Universitat Oberta de Catalunya (UOC)

Master's degree in Data Science

MASTER'S FINAL PROJECT

Data Mining and Machine Learning

Evolutionary analysis of *Bordetella pertussis*: understanding the recent resurgence of Pertussis

Author: Pol Major i Munich

Supervisor: Laura Ruiz Dern

Professor: Jordi Casas Roma

Barcelona, June 6, 2019

Copyright

Final project fiche

Project title:	Evolutionary analysis of *Bordetella pertussis*: understanding the recent resurgence of Pertussis
Author:	Pol Major i Munich
Supervisor:	Laura Ruiz Dern
Professor:	Jordi Casas Roma
Date (mm/yyyy):	06/2019
Degree:	Master in Data Science
Final project area:	Data Mining and Machine Learning
Language:	English
Key words	*Bordetella pertussis*, Data Mining, Evolutionary analysis

Acknowledgements

First, I would like to thank Dr. Juanjo González and Dr. Anna Fàbrega from the Vall d'Hebron Hospital, who provided us the data of Bordetella pertussis. They were also part of the definition of the proposal and the research options, following the work that they are doing in the hospital. This master thesis would not have been possible without their contribution.

I would also like to thank my thesis supervisor, Dr. Laura Ruiz Dern, for her continuous support and guidance throughout this project. While she let me do the work in my own way, she was always there whenever I needed it.

Finally, my very profound gratitude to my mother for providing me with continuous encouragement since ever. Nothing would have been possible without her. Thank you.

Pol

Abstract

The whooping cough (Pertussis) is one of the main causes of vaccine preventable deaths worldwide. In recent years, it was observed a resurgence of the disease in countries with a high vaccination coverage. In this work we analyzed 339 strains of Bordetella pertussis, isolated in Catalonia between the years 1986 and 2015. From each strain we knew the dominant allele of the main antigens (ptxA, ptxP, prn and fim3) and their epidemiological profiles (MLVA, PFGE). For the analyses we used data exploration methods, prediction models, a Bayesian network and an autoregressive model.

The results indicate a specialization of Bordetella pertussis, initiated during the transitional period from the whole-cell vaccine to the acellular vaccine, in which the MLVA type 27 became dominant along with the alleles ptxP3 and prn2. Similar results were obtained in studies carried out in other countries. Moreover, the last outbreak of Pertussis in Catalonia (2011) might be related to a change in the fimbrial 3 allele, from number 2 to 1. Finally, from Pertussis weekly incidence in Catalonia, we determined that 2020 might be the next epidemic year, specifically at the end of the summer. The results of this project may help to develop more effective vaccines and to improve the preventive measures for Pertussis.

Key words: Bordetella pertussis, Data Mining, Evolutionary analysis

Resum

La tos ferina és una de les malalties evitables per vacunació que causa més morts arreu del món. En els últims anys, s'ha observat un ressorgiment de la malaltia en països amb alta cobertura de vacunació. En aquest treball s'han analitzat 339 soques de la Bordetella pertussis, l'agent causant de la tos ferina, isolades a Catalunya entre els anys 1986 i 2015. De cada soca se'n coneixien els al·lels dominants dels antígens principals (ptxA, ptxP, prn i fim3) i els seus perfils epidemiològics (MLVA, PFGE). Per els anàlisis s'han utilitzat mètodes d'exploració de dades, models predictius, una xarxa bayesiana i un model autoregressiu.

Els resultats indiquen una especialització de la Bordetella pertussis, iniciada durant el període de transició de vacuna completa a vacuna acel·lular, i en la qual el perfil MLVA 27 ha esdevingut dominant, juntament amb els al·lels ptxP3 i prn2. S'han obtingut resultats similars en estudis fets en altres països. A més, l'últim brot de tos ferina a Catalunya (2011) podria estar relacionat amb un canvi en l'al·lel de la fímbria 3, del número 2 al 1. Finalment, a partir de la incidència setmanal de tos ferina a Catalunya, s'ha determinat l'any 2020 com a potencialment epidèmic, concretament a finals d'estiu. Els resultats d'aquest treball poden ajudar a desenvolupar vacunes més efectives i a millorar les mesures preventives per a la tos ferina.

Paraules clau: Bordetella pertussis, Mineria de dades, Anàlisi evolutiu

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General description of the problem

The Bordetella pertussis [5], a Gram-negative bacteria, is the causative agent of the whooping cough, also known as Pertussis. It infects the human respiratory system and it is highly contagious. Although it affects people of all ages, it is especially dangerous in children and potentially life-threatening in babies less than one year old.

Over the last few years there has been an alarming increase in the number of Pertussis cases [62]. This resurgence of Pertussis is causing public health concerns and has lead to a renewal interest in research. There has always been some controversy regarding Pertussis vaccines efficacy and risks [11, 22]. In result, many changes in the vaccination programs were applied along the years. Nevertheless, Pertussis is still nowadays one of the main causes of vaccine preventable deaths worldwide.

Apparently, the protective effect of vaccines could have decreased due to B.pertussis evolutionary adaptation to vaccine immunity [8, 30]. In this project we studied the evolution of the B.pertussis bacteria, as well as its preventive measures, in order to explain and determine the causative factors of Pertussis resurgence. The ultimate goal was to predict B.pertussis trajectory, thus providing information that could lead to developing more effective preventive treatments or strategies for Pertussis.

1.2 About the data

The data to address this project were provided by the microbiology unit of the Vall d'Hebron Hospital (Barcelona). They contained information about 339 strains of B.pertussis bacteria, isolated from patients between 1986 and 2015, both included.

About the patient where a given strain was isolated, it was known the gender, age and vaccine doses. Of each strain, there was information about its epidemiological profiles [14] (PFGE, MLVA) and clade, the year in which it was isolated and its main antigen alleles. These last ones were the pertussis sub-unit A and promoter toxins (ptxA and ptxP) and the pertactin and fimbrial proteins. A more detailed explanation can be found in chapter 2.

Finally, data from public sources [29, 39, 42] were gathered to improve some of the results and to compare them to other similar studies.

1.3 Personal motivation

Data Science applications are demonstrating their utilities in almost every field. It is a fact that we are generating more data than ever, opening the doors to a whole new world of knowledge in research. Biology and Medicine, fields that have always interested me, are not an exception. Thus, I decided to undertake this project with the aim of learning more about these areas, and contributing from the data science point of view. Also, the results of this project could lead to improvements and discoveries in benefit of the society.

1.4 Project objectives

This project sought to deepen in the current knowledge of the B.pertussis bacteria. Its main objective was to understand the evolution of Pertussis. Specifically, how and why the mentioned variables (antigen alleles, epidemiological profiles, etc.) changed over time. For that, the first thing to do was to understand the current situation and what led to it. That is, study the state of the art. Then, these are some questions that were expected to be answered:

- How did strain antigens change over time? How does it relate to vaccine type?
- How do vaccine doses, strain antigens and/or epidemiological profile relate?

- Are gender and age something important (in this context)?
- Is it possible to cluster by epidemiological profile and/or other variables? If so, how do the variables change over time?
- Can the evolution be graphically or statistically modelled?
- Which insights can be found about the resurgence of Pertussis?
- Is it possible to predict (and thus, explain), the evolution of Pertussis bacteria?

Notice that these objectives were subject to change based on the project evolution. Having that in mind, the work planning was designed with the addition of some short periods to, if needed, redefine the objectives or the planning. Also, there was the possibility to gather new public data and to mark new objectives.

1.5 Approach adopted

First we will explain the state of the art to understand the current situation. For that, a research of information about the field in articles, thesis, books and other scientific papers was done. Next, we will proceed to analyse the data about the B.pertussis strains, starting with an Exploratory Data Analysis (EDA) to understand them better. That includes, statistical and data analyses, data cleaning, feature engineering and data visualization. After that, different but related ideas were considered and implemented:

- Studying the evolution by first clustering the data and seeing what variables change over time. Then, making a dynamic probabilistic model (a Bayesian network) to study how the B.pertussis population changed. This model allowed to validate what would happen when something (some variable) changes given a probability.
- Predicting the evolution of B.pertussis with an autoregressive model (AR) [33], using Pertussis incidence data to predict the next epidemic year. In addition, we tried to add information about the B.pertussis strains to improve the model.
- Gathering more data from public sources, either to generate better models or to compare the results with other similar studies. For instance, data of weekly incidence of Pertussis in Catalonia were collected to improve the AR model.

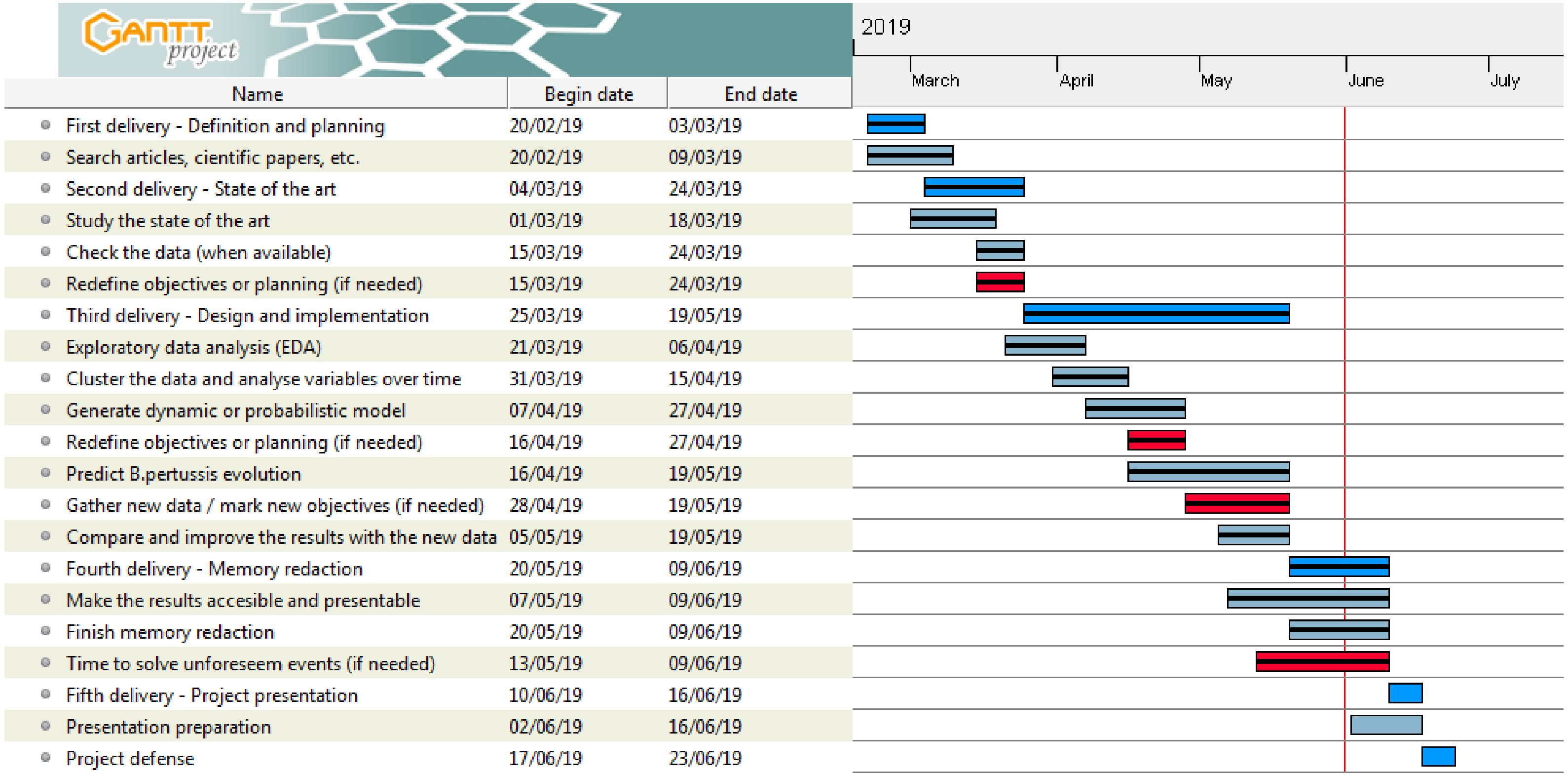
Regarding the chosen tools, we decided to mainly use the programming language Python, which is very popular in the Data Science field. Its packages contain everything needed for this project. In spite of that, Tableau was used to better visualize the data.

1.6 Work planning

This project was divided in six deliveries, starting on February 20 and ending on June 23, 2019. A Gantt Diagram [10] was used to make the planning, as shown in figure 1.1. As mentioned before, this was a research project and therefore it was subject to changes based on its evolution. Regarding that, the red lines of the diagram represented programmed but not exclusive redefinition periods. The six deliveries were in dark blue, while all the main tasks were in light blue.

1.7 Brief summary of the results

The results of this project are many. In first place, there are the visualizations that explain the evolution of the antigen alleles, as well as that of epidemiological profiles. Secondly, the gradient boosting models can predict the antigen alleles or the clade of a given strain, while the Bayesian network gives the conditional probabilities of any missing value. Third, there is the interactive forecasting of Pertussis incidence, obtained from the SARIMAX model, which can be used to anticipate a possible epidemic. Last but not least, the final products are this document itself with all the explanations and the restful API deployed to Heroku [52], which makes the results interactive and accessible anywhere.



Name	Begin date	End date
First delivery - Definition and planning	20/02/19	03/03/19
Search articles, cientific papers, etc.	20/02/19	09/03/19
Second delivery - State of the art	04/03/19	24/03/19
Study the state of the art	01/03/19	18/03/19
Check the data (when available)	15/03/19	24/03/19
Redefine objectives or planning (if needed)	15/03/19	24/03/19
Third delivery - Design and implementation	25/03/19	19/05/19
Exploratory data analysis (EDA)	21/03/19	06/04/19
Cluster the data and analyse variables over time	31/03/19	15/04/19
Generate dynamic or probabilistic model	07/04/19	27/04/19
Redefine objectives or planning (if needed)	16/04/19	27/04/19
Predict B.pertussis evolution	16/04/19	19/05/19
Gather new data / mark new objectives (if needed)	28/04/19	19/05/19
Compare and improve the results with the new data	05/05/19	19/05/19
Fourth delivery - Memory redaction	20/05/19	09/06/19
Make the results accesible and presentable	07/05/19	09/06/19
Finish memory redaction	20/05/19	09/06/19
Time to solve unforeseem events (if needed)	13/05/19	09/06/19
Fifth delivery - Project presentation	10/06/19	16/06/19
Presentation preparation	02/06/19	16/06/19
Project defense	17/06/19	23/06/19

Figure 1.1 Gantt diagram.

1.8 Brief description of the other chapters

Chapter 2 starts with an introduction of the whooping cough and its causative bacteria, the Bordetella pertussis. There is an explanation of the state of the art where the current situation, known problems, possible solutions and lines of research will be discussed. The main objective of this chapter is to introduce the problem and explain its importance, as well as giving a general understanding of the topic.

Chapter 3 contains the exploratory data analysis (EDA), which starts by describing and cleaning the data. Next, they are deeply explored and visualized, fulfilling some of the basic objectives of the project. The evolution of the epidemiological profiles is explained, as well as the relationship between the PFGE profiles and the MLVA types.

Chapter 4 is divided in three sections. Firstly, gradient boosting classifiers are trained to predict strain variables. In second place, a Bayesian network is built, which gives the conditional probabilities of any missing value of the strains variables. Thirdly, autoregressive models are used to forecast the Pertussis incidence.

In chapter 5 the results of this project are compared with other similar studies. Also, the classifiers are applied to external datasets and the obtained results are discussed. In the second section of the chapter, a RESTful API and its website interface are presented with the aim of making the results accessible.

Finally, in chapter 6 the results and methodologies of the project are discussed, as well as its future lines.

Chapter 2

State of the art

2.1 The whooping cough

The whooping cough [47], also known as Pertussis, is a disease caused by the bacteria Bordetella pertussis, of the genus Bordetella. There are many other Bordetella species, such as B.bronchiseptica, B.holmesii and B.parapertussis. The latter can also cause Pertussis, although a less severe version. B.parapertussis is responsible for only a minority of the cases, while B.pertussis is the main causative agent.

Pertussis is one of the main causes of vaccine preventable deaths worldwide. Practically all of them occur in undeveloped countries, where children are especially vulnerable due to the lack of vaccinations. Its incubation period is commonly 7 to 10 days and the disease can last for 4 months. The first symptoms are mild and very similar to the common cold, thus hindering the diagnosis. After a couple of weeks, however, the symptoms evolve to uncontrollable cough attacks that can cause vomits and breathing difficulties. After that stage (up to 10 weeks), the disease remains convalescent for 2 to 6 more weeks. Even in that last stage, the infected ones can still transmit the disease to others. Pertussis spreads through air from the mucous membranes, being thus especially contagious during the catarrhal period [15].

