Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230691

Explainable Machine Learning Models for Predicting COVID-19 Cases in Catalonia Based on Wastewater Monitoring Data

Marc RIBALTA[a, b,1], Josep PUEYO-ROS[c], Lluis COROMINAS[c,d], Ramón BÉJAR[b], Carles MATEU[b], Edgar RUBIÓN[a]
[a] *Eurecat, Technology Centre of Catalonia, Unit of Applied Artificial Intelligence, Bilbao 72, Barcelona, 08005, Spain*
[b] *GREiA - UdL, Jaume II, 69, Lleida, 25001, Spain*
[c] *Catalan Institute for Water Research (ICRA-CERCA), Emili Grahit 101, Girona, 17002, Spain*
[d] *University of Girona, Plaça de Sant Domènec, 3, Girona, 17004, Spain*

Abstract. Surveillance of infectious diseases that can spread quickly among the population has become a critical goal nowadays due to the dramatic effect of diseases like SARS-CoV-2. One promising method to be able to monitor the spread of such diseases among the population of a city is the analysis of biological compounds in the sewage network of different cities. In this paper, we summarize the results of training a prediction model for SARS-CoV-2 cases based on historical biological data collected from different wastewater treatment plants (WWTP) located in different parts of Catalunya. We consider different approaches for the prediction problem and develop models based on extreme gradient boosting. Finally, we evaluate the quality of the results of our model and study the relevance of the different variables of the model using SHapley Additive exPlanations (SHAP) analysis.

Keywords. Explainable AI, Machine Learning, Wastewater-based epidemiology

1. Introduction

During the COVID-19 pandemic, the research community found different ways of monitoring the SARS-CoV-2 within the population to estimate the trend of infection cases. One of them is sewage surveillance to understand the prevalence of COVID-19 within the population. Different gene targets can be used to quantify the prevalence, such as the N1, N2, or E [1]. The prediction of SARS-CoV-2 cases using machine learning (ML) models is a difficult task due to its uncertainty in population behavior and unknown virus evolution, but gene target values can be processed and used to enhance them.

In this paper we consider the predictive problem in Catalunya, using historic data for COVID-19 cases plus biological indicators obtained from a total of 56 WWTPs across different points of Catalunya [2]. The general goal is to obtain predictive models for covid cases in the next period based on previous COVID-19 cases and the current presence of these biological markers in wastewater. We consider different attributes related to either the water samples or the population linked with the WWTP where the

[1] Corresponding Author: Marc Ribalta, E-mail: marc.ribalta@eurecat.org

water comes from. As ML algorithms, we use ensemble trees obtained with extreme gradient boosting. Given the challenging nature of this problem and that some attributes have a high extraction cost, we not only consider the performance of the models obtained, but we also study the relevance of the different attributes, using SHapley Additive exPlanations (SHAP) analysis. Our final aim is to provide relevant conclusions that improve the data acquisition procedure for a more reliable and sustainable monitoring network for future similar infectious diseases.

2. Materials and methods

2.1. Sampling points and collected data

The biological data from wastewater data was obtained from a total of 56 WWTPs selected based on two criteria: 80% coverage of the served population (about 6M inhabitants) and high territorial coverage (41 out of 42 regions covered). The sampling frequency was 1 week for some WWTP stations and biweekly for others, amounting to a total of 2000 samples. Each sample was obtained from a flow-proportional, 24h-composite set of INLET subsamples, where each subsample was obtained every 20 minutes[2]. Regarding the number of covid cases, we used the public data for municipal covid cases publicly reported[3].

The attributes used from each sample in the ML algorithm can be divided into four different groups: 1) parameters linked with the characteristics of the water flow including quality parameters such as TSS, COD and BOD, 2) a scaled value of the wastewater gene targets N1/N2, 3) WWTP population data such as inhabitants assisted and estimated citizens mobility, and 4) number of total covid cases in Catalunya in the last 14 days.

