Dauphine | PSL
UNIVERSITÉ PARIS

Big data and health

Digital Economics

Predicting the covid incidence at weekly and regional level in France

CHENG Yichen,
ZHANG Yanming

supervised by
GODARD Mathilde

Contents

1 Introduction

A search engine query is a request for information that is made using a search engine. Every time a user puts a string of characters in a search engine and presses "Enter", a search engine query is made. The string of characters (often one or more words) act as keywords that the search engine uses to algorithmically match results with the query. Every search engine query adds to the mass of analytical data on the Internet. The more data search engines collect, the more accurate the search results become – and that's a good thing for Internet users.

Text analysis is a machine learning technique that leverages text data to extract valuable insights, with various applications such as understanding public sentiment or detecting fraudulent activities. A previous study demonstrated the use of text analysis to predict the hygiene situation of a restaurant by correlating customer reviews with inspection results.

In this paper, we focus on the application of text analysis in the healthcare field, particularly in predicting the incidence of COVID-19 across different regions in France. The sudden outbreak of the pandemic has significantly impacted the world, particularly in terms of medical resources. If the government can predict potential outbreaks in advance, they can allocate medical resources to the affected regions efficiently. To achieve this, we leverage query data from Google Trends as people tend to search for COVID-related information online, especially in regions with high internet usage.

Since we do not know which query is highly correlated with COVID incidence, we selected 100 popular keywords related to COVID from 2020 May to 2022 May on Google Trends. In the second part of data processing, we explain how we processed these keywords to create our final dataset for correlation analysis. We then perform simple statistical analysis and construct linear and Random Forest Regression models to provide quantitative results. Finally, we discuss the limitations of our study and models.

2 Related Work

The article "Detecting influenza epidemics using search engine query data" explores the potential of using search engine query data to predict influenza epidemics. The authors collected search engine data from Google using a set of selected keywords and used it to train a model to predict influenza activity in the United States. They found that their model was able to predict influenza epidemics up to two weeks in advance of traditional methods. This research demonstrates the usefulness of search engine query data for epidemiological surveillance and provides a valuable tool for predicting the spread of influenza in real-time.[1]

[1]Detecting Influenza Epidemics UsingSearch Engine Query Data." Nature 457(7232): 1012–14.

3 Data Processing

We obtained 100 popular keywords related to Covid by downloading several CSV files from Google Trends. Using the *Pytrends* package in Python, we directly extracted the popularity index of each keyword on a weekly basis from GoogleTrends.[2]

To obtain information on Covid incidence in France, we searched the France government website[3] and downloaded a dataset containing the number of positive Covid cases and population for each department in France on a weekly basis from May 2020 to May 2022 (From May 11th 2020 to May 23rd 2022). We calculated Covid incidence at the regional level by dividing the number of positive Covid cases by the population and multiplying the result by a certain factor (100000). The resulting dataset was merged with the dataset containing keyword index from *Google Trends*, using the common variable of week. **Pre-selection**:From the insights of Ginsberg et al.(2009), an evaluation of how many top-scoring queries should be included in the ILI (influenza-like illness) related query fraction.The study found that the maximal performance at estimating out-of-sample points during cross-validation was achieved by summing the top 45 search queries. However, there was a steep drop in model performance after adding query 81, which happened to be 'oscar nominations'. This suggests that the choice of search queries can have a significant impact on the performance of the model in estimating ILI-related activity. We then calculated the correlation between each keyword and Covid incidence at the national and weekly level, and found that the keyword *"Auto-test"* had the highest correlation with Covid incidence (0.91), followed by 7 and 39 keywords with correlations greater than 0.6 and 0.2, respectively. We also plotted the distribution of correlation for all keywords and found that 70% of keywords had a correlation below 0.2.

Keywords	Correlation with Covid incidence
Auto-test	0.914843
Isolement	0.870648
Symptômes	0.814776
Contagion	0.753281
Fièvre	0.743743
Contact tracing	0.619361
Autotests COVID	0.610457
Essoufflement	0.542841
Maux de tête	0.534540
Toux	0.530234

Table 1: Correlation between each keyword and Covid incidence at the national and weekly leve

[2]100 keywords based on Google trend query data
[3]Données de laboratoires pour le dépistage

To obtain the popularity index of each keyword at both weekly and regional levels, we created two loops in Python. However, some keywords had a popularity index that was too low in some regions to be extracted from Google Trends, and some keywords could not be extracted for certain regions. We removed these keywords, but given their low correlation at the national level, we believe their removal will not significantly affect the model results.

