number of Tests of the Day Before* as an offset. It was assumed that the expected value of tests performed on day t was given by:

$$\mathbb{E}(N_t) = N_{t-1} e^{\alpha + \boldsymbol{\beta}\mathbf{c}_t + \boldsymbol{\gamma}\mathbf{w}_t}$$

where:

- N_t: Number of tests performed on day t.
- α: Intercept.
- $\boldsymbol{c}_t$: Weather variables with corresponding vector of coefficients $\boldsymbol{\beta}$, on day t.
- $\boldsymbol{w}_t$: Week variables with corresponding vector of coefficients $\boldsymbol{\gamma}$, on day t.
- t: Index running over all observations, e.g. the days for the 30 counties.

The model was trained with different predictors for the period from May 1, 2020, to Nov 1, 2021 [3], for the 30 most populated counties. Dummy variables were used for the days of the week, denoted as $mon(t) = 1$ if the day t is Monday and 0 otherwise. The predictor $\delta(t)$ represents the variation in millimeters of rain between day t and $t-1$. Additionally, δ^+ and δ^- were introduced as follows:

$$\delta^+(t) = \begin{cases} \delta(t) & \text{if } \delta(t) > 0 \\ 0 & \text{otherwise} \end{cases} \qquad \delta^-(t) = \begin{cases} \delta(t) & \text{if } \delta(t) < 0 \\ 0 & \text{otherwise} \end{cases}$$

These cases were separated in order to understand what is the response in testing if today is raining more than yesterday, but also if there is a recovery of the tests "lost" yesterday when today's rainfall is less than yesterday's. This separation allows understanding any symmetry between lost tests during rainfall and the recovered ones. Finally, the weather variables of $T_{\min}$, $T_{\max}$, and $Prec$ were categorized using the categories described in 3.1, and their dummy variables were employed as predictors. For instance, if $Prec_A(t) = 1$, it indicates that an extreme rainfall event has occurred on day t; otherwise, it is set to zero.

The following *Table* 4.3 lists the values of Δ_N, defined as the percentage change in the number of tests compared to the previous day, for each predictor. For the binary variables, Δ_N was computed as $\Delta_N = e^b - 1$ where b is the corresponding coefficient of the binary variable resulting from the regression. While, for the predictors accounting for the variation in millimeters of rain, Δ_N was computed considering a 25mm of variation [4], meaning that in this case Δ_N refers to the percentage change of tests with respect to the day before when it rained 25mm more rain. Hence, $\Delta_N = e^{r \cdot 25} - 1$ where r is the corresponding coefficient of the rain variable resulting from the regression. [5]

From the results in the table, some considerations can be made. For instance, the 'Mon' predictor has a Δ_N value of approximately 33.3%, indicating that on Mondays, there were 33.3% more tests performed compared to Sundays, or that in case of extreme rainfall $Prec_A$ the test were reduced by $\approx 12.7\%$ with respect the former day.

[3] This period, which is shorter than the period used for anomaly detection, was chosen to avoid the later peak caused by the Omicron variant, which could result in totally different behavior.

[4] 25mm of rain was chosen as the unit of measure because, from the study of the precipitation time series, it emerged as a mid-point between extreme and light rainfall

[5] A p-value is the probability of obtaining test results at least as extreme as the observed results, assuming the null hypothesis (H_0) is true. The null hypothesis is the assumption that there is no effect, difference, or relationship between variables. It represents the idea that any observed differences or effects in the data are due to random chance rather than a true underlying effect. A low p-value (≤ 0.05) suggests rejecting the null hypothesis, while a high p-value (> 0.05) suggests failing to reject it.

	Intercept	Mon	Tue	Wed	Thu	Fri	Sat	δ
Model 1	2.5% (2e-16)	33.5% (2e-16)	23.7% (2e-16)	7.4% (2e-16)	1.3% (2e-16)	-10.4% (2e-16)	-24.5% (2e-16)	-3.0% (2e-16)
Model 2	-2.4% (2e-16)	33.5% (2e-16)	23.7% (2e-16)	7.5% (2e-16)	1.3% (2e-16)	-10.4% (2e-16)	-24.4% (2e-16)	
Model 3	-2.4% (2e-16)	33.1% (2e-16)	23.6% (2e-16)	7.4% (2e-16)	1.1% (2e-16)	-10.6% (2e-16)	-24.4% (2e-16)	
Model 4	-2.7% (2e-16)	33.3% (2e-16)	23.6% (2e-16)	7.5% (2e-16)	1.1% (2e-16)	-10.6% (2e-16)	-24.6% (2e-16)	

	δ^-	δ^+	$Prec_A$	$Prec_B$	$T_{max}A$	$T_{max}B$	$T_{min}A$	$T_{min}B$
Model 1								
Model 2	2.7% (2e-16)	-3.3% (2e-16)						
Model 3			-12.5% (2e-16)	2.7% (2e-16)				
Model 4			-12.8% (2e-16)	2.7% (2e-16)	-4.0% (2e-16)	2.3% (2e-16)	0.5% (0.0327)	-10.9% (2e-16)

Table 4.3: The table shows the values of Δ_N for the different predictors. The values in the brackets refer to *p-value* of the coefficients.

However, the results of the models (3-4) with categorical variables for weather conditions are not straightforward to interpret. Indeed, it is necessary to verify for each county under analysis what extreme rainfall or a temperature much lower or much higher than the seasonal average means.

Moreover, the performance of these models appears to be poor compared to models 1 and 2, as indicated by their higher AIC [6] values. Hence, for the rest of the analysis it was decide to focus on quantifiable predictors, i.e. not categorical, for the weather variables.