We used gradient boosting to learn ensembles of prediction trees as the ML model. We used different metrics to evaluate the error: the mean absolute error (MAE) and the coefficient of determination (R^2). The parameters used with xgboost are the number of trees: 500, maximum depth: 15 nodes, learning rate: 0.01, and subsample ratio: 0.5.

2.2. Predictive exercises

The problem of predicting future covid cases from biological samples in the setting considered here is very challenging, so we have considered four different specific variants of the problem, which predict the covid cases for each municipality:

1. In our first approach, we want to predict the number of cases in the next week. We use the wastewater data (N1/N2 gen presence) plus the number of covid cases in the two previous weeks, the number of inhabitants, and estimated mobility of citizens, just to know how many people leave or enter the municipality.
2. A second approach that we had is instead of predicting the total amount of cases, we use the number of cases for each hundred thousand inhabitants, given the same input data used in the first problem.

[2] The final data set can be found at Zenodo: https://zenodo.org/record/4955582#.ZGIwONJBxBW

[3] The COVID-19 cases data can be found at: https://analisi.transparenciacatalunya.cat/Salut/Incid-ncia-de-la-COVID-19-a-Catalunya/623z-r97q/data

3. Given that we have found that depending on the analysis lab, the accuracy of the measured viral load differs, we have also considered learning a model for each analysis lab (three in total) to predict the average cases for every 100k inhabitants. Therefore, one of the three models is used depending on the municipality.
4. Finally, we have considered the hardest problem: predicting covid cases using only wastewater data, without using the data from past covid cases count.

3. Results

3.1. Model predictions

Table 1 shows the results of evaluating the models using a test set of cases from November 2020 to April 2021. The R^2 explains the total variation of the model predictions, and the MAE is the average number of wrongly predicted cases. Combining them, we can understand the robustness of the models' errors. The *total cases prediction* model shows a high R^2 which indicates the reliability of the MAE value of 161. In contrast, the *cases per 100k inhabitants* model shows a low R^2, indicating that the 149 MAE is mixed by predictions with no error and predictions with a high error. The same happens for the *cases per laboratory* model, and this is because the models learn to predict only for some of the municipalities. Figure 1 shows a clear example of low and high-error municipalities. The *total cases prediction without past cases information* model is different since the R^2 is low and the MAE is high, meaning that the model error may be represented across the municipalities.

Predictive exercise	Coefficient of determination (R^2)	Mean Absolute Error (MAE)
Total cases prediction	0.90	161
Cases per 100k inhabitants	0.45	149
Cases per laboratory	0.28	226
Total cases prediction without past cases information	0.65	333

Table 1. Scores of the models created for each predictive exercise

Figure 1 shows the result of predicting using the second model for two municipalities. While both cases are standardized, their population density and local culture are different. For instance, Sant Feliu de Llobregat holds a population of 44.474 (2018) and an area of 11.8 km^2, and Vielha e Mijaran has a population of 5493 (2018) and an area of 211.7 km^2. Furthermore, they also differ in being an urban and a rural area. In this case, the model predicts better the cases for the higher-density municipality.

Figure 1. Predictive exercise nº2 results. Left: Weekly predicted cases for the municipality Sant Feliu de Llobregat. Right: Weekly predicted cases for the municipality Vielha e Mijaran.

3.2. Explainability

In this subsection, we identify the most relevant features of the prediction models we have developed. Our analysis is based on SHapley Additive exPlanations (SHAP). At a general level, and across all the ML models developed, the most important variables are the past number of cases and the wastewater indicator. Figure 2 shows the relevance of the past cases, being the most impactful feature in both cases and the importance of the wastewater indicator "N", which is higher when predicting the cases per 100k inhabitants. The other features vary in importance depending on the predictive exercise. To predict the total cases, the inhabitants assisted by the WWTP, the citizens' mobility, and the wastewater quality parameters are the most relevant, but to predict the cases per 100k inhabitants, only the wastewater quality ranks high.