To merge the datasets, we used a supplementary dataset to find the correspondence between departments and regions in France used by Google Trends, and transformed the type of department code in both datasets to enable successful merging. We then created a new dataset that sums the positive Covid cases and population at the weekly and regional level and added Covid incidence information to it by merging it with the dataset from Google Trends, and this one is our final data set.[4]
The final data set contains **22 regions**:*Nord Pas de Calais, Picardie, Haute Normandie, Ile de France, Corse, PACA, Languedoc Roussillon, Midi Pyrénées, Aquitaine, Rhône Alpes, Auvergne, Limousin, Poitou Charente, Basse Normandie, Bretagne, Pays de la Loire, Alsace, Lorraine, Champagne Ardennes, Centre Val de Loire, Bourgogne, Franche Comté* and **39 keywords** after the pre-selection:*Autotest, Isolement, Symptômes, Contagion, Fièvre, Contact tracing, Autotests COVID, Essoufflement, Maux de tête, Toux, Système immunitaire, Diarrhée, Urgences, Fatigue, Dépression, Hôpital, Variant omicron, Anxiété, Gamma, Soins à domicile, Voyages, Green Pass,Télétravail, Coronavirus, Soins intensifs, Tests PCR, Tests antigéniques, Certificat sanitaire, Ventilation, Perte de goût, Frontières, Quarantaine, Restrictions, Distanciation sociale, Alpha, Chômage, Gel hydroalcoolique, Réanimation, aéroports* and with *date, region, region name, week, p, pop, incidence* these 7 columns.
We calculated the mean correlation of each keyword with Covid incidence by calculating the correlation for each region and then averaging it, resulting in a ranked list of keywords based on their correlation. In the next section of our analysis, we will use this information to determine how many keywords should be included in our model.Big Data and health Github Repository

4 Statistic analysis

From the official website statistician statista, we downloaded the Cumulative number of confirmed cases of coronavirus in France from February 14, 2020 to February 26, 2023 (Figure 1). Based on our existing dataset at the regional level, we plotted the weekly number of positive cases in each region from May 11, 2020, to May 23, 2022 (Figure 2). The overall trend was similar, with four small peaks and two large peaks, one of which was particularly prominent. The duration of each peak varied. For example,

[4]final data set

the first small peak occurred in November 2020 and lasted for about a month. We speculate that this was because the French government had implemented a series of restrictive measures, such as closing schools and prohibiting large gatherings, in the preceding months but began to relax these measures in early November, leading to increased contact between people and more opportunities for the virus to spread. With the arrival of autumn and winter, the weather became colder, and people spent more time indoors, reducing air circulation and increasing the risk of virus transmission. Additionally, as the pandemic had been ongoing for a long time, people may have become fatigued with the situation and started to relax their vigilance, ignoring preventive measures, which resulted in an increase in new cases.

The second small peak had a short duration, but its trend was very rapid, occurring at the end of 2020. We believe this was due to the following reasons: during the Christmas and New Year holidays, people's social activities increased, leading to increased virus transmission; in some regions, especially in the Paris region, new virus variants appeared, which were more contagious and may have led to faster transmission and higher infection rates; during the end-of-year period, some people may not have followed social distancing and protective measures, which could have led to virus transmission; the French government relaxed some restrictions in November and December, such as opening stores and restaurants, which may have led to more social activities and increased the risk of infection; France's testing capacity has been improved, which may have led to more cases being diagnosed.

The third small peak lasted longer and occurred in March and April 2021. The emergence of new virus variants, such as those from the UK, South Africa, and Brazil, resulted in faster and easier virus transmission, leading to an increase in the number of cases. Additionally, there was a lag in vaccination in France compared to other European countries. The slow progress of the vaccination program may have resulted in insufficient herd immunity, allowing the virus to continue spreading. Due to the complexity and intertwined effects of these factors, the peak of the epidemic lasted longer.

The fourth peak also has unique characteristics. In June 2021, the number of infections and infection rate in France reached the lowest point due to the effectiveness of vaccination and social distancing measures. However, in July, there was an increase in cases, and some areas had very high incidence values. The reasons for the increase in the number of infections in July may be related to various factors, including the emergence and spread of variant viruses, people's distrust of vaccines, and the relaxation of preventive measures. In addition, the higher infection rates in some areas may be related to population density, increased social activities, tourism, and other factors. These factors may lead to rapid transmission and a high incidence rate, resulting in an extended peak duration.

Within our specified time frame, there was a significant increase in new cases in France in January and

February of 2022, as can be seen in Figure 1. This peak and trend are the result of multiple factors. One major contributor is the spread of the new Omicron variant. The Omicron variant is more infectious than the previous Delta variant, leading to more rapid transmission of the virus. Additionally, France's winter climate may have exacerbated the spread of the virus, as people spend more time indoors and air circulation is limited, making it easier for the virus to spread in enclosed spaces. Holiday travel and family gatherings may have also contributed to the spread of the virus.