	AIC
Model 1	4834078
Model 2	**4833943**
Model 3	4843275
Model 4	4836712

Table 4.4: The table lists the AIC value for each model trained, using the *Number of Tests of the Day before* as offset

Model 2, which achieves the best AIC value, shows the most interesting results. It suggests that the number of tests decreases by 3.3% if there is 25mm more rain compared to the previous day, with the regression coefficient of δ^+ being $-1.343 \cdot 10^{-3}$. Conversely, the number of tests increases by 2.7% if there was 25mm more rain the previous day compared to the present day, with the regression coefficient

[6]The Akaike information criterion (AIC) is an estimator of the quality of the statistical model. It is defined as $AIC = 2k - 2ln(\mathbb{L})$, where k is the number of estimated parameters and $\mathbb{L}$ is the maximized likelihood function of the model. The preferred model is the one with the minimum AIC value.

of δ^- being $1.065 \cdot 10^{-3}$. Hence, the recovery of tests and the loss of tests are not symmetric, namely the recovery is smaller then the loss meaning that some tests are just lost. The lower panel of *Figure* 4.6 displays a line with a slope representing the Pearson correlation between the predicted counts [7] and the actual counts. The Pearson correlation coefficient can be interpreted as a parameter that accounts for the test performance. Even if the Pearson correlation appears high, taking as offset the *Number of tests of the day before* can lead to overfitting and diminish the effectiveness of the coefficients. For instance, if it rained yesterday and also today, the number of tests conducted yesterday might have already been influenced by that rainfall. Consequently, using this data as a reference for today's predictions could result in erroneous conclusions due to the fitting of the yesterday's rain in the offset.

4.2.2 Trends of tests

To avoid overfitting and remove the influence of weekly patterns and weather conditions on the testing rate, it was convenient to extract the underlying trends from the time series of tests. In particular the trends listed in 3.1 were used as offset for the Poisson regression models. Due to the simpler interpretation of changes in the number of tests in relation to millimeters of rain, the analysis was restricted to the weather variable related to precipitations. Two simple models were trained, using as offset the three trends T_1, T_2, T_3.

- Model A:

  $$N_t \sim \boldsymbol{w_t}$$

  the number of tests N_t depends only on the day of the week, where $\boldsymbol{w_t}$ is the vector of dummy variables representing the days of the week;

- Model B:

  $$N_t \sim r_t + \boldsymbol{w_t}$$

  the number of tests N_t depends on both the day of the week and the rainfall r_t in millimeters on day t.

Redefining Δ_N, which now represents the percentage change in the number of tests relative to the trend for 25mm of rain, the results of the two models are displayed in *Table 4.5*.

[7] It's important to note that the model's training and prediction were performed on separate datasets: training on a randomly selected 80% of the total dataset, and testing or prediction on the remaining 20%.

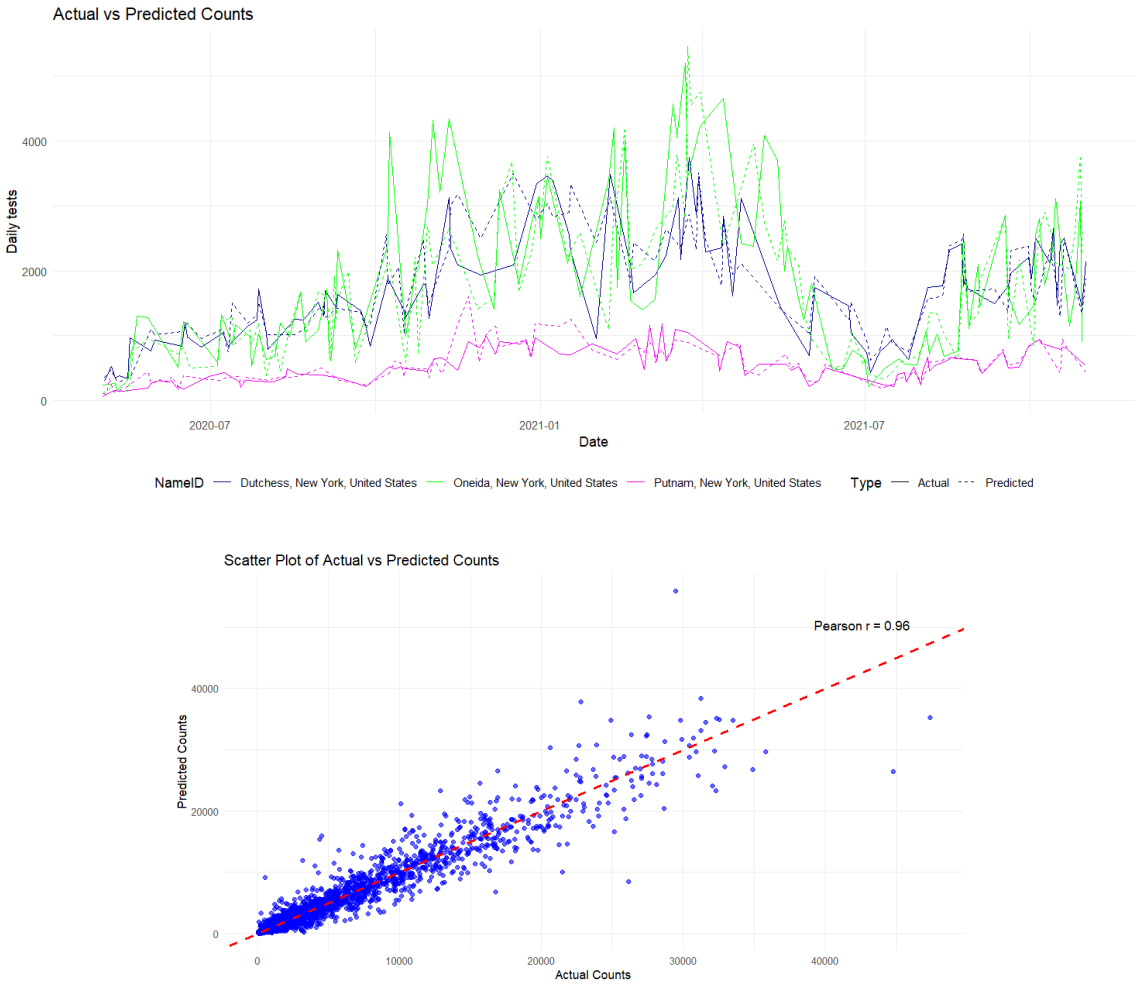

Figure 4.6: The upper panel illustrates the time-series data, plotting together the actual test values (solid line) with those predicted by **Model 2** (dashed line) across three distinct counties. The lower panel presents a scatter plot, mapping the relationship between predicted and actual data points across all 30 counties analyzed.