In order to prevent the disease, in most developed countries, including Spain, three vaccine doses are applied at early ages (less than 18 months), even though a fourth and fifth doses are recommended during infancy to ensure complete protection. Nevertheless, vaccine effectiveness has recently been questioned due to the recent worldwide resurgence of the disease. It is true that diagnosis techniques have improved, but this is not enough to explain the increase of cases. Using data from the Spanish epidemiological bulletin [42], figure 2.1 shows the incidence of Pertussis in Spain between 1998 and 2016.

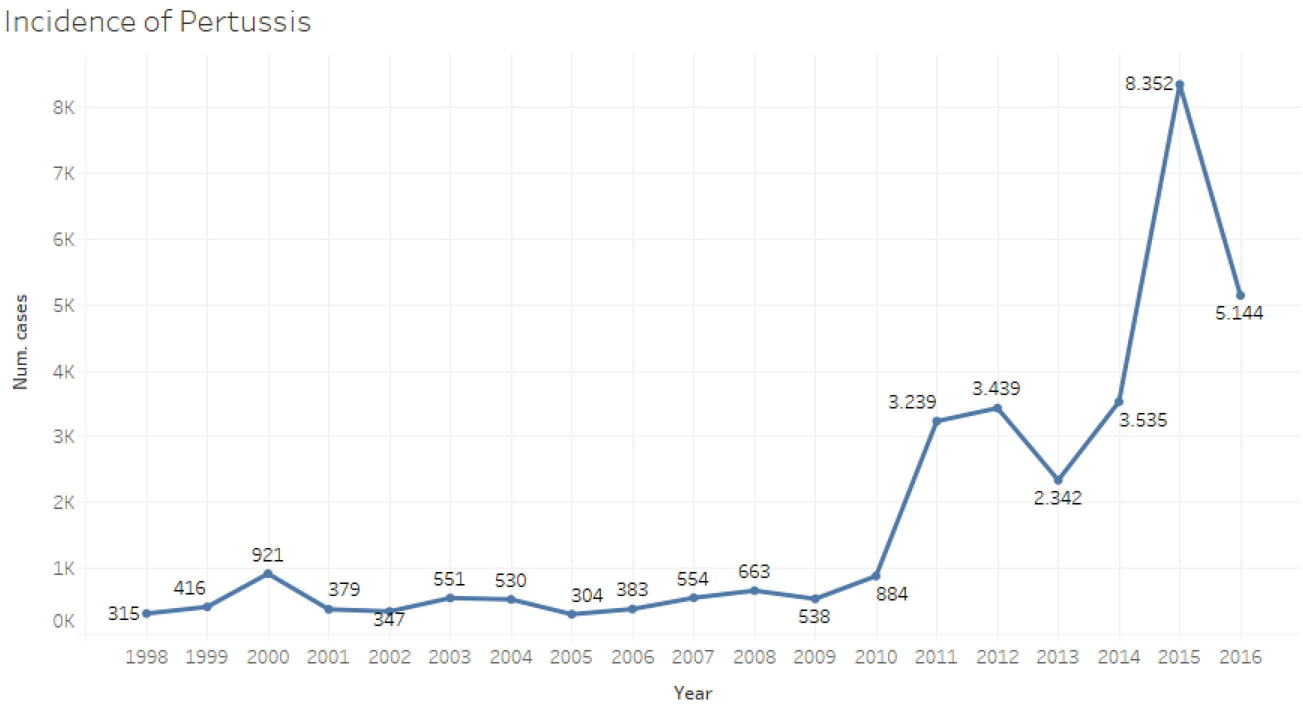


Figure 2.1 Pertussis incidence in Spain. Data from [42]

There seems to be a cyclic epidemic pattern every 3 to 5 years, with its peaks (2000, 2003, 2008, 2012 and 2015) and descents. Since 2010, this cyclical pattern has achieved and maintained higher ranges.

2.1.1 Pertussis vaccine history and problems

The first vaccine for the whooping cough was introduced in the 40s [11, 22] as a Whole-Cell Vaccine (WCV). This vaccine contained a huge amount of antigens and did completely inactivate the pathogen. However, it also had undesired secondary effects, such as low fever, soreness, redness and swelling. In response, Acellular Vaccines (ACVs) were developed and first introduced in the 80s. These contained purified proteins (antigens) of the B.pertussis bacteria such as pertussis toxin, pertactin, filamentous hemagglutinin, as well as fimbrial proteins (types 2 and 3). All these components could induce immunity responses without having adverse effects. Since various studies showed the efficacy of the ACVs, the majority of developed countries adopted them around the 90s. Nowadays, Pertussis ACV is usually administered in combination with Diphtheria and Tetanus vaccines (DTPa). In Spain, the DTPa vaccine was introduced in 2005.

However, there are countries that still use the WCV. The full country list can be consulted at [13].

Nevertheless, recent studies indicate that the use of ACV is strongly correlated to the resurgence of Pertussis. It has been found that ACV-induced immunity last less than that induced by WCV, thus increasing the risk of infection in older children. Also, there are clear differences between the immunity induced by each of the two vaccine types. For instance, even though both block the disease, ACV may permit transmission to non-vaccinated individuals. ACV-induced mucosal immunity seems to be absent, thus allowing carriage. An ideal vaccine should grant both direct and indirect protection.

Apparently the B.pertussis pathogen adapted to avoid the ACV-induced immunity. In fact, the protective role of each of the ACV components is still not clear, since different combinations have shown similar results, being the pertactin the only component used in all ACVs. This caused a spread of pertactin resistant strains. Current research lines are trying to solve these problems. For example, adding more amounts of fimbrial proteins may increase vaccine efficacy [53]. Also, it seems that the best ACVs protection is achieved using at least 3 components [25].

Among the solutions, changes on the prevention strategies have been discussed. Being the children and especially the newborns the most vulnerable group, vaccinating their parents or at least the pregnant women might potentially help. For instance, a recent maternal vaccination programme has successfully reduced Pertussis cases in infants [56], thus supporting maternal immunization as a preventive method. Moreover, in Spain mothers are being vaccinated in their third trimester of pregnancy since 2016, which has also reduced the number of cases among newborns [24].

Finally, the resumption of the WCV or the development of novel Pertussis vaccines [35] are being considered, as well as some possible boosters for the ACVs. Therefore, having more knowledge about Pertussis would certainly help.

2.1.2 Antigens of the ACVs

There are many different types of ACVs for Pertussis, each one containing a different combination of antigens. As mentioned before, the best results are obtained by combining at least 3 of the 5 components. These are described below:

- **Pertactin:** is one of the most virulence factors of the B.pertussis bacteria. It promotes the adhesion to tracheal epithelial cells. The purified pertactin is the most used component in the ACV. This fact has caused the spread of pertactin resistant strains. There are 7 types of pertactin in the ACVs, numbered as PRN 1 to 7.

- **Pertussis toxin:** is the toxin that B.pertussis produces. The promoter for the protein synthesis is called ptxP and it has many alleles. The toxin is divided in two parts, A and B. The sub-unit A (ptxA), enzymatically active, is weakened and used as a component of the ACVs (alleles 2 and 4). The sub-unit B is the cell binding.

- **Fimbrial proteins (2/3):** are appendages of B.pertussis that enhances its attachment abilities. There are two main types used in the ACVs: fimbrial 2 and fimbrial 3.

- **Filamentous hemagglutinin:** is a protein that B.pertussis uses to adhere into the respiratory tract, concretely in the ciliated epithelial cells. It is another type of fimbrial that is used on the ACVs.

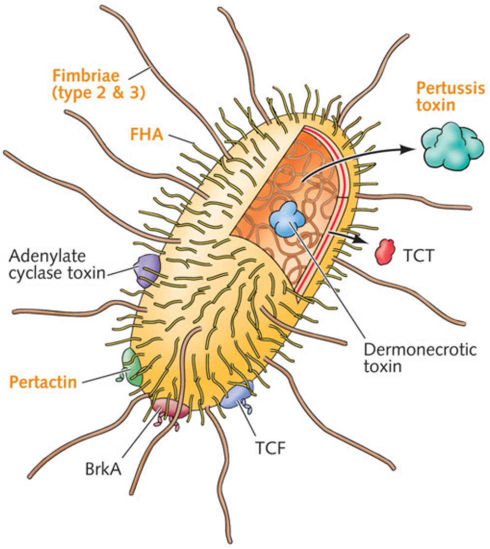


Figure 2.2 B.pertussis organism and its antigens. Image from [16].

The data provided for this project contains the allele types of the ptxA, ptxP, prn and fim3 antigens of each isolated strain. These are considered the most virulent ones, thus being the most studied.

2.2 B.pertussis genomic evolution

The latest technological advances in the genomic field have given a huge amount of valuable information about B.pertussis. The B.pertussis bacteria evolved as species by losing DNA and by intragenomic recombination. Specifically, IS481 element is highly repeated in B.pertussis genome [26]. An insertion sequence (IS) is the simplest transposon type [2]. These are genetic elements that can be transposed and thus integrate with the DNA, causing mutations. The comparison among recent and old B.pertussis strains suggest that the gene loss is still ongoing.

An analysis of 343 B.pertussis strains genome showed how the organism has emerged within the last 500 years. Also, it warned about the rapid evolution of the species and why this can enable a vaccine escape [21]. Figure 2.3 shows this worldwide evolution.

It is clear that the introduction of the WCV and ACVs hugely shaped the B.pertussis species. The ACV introduction lead to an expansion of ptxP3 lineage. This supports the theory of genetic adaptation due to vaccination being the cause of Pertussis resurgence, even though there is no definitive evidence yet. Either way, this raises more concerns about ACV efficacy against Pertussis.

The next step should be to use all these new data to understand and, maybe, predict the trajectory of B.pertussis. Functional genomic analysis, which try to describe the functions and interactions between genes, might provide a powerful approach to monitor active processes. That is, changes in gene expression levels during the course of Pertussis infection, from the initial colonization of the bacteria to the disease evolution and transmission.

2.3 B.pertussis epidemiological profiles

Strain typing has become an important tool for epidemiological surveillance of bacteria, especially for the most resistant and virulent ones [44]. B.pertussis strain typing has been very important to study the causes of Pertussis resurgence, revealing a huge change in the population after the introduction of the vaccines.

Although there are several methods for strain typing, the following two are the most commonly used for B.pertussis: Pulsed-Field Gel Electrophoresis (PFGE) and Multiple-Locus Variable number of tandem repeat Analysis (MLVA) [14].

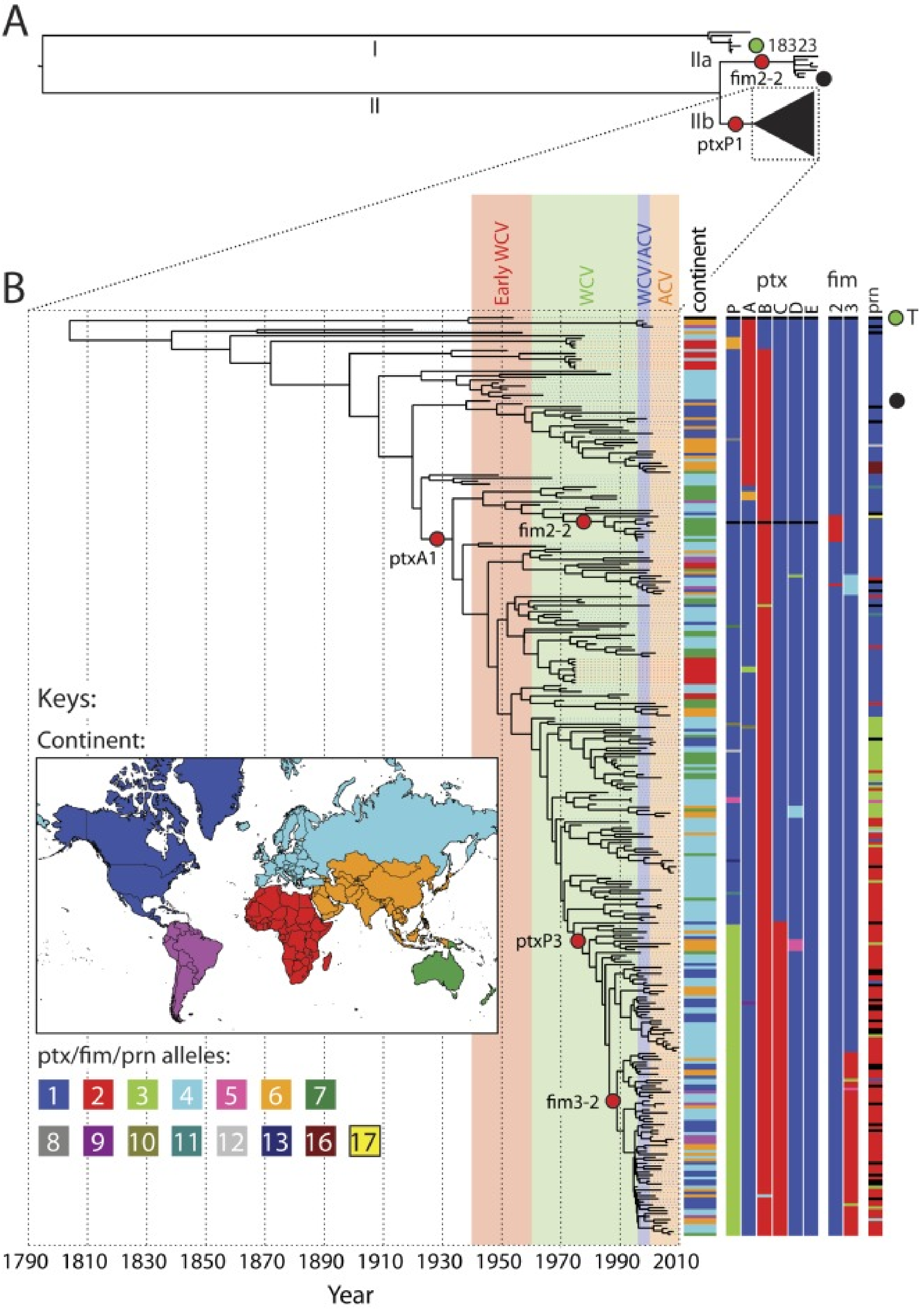


Figure 2.3 Bayesian phylogenetic tree of B.pertussis. Image from [21] (figure 1 of the article).

Worldwide evolution of the B.pertussis. Notice that WCV and ACV periods are shown in the background, while each red dot point to a significant change in the alleles of the antigen.

2.3.1 PFGE

Nucleic acid electrophoresis is used to fragment DNA or RNA. It uses a viscous medium (gel) where the nucleic acid molecules are set. Since these are negatively charged, applying an electric field makes them move towards the positive charge. The separation of the molecules takes place due to the properties of the gel. The size of the molecules determine the pace at which them will pass through it (the high-sized will encounter more resistance).

The previous technique is unable to separate very large DNA molecules. To solve that, an alternative technique called Pulsed-Field Gel Electrophoresis was introduced. It uses an alternating voltage gradient (the voltage periodically switches directions) that allows to analyse the large sequence in fragments. This technique improves the resolution of these larger DNA molecules. PFGE is sometimes used as a genetic fingerprint technique.

2.3.2 MLVA

Before explaining this technique, it is necessary to introduce some terminology. A tandem repeat (TR) in DNA happens when a certain pattern of nucleotides is repeated consecutively. A variable number tandem repeat (VNTR) is the location in a genome (loci) where a tandem repeat occurs, which is usually used as a DNA fingerprint. Then, a Multiple-Locus VNTR Analysis (MLVA) is a technique that takes advantage of the polymorphism of VNTRs and is used to perform molecular typing.

MLVA assesses the number of repeats in a selected group of VNTR loci. It uses a polymerase chain reaction (PCR) to amplify and measure each of the VNTR. The resulting numbers of repeats is referred as the MLVA profile. Each profile is unique and can be used to compare with other data sources. Also, it might be of utility to cluster data by profile.

2.3.3 Comparison of PFGE and MLVA

PFGE is highly discriminative and it is considered the standard among the molecular typing techniques. However, this method is expensive and high time consuming. It does not discriminate between all unrelated isolates and its results may vary slightly depending on the technician that performs it [6]. Therefore, interlaboratory comparison of results becomes difficult.

The MLVA technique is less discriminative but it may be able to differentiate fast-evolving strains that might seem equal when using PFGE. Also, it gives full reproducibility of results. MLVA is commonly used as a complement of PFGE to get more details about the bacteria that have similar PFGE patterns [3]. It seems that the correlation of the results obtained with both techniques is around 73% [46].

2.4 Data science applications in biology

Data Science, as shown in figure 2.4, is the intersection of computer science, mathematics and an expert domain. Being this domain Biology, it becomes what it is called bioinformatics [31], the purpose of which is to understand biological data. Its applications encompasses, among other things, cancer detection, genome exploration, personalized medicine and nutrition, treatment outcome predictions and vaccine developments.