In March 2022, the number of new cases in France continued to increase, but the incidence value was decreasing. This may be due to factors such as an increase in population immunity or a decrease in the degree of virus mutation. Additionally, it may also be related to the effectiveness of prevention and control measures, such as restrictions on gatherings and strengthened vaccine administration. However, in April, the incidence value started to increase again, possibly due to the emergence of new variants of the virus, relaxation of prevention and control measures, and increased gathering activities.

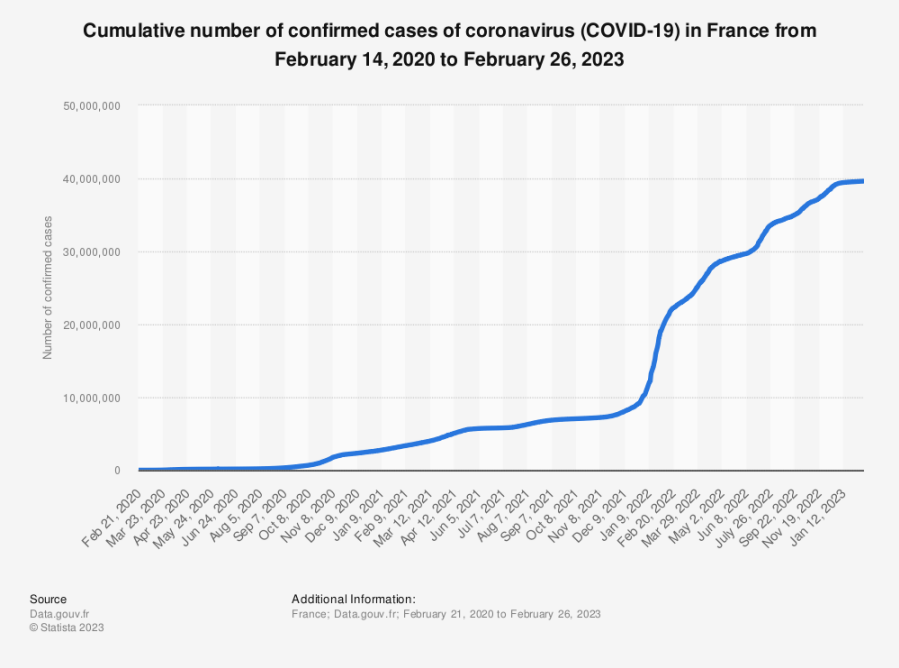

Figure 1: Cumulative number of confirmed cases of coronavirus (COVID-19) in France from February 14, 2020 to February 26, 2023

We conducted an analysis of the most searched keywords for each region during the three most prominent peaks of the COVID-19 pandemic. To do so, we calculated the average popularity index of each keyword and each region for the period of three peaks. The resulting figures, denoted as 3, 4, and 5, respectively, represent the average popularity index of each keyword for each region in November 2020, January and February 2022, March 2022. This approach allowed us to gain insights into the topics and concerns that were of greatest interest to individuals in different regions during the most critical periods of the pandemic.