Trend	Model A	Model B
T_1	AIC=4393444	$\Delta_N = -5.4\%$ (2e-16), AIC= **4365417**
T_2	AIC=3669219	$\Delta_N = -5.1\%$ (2e-16), AIC= **3647683**
T_3	AIC=3176207	$\Delta_N = -4.8\%$ (2e-16), AIC =**3157658**

Table 4.5: Table of results for Model A and Model B. The values in round brackets refer to *p-value* of the precipitation coefficient.

As seen from the results of the models' training, Model B scores the best AIC values across all trends. This suggests that the amount of rainfall on a given day in some way explains the number of tests performed on that day. In particular the response of the daily performed tests to $25mm$ of rain results in a reduction of $\approx 5\%$ compared to the selected trend.

Up to this point, it was implicitly assumed that each county have the same behaviour in response to the rain or other weather condition. This is of course a limitation and can be overcome using a *a Mixed Effects regression model*, which takes into account the possibly existing differences in the response for different counties.

4.2.3 The county effect

This last subsection aims to incorporate in the previous models the variability in the response to weather conditions across different counties. To achieve this, a *Generalized Linear Mixed Model (GLMM)* was used. In the GLMM framework, both fixed effects and random effects are considered, which allows to capture the differences between counties.

Expanding on the previous results, the fixed effects include as predictors the day of the week and the daily millimeters of rain. To address the unique characteristics of each county, also random effects are introduced.

While the number of tests performed can largely vary from one county to another, this baseline variability is already taken into account by the offset, which is chosen to be the testing trends independently extracted for each county. Hence, no random intercept was considered. Instead more interestingly, the variability in the response to precipitation can be modeled by a random slope for the precipitation predictor. This allows the effect of precipitation on the number of tests to differ from one county to another, reflecting that some counties might be more sensitive to extreme or unusual rainfall than others. Finally, one can define the last model as:

- Model C:

$$N_t \sim \boldsymbol{w_t} + r_t + (0 + r_t|ID)$$

the number of tests N_t depends on both the day of the week and the amount of rainfall r_t. Additionally, the model includes a random slope for r_t to account for variability in the effect of rainfall across different counties. Here, the syntax of R is used to formalize the random effects, referring to the term $(0 + r_t|ID)$, where 0 represents the absence of random intercept for the reasons above, while the second term represents the random slope for r_t on the grouping variable "ID" which is the geographical ID for the different counties.

Using the definition for Δ_N of the previous subsection, the results of the Model C are listed in the *Table 4.6*.

Trend	Model C
T_1	$\tilde{\Delta_N} = -6.2\%$ (2e-7), AIC= 4306739
T_2	$\tilde{\Delta_N} = -4.8\%$ (5e-6), AIC= 3615250
T_3	$\tilde{\Delta_N} = -4.0\%$ (6e-5), AIC =3139788

Table 4.6: Table of results for Model C, where $\tilde{\Delta_N}$ denotes the mean of Δ_N over all the studied counties. The values in round brackets refer to *p-value* of the precipitation coefficient.

From the *Table 4.7* one can observe how the response to 25 millimeters of rain varies significantly across counties. While most counties exhibit a reduction in the number of tests, the magnitude of this reduction varies widely among them. Additionally, some counties' responses are influenced by trends, whereas others demonstrate coherent results across all three trends. See *Figure* 4.7. In order to understand the nature of this variability across counties, an analysis on the correlation between the Δ_N's and the socio-demographic parameters was conducted. Defining P_{Δ_N} the Pearson Correlation between the Δ_N and the socio-demographic parameters, the *Table 4.8* below presents the significant values of P_{Δ_N}.

The significance of the correlation results is highly dependent on the trend. However, the results suggest that higher is the median income of a county, then smaller is the reduction in tests due to precipitation. Conversely, poorer health conditions, such as a higher prevalence of diabetes and/or obesity, in a county correlate with a larger reduction in tests. Nonetheless, these considerations do not fully explain variability between counties, leaving this an open question.

County Code	**County Name**	$T_1(\Delta_N)$	$T_2(\Delta_N)$	$T_3(\Delta_N)$
US36001	Albany County	-7.4	-6.2	-5.6
US36005	Bronx County	-7.7	-7.3	-7.0
US36007	Broome County	-3.1	-1.8	-1.0
US36013	Chautauqua County	-13.5	-8.7	-4.6
US36027	Dutchess County	-1.6	-1.4	-1.3
US36029	Erie County	-6.4	-5.2	-4.8
US36045	Jefferson County	-9.9	-7.3	-5.4
US36047	Kings County	-5.8	-5.2	-4.8
US36055	Monroe County	-5.2	-3.1	-2.7
US36059	Nassau County	-6.1	-5.6	-5.5
US36061	New York County	-2.9	-2.4	-2.3
US36063	Niagara County	-6.3	-3.6	-2.0
US36065	Oneida County	-8.1	-5.1	-3.8
US36067	Onondaga County	-1.2	1.4	2.2
US36069	Ontario County	-7.6	-6.6	-6.3
US36071	Orange County	-4.4	-4.0	-4.0
US36075	Oswego County	-15.1	-11.4	-9.3
US36079	Putnam County	-5.1	-5.0	-5.1
US36081	Queens County	-8.0	-7.7	-7.5
US36083	Rensselaer County	-1.2	1.8	4.3
US36085	Richmond County	-10.8	-10.1	-9.7
US36087	Rockland County	-3.6	-3.1	-2.9
US36089	St. Lawrence County	-23.3	-22.6	-22.2
US36091	Saratoga County	-6.1	-4.9	-4.6
US36093	Schenectady County	-0.7	-0.2	0.2
US36101	Steuben County	-11.1	-10.4	-10.6
US36103	Suffolk County	-5.6	-5.3	-5.1
US36109	Tompkins County	-10.0	-0.8	8.9
US36111	Ulster County	-6.9	-6.4	-6.1
US36119	Westchester County	-6.2	-5.9	-5.7

Table 4.7: The table shows the values of Δ_N for each studied county, for the three trends T_1, T_2, T_3.