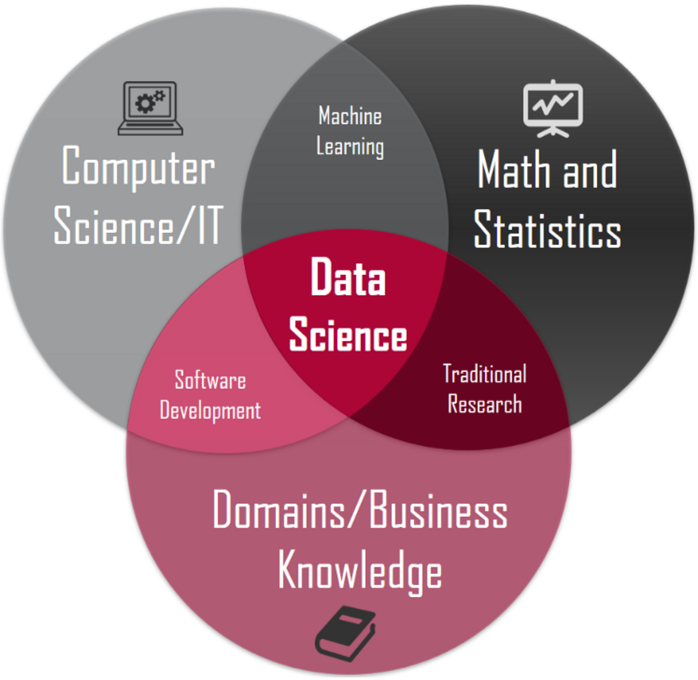


Figure 2.4 Data Science. Image from [28] (figure 1 of the article).

As mentioned before, massive amounts of new biological data are being generated, especially in the genomic sequencing field. Therefore, this is an ideal scenario to deepen and search for hidden knowledge using data mining, machine learning or mathematical modelling techniques [37]. A good example is systems vaccinology (application of systems biology [18] in vaccine development), which are being used to better understand immune responses from candidate vaccines [55].

2.4.1 Machine learning techniques

There are lots of different machine learning algorithms each one with its particular characteristics. Thus, choosing the right ones may become a difficult task. Nowadays, deep learning [32] (a machine learning subfield) is showing promising results in many areas. Some example techniques are convolutional neural networks (CNNs) to classify image-like data, long-short term memory neural networks (LSTMs) to deal with sequence data and generative models to generate synthetic new data. The possible applications of all this are huge, but large amounts of data are required to train the models.

As for this project concerns, a simpler approach (which require less data) is more indicated. A more traditional way to model time series data is to use autoregressive (AR) methods [33]. The idea behind it is to take input variables from previous time steps (lag variables) to predict the next ones. Also, visualization and graph techniques are useful to understand the data and how they relate to each other. For the sole purpose of serving as an example, figure 2.5 shows how different diseases relate with each other from a gene perspective. The data are clustered by disorder class.

Since the data are not always labelled, clustering techniques are used to find similar patterns, creating clusters with similar characteristics. One of the most popular clustering technique is called k-means [63], which divides the data in a number 'k' of groups. Every data point is assigned to its closest group. For that, it is common to use the Euclidean distance and the group mean.

2.5 Summary of the state of the art

Putting all the pieces together, the causes for the resurgence of Pertussis are still not clear. However, the majority of the studies coincide in some factors:

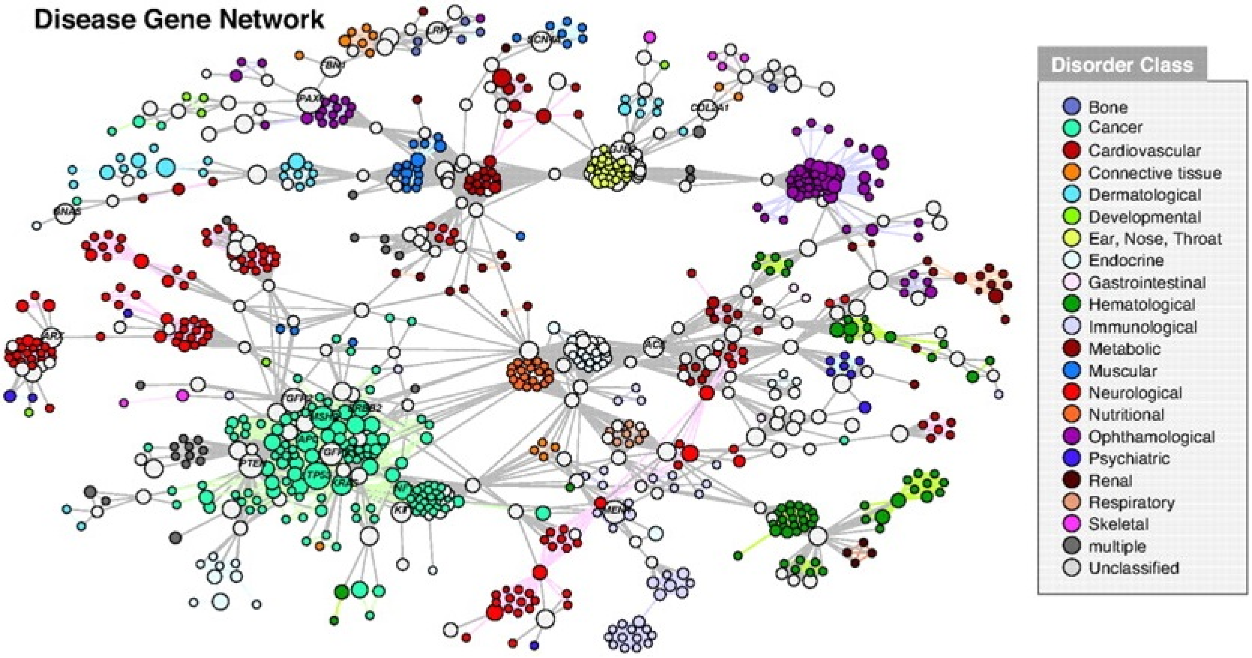


Figure 2.5 Gene network by disorder class. Image from [40] (figure 2 of the article).

Example of how visualization and graph techniques can be used to explain the relations of the data. Each point represents a gene and its color identifies the associated disorder class.

- Pertussis resurgence is causing health concerns. It is one of the main causes of vaccine preventable deaths worldwide.
- ACVs are safe and effective, but its protection is suboptimal and last less.
- Vaccine protection decrease with time and more booster doses are recommended.
- To protect children (the most vulnerable group), parent vaccination strategies are considered.
- Current ACVs allow B.pertussis carriage, probably due to the lack of mucosal-immunity.
- The protective role of each of the ACVs components is still not clear, but using 3 or more components show better results.
- The ACV components are: prn, fim 2 and 3, FHA and ptxA.
- Pertactin is used in all (or almost) of the ACVs, which have caused the spread of pertactin resistant strains.
- B.pertussis has evolved in adaptation to vaccines, emerging from the more resistant strains.

- Genomic analyses might be key to understand and predict B.pertussis trajectory.
- Strain typing techniques are essential for epidemiological surveillance. PFGE and MLVA are two typing techniques and are used to type B.pertussis strains.
- MLVA is usually used as a complement of PFGE to get more details about similar PFGE profiles.
- The applications of Data Science techniques in the Biology field are promising, but it is important to choose the best techniques for each problem. The amount of data is also key to determining which technique to use.
- Autoregressive methods are a good approach to deal with small time series data. Also, visualization and graph techniques are useful to understand the relationships between the data. Finally, clustering techniques are used to group the data by similarity.

Chapter 3

Exploratory data analysis

3.1 Data description

The data provided by the Vall d'Hebron Hospital have information from 339 B.pertussis strains, isolated from patients between 1986 and 2015. The variables and their meaning are the following:

Ref.Strain: name that identifies the strain.
Date: year of isolation.
Gender: gender of the patient from where the isolation was taken.
Age: age of the patient from where the isolation was taken.
Vaccine doses: number of vaccine doses of the patient from where the isolation was taken.
PFGE profile: profile obtained after performing strain typing with the PFGE technique.
Clade: clade of the isolation calculated from the similarity of PFGE profiles.
MLVA type: type obtained after performing strain typing with the MLVA technique.
ptxA allele: dominant allele of the pertussis toxin sub-unit A.
ptxP allele: dominant allele of the pertussis toxin promoter.
prn allele: dominant allele of the pertactin.
fim 3 allele: dominant allele of the fimbrial protein type 3.

3.1.1 Data cleaning and missing values

Table 3.1 shows a basic description of the data. There are a lot of missing (unknown) values but only a few of them can be treated properly. For instance, vaccine doses may be inferred from age, since there is a high correlation between both variables. However, missing values

Table 3.1 Data summary.

```
REF.STRAIN       DATE      GENDER        AGE      VACCINE.DOSES fim3.allele
BP1     :  1   2000    : 48  F  :166   1 m     : 66  0  : 63      1  :129
BP10    :  1   2011    : 44  M  :163   2 m     : 44  1  : 24      2   :102
BP100   :  1   1989    : 34  M  :  1   unk     : 41  2  : 12      NA's:108
BP101   :  1   2007    : 27  unk:  9   3 m     : 28  3  : 25
BP102   :  1   2012    : 23            1 y     : 19  4  : 24
BP103   :  1   2008    : 19            4 m     : 13  5  : 29
(Other):333   (Other):144             (Other):128   unk:162


PFGE.profile  Clade        MLVA.type   ptxA.allele ptxP.allele prn.allele
VH19    :83   1    : 26   27      : 41  1   :231   1    : 37    1    : 19
VH2     :54   2    : 16   28      :  5  NA's:108   3    :187    2    :200
VH26    :43   3    :116   60      :  5             5    :  1    3    : 10
VH20    :38   4    :141   70      :  5             10   :  1    9    :  2
VH8     :17   5    :  4   16      :  3             11   :  1    NA's:108
VH22    :15   NA's: 36    (Other): 13              15   :  4
(Other):89                NA's   :267              NA's:108
```

from age, gender, epidemiological profiles, clades or pertussis alleles can not be filled.

Regarding the antigen alleles, it is important to point out that ptxA has never changed, so it does not provide useful information. Also, notice how the age is conformed by a number plus a character (y = year, m = month, d = day). For that, the age variable was transformed to a numerical scale by month. Finally, some writing errors were corrected, such as the extra spaces in the variable gender. In numerical variables, the unknown values were set to -1.

3.1.2 Basic feature engineering

As said before, for those patients whose age is known but their vaccine doses are not, a simple inference from age was applied. The purpose of doing so is to seek correlations between the vaccine doses and the strain characteristics. In addition, the type of vaccine administered (WCV, ACV or none) can be determined from the date, the patient's age and the number of vaccine doses. In order to do that properly, the data were clustered by vaccine doses and the outliers (more than 3 standard deviation apart) were eliminated. Then, ranges from averages were created to classify the unknown data points. For a final and more accurate decision, the obtained results were compared to the recommendations (which have not changed much over time) of the Spanish vaccine calendar [50], as seen in table 3.2.

According to [36], the coverage of the Pertussis vaccine (DTPa) in Spain is over 97%. Therefore, it was decided to apply the official recommendations to fill the missing values of vaccine doses variable. That is, we assumed that every newborn older than 2 months was vaccinated at least once. Figure 3.1 shows the final distribution of vaccine doses.

Table 3.2 Comparison of vaccine doses and age (months).

Doses	1	2	3	4	5
Recommended month	2	4	6	18-36	48-72
Calculated month	2	4	8	27	-

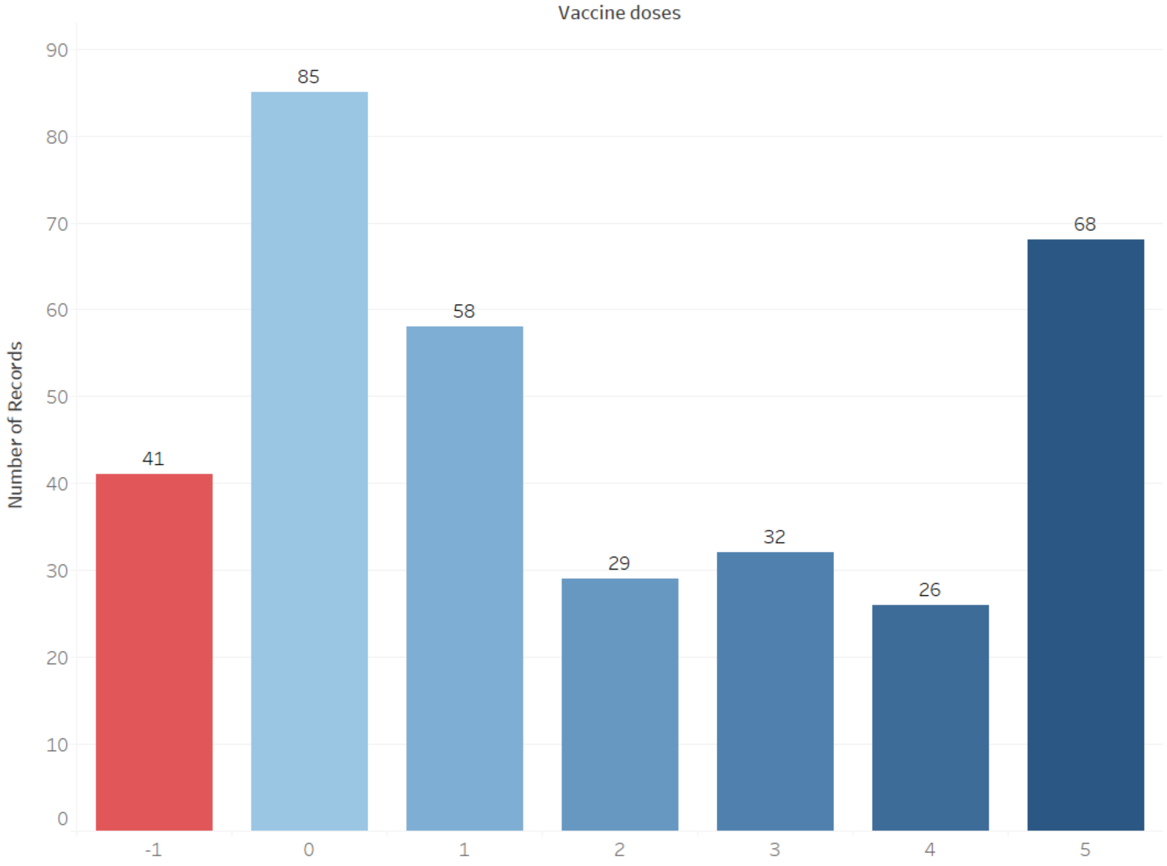


Figure 3.1 Number of isolated strains by vaccine doses.

The majority of the strains were isolated from patients with 0 or 1 vaccine dose, or from those with 5 vaccine doses. The -1 value represent those whose age and vaccine doses were unknown.

Furthermore, a new variable (vaccine type) was created to represent what kind of vaccine was the one administered to each patient. In Spain the WCV was used until 1997, when the ACV started replacing it. The year of birth of a patient can be calculated using the date of isolation and the age of the patient at that time. The inferred vaccine doses variable can be used to assign the unknown values. Figure 3.2 shows the result of this.

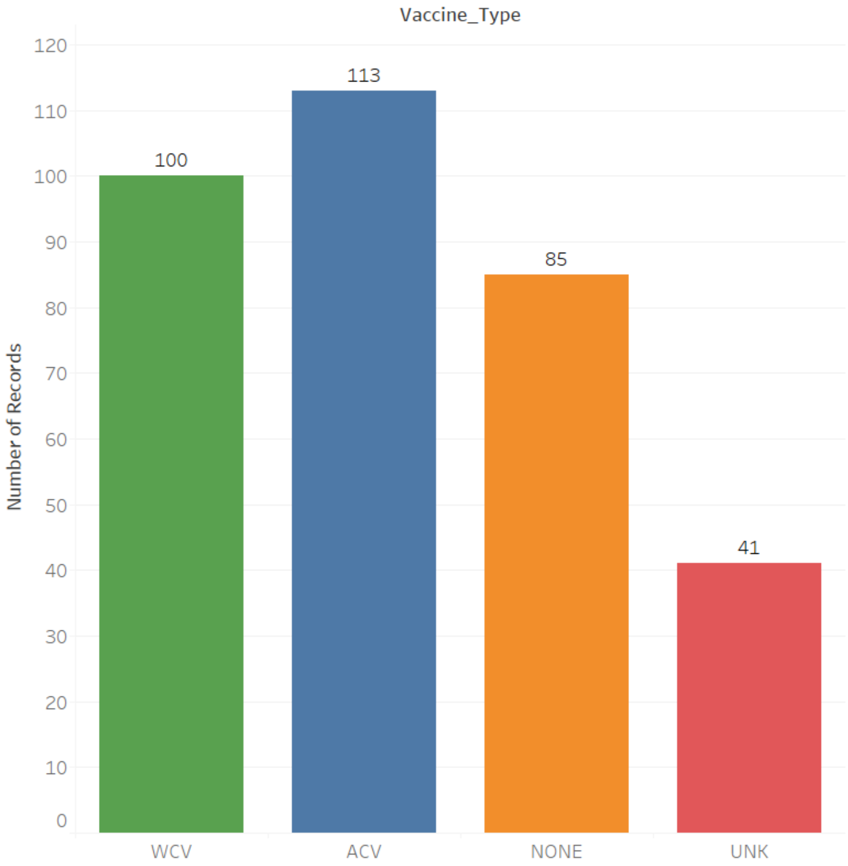


Figure 3.2 Number of isolated strains by vaccine type.