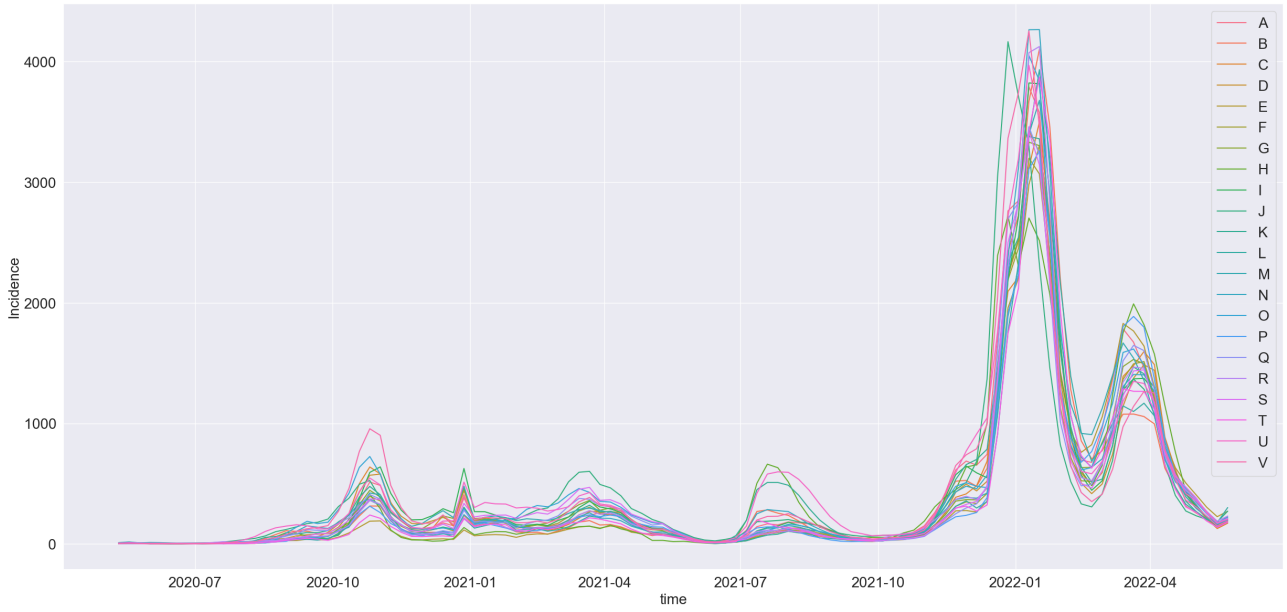

Figure 2: Time series of covid incidence for each region in France

From Figure 3, we can observe that the top few searched keywords in almost all regions are depression, test PCR, coronavirus, and soins intensifs, with a popularity index greater than 50. In November 2020, France was in the early stage of the COVID-19 pandemic, and the number of cases and deaths was rapidly increasing, which had a significant impact on the public's mental and physical health. Firstly, "depression" is a common psychological issue characterized by low mood, especially in situations of increased stress and uncertainty. In the context of COVID-19, many people faced multiple factors such as economic, social, and emotional stress, making depression one of the popular search terms. Secondly, "test PCR" (PCR testing) and "coronavirus" (coronavirus) are directly related to COVID-19 search terms. In the early stages, people were highly interested in information on virus transmission and testing methods, making these terms popular search terms. Lastly, "soins intensifs" (intensive care) refers to medical care required in severe cases, which is related to the severe situation of the COVID-19 pandemic. In the early stages, due to the lack of effective treatment methods, many infected individuals required intensive care. Therefore, this term became one of the popular search terms.

During the second peak period (Figure 4), which was in January and February 2022, the highly searched keywords with a popularity index above 60 were: fatigue, système immunitaire (immune system), restrictions, and chômage (unemployment). At that time, France was experiencing a significant increase in COVID-19 cases, and many people may have felt tired and weak, which could be the reason for searching for "fatigue". In addition, due to the tense situation of the pandemic during this period, people's interest in the immune system has also increased, and they may have searched for information related to improving immunity. Furthermore, as the French government implemented various restric-

tive measures, such as limiting indoor gatherings, people may have searched for information related to "restrictions". Additionally, the pandemic has had an impact on the economy, and many people may have lost their jobs, leading them to search for information related to "unemployment".

During the third peak (Figure 5), the highly searched keywords with significant search volume are "auto-test" and "voyages". In some regions, their search volume even exceeded 80. In March 2022, French people's searches regarding COVID-19 mainly focused on self-testing and travel. Self-testing became a hot topic as self-testing kits became widely available and people increasingly paid attention to personal diagnosis and preventive measures. Self-testing was seen as a security measure to detect whether one had contracted the virus. As the pandemic situation came under control, people started to reconsider the possibility and safety of traveling, thus making searches related to travel a popular topic.