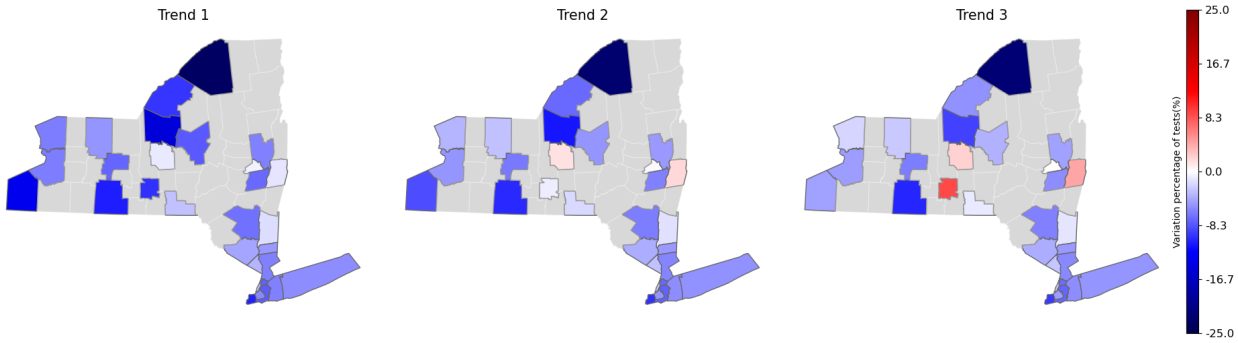

Figure 4.7: The figure shows the selected counties in New York State used for the analysis. The color map indicates the percentage change in daily tests per 25 mm of rain, for the three trends.

Trend	Median Income	Percent of people with Diabete	Sex ratio	Percent of people Obese
T_1	P_{Δ_N}= 0.36 (0.05)	P_{Δ_N}= -0.39 (0.03)	P_{Δ_N}= -0.42 (0.02)	
T_2		P_{Δ_N}= -0.38 (0.04)		P_{Δ_N}= -0.36 (0.05)
T_3				P_{Δ_N}= -0.36 (0.05)

Table 4.8: Table with significant values of P_{Δ_N}. The values in brackets refer to *p-value* of P_{Δ_N}.

Chapter 5

Conclusions

This work aimed to understand the reaction of the surveillance on respiratory disease in case of an extreme or anomalous weather conditions, with particular focus on COVID-19. The level of surveillance was quantified in terms of daily performed tests. The anomaly detection carried out on the weather variables T_{min}, T_{max} and $Prec$ to discern anomalous weather conditions from the normal one, allowed to get an insight about the testing rate during the anomaly. The results showed roughly a reduction in the number of tests performed the day of the anomaly. However, these results were challenging to interpret and mostly to quantify. Hence, once ensured the presence of an effective modulation of the surveillance in case of unusual weather condition, some regression models were used to weight this reduction. To model the daily counts of tests, three main factors were considered: the weather conditions; the day of the week; the public awareness about the epidemic status (incorporated as an offset). The results for models incorporating categorical weather variables, describing the magnitude of the anomaly, proved challenging to demystify, as the definitions of extreme weather events or deviations from seasonal norms may differ across counties. In this context the **Model 2** scored the best AIC value, proving that for 25mm of more rain the $\approx 3.3\%$ of tests are lost and only the $\approx 2.7\%$ of tests are recovered for 25mm of less rain. Despite the high performance of the **Model 2** the use of *the Number of Tests of the Day before* as offset can lead to over-fitting and reduce the effect of the weather conditions. Therefore others regressions models were trained, using as offset the underlying trend of tests, which encodes in some way the epidemic curve. The simple **Model A** was trained in order to see if the only week effect could explain the daily number of tests. This is not the case, given that the **Model B**, according to which for 25 mm of rain the tests are reduced by $\approx 5\%$, scored a better AIC values compared to **Model A** and **Model 2**. This suggests

that the level of precipitations is relevant to explain the daily tests. However, until this point a big approximation was done, i.e the response to the precipitations was considered the same across all the counties. This simplification does not respect the reality. The different counties can be more or less used to a certain amount of precipitations and/or can be more or less ready to face a weather anomaly. Thus to take into account the variability in the reaction of the counties, a *GLMM* model was trained, using the identity of each county as grouping factor. From the results of the **Model C** it was observed, that apart for some exceptions, most of the counties react to 25 mm of rain reducing the number of tests performed during that day. However the magnitude of the response varies widely going from $\approx 1.8\%$ to $\approx -22.6\%$ for the *mid sensitive trend* T_2, enforcing the idea that the actual people response to 25 mm of precipitations can be different from counties to counties. To explore the nature of these differences the correlation between some socio-demographic parameters and the response to 25 mm of rain was studied. Still the results were not very explanatory leaving the question open. The only conclusion that can be made is about the poor health conditions of a county, which can increase the reduction in the number of tests in case of large precipitations. Overall, the results presented can be useful to the public health to understand how the surveillance on a disease can be sensitive to the anomalous weather conditions. These insights can help identify the weak points of the surveillance system and guide improvements to ensure accurate disease monitoring even during adverse weather.

This study has some limitations, primarily due to the resolution of the data, as all analyses were conducted at the county level using mean weather values. This approach did not allow to dig more about the relationship between weather and surveillance, preventing the identification of other confounding factors. Future work could extend this analysis to other countries, to compare responses and understand whether different States exhibit similar behaviors or substantial differences. This comparative approach could help uncover underlying reasons for these variations and inform more robust public health strategies globally.