There were 100 isolated strains taken from patients vaccinated with WCV and 113 with ACV. The 85 NONE are mostly from babies less than 2 months that were not vaccinated yet. The 41 UNK are from patients whose vaccine doses are unknown. This was calculated from the inferred vaccine doses variable (NONE = 0, UNK = -1).

To make a comparison, figure 3.3 shows the difference between the use of the original and the inferred data to assign the vaccine type. Basically, the most incomplete data is also the oldest. Therefore, it is correct to assume that most of the unknown values should be WCV.

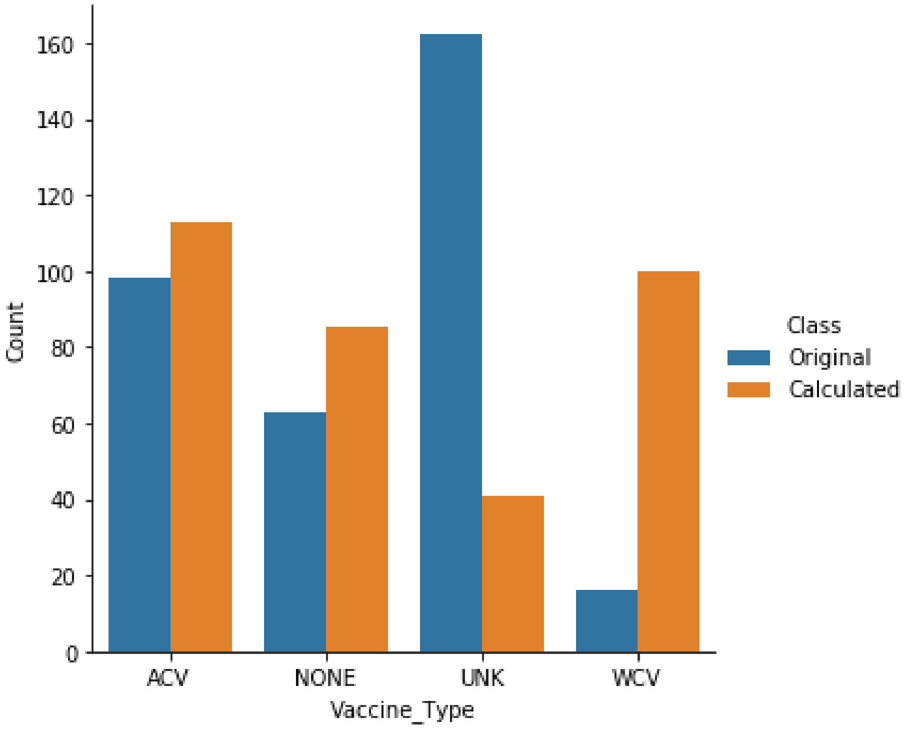
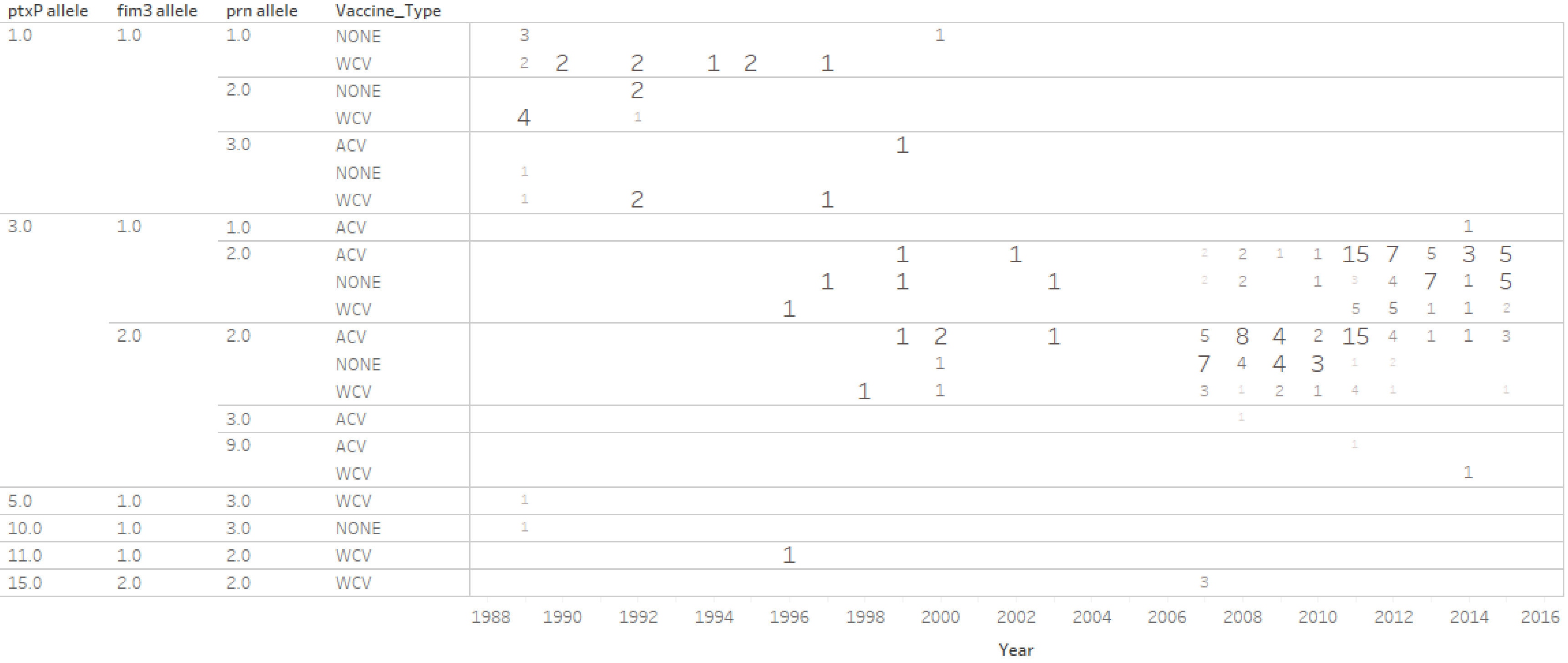


Figure 3.3 Comparison of vaccine type assignation methods.

3.2 Data exploration and visualization

In this section the cleaned data will be explored more deeply and some of the basic objectives of this project will be achieved. For that, both the original and the inferred data were used. However, the results were similar for both data sets, since no statistical relationships were found between the characteristics of the strain and the inferred variables (vaccine doses and vaccine type).

Figure 3.4 shows the evolution of the antigen alleles between 1989 and 2015. The main ptxP allele changed from 1 to 3 with the introduction of the ACV. Also, the pertactin allele number 2 became dominant. Moreover, the last big Pertussis epidemic (2011-2015) might be explained by a change in the fim3 allele from 2 to 1. Notice that ptxA was not analysed, since all the samples contained the same allele (ptxA1).



ptxP allele	fim3 allele	prn allele	Vaccine_Type	1989	1990	1992	1994	1995	1996	1997	1998	1999	2000	2002	2003	2007	2008	2009	2010	2011	2012	2013	2014	2015
1.0	1.0	1.0	NONE	3									1											
			WCV	2	2	2	1	2		1														
		2.0	NONE			2																		
			WCV	4		1																		
		3.0	ACV									1												
			NONE	1																				
			WCV	1		2				1														
3.0	1.0	1.0	ACV																				1	
		2.0	ACV									1		1		2	2	1	1	15	7	5	3	5
			NONE							1		1			1	2	2		1	3	4	7	1	5
			WCV						1											5	5	1	1	2
	2.0	2.0	ACV									1	2		1	5	8	4	2	15	4	1	1	3
			NONE										1			7	4	4	3	1	2			
			WCV								1		1			3	1	2	1	4	1			1
		3.0	ACV														1							
		9.0	ACV																	1				
			WCV																				1	
5.0	1.0	3.0	WCV	1																				
10.0	1.0	3.0	NONE	1																				
11.0	1.0	2.0	WCV						1															
15.0	2.0	2.0	WCV											3										

Figure 3.4 Antigen evolution by vaccine type.

The numbers correspond to the isolations of a given combination in a given year. During the WCV vaccine period (1989-1997), the most frequent combination of alleles was the ptxP1, with the fim3-1 and a balanced prn. This situation changed rapidly with the introduction of the ACV (1998-2015) and the main allele combination became the ptxP3, with a balanced fim3 and a dominant prn-2. Finally, between 2007 and 2010, fim3-2 became dominant. However, during the last and most epidemic years (2011-2015) there was a clear change from the fim3-2 to the fim3-1. This might relate to the cyclic pattern of Pertussis, which achieved the largest values of the registered data.

The Pearson correlation coefficient was used to generate a heatmap, shown in figure 3.5. These correlations confirm the hypotheses made in the previous paragraph, as well as showing how the antigen alleles are related, except for the fim3-1. This is because it was present during the 90s, then almost disappeared in the 00s to come back in 2011. Figure 3.6 visualizes the fim3 evolution.

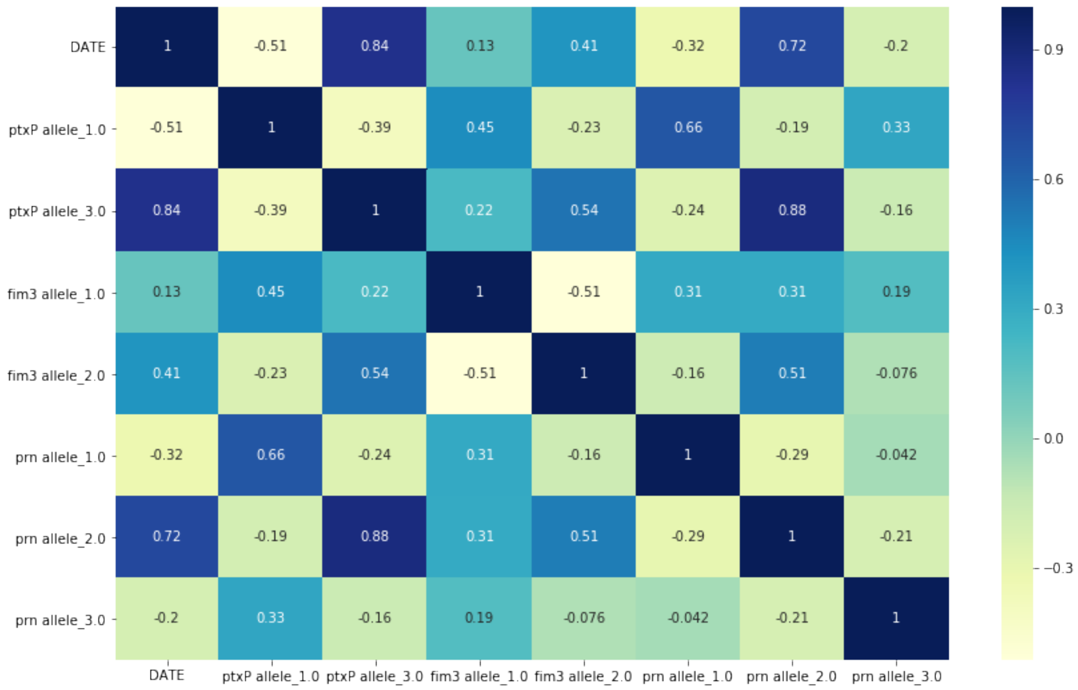


Figure 3.5 Antigen correlation heatmap.

Date is highly correlated to ptxP3 and to prn2, as well as quite correlated to fim3-2. On the other hand, ptxP1, fim3-1 and prn(1,3) constitute another correlated group.

Regarding the vaccine type, figure 3.7 shows that it has no clear relation with the antigen alleles (apart from the date). Notice in the previous figure 3.4 that there are always few (or a single one) dominant combinations of alleles at a given period, regardless of the vaccine type. Moreover, the number of vaccine doses also shows no statistical relation with them. There is no information about the decision criterion between isolating a strain from one patient or another. Therefore, the number of samples of each vaccine type is not a good indicative of the total infected population. For instance, having more samples of patients with the ACV does not mean that ACVs are less effective. That is, there is not enough information within the data to extract proper conclusions about that.

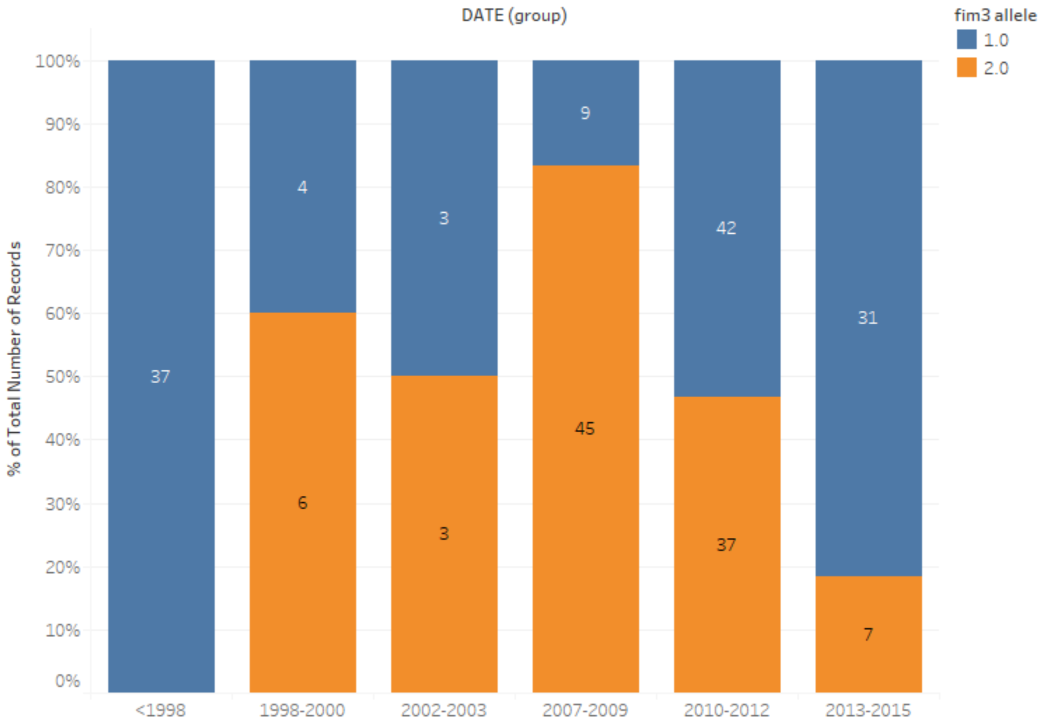


Figure 3.6 Fim3 evolution.

The fim3-1 allele was dominant until the 00s. Then it slowly changed to the type 2, which became dominant in 2008. In 2011, fim3-1 took over the relief, coinciding with the last and biggest epidemic cycle.

3.3 Evolution of PFGE profiles

The PFGE profile is available for all 339 strains. Also, the clade variable clusters these profiles in groups of at least 82% of similarity (information provided by the Vall d'Hebron Hospital). This can be used to model the evolution of B.pertussis over the years, as well as to determine which profiles induced the changes in the antigen alleles. As explained in the state of the art, PFGE typing technique is highly discriminative but its results may fail to differentiate fast-evolving strains, which could be the case of B.pertussis. MLVA variable might be used as a complement to differentiate similar PFGE patterns, even though our sample contains a lot of missing values.