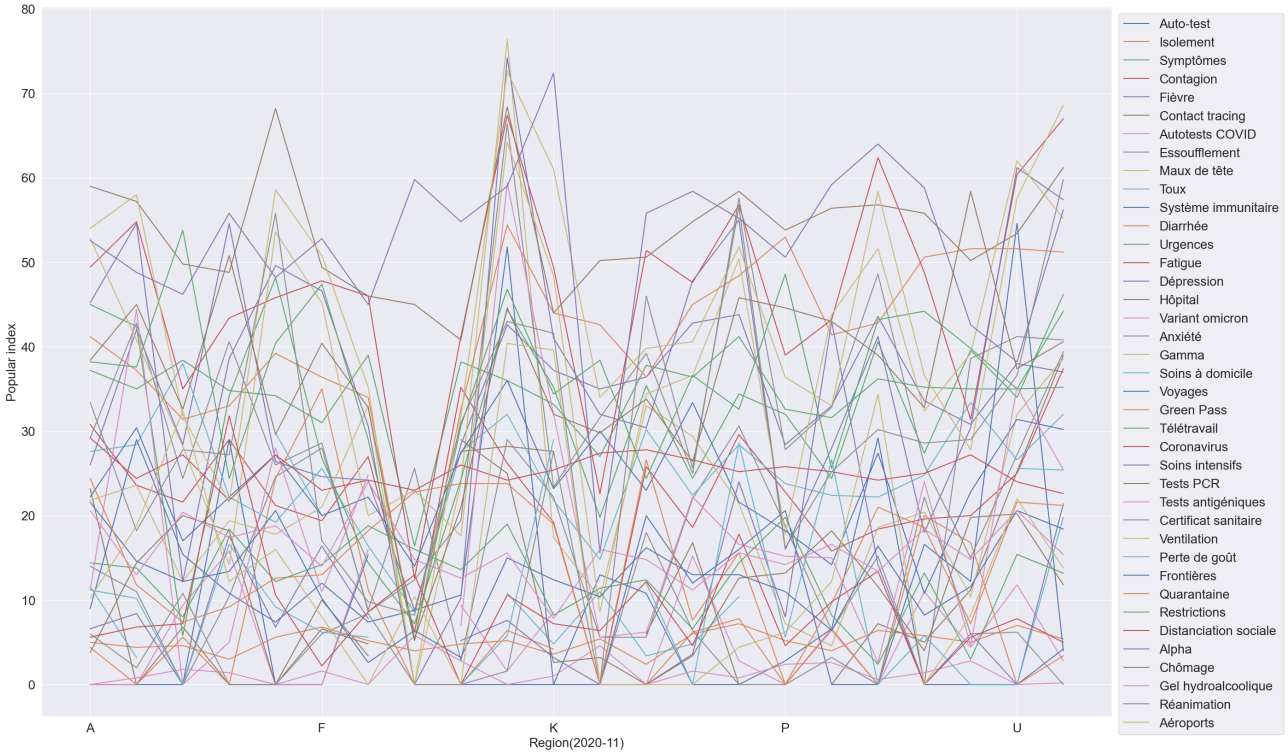

Figure 3: The popularity index of each keyword for each region in November 2020

5 Model analysis

In this task, the goal is to build a model that can predict the incidence of COVID-19 cases. The dataset contains 39 keywords that may be relevant to predicting the incidence of COVID-19.To determine the optimal number of keywords to include in the model, a training set, a validation set, and a test set have been created. A loop has been set up to add one additional keyword to the model at a time, based

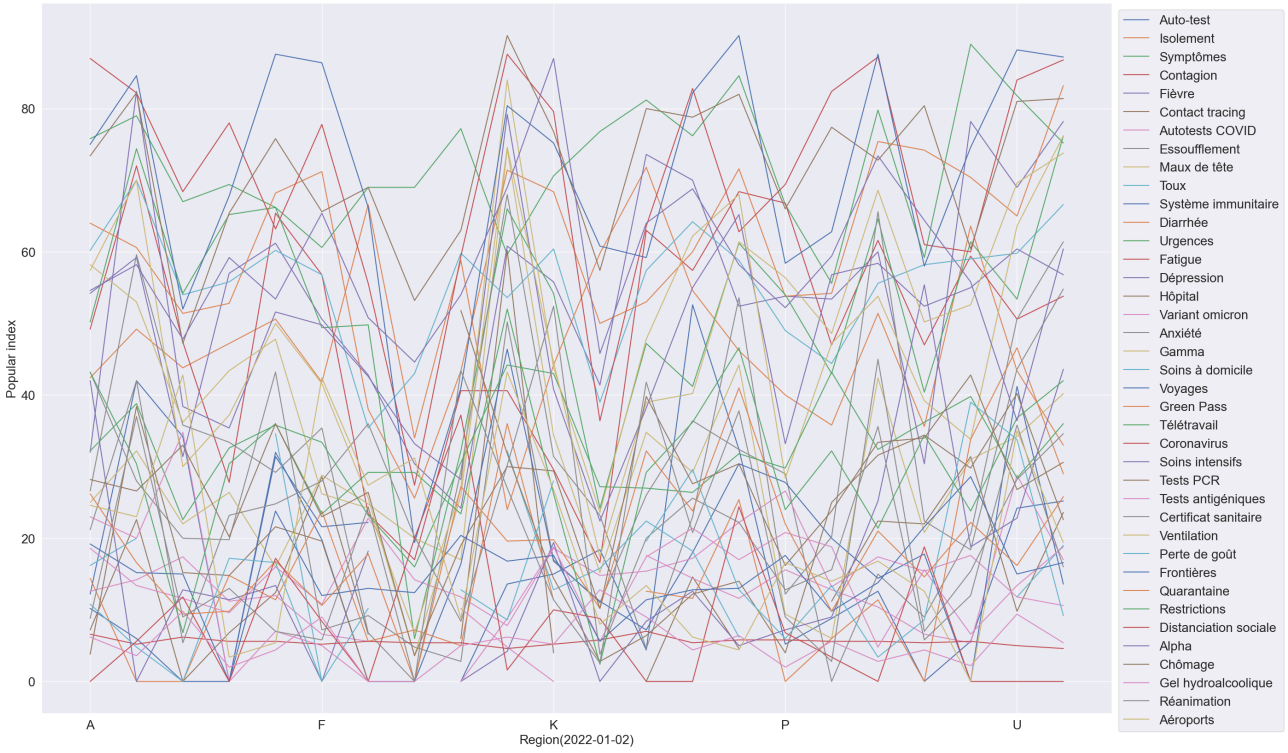

Figure 4: The popularity index of each keyword for each region in January-Feburary 2022