Appendix A

Trend function in Prophet

In the Prophet algorithm two trend models are implemented: the saturating growth model; the piece-wise linear model. For the purpose of this thesis only the second model was used. Specifically the trend function $g(t)$ of the additive model in 3.1 has the following form:

$$g(t) = (k + \boldsymbol{a}(t)^T \boldsymbol{\delta})t + (m + \boldsymbol{a}(t)^T \boldsymbol{\gamma})$$

where:

- k is the base growth rate;
- $\boldsymbol{\delta}$ is the vector of rate adjustment. Suppose that there are S changepoints which occur at times s_j, with $j = 1, ..., S$. The element δ_j represents the change in rate that occurs at time s_j;
- m is the offset parameter;
- $\boldsymbol{a(t)} \in \{0, 1\}^S$ is a vector indicating the presence of changepoints, defined as:

$$a_j(t) = \begin{cases} 1, & \text{if } t \geq s_j; \\ 0, & \text{otherwise,} \end{cases}$$

- $\boldsymbol{\gamma}$ is another parameter to adjust the offset parameter m in order to ensure the continuity of the function. Then, the correct adjustment at changepoint j is easily computed as: $\gamma_j = -s_j \delta_j$

The changepoints s_j can either be specified by the analyst, or selected automatically from a set of candidates. Automatic selection is achieved by placing a sparse prior on $\boldsymbol{\delta}$. Typically, the prior is set to be $\delta_j \sim \text{Laplace}(0, \lambda)$, where λ controls the

model's flexibility in altering its rate. In particular for the analysis of this thesis the parameter λ was tuned to select the trends T_1, T_2, T_3 with different adherence to the data. [23]

Appendix B

An extension: France

Building on the positive results obtained in Sections 4.2.2 and 4.2.3, the same methodology was applied to data from France. As before, the trend of tests was extracted from the data Santé publique France [21] using the *Prophet* algorithm, identifying three different trends based on flexibility. These trends are illustrated in *Figure B.1.*

The same models **A** and **B** were trained using the weather data obtained by Météo France [15]. Once again the results of *Table* 4.5 show that the **Model B**

Trend	Model A	Model B
T_1	AIC=12407084	$\Delta_N = -11.4\%$ (2e-16), AIC= **12365053**
T_2	AIC=11751103	$\Delta_N = -11.3\%$ (2e-16), AIC= **11710094**
T_3	AIC=10082328	$\Delta_N = -9.6\%$ (2e-16), AIC =**10053102**

Table B.1: Table of results for Model A and Model B. The values in the brackets refer to *p-value* of the precipitation coefficient.

scores a better performance than the **Model A**, suggesting that also in France, the daily number of tests performed can be explained by the millimeters of rainfall. Furthermore the **Model C** was trained to account for potential variability in the response of the number of daily tests to 25mm of rain across different departments.

As for the New York case, the results of **Model C**, see *Table B.2*, show a better performance compared to **Model B** for the corresponding trends. This indicates that different departments exhibit varied responses to extreme rainfall events. The *Figure B.2* illustrates that most of the departments studied display a blue shade, proving that per 25mm of rain there is a reduction in the number of test compared to the

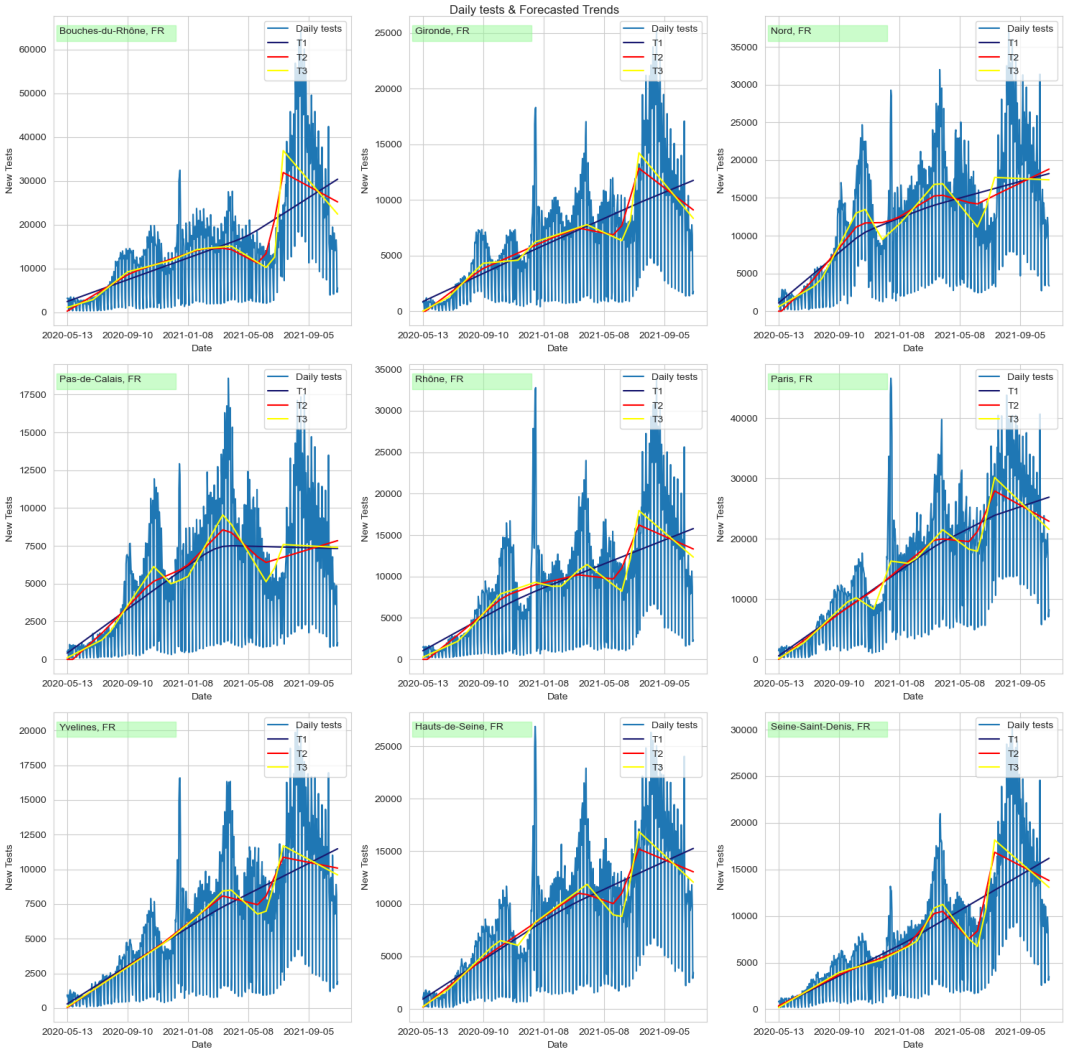

Figure B.1: The nine panels represent daily tests performed in the period between 2020-05-01 and 2021-11-01 of the nine most populated departments of France. Each panel also displays the different trends chosen for the analysis T_1, T_2, T_3.