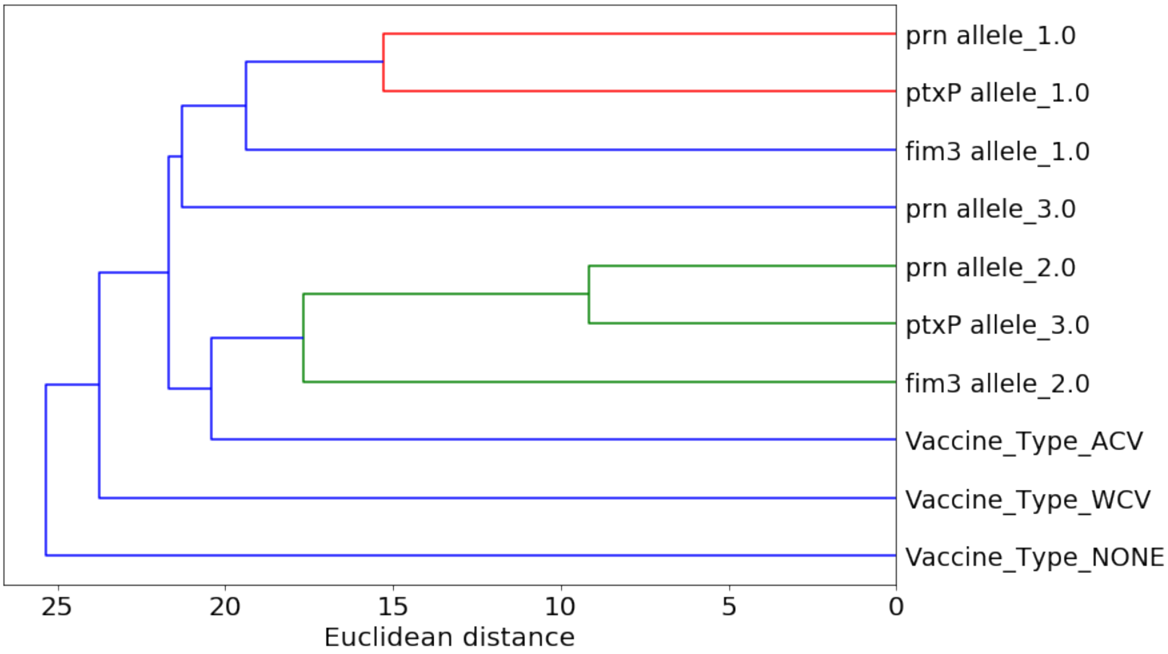
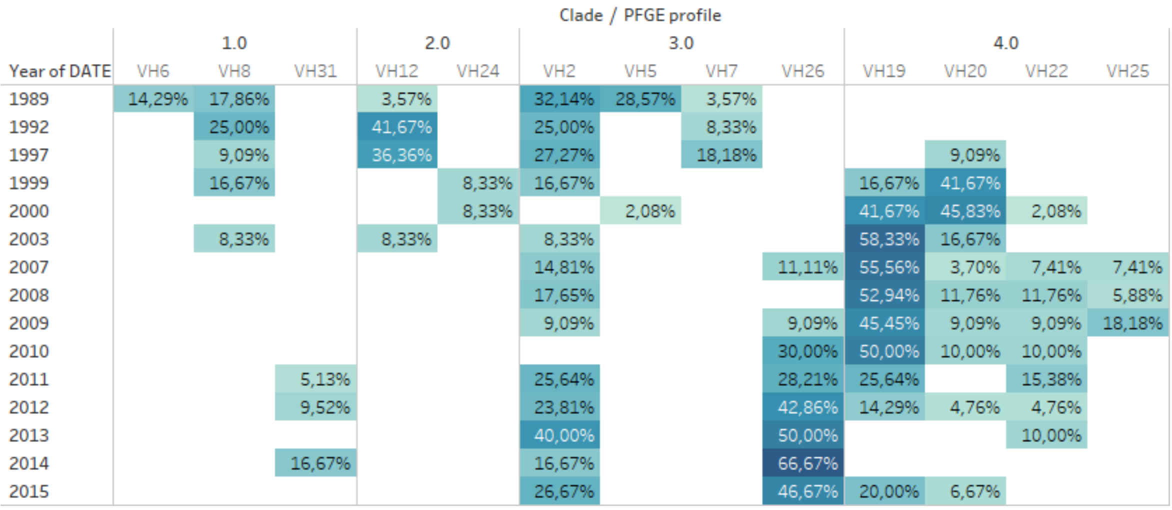


Figure 3.7 Dendrogram of antigens and vaccine types.

Each color represents a cluster based on the similarity of the data points that contain each one of the attributes.

Figure 3.8 shows the evolution of PFGE profiles over time, grouped by clade. The percentages were calculated from the number of isolated strains in a given year. Regarding the clades, figure 3.9 shows no clear patterns in terms of antigen alleles for clades number 1 and 2 (in this case, percentages were calculated by PFGE profile). Furthermore, the clades 3 and 4 are clearly differentiated by the presence of the fim3-1 and fim3-2 alleles, respectively.

The evolution of the PFGE profiles is highly correlated with the changes in the fim3 alleles, as figure 3.10 indicates. The year 2000 was the peak of an epidemic cycle, coinciding with the appearance of the VH19 and VH20 and the first cases of fim3-2 and ptxP3 alleles. Between the years 2000 and 2010, the VH19 became dominant along with some other minor profiles. After that, during the year 2011 there was a clear balance among clades 3 and 4. In fact, clade number 3 is basically defined by the dominance of fim3-1 allele, while clade number 4 is characterized by the fim3-2 allele. Finally, after the year 2011 it was when Pertussis cases soared along with the VH2 and VH26 profiles.



Clade / PFGE profile

	1.0			2.0		3.0				4.0			
Year of DATE	VH6	VH8	VH31	VH12	VH24	VH2	VH5	VH7	VH26	VH19	VH20	VH22	VH25
1989	14,29%	17,86%		3,57%		32,14%	28,57%	3,57%					
1992		25,00%		41,67%		25,00%		8,33%					
1997		9,09%		36,36%		27,27%		18,18%			9,09%		
1999		16,67%			8,33%	16,67%				16,67%	41,67%		
2000					8,33%		2,08%			41,67%	45,83%	2,08%	
2003		8,33%		8,33%		8,33%				58,33%	16,67%		
2007						14,81%			11,11%	55,56%	3,70%	7,41%	7,41%
2008						17,65%				52,94%	11,76%	11,76%	5,88%
2009						9,09%			9,09%	45,45%	9,09%	9,09%	18,18%
2010									30,00%	50,00%	10,00%	10,00%	
2011			5,13%			25,64%			28,21%	25,64%		15,38%	
2012			9,52%			23,81%			42,86%	14,29%	4,76%	4,76%	
2013						40,00%			50,00%			10,00%	
2014			16,67%			16,67%			66,67%				
2015						26,67%			46,67%	20,00%	6,67%		

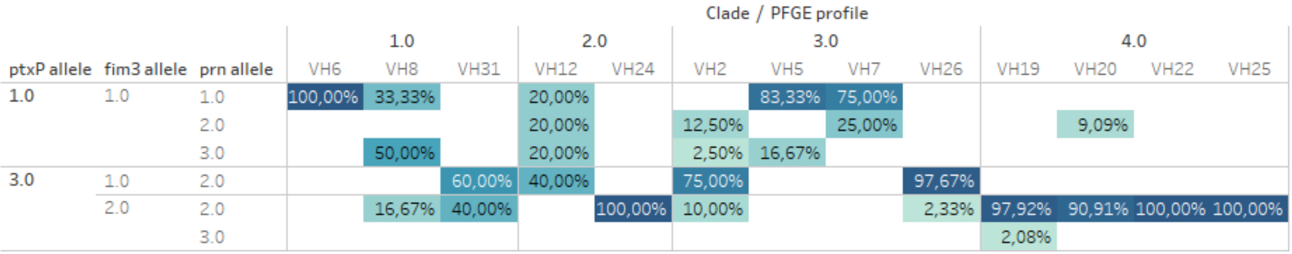
Figure 3.8 PFGE profiles over time.

The percentages were calculated from the number of strains per year (the higher the percentage, the darker). The PFGE profiles have been evolving over time: some of them disappeared while other new ones appeared. For instance, clades 1 and 2 mostly lived in the WCV period. Notice how VH2 (clade 3) was one of the main profiles before 1998, then it almost faded, just to return in 2011. Also, VH19 and VH20 were dominant during the 00s, being replaced later for the VH2 and VH26. These patterns are very similar to those of the fim3 antigen alleles.

3.3.1 Relationship between PFGE profiles and MLVA types

The evolution of the MLVA types has similar patterns to those of the PFGE profiles. During the WCV period there was a balance between many different MLVA types, as shown in figure 3.11. However, during the transition period (1998-2003), most of them disappeared and the MLVA-27 quickly became dominant.

A very interesting aspect is that the most aggressive PFGE profiles seem to be related with the MLVA-27, as shown in figure 3.12. In fact, this MLVA type has been present since the WCV period, which was dominated by the PFGE profiles VH2, VH7 and VH8. These profiles changed to the VH19, VH20 and VH22 during the transition period (clade 4), causing the



			Clade / PFGE profile												
			1.0			2.0		3.0				4.0			
ptxP allele	fim3 allele	prn allele	VH6	VH8	VH31	VH12	VH24	VH2	VH5	VH7	VH26	VH19	VH20	VH22	VH25
1.0	1.0	1.0	100,00%	33,33%		20,00%			83,33%	75,00%					
		2.0				20,00%		12,50%		25,00%			9,09%		
		3.0		50,00%		20,00%		2,50%	16,67%						
3.0	1.0	2.0			60,00%	40,00%		75,00%			97,67%				
	2.0	2.0		16,67%	40,00%		100,00%	10,00%			2,33%	97,92%	90,91%	100,00%	100,00%
		3.0										2,08%			

Figure 3.9 Antigen alleles by PFGE profile.

The percentages were calculated by PFGE profile (the higher the percentage, the darker). While clades 1 and 2 show a high variety of antigen allele combinations, profiles from clades 3 and 4 are clearly distinguished by their respective fim3 allele.

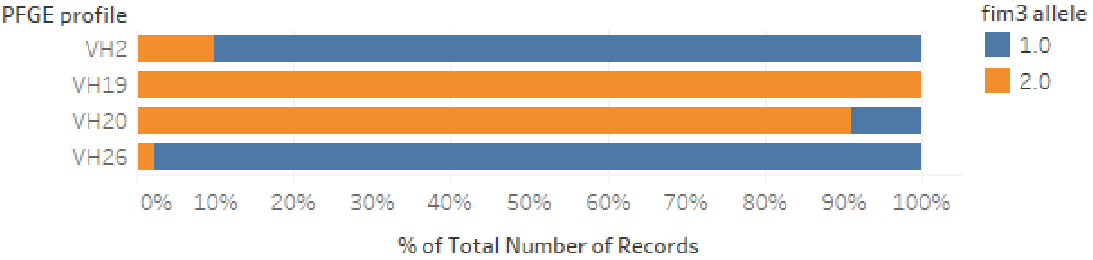


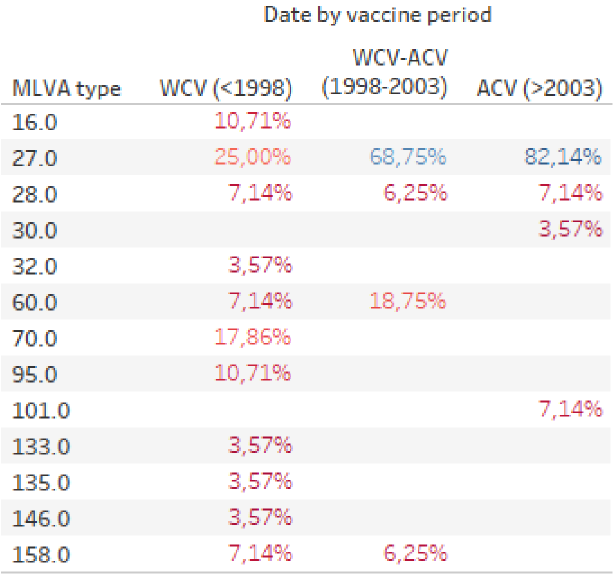
Figure 3.10 Correlation between fim3 alleles and PFGE profiles.

spread of the ptxP3 and fim3-2 combination. Moreover, the newest profile VH26, which seem to have caused the last and biggest epidemic (fim3-1 allele) along with the VH2, did also evolve from MLVA-27 type profiles (probably from VH2).

3.4 EDA summary and conclusions

Although there are a lot of missing values within the data, the exploratory data analysis has been quite successful. First, we statistically proved that the patients were (mostly) correctly vaccinated, according to the Spanish vaccine calendar. This also allowed the inference of the missing values of vaccine doses, as well as the vaccine type, of each patient.

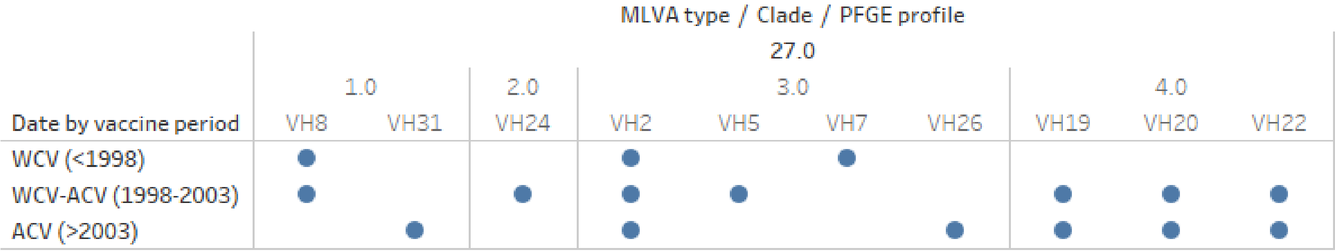
Second, the relation between the antigen alleles and the vaccine type was analysed. Some



	Date by vaccine period		
MLVA type	WCV (<1998)	WCV-ACV (1998-2003)	ACV (>2003)
16.0	10,71%		
27.0	25,00%	68,75%	82,14%
28.0	7,14%	6,25%	7,14%
30.0			3,57%
32.0	3,57%		
60.0	7,14%	18,75%	
70.0	17,86%		
95.0	10,71%		
101.0			7,14%
133.0	3,57%		
135.0	3,57%		
146.0	3,57%		
158.0	7,14%	6,25%	

Figure 3.11 Evolution of MLVA types.

With the transition from WCV to ACV, a specialization of the MLVA types took place. At first, MLVA-27 became dominant along with other minor types such as MLVA-60. However, this latter ended up disappearing in the ACV period.



	MLVA type / Clade / PFGE profile									
	27.0									
	1.0		2.0	3.0				4.0		
Date by vaccine period	VH8	VH31	VH24	VH2	VH5	VH7	VH26	VH19	VH20	VH22
WCV (<1998)	●			●		●				
WCV-ACV (1998-2003)	●		●	●	●			●	●	●
ACV (>2003)		●		●			●	●	●	●

Figure 3.12 MLVA-27 relation with PFGE profiles.

insights found about the resurgence of Pertussis were:

- The change from WCV to ACV, around 1998, caused a major change of the antigen alleles (from ptxP1 to ptxP3 and prn2), which confirms the B.pertussis adaptation to avoid ACV-induced immunity. Another major change, from fim3-2 to fim3-1, would explain why Pertussis cases soared in 2011.

- The vaccine type shows no correlation with the antigens found in the isolated strains, probably because the vaccine effects vanish over time, thus allowing the infection of B.pertussis pathogen regardless of the vaccine. Despite this, no further clear conclusions can be done here on this aspect.

- Fim3-1 was dominant during the WCV period, but it slowly changed to fim3-2 during the 00s. After that, the fim3-1 recovered the dominance (2011).

Finally, the analyses of the PFGE profiles, the clades, and the MLVA types have determined the causative profiles for the changes in the antigen alleles. Summarizing:

- Profiles from clades number 1 and 2 are older and are almost extinct. They showed balance and diversity between different combinations of antigen alleles.
- VH2 profile (clade 3) is a survivor from the WCV period that reappeared around 2010, causing the spread of the fim3-1 allele together with the new profile VH26.
- Profiles from clades number 3 and 4 are characterized by their fim3 allele. Also, they show almost no variety regarding the antigen allele combinations.
- The conclusion here is that the newest profiles (clade 3 and 4) are more specific, probably due to the adaptation to the ACVs. The chances are that the ACV immunity resistant strains have proliferated.
- In particular, strains with the fim3-1 allele are directly correlated to the last and worst epidemic cycles. In this respect, the two main causative PFGE profiles are the VH2 (which has resurged from the WCV period) and the newest VH26.
- There were many MLVA types during the WCV period, but during the transition to the ACV most of them disappeared, similarly to what happened with the PFGE profiles.
- MLVA-27 is directly related to the most aggressive and recent PFGE profiles. Determine what differentiates the MLVA-27 might, thus, be key to understand the Pertussis resurgence.

Chapter 4

Prediction models

In the previous section we explained, among other things, how the clades and the PFGE profiles are related to the antigen alleles, year of isolation and MLVA types. Using this information, the purpose of this chapter is to consider different statistical models to make predictions of these variables. For that, we considered different classification and regression techniques. Given the kind of data and the objectives, the following three were selected:

- **Gradient boosting:** to predict the characteristics of a strain, given the year of isolation and/or other variables.

- **Bayesian network:** to model, as conditional probabilities, the relationships between the clades, the antigen alleles and the periods in which a given strain was isolated.

- **Autoregressive model:** to forecast the Pertussis incidence from past data.