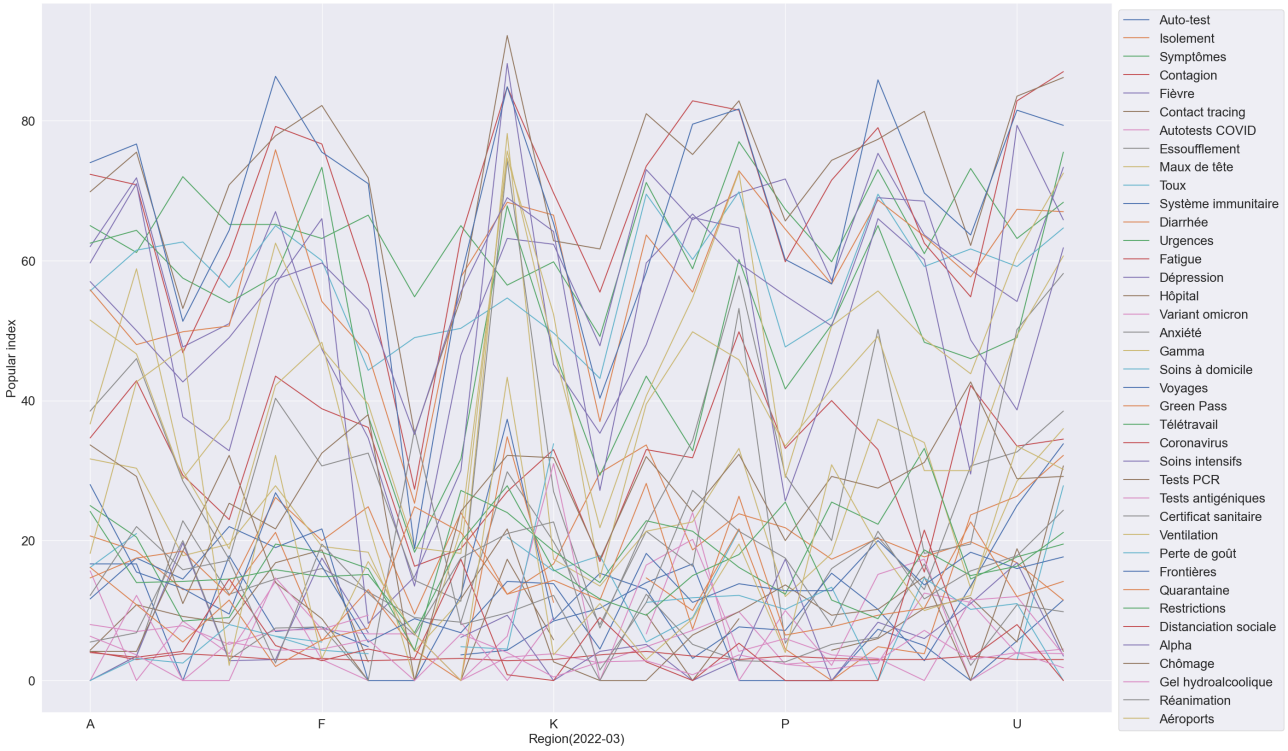

Figure 5: The popularity index of each keyword for each region in March 2022

on their correlation with the target variable.For each iteration of the loop, the R^2 score of the model is calculated on the validation set. The number of keywords that leads to the highest R^2 score on the validation set is chosen as the optimal number of keywords to include in the model. This process allows for the selection of the most relevant keywords to include in the model, and helps to avoid overfitting by selecting only the most informative features. The ultimate goal is to build a model that accurately predicts the incidence of COVID-19 cases, based on a subset of the most relevant keywords in the dataset.

5.1 Linear regression

Forecasting GDP growth is useful for policy makers to assess macroeconomic conditions in real time. In Ferrara et al.(2022), the researchers aim to use a large database of Google search data to make forecasts of Euro-zone GDP. In this regard, we estimate reduced bridge regressions that integrate Google data filtered by the targeting approach, which we empirically show provides some gains in pseudo-real-time forecasts of quarterly euro area GDP growth. Moreover, a true real-time analysis confirms that Google data constitute a reliable proxy when official data are scarce. In an analysis focused on Covid confirmed cases, regression analysis is the underlying method.