Trend	Model C
T_1	$\tilde{\Delta_N} = -13.0\%$ (1e-11), AIC= 12323019
T_2	$\tilde{\Delta_N} = -13.0\%$ (9e-12), AIC= 11670244
T_3	$\tilde{\Delta_N} = -11.1\%$ (4e-10), AIC =10019578

Table B.2: Table of results for Model C, where $\tilde{\Delta_N}$ denotes the mean of Δ_N over all the studied counties. The values in round brackets refer to *p-value* of the precipitation coefficient.

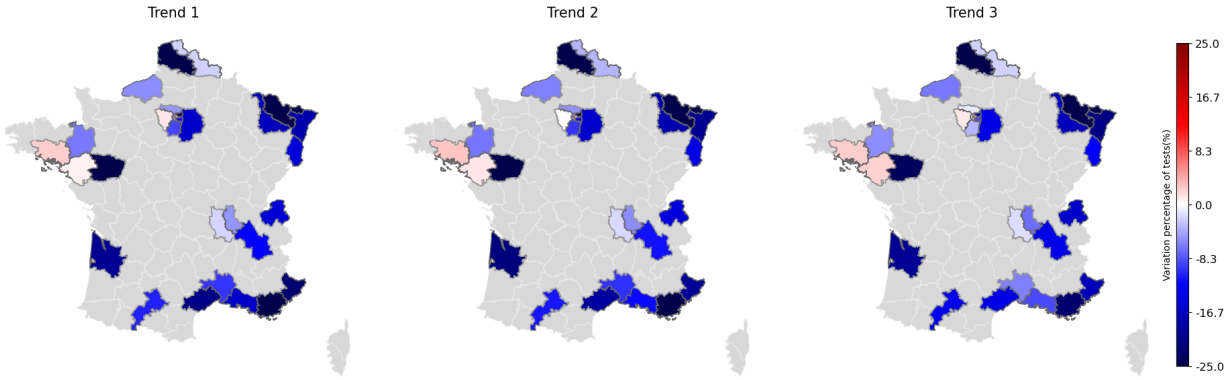

Figure B.2: The figure shows the selected departments of France used for the analysis. The color map indicates the percentage change in daily tests per 25 mm of rain, for the three trends.

trend. This reduction ranges from -31.1% for the department of *Moselle* to -1.9% of *Loire*, considering the mid-flexible trend T_2, see *Table B.3*. Other departments, instead, don't follow this behaviour, for example *Paris* which is pink coloured.

Analyzing the correlations between the changes (Deltas) and socio-demographic parameters can provide some insights into the nature of this variability among France's departments. The Pearson Correlation between the Δ_N and the socio-demographic parameters is denoted as P_{Δ_N}. *Table B.4* below presents the significant values of P_{Δ_N}, including the following metrics:

- *EI* is the median value of "niveau de vie" which is a measure used by INSEE [1] to assess economic well-being. It is the disposable income of a household divided by the number of consumption units, which adjusts for household size

[1]Institut National de la Statistique et des Études Économiques

Department Code	**Department Name**	$T_1(\Delta_N)$	$T_2(\Delta_N)$	$T_3(\Delta_N)$
6	Alpes-Maritimes	-22.3	-19.8	-17.3
13	Bouches-du-Rhône	-16.3	-12.1	-8.8
30	Gard	-9.8	-10.1	-6.2
31	Haute-Garonne	-10.8	-11.5	-13.7
33	Gironde	-20.1	-21.5	-19.8
34	Hérault	-21.0	-19.2	-14.2
35	Ille-et-Vilaine	-6.6	-6.8	-5.5
38	Isère	-12.4	-11.5	-13.9
42	Loire	-2.0	-1.9	-1.7
44	Loire-Atlantique	0.6	1.3	2.2
49	Maine-et-Loire	-25.1	-26.0	-23.7
54	Meurthe-et-Moselle	-17.3	-18.1	-17.1
56	Morbihan	2.5	2.8	2.5
57	Moselle	-31.1	-31.1	-28.6
59	Nord	-2.3	-3.6	-2.3
62	Pas-de-Calais	-26.1	-26.3	-25.8
67	Bas-Rhin	-18.2	-20.1	-21.0
68	Haut-Rhin	-15.1	-14.2	-12.9
69	Rhône	-5.1	-5.5	-7.2
74	Haute-Savoie	-15.2	-14.6	-16.0
75	Paris	5.5	4.9	5.8
76	Seine-Maritime	-5.5	-6.1	-6.5
77	Seine-et-Marne	-16.3	-15.9	-14.3
78	Yvelines	1.3	0.2	1.1
83	Var	-27.5	-27.9	-22.6
91	Essonne	-9.1	-9.9	-3.6
92	Hauts-de-Seine	-4.5	-4.8	-1.1
93	Seine-Saint-Denis	-26.7	-26.5	-17.4
94	Val-de-Marne	-12.1	-12.6	-8.7
95	Val-d'Oise	-5.1	-5.3	-1.1

Table B.3: Department Data with Delta Values.

Trend	EI	Prevalence of diabetes	Poverty rate
T_1	P_{Δ_N}= 0.46 (0.01)		P_{Δ_N}= -0.35 (0.05)
T_2	P_{Δ_N}= 0.45 (0.01)	P_{Δ_N}= -0.36 (0.05)	
T_3	P_{Δ_N}= 0.40 (0.03)		

Table B.4: Table with significant values of P_{Δ_N}. The values in brackets refer to *p-value* of P_{Δ_N}.

and composition [12].

- *Poverty rate* is the percentage of people living in households where the equivalized disposable income is below a certain threshold. This threshold is commonly set at 60% of the median equivalized disposable income, which is the household's total income after taxes and social contributions, adjusted for household size and composition using a standard equivalence scale, defined by INSEE.