4.1 Gradient boosting classifier

A gradient boosting machine (GBM) is a machine learning technique that combines various weak models to achieve stronger learning [59]. Typically, it uses an ensemble of decision trees to make predictions. The models are trained in a sequential and additive manner, using gradients in the loss function to identify the shortcomings. Each subsequent tree takes these gradients to improve the predictions. The final result of the model is the weighted sum of the predictions made for each tree.

The advantages of this method are many. It allows the model to learn from different perspectives, resulting in more accurate predictions, especially in the most difficult observations to classify. Also, as it uses decision trees, it is possible to determine the importance that the model gives to each feature. This technique will allow us to see what the model considers most relevant to make the classification, and which is the degree of certainty of the result thanks to the probabilities (of belonging to each class) given by the model. This information will then be compared with what we know from the previous chapter.

As for the disadvantages, the GBM training process is usually longer due to the fact that trees are built sequentially, but this won't be an issue since our dataset is relatively small. This technique is also prone to overfitting, being thus important to look for a more generalized model using different combinations of the hyperparameters (learning rate, depth of the tree, etc.)

The variables we aim to predict in this chapter are the clades, the PFGE profiles, and the antigen alleles. For instance, for the fim3 antigen the classes to classify were the alleles fim3-1 and fim3-2. For that, we took care to maximize the accuracy while using the minimum possible information.

4.1.1 Clade classifier

Unfortunately, there are not enough observations from clades 1 and 2 to build a classifier. Figure 4.1 shows the distribution of the samples by clade. However, as it was observed in chapter 3, clades 1 and 2 belong mainly to the WCV period, while clades 3 and 4 are more recent. Thus, we decided to build the classifier based on these last two.

The data was divided into 142 examples for training and 36 for testing. The accuracy of the model is over 97%. Figure 4.2 shows the confusion matrix of the model. There is a single misclassified example, where the real clade is 3 but the model predicted 4. The reason is that while almost all the observations of clade 3 have the allele fim3-1, the misclassified observation is an exception that has fim3-2. Entering more in detail, there were some VH2 strains around 2007 that showed a fim3-2 allele dominance, but they did not last.

One of the most interesting aspects is the feature importance of the model. It seems that the determining factor is the fim3 allele, as shown in figure 4.3. Also, the year of isolation

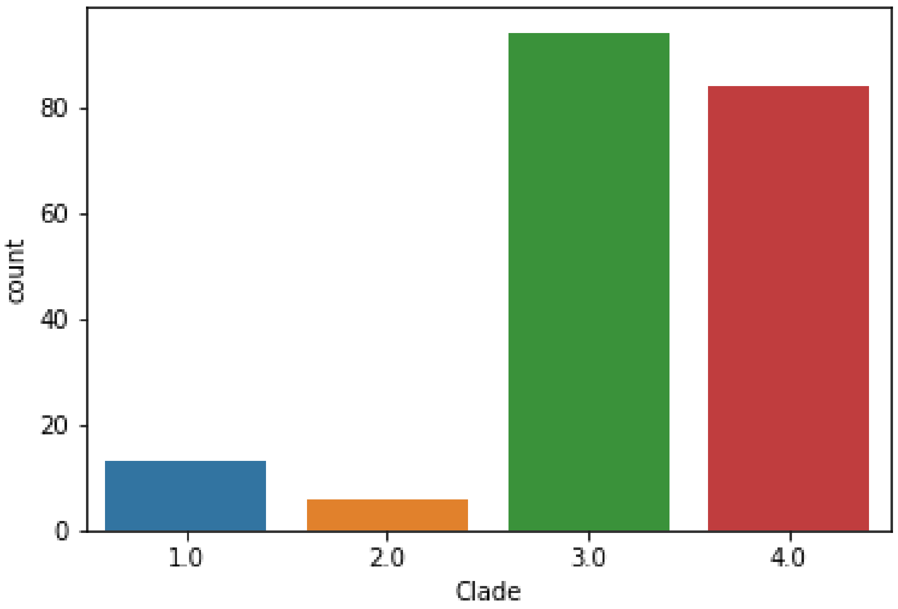


Figure 4.1 Clade distribution.

is somewhat important. All this reaffirms what was found in the exploratory analysis. While clade 4 is very correlated to the fim3-2 allele dominance that took place between 1998 and 2010, clade 3 is characterized by the fim3-1 allele.

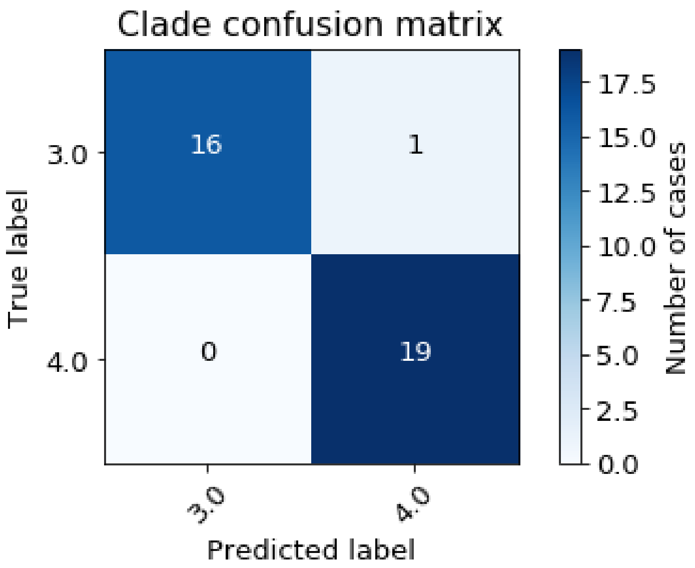


Figure 4.2 Clade classifier confusion matrix.

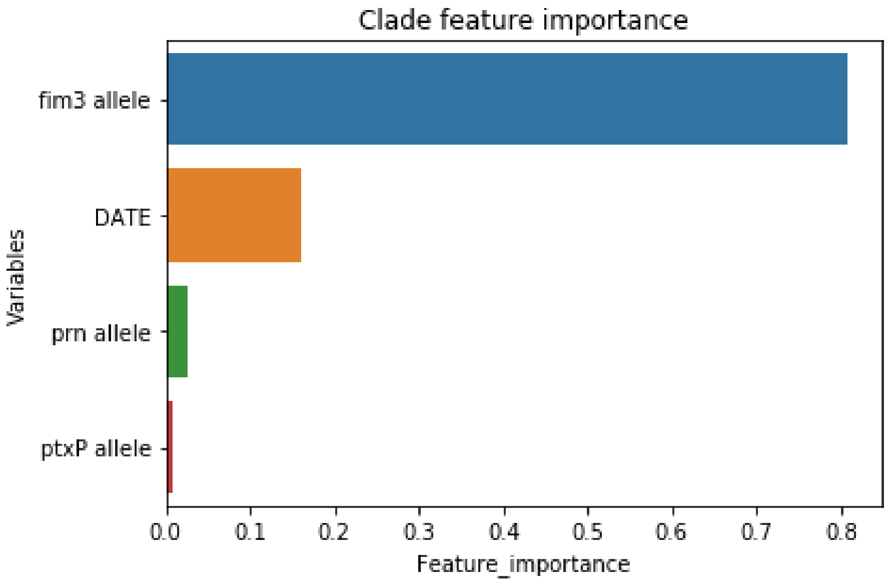


Figure 4.3 Feature importance of the clade classifier.

4.1.2 PFGE profile classifier

While the clade can be easily predicted from the alleles and the year of isolation, predicting the specific PFGE profile within the clade does not seem possible, as figure 4.4 indicates. The model is not able to differentiate the profiles within a clade, since there are no clear differences in terms of alleles or even year of isolation.

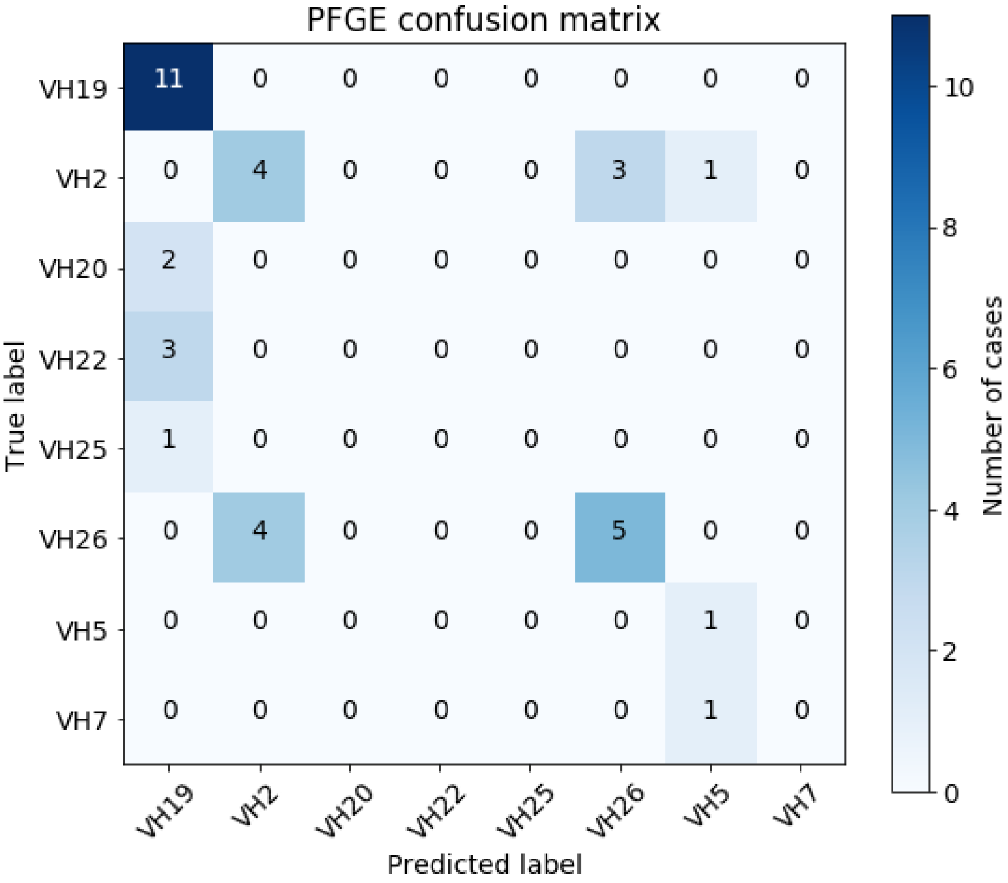


Figure 4.4 PFGE profile classifier confusion matrix.

Within clade 4 (VH19, VH20, VH22, VH25), the model just assigns all the samples to VH19. For clade 3, the model randomly assigns the observations to VH2 and VH26.

Also, the MLVA type does not help at all in this task, since as it was explained all the PFGE profiles relate to the MLVA-27. Therefore, the conclusion is that PFGE profiles can not be predicted with these data.

4.1.3 Predicting antigen alleles

Finally, by using the Gradient Boosting classifier we can also predict the antigen alleles, with the minimum amount of input information.

4.1.3.1 ptxP allele

Predicting the ptxP allele is totally possible using only the year of isolation, as shown in figure 4.5. The accuracy is higher than 98% using 121 examples for the training and 52 for testing. That makes sense since ptxP allele simply changed from ptxP1 to ptxP3 during the transition from WCV to ACV.

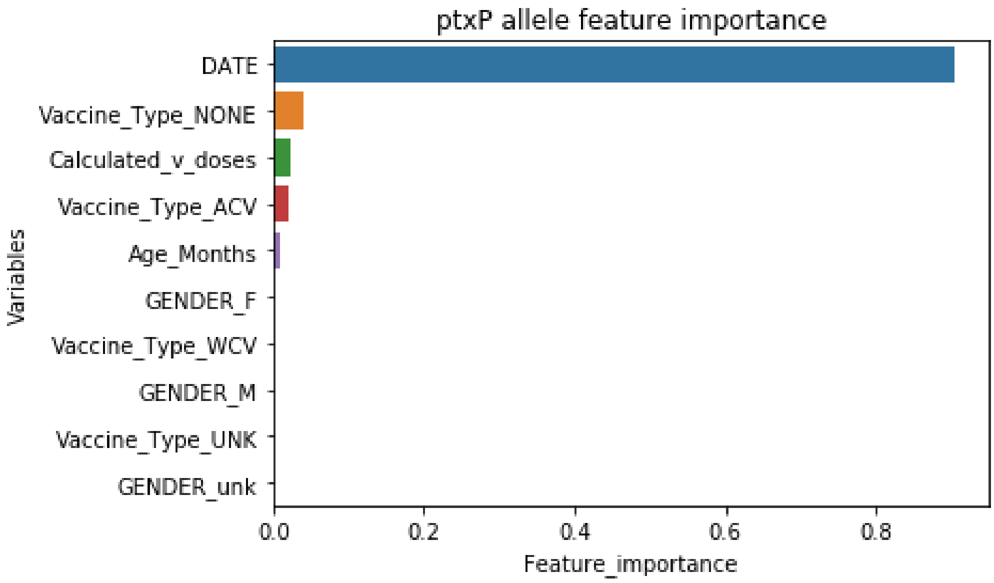


Figure 4.5 Feature importance of the ptxP allele classifier.

4.1.3.2 prn allele

Predicting the prn allele is, maybe, not very useful. The fact is that since the 00s the prn2 allele has been completely dominant. The confusion matrix in figure 4.6 shows how the model can not classify correctly the prn(1,3), since there are only a few cases within the data. As seen in chapter 3, prn(1,3) disappeared after the WCV period.

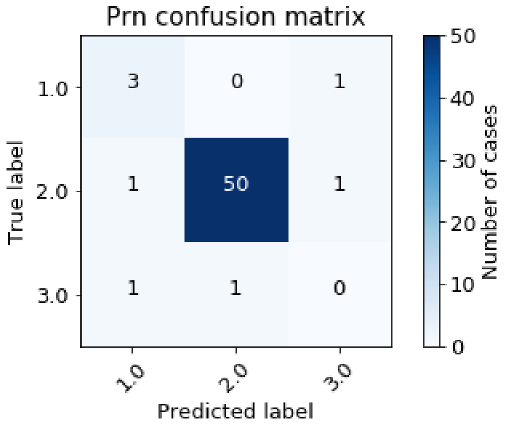


Figure 4.6 Confusion matrix of the prn allele classifier.

4.1.3.3 fim3 allele

The most interesting antigen within the data is the fimbrial 3. Changes in the alleles of this antigen might be one of the main reasons of the resurgence of Pertussis. Using only the year of isolation, the resulting model achieves around 75% of accuracy. However, adding the PFGE profile (or the clades) to the model hugely increases this percentage, achieving around 93% of accuracy, as shown in figure 4.7.

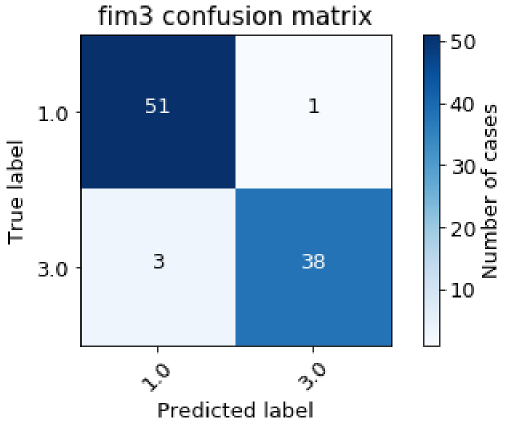


Figure 4.7 Confusion matrix of the fim3 allele classifier.

4.1.3.4 Conclusions of allele predictions

The alleles of the antigens ptxP and prn can be easily predicted using only the year of isolation, reaffirming everything found during the exploratory data analysis done in the previous chapter.

For instance, that ptxP3 and prn2 alleles became dominant more recently, while ptxP1 and prn(1,3) dominated the WCV period. Furthermore, although fim3 alleles are a bit harder to predict, a simple model can achieve up to 75% of accuracy. That is, the variance of the antigen alleles is strongly correlated with the year of isolation, especially in recent times, where the B.pertussis strains have specialized.

The data contain many missing values in the antigen allele variables. These can now be filled, with high accuracy, knowing exclusively the year of isolation. Moreover, adding information about the PFGE profiles into the model, improves the predictions assuring an accuracy of at least 90%. Finally, all the other variables (gender, age, vaccine doses, vaccine type, etc.) were excluded from the models since they did not add value, but only noise.

4.2 Bayesian network

A Bayesian network was built to explain the probabilistic relations between date, clade and the antigen alleles. For instance, it would allow to determine which are the probabilities for a strain to belong to each clade, if it was isolated in 2011 and its main antigen alleles were ptxP3, prn2 and fim3-2. Now, if this same strain would have been, instead, isolated in 1996, how would the probabilities change? To answer these questions, it is important to understand how a Bayesian model works.

The model is basically a directed acyclic graph that relates each variable with a conditional probability. That is, the probability of 'A' given 'B'. These probabilities were directly calculated from the data during the training process and were used to quantify the dependencies between variables. Notice that a Bayesian network assumes independence between non related variables. Therefore, it is mandatory to make a good design of the network that correctly represents the relatives.