We have used linear regression as our first model to predict the incidence of COVID-19 in different regions. We have created a function that takes the code of a region as input and automatically predicts the incidence using the best-fitted number of keywords. We have divided the regional dataset into three parts, with 60 weeks for training, 23 weeks for validation, and 24 weeks for testing.To select the best-fitted number of keywords, we have used the validation set and then applied it to the test set. We defined three functions here.

- plot1: This function is used to plot a line chart of the true and predicted values for the **training set**. *ytrain* is the true values of the training set, and *LR_pred_train* is the predicted values of the training set. The function uses *Seaborn* to plot the line chart.

- plot2: This function is used to plot a line chart of the true and predicted values for the **test set**. *ytest* is the true values of the test set, and *LR_pred* is the predicted values of the test set. The function uses *Matplotlib* to plot two subplots, where the left subplot displays the true values of the test set, and the right subplot displays the predicted values of the test set.

- Prediction_LR: This function predicts the housing prices based on the specified Region_Code parameter and outputs the R^2 score of the model on the test set, the optimal number of variables, and the prediction charts for the training and test sets. The function first selects a data subset

that matches the Region_Code from df_all, and then splits this subset into training, validation, and test sets. Next, the function fits the training set using a linear regression model and selects the optimal number of variables using the validation set. Finally, the function predicts the test set using the selected number of variables and outputs the R^2 score of the model on the test set and the prediction charts for the training and test sets. For instance, for region B (Aquitaine), the best number of keywords was found to be 39. The performance of the model was evaluated using the R^2 score, which was -5.79 for the test set. We have plotted the actual incidence and our predictions for both the training and test sets. The results show that our model has a great prediction for the training set, with the two lines almost coinciding. However, the prediction of incidence for the test set does not have the same magnitude as the actual incidence, although it can still predict the trend of COVID-19 incidence well.

Figure 6: The comparison of the prediction of incidence on training set in region "Aquitaine"

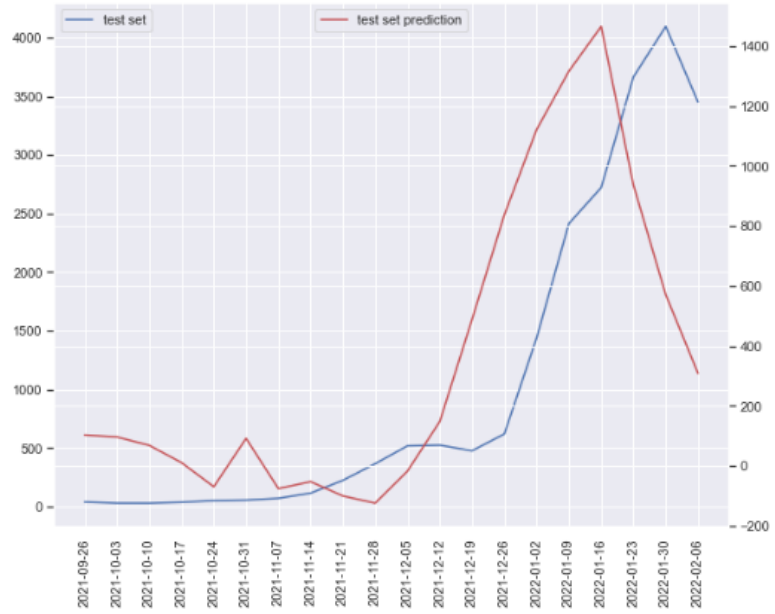

Figure 7: The comparison of the prediction of incidence on test set in region "Aquitaine"

5.2 Random Forest

In the selection of another model, there may be a nonlinear relationship in the dataset. If such a relationship exists, linear models may not be able to capture it, whereas decision trees and random forests can handle such nonlinear relationships. Considering the possibility of noise or missing values in the dataset, random forests can handle these situations well because they are not influenced by individual data points or features. Additionally, even after pre-selection, the dataset may still have a high-dimensional feature space. Random forests perform well in handling high-dimensional feature space and can automatically select important features. Random forests are an ensemble learning method that can use the results of multiple trees to arrive at the final prediction. This approach is generally more accurate than a single decision tree or linear model.

We define three functions for running a random forest regression model on COVID incidence data for a specific region.

- plot1_rf(): creates a line plot of the training set and the predicted values of the training set using random forest. The function returns the plot.
- plot2_rf(): creates a line plot of the test set and the predicted values of the test set using random forest. The function returns the plot.
- Prediction_rf(): Filters the data for the specific region indicated by Region_Code, splits the data into training, validation, and test sets, runs a loop to determine the optimal number of variables to use in the random forest model using the mean squared error (MSE) on the validation set as the criterion for selection, fits the random forest model on the training set using the optimal number of variables and predicts the COVID incidence for the test set, calculates the MSE and R^2 score of the random forest model on the test set, plots the training set and test set predictions using the plot1_rf() and plot2_rf() functions.The function returns the MSE, R^2 score, the number of variables used in the optimal model, and the plots of the training set and test set predictions.