- *Prevalence of diabetes* refers to the proportion of the population diagnosed with diabetes and undergoing pharmacological treatment, adjusted for age difference and other demographic factors [20].

Simply studying the correlation alone is not enough to fully explain the variability across departments. However, understanding the nature of this variability can turn into a strength of public health, both to have a more accurate monitoring system and to improve the weak points of surveillance even in emergency climatic situations.

List of Figures

1.1 Color map according to the percentage of COVID-19 total cases that highlights the countries that contribute the most to the total number of COVID-19 cases worldwide. 2

2.1 The left panel represents the mean number of tests conducted per day of the week in the metropolitan area of NYC by ZCTA. It shows an unusual increase in the number of tests as the weekend approaches. This trend can be attributed to the fact that many tests recorded on Sunday show an higher reporting delay, referring to the right panel. In detail, there are missing entries on the previous days, and it is likely that cumulative data for these missing entries were logged on Sunday, leading to the observed spike. 5

3.1 The nine panels represent daily tests performed in the period between 2020-05-01 and 2021-11-01 of the nine most populated counties of New York. Each panel also displays the different trends chosen for the analysis T_1, T_2, T_3. 9

4.1 The figure illustrates the daily test time series for Kings County. The colored dots indicate when the weather anomalies occurred identified using the Isolation Forest algorithm on the residuals of weather variables after seasonal decomposition. 15

4.2 Boxplot for the distribution of ρ's for the Kings county. As expected in case of no anomalies the distribution of ρ_0 is almost centered on zero, given that all the days of the week are considered. Instead the distributions of ρ_{tmax} and ρ_{prec}, are slightly shifted to more negative values, indicating a reduction of the number of test compared to the baseline (no anomalies). The black line in each box represents the median. 16

4.3 Boxplot for the distribution of ρ's for the Kings county. The figure displays a shift in the distribution towards smaller values of ρ, both in case of an unusual temperature and in case of an extreme rainfall. The black line in each box represents the median. 17
4.4 Mean number of tests performed for each day of the week, for the 35 most populated counties. 18
4.5 Boxplot for the distribution of ρ's for the 35 most populated counties in New York State. The left panel displays the distributions of ρ for anomalies in T_{min} (extreme cold) and T_{max} (extreme heat) separately. The right panel shows the distribution of ρ_T, which combines the temperature anomalies. The temperature anomaly seems to have an amplifying behaviour on the reduction in tests of Wednesday compared to Tuesday. While the effect of precipitations appears less clear, and the distribution of ρ_{prec} is wider. The black line in each box represents the median value. 19
4.6 The upper panel illustrates the time-series data, plotting together the actual test values (solid line) with those predicted by **Model 2** (dashed line) across three distinct counties. The lower panel presents a scatter plot, mapping the relationship between predicted and actual data points across all 30 counties analyzed. 24
4.7 The figure shows the selected counties in New York State used for the analysis. The color map indicates the percentage change in daily tests per 25 mm of rain, for the three trends. 28

B.1 The nine panels represent daily tests performed in the period between 2020-05-01 and 2021-11-01 of the nine most populated departments of France. Each panel also displays the different trends chosen for the analysis T_1, T_2, T_3. 34
B.2 The figure shows the selected departments of France used for the analysis. The color map indicates the percentage change in daily tests per 25 mm of rain, for the three trends. 35

List of Tables

4.1 The table shows the medians of the distribution of ρ's. Except for ρ_{tmin}, referring to an event of extreme cold, the medians of ρ_{tmax} and ρ_{prec} are smaller than the one of ρ_0, indicating a reduction of tests performed. 15

4.2 The table shows the medians of the distribution of ρ's. Both the medians of ρ_T and ρ_{prec} are smaller than the one of ρ_0, indicating a reduction of tests performed when a weather anomaly occurs. 18

4.3 The table shows the values of Δ_N for the different predictors. The values in the brackets refer to *p-value* of the coefficients. 22

4.4 The table lists the AIC value for each model trained, using the *Number of Tests of the Day before* as offset 22

4.5 Table of results for Model A and Model B. The values in round brackets refer to *p-value* of the precipitation coefficient. 25

4.6 Table of results for Model C, where $\tilde{\Delta_N}$ denotes the mean of Δ_N over all the studied counties. The values in round brackets refer to *p-value* of the precipitation coefficient. 26

4.7 The table shows the values of Δ_N for each studied county, for the three trends T_1, T_2, T_3. 27

4.8 Table with significant values of P_{Δ_N}. The values in brackets refer to *p-value* of P_{Δ_N}. 28

B.1 Table of results for Model A and Model B. The values in the brackets refer to *p-value* of the precipitation coefficient. 33

B.2 Table of results for Model C, where $\tilde{\Delta_N}$ denotes the mean of Δ_N over all the studied counties. The values in round brackets refer to *p-value* of the precipitation coefficient. 35

B.3 Department Data with Delta Values. 36

B.4 Table with significant values of P_{Δ_N}. The values in brackets refer to *p-value* of P_{Δ_N}. 37

Bibliography

[1] New York State Community Health Indicator Reports (CHIRS). $chir_{c}urrent_{d}ata$. 2021. URL: https://apps.health.ny.gov/public/tabvis/PHIG_Public/chirs/#dataexport.

[2] Hamada S. Badr et al. "Unified real-time environmental-epidemiological data for multiscale modeling of the COVID-19 pandemic". In: *medRxiv* (2021). DOI: 10.1101/2021.05.05.21256712. eprint: https://www.medrxiv.org/content/early/2021/05/07/2021.05.05.21256712.full.pdf. URL: https://www.medrxiv.org/content/early/2021/05/07/2021.05.05.21256712.

[3] Aishwarya Bhaskaran. "Likelihood Theory and Methods for Generalized Linear Mixed Models". Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Mathematics under the supervision of Prof. Matt P. Wand and Dr. Joanna Wang. PhD Thesis. University of Technology Sydney, Faculty of Science, Sept. 2022.