4.2.1 Bayesian model design

The conditional probabilities are represented in tables, one for each variable. Figure 4.8 shows a simple example of a binary network with its tables. The maximum size of the table can be calculated as the number of values that the variable can take (d), raised to the amount of parents it has (k) plus one [27]. That is, an exponential complexity of $\mathrm{O(d^{(k+1)})}$. Therefore, the number of parents must be low.

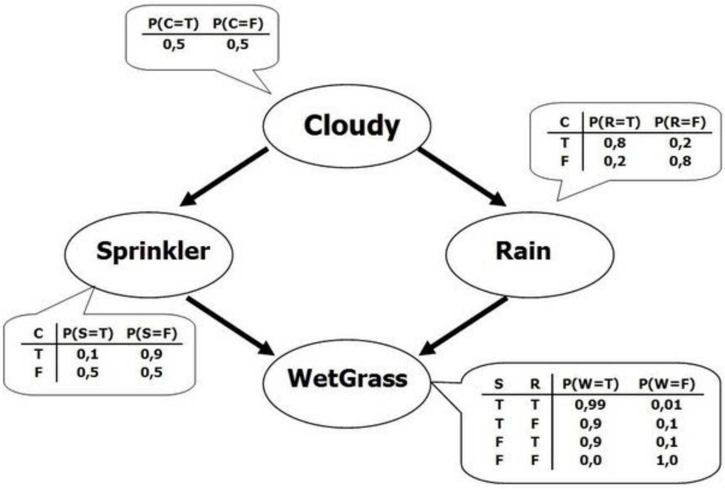


Figure 4.8 Simple Bayesian network example. Image from [60] (figure 3 of the article).

In order to make a good design, the first thing to do is to determine the relations between the variables that will form the network. From what was found in the previous section and also in chapter 3, it seems that the year of isolation is directly related to the antigen alleles and clades. Also, the antigen alleles were determined by the clade. Having this in mind, figure 4.9 shows the final design.

Next, the data were adapted to simplify the network. Classes with too few values (less than 2% of the total) were deleted. In addition, the dates were grouped by vaccine periods (WCV, transition and ACV), taking also into account all the information found in the previous chapter. Finally, it is important to select the most appropriate training data. Different Bayesian models were trained using various methods. First, with the whole original data. After that, all the missing values of the antigen allele variables were predicted to train another model using extra data. Third, clades 1 and 2 were fused, since they show no differences regarding the other variables. However, the period class was very unbalanced in number, so a new network was trained to fix it (with less data). Table 4.1 summarizes the results in terms of the accuracy of the predictions made by the models.

In terms of prediction accuracy, the network trained with the augmented data show superior performance in the period predictions at the expense of the clade prediction. Also, fusing the clades 1 and 2 increased the accuracy of the clade and the allele predictions, but decreased that of the period prediction. Finally, the small but balanced data resulted in a poor network

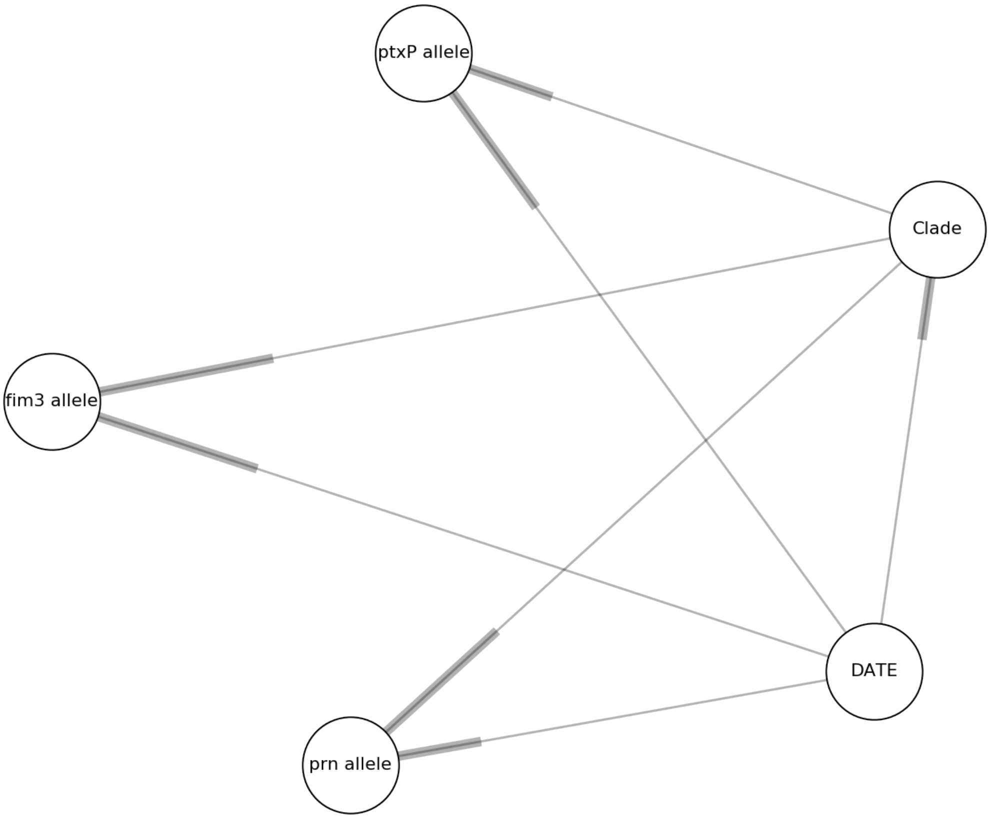


Figure 4.9 Design of the Bayesian network.

in terms of prediction accuracy. However, the main objective of the Bayesian model is to determine fair probabilities, not to make the most accurate predictions. In terms of probability distributions, this last model is the one that works best. The reason behind it is that the conditional probabilities are calculated directly from the data. To ensure fair probabilities it is mandatory to have the same amount of samples for each period (group of years). Otherwise, the model would be biased. Therefore, this last network was selected as the definitive model. The results will be presented in chapter 5.

4.3 Autoregressive models

The objective of this section is to predict the B.pertussis evolution. Specifically, to determine the tendency and the seasonality (epidemic cycles) of the Pertussis cases. To achieve that, an autoregressive (AR) model was used. This kind of model is characterized by the fact that the

Table 4.1 Prediction results by network training method.

Method/Variable	fim3	prn	ptxP	Period	Clade
Original data	95%	90%	98%	75%	86%
+ Predicted data	96%	89%	98.5%	85%	79%
Fused clades (1,2)	96.5%	91%	99.3%	78.5%	85%
Small balanced data	94%	88%	99%	81%	74.5%

predictions are made from past values within a time series data [61]. In other words that is, it uses the correlation between past and current values to predict futures.

The order of the model (p) defines the amount of time steps related to a specific time point. For instance, for monthly data with order 2, the model would use the last two months as lag variables. In addition to the AR factor, it is common to use the moving average (MA) which refers to the lags of the forecast errors. It is obtained doing a linear prediction from the past and current values. The ARMA(p,q) model is the result of combining both.

Finally, adding a (d)-order differencing to the ARMA model is called ARIMA(p,d,q), where the I stands for integrated. The parameter (d) indicates how many times the data have to be differenced to become stationary. The data are stationary when neither the mean, the variance nor the covariance are functions in time. This is important because linear regression methods assume independence between observations, but time series data is time-dependant [58]. Also, there is one last term to take into account: the seasonality.

4.3.1 SARIMA model

A Seasonal Autoregressive Integrated Moving Average (SARIMA) [34] is basically an ARIMA model with the addition of seasonality. Figure 4.10 shows the decomposition of the annual Pertussis cases. There is a seasonality of 4 years, which means that there actually exist a cyclic epidemic pattern of Pertussis. Also, the trend confirms its resurgence in recent years.

Before making any prediction, it is necessary to check the stationarity of the data. The statistical Dickey-Fuller test is used precisely for that. Also, figure 4.11 shows the rolling mean and standard deviation of the data, within a window of 4 years. The results indicate that the data are not stationary and, therefore, some transformations will be needed in order to find the optimal parameters for the SARIMA model.

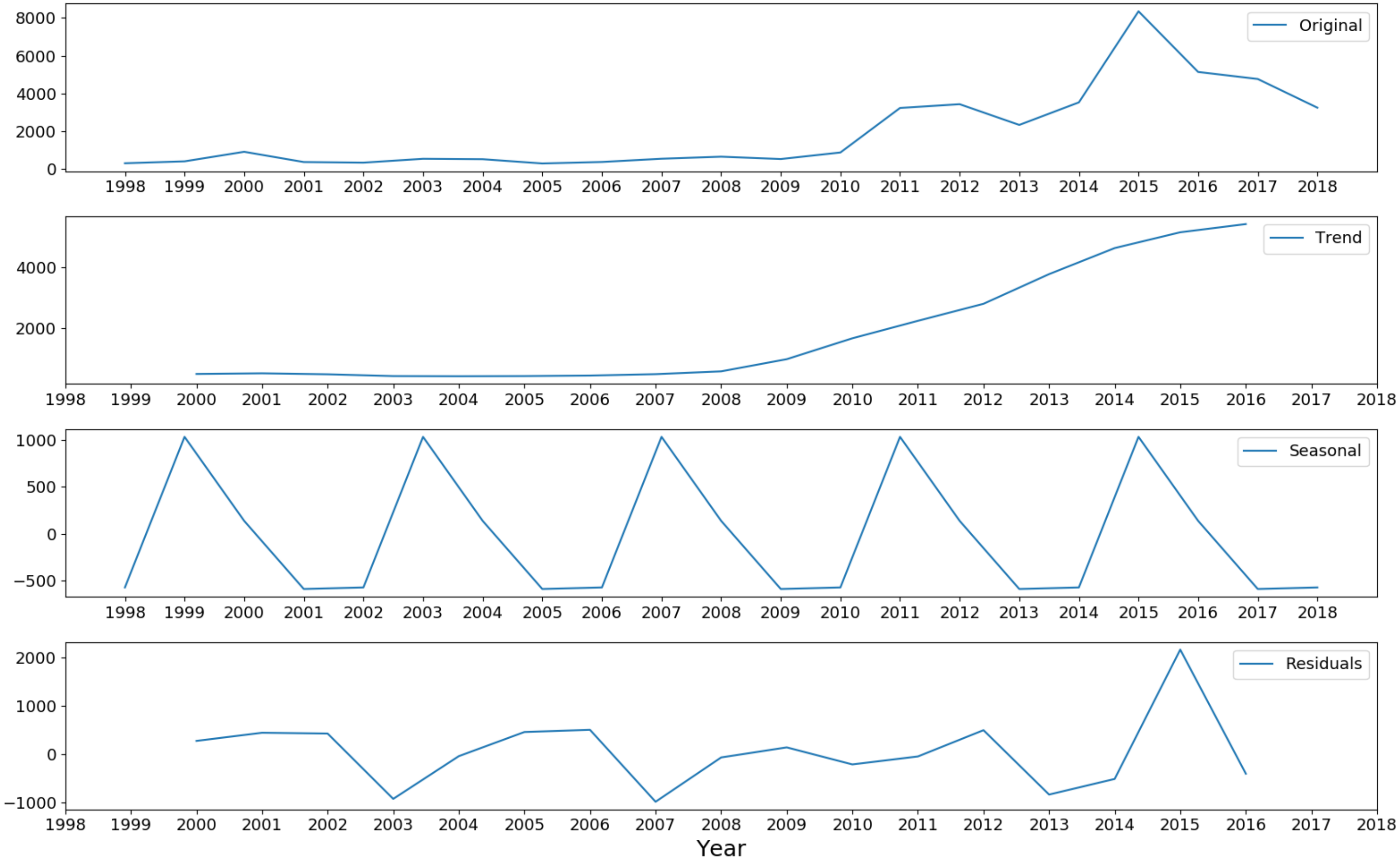


Figure 4.10 Decomposition of the number of Pertussis affectations in Spain by year.

There is a clear 4-year cyclic epidemic pattern (a large peak followed by 3 years of gradual decline). Also, the trend confirms the recent resurgence of Pertussis.

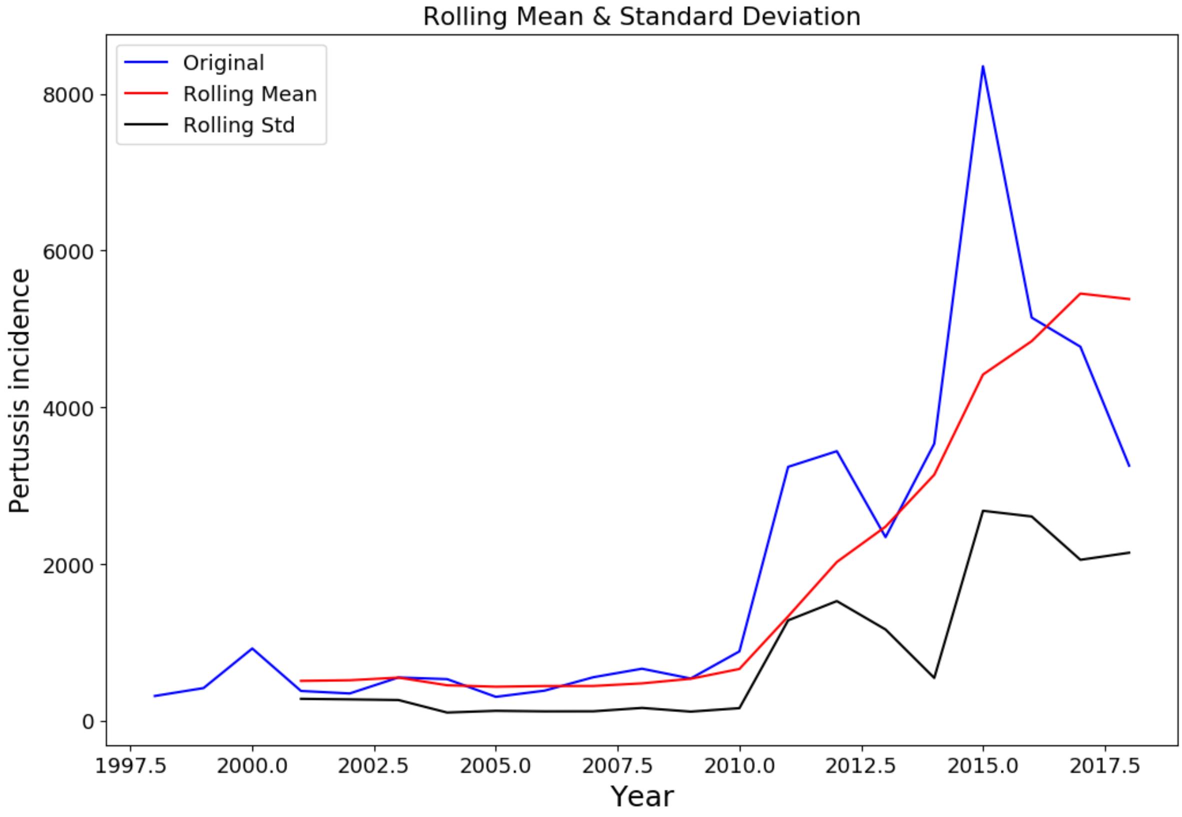


Figure 4.11 Rolling mean and standard deviation of Pertussis cases.

The data is clearly not stationary. Also, the Dickey-Fuller test has resulted in a p-value of 0.95, which confirms the non stationarity.

4.3.2 Optimal parameters

The first thing to do is to transform the data to make them stationary. For that, a season difference was applied, as shown in figure 4.12. The Dickey-Fuller test after this process resulted in a p-value of 0.02, which means the data are now stationary. However, the trend in recent years is too high and there are probably too few data points to get optimal the desired results.

To find the optimal parameters (p,q,d), it is necessary to analyse the autocorrelation of the data. Figure 4.13 shows the autocorrelation function (ACF) and the partial autocorrelation function (PACF) charts. The model has an AR term since the first autocorrelation lag is positive [17, 57]. Its value will be AR(2) because the PACF cuts off at lag 2. For the same reason, the MA term is 0. In addition, since no differing was applied, the I term is also 0. That is, ARIMA(2,0,0).