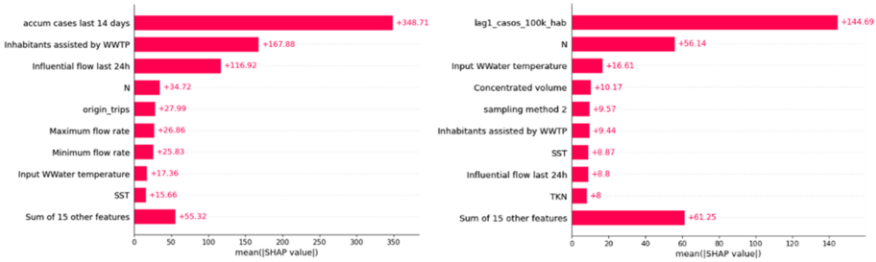

Figure 2. SHAP value for the most important variables. Left: Total cases prediction. Right: Cases per 100k inhabitants.

To improve future feature engineering, we study the impact of the individual feature values for the model. Figure 3 shows the impact of past cases and the wastewater indicator for the model prepared to predict the cases per 100k inhabitants. The SHAP for both features indicates that the bigger the value, the higher the importance to the model. When the value is low, the importance is near zero and in some cases, the feature value has a negative SHAP value, meaning it confuses the model. Those negative values demand future feature engineering to ensure stable behavior.

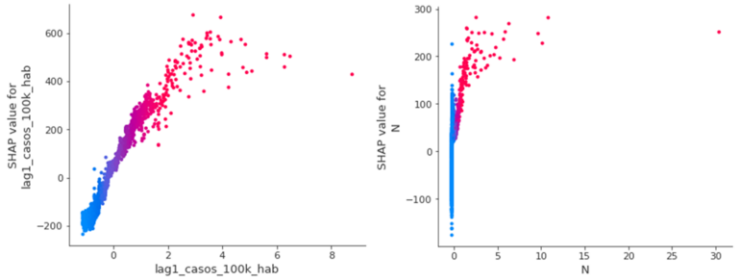

Figure 3. SHAP value for the standardized COVID-19 cases variable (left) and the N value (right). The X axis indicates the value of the feature, and the Y axis indicates the SHAP value for the feature.

4. Conclusions

The machine learning model is good at predicting how COVID-19 cases will change over time, but it struggles to predict the highest number of cases during a peak. The SHAP importance of the features varies depending on the objective variable to be predicted, but the number of people living in an area and the N indicator are always

relevant. To improve the model, samples should be collected more often. Right now, the frequency is one per week, but if it increases to three to seven samples, the model can better understand how wastewater samples relate to the number of COVID-19 cases daily.

Funding

This research has received funding from the Innovation and Technological Development Center (CDTI) and the European Regional Development Fund (ERDF) with the corresponding consolidation (COI-20201289). Marc Ribalta also acknowledges funding from AGAUR DI-2019-066. Ramon Bejar was partially funded by Spanish Project PID2019-111544GB-C22/AEI/10.13039/501100011033 (MINECO / FEDER). Ramon Bejar and Carles Mateu also ackowledge the funding by 2021 SGR 01615.

References

[1] Medema G, Heijnen L, Elsinga G, Italiaander R, Brouwer A. Presence of SARS-coronavirus-2 RNA in sewage and correlation with reported COVID-19 prevalence in the early stage of the epidemic in the Netherlands. Environ Sci Technol Lett. 2020;7:511–516 doi: 10.1021/acs.estlett.0c00357

[2] Guerrero-Latorre L, Collado N, Abasolo N, et al. The Catalan Surveillance Network of SARS-CoV-2 in Sewage: design, implementation, and performance. Sci Rep. 2022;12(16704), doi: 10.1038/s41598-022-20957-3