[4] U.S. Census Bureau. *AGE AND SEX*. U.S. Census Bureau. Accessed on 24 March 2024. URL: https://data.census.gov/table/ACSST5Y2020.S0101.

[5] U.S. Census Bureau. *HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE*. U.S. Census Bureau. Accessed on 24 March 2024. URL: https://data.census.gov/table/DECENNIALDHC2020.P9?q=DECENNIALDHC2020.P9&g=040XX00US36$0500000.

[6] U.S. Census Bureau. *INCOME IN THE PAST 12 MONTHS (IN 2020 INFLATION-ADJUSTED DOLLARS)*. U.S. Census Bureau. Accessed on 24 March 2024. URL: https://data.census.gov/table/ACSST5Y2020.S1901?t=Income%20and%20Poverty&g=040XX00US36$8600000.

[7] U.S. Census Bureau. *POVERTY STATUS IN THE PAST 12 MONTHS*. U.S. Census Bureau. Accessed on 24 March 2024. URL: https://data.census.gov/table/ACSST5Y2020.S1701?q=Poverty&t=Income%20and%20Poverty&g=040XX00US36$8600000.

[8] U.S. Census Bureau. *RACE*. U.S. Census Bureau. Accessed on 24 March 2024. URL: https://data.census.gov/table/ACSDT5Y2020.B02001?q=B02001:%20Race&g=040XX00US36$8600000.

[9] U.S. Census Bureau. *TOTAL POPULATION*. U.S. Census Bureau. Accessed on 24 March 2024. URL: https://data.census.gov/table/DECENNIALDHC2020.P1?g=040XX00US36$8600000.

[10] HHS Office of the Chief Data Officer. *New York State Statewide COVID-19 Testing (Archived)*. https://health.data.ny.gov/api/views/xdss-u53e. 2023.

[11] Centers for Disease Control and Prevention. *PLACES: Local Data for Better Health, ZCTA Data 2023 release*. 2023. URL: https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-ZCTA-Data-2023/qnzd-25i4/data_preview.

[12] Institut National de la Statistique et des Études Économiques. 2024. URL: https://www.insee.fr/fr.

[13] A. Liveris et al. "When New York City was the COVID-19 pandemic epicenter: The impact on trauma care". In: *The Journal of Trauma and Acute Care Surgery* 93.2 (2022), pp. 247–255. DOI: 10.1097/TA.0000000000003460. URL: https://doi.org/10.1097/TA.0000000000003460.

[14] Hannah McClymont and Wenbiao Hu. "Weather Variability and COVID-19 Transmission: A Review of Recent Research". In: *International Journal of Environmental Research and Public Health* 18.2 (2021). ISSN: 1660-4601. DOI: 10.3390/ijerph18020396. URL: https://www.mdpi.com/1660-4601/18/2/396.

[15] Météo France. *Données météorologiques - Météo France*. 2024. URL: https://meteo.data.gouv.fr/form.

[16] National Centers for Environmental Information. *Web-Accessible Folder of nClimGrid-Daily's Gridded Fields and Area Averages, v1.0.0*. https://www.ncei.noaa.gov/data/nclimgrid-daily/. 2022.

[17] World Health Organization. *Coronavirus disease (COVID-19)*. URL: https://www.who.int/health-topics/coronavirus/coronavirus#tab=tab_1.

[18] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[19] Marcus R. Pereira et al. "COVID-19 in solid organ transplant recipients: Initial report from the US epicenter". In: *American Journal of Transplantation* 20.7 (2020), pp. 1800–1808. DOI: https://doi.org/10.1111/ajt.15941. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajt.15941. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ajt.15941.

[20] Santé Publique France. *GéoDÉPO - Indicateurs de santé*. 2024. URL: https://geodes.santepubliquefrance.fr/#c=indicator&view=map2.

[21] Santé publique France. *Données de laboratoires pour le dépistage à compter du 18/05/2022 - SI-DEP*. 2022. URL: https://www.data.gouv.fr/fr/datasets/donnees-de-laboratoires-pour-le-depistage-a-compter-du-18-05-2022-si-dep/#/resources/2f8fd565-2691-4c83-8a5f-376e9da09ae5.

[22] *Statistical Modeling and Forecasting*. https://timeseriesreasoning.com/.

[23] Letham B. Taylor SJ. "Forecasting at scale." In: *PeerJ Preprints* (2017). URL: https://doi.org/10.7287/peerj.preprints.3190v2.

[24] Zhi-Wei et al. Xu. "Global dynamic spatiotemporal pattern of seasonal influenza since 2009 influenza pandemic". In: *Infectious diseases of poverty vol. 9,1 2.* (2020). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6942408/.

[25] Xinxuan Zhang et al. "The impact of weather condition and social activity on COVID-19 transmission in the United States". In: *Journal of Environmental Management* 302 (2022), p. 114085. ISSN: 0301-4797. DOI: https://doi.org/10.1016/j.jenvman.2021.114085. URL: https://www.sciencedirect.com/science/article/pii/S0301479721021472.

Acknowledgements

I would like to express my endless gratitude to everyone who has supported and tolerated me during my master's program.
Firstly, I extend my sincere thanks to my supervisor, Eugenio, whose kindness and guidance have continually fueled my passion for applied research.
I also gratefully acknowledge Professor Gamba for their support and mentorship throughout this thesis.
A special thanks goes to my family for their unwavering support, especially during the most challenging times. I also extend my heartfelt gratitude to the D'Agostino, Laudani, and Nalin families for making me feel loved and accepted at all times.
I'm thankful to all my classmates for sharing this crazy master's journey, filled with laughter and sacrifices.
To Claudio, Suprabhath, Boxuan, Nayeon for making my internship more than enjoyable.
To all the girls of *Confraternita* for empowering me and always sending good vibes.
To my dear friends Sara, Ussi, Annamaria, and Ale for listening to my complaints and my "I can't"s.
To my best friend Peppi for always walking with me, from high school onward, believing in me, and sharing the university experience despite the physical distance.
Lastly, I want to thank Caramellino for being with me through it all, visiting me in every country, making me a better person, healing my wounds, drying my tears during stressful times and showering me with love every day.
I could never have done it without you.