Using region B (Aquitaine) as an example as the linear model section, it was found that the optimal number of keywords for the model was 13. Figure 8 illustrates the comparison between the actual incidence and the predicted incidence by the model for the training set in region B. The plot indicates a close agreement between the two lines, suggesting that the model has achieved high accuracy for the training set. However, the R2 score for the test set is -56, indicating that the model did not perform well for the test data. Figure 9 also displays the comparison between the actual and predicted incidence for the test set. It shows that the random forest model performed worse than the linear regression model for the test set.

Figure 8: The comparison of the prediction of incidence on training set in region "Aquitaine"

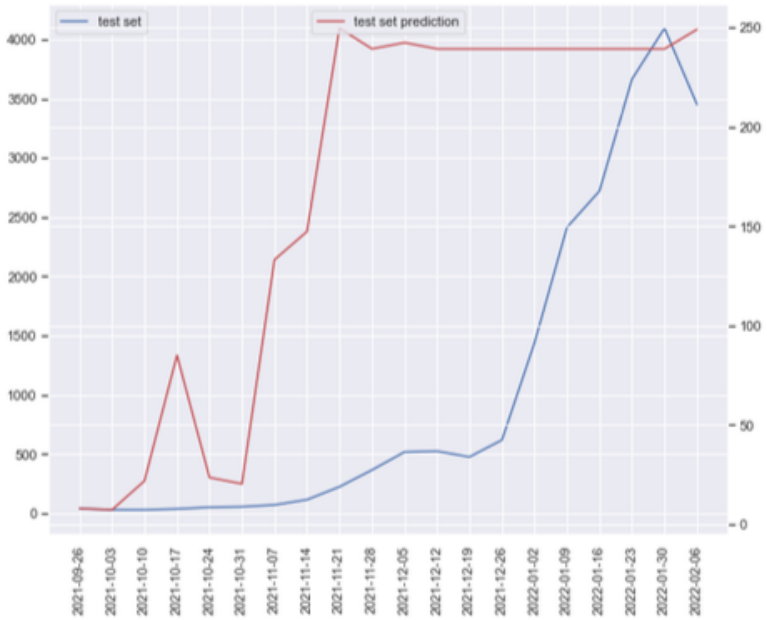

Figure 9: The comparison of the prediction of incidence on test set in region "Aquitaine"

6 Conclusion

In this study, we used the popularity index of COVID-related keywords in Google Trends to predict the incidence rate of COVID on a weekly and regional basis. We built models for 22 regions distributed across 107 weeks. In the report, we only showed results for region "B" (Aquitaine), where we used linear regression as our first model. We found that the training set for all regions had very high R2 scores. For the test set, our model was able to predict the trend of the incidence rate, but with a different amplitude compared to the actual incidence rate, indicating poor predictive ability. Our second model used random forest regression, which also showed good performance on the training set, but suffered from overfitting. Our findings suggest that linear models can predict the trend of incidence rate better than random forest regression.

We had several thoughts on the overfitting problem in our models:

1. Keyword selection problem: after pre-selection, we reduced the number of keywords from 100 to 39, which is far less than what Google Trends can provide, but we cannot extract all keywords online. We may have selected too many features or features that are not relevant to the target variable, but this limitation seems to have no solution.

2. Model complexity problem: when the model is too complex, it can easily capture all noise and details in the training data, which leads to the model being unable to generalize to new data.

3. Insufficient data problem: we had limited training data, which resulted in the model overfitting to the data and failing to generalize to new data.

4. The impact of psychological factors on Google data: people search for information related to COVID online even if they have not been diagnosed, especially during times of panic. This behavior increases the number of keywords searched online, thereby affecting the model results.

7 References

Ferrara, Laurent, and Anna Simoni. "When are Google data useful to nowcast GDP? An approach via preselection and shrinkage." Journal of Business & Economic Statistics (2022): 1-15.

Ribeiro, Matheus Henrique Dal Molin, Ramon Gomes da Silva, Viviana Cocco Mariani, and Leandro dos Santos Coelho. "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil." Chaos, Solitons & Fractals 135 (2020): 109853. Ribeiro, M.H.D.M., d

Maleki, Mohsen, Mohammad Reza Mahmoudi, Darren Wraith, and Kim-Hung Pho. "Time series

modelling to forecast the confirmed and recovered cases of COVID-19." Travel medicine and infectious disease 37 (2020): 101742.