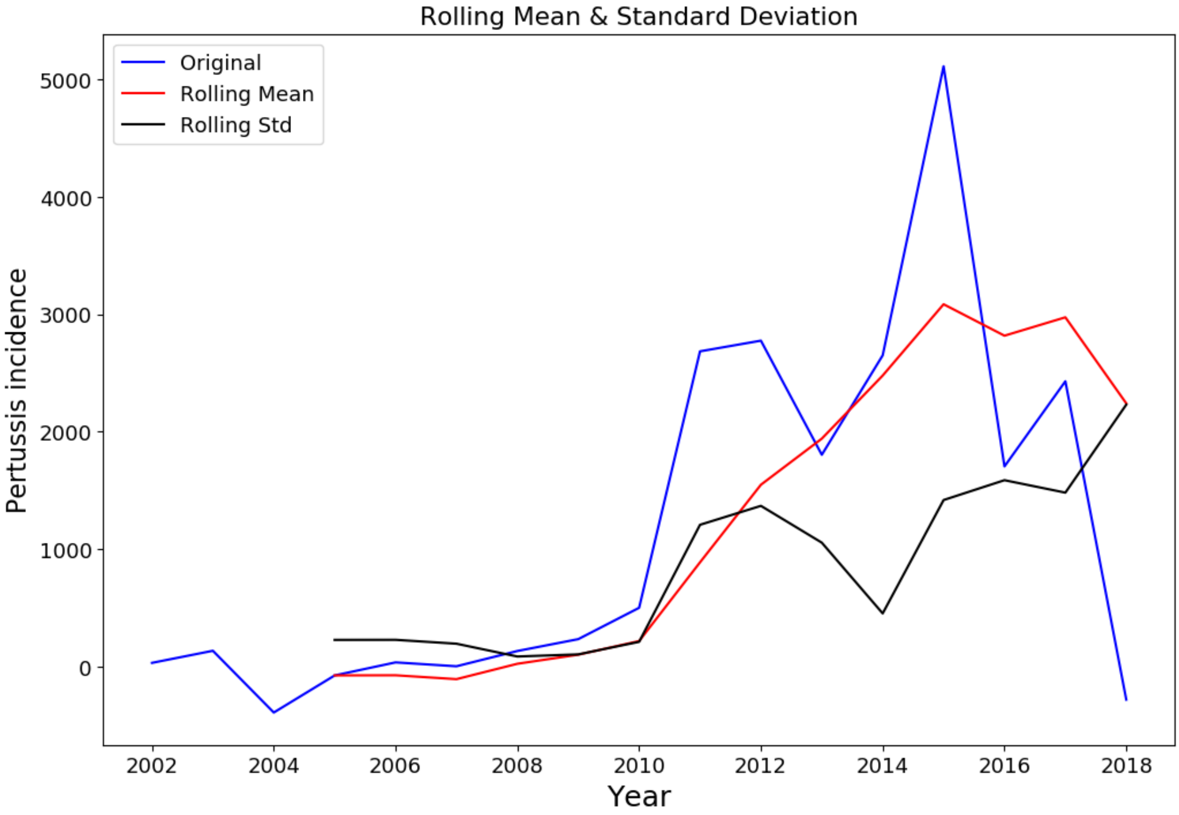


Figure 4.12 Rolling data after the seasonal difference.

Furthermore, since the data have a consistent seasonal pattern, a seasonal differing was needed to make the data stationary. The autocorrelation is positive at lag 4, which is the seasonal number, indicating the need of a SAR term. Combining all the results, the final model should be a SARIMA(2,0,0)(1,1,0,4). Nevertheless, there is no definitive method to choose the optimal values for the terms, so there might be other acceptable combinations. For instance, using AR(4) instead of AR(2) may improve the short term forecasting. In this case this is interesting because the data are annual. Predicting a couple of periods means to predict 2 years to the future.

4.3.3 SARIMA results and limitations

Figure 4.14 shows the predictions made by the SARIMA model. The forecasts follow the seasonality pattern and mark this year (2019) as a potentially epidemic one. However, this model was built with a single variable (annual Pertussis cases in Spain) and very limited data. Moreover, the model completely ignores external factors, such as new preventive methods against

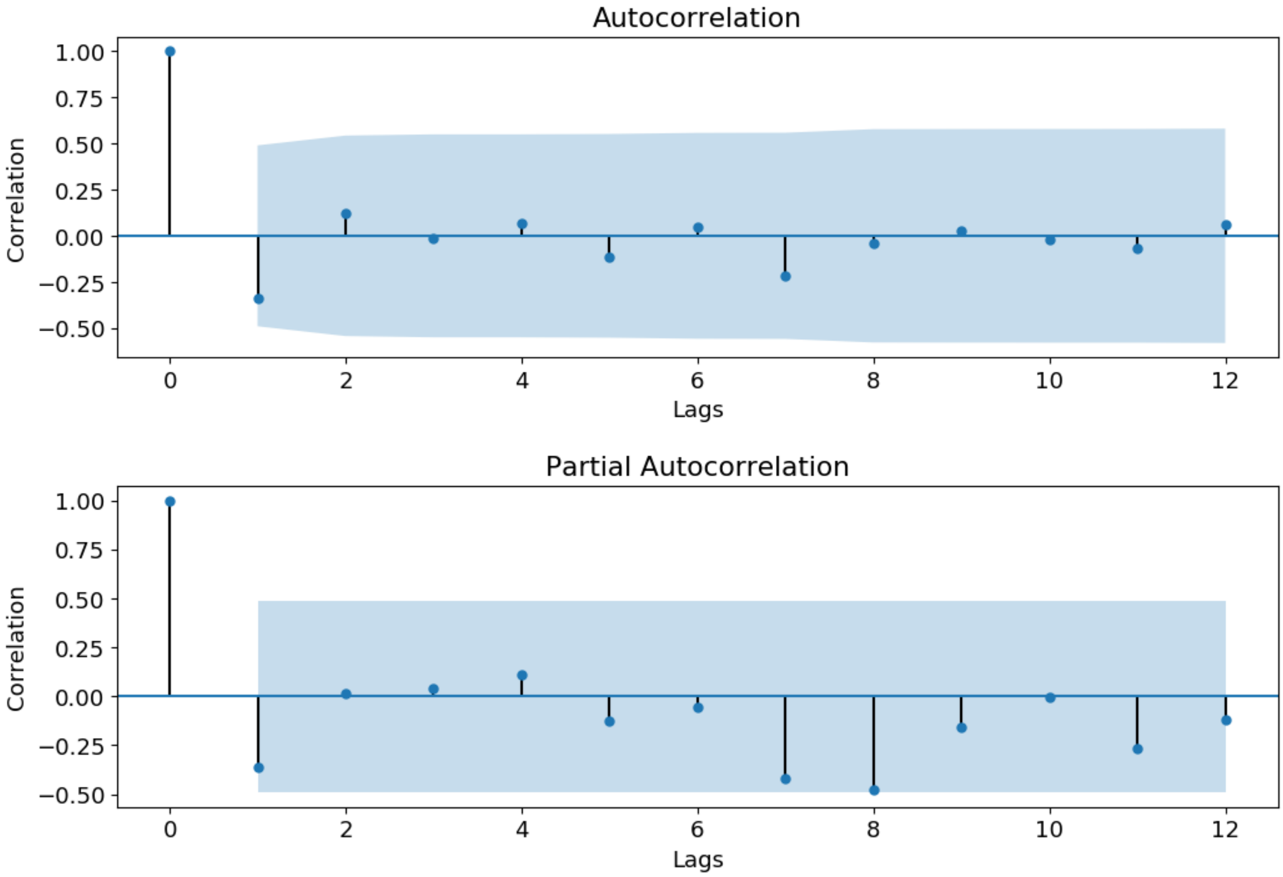


Figure 4.13 Autocorrelation and partial autocorrelation charts.

The ACF and PACF charts are used to determine the optimal parameters. Since the lag 1 in the ACF is positive, the cut off in PACF lag 2 indicates the AR term. In addition, the positive correlation at lag 4 and the strong seasonality recommend both SAR and seasonal differing terms.

Pertussis or the B.pertussis evolution itself. Notice how the prediction for the year 2018 is already slightly off. Besides, according to the Spanish epidemiological bulletin [42], this year so far (June) is not being especially epidemic. There are also some negative values in the first years, which only makes sense within the model but not in real life. Therefore, these results must be taken as purely orientative.

In order to improve the forecasting, it would be interesting to add information about the B.pertussis strains, such as the antigen alleles tendencies or the PFGE profiles evolution. In addition, it might be more convenient to gather and use weekly or monthly data from the last 8-10 years, which was when the resurgence of Pertussis started.

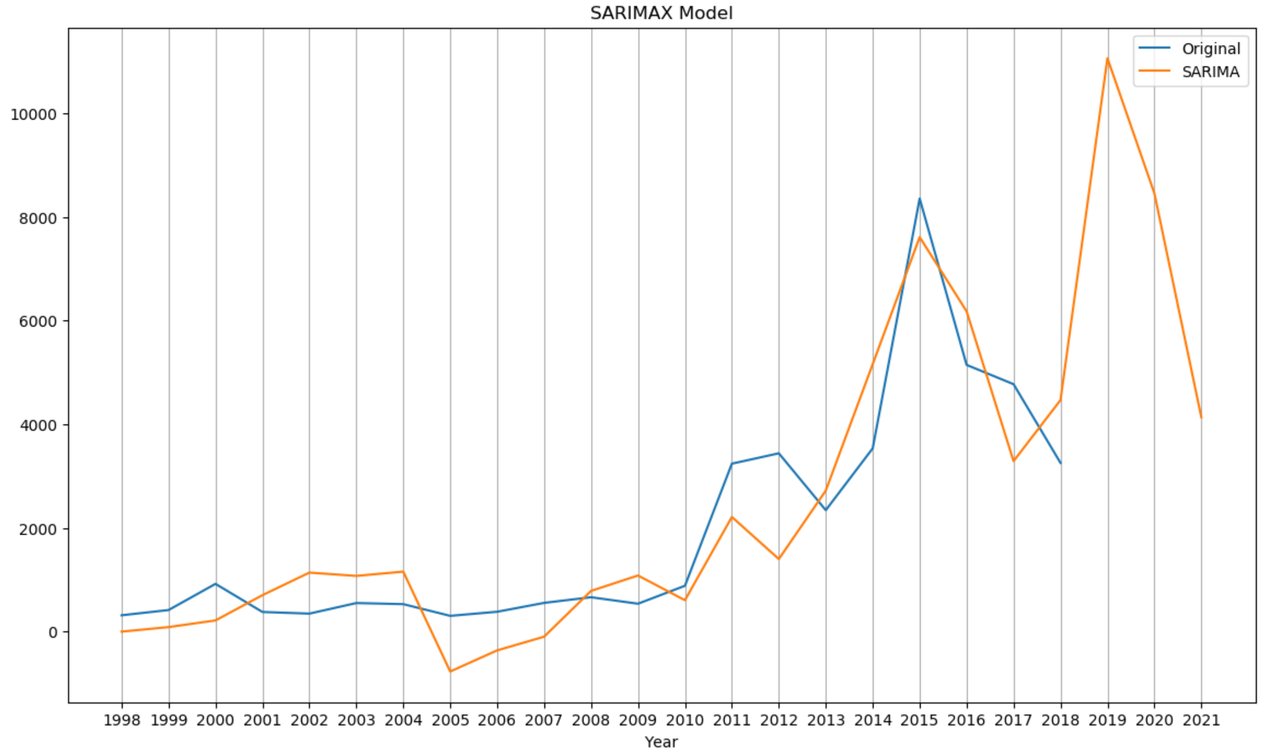


Figure 4.14 SARIMA forecasts of Pertussis cases.

Taking into account the seasonality, the next epidemic will be likely between 2019 and 2020. However, the results obtained with this model are just orientative, since it does not contemplate any external factor.

4.3.4 Scraping more data

Instead of using the annual incidence of Pertussis in Spain, it would be more appropriate to gather data from the Spanish weekly epidemiological bulletin [42]. The problem is that these data are encoded in tables from multiple portable document format (pdf). The purpose of this type of file is to be a reliable way to visualize and share documents, not a source to extract structured data. There exist, however, many tools that try to solve this problem [51].

The first thing is to get all the document links and download them. A web scraper was designed to automatize this work. After that, the binary files were converted to text using a Python package called Tika [48]. Now came the hard part: detecting the desired value in an unstructured raw text. Since the strain data came from Catalonia, we decided to extract, from each file, the weekly incidence of Pertussis in this region instead of that of Spain. After some complicated text parser, to handle missing columns and various formats of the tables, the final result is shown in figure 4.15.

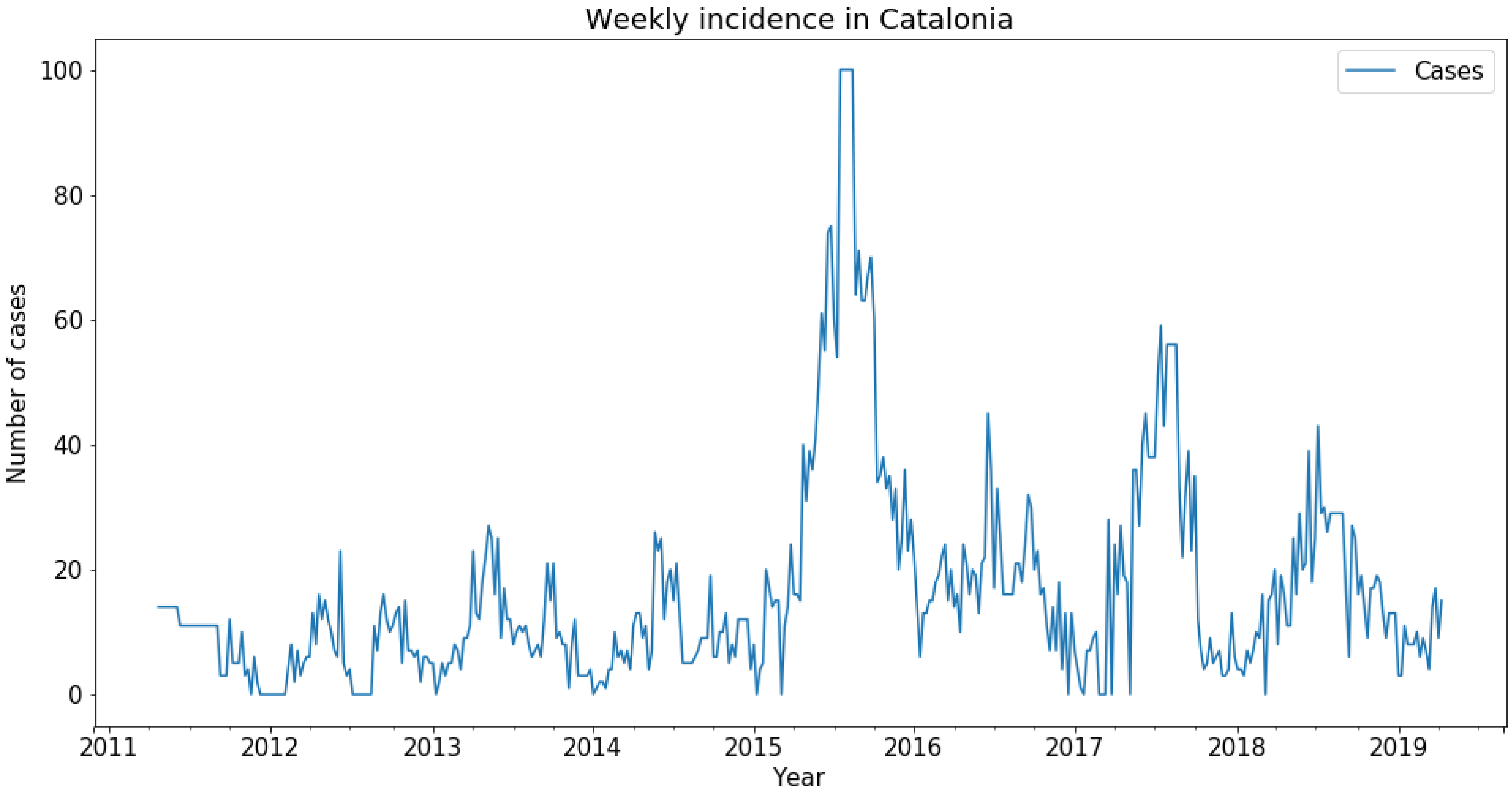


Figure 4.15 Weekly Pertussis incidence in Catalonia.

Similar to Spain, in Catalonia the Pertussis cases soared in 2015. There is an annual seasonal pattern, being the end of summer the most epidemic season [4]. Having weekly data, instead of annuals, clearly added informative value.

4.3.5 SARIMAX: adding exogenous variables

With the new data collected, it is time to improve the SARIMA model. In addition, we added exogenous variables to provide more input information to the model. That is, adding external predictors, usually known as 'x' variables, to improve the endogenous ('y' variable) forecast [43, 45]. This procedure gives the SARIMAX model.

As it was explained before, there exist a cyclic epidemic pattern of 3 to 5 years. Also, in the previous section it was discovered an annual seasonality. Using both would certainly improve the model and the way to do that is through exogenous variables. The only requirement of these new variables is that they must be available for the date to forecast. Since the seasonality can be easily extrapolated, this will not be a problem. Figure 4.16 shows the seasonal patterns for 3, 4 and 5 years respectively. All these were used as external predictors, while the annual seasonality was already within the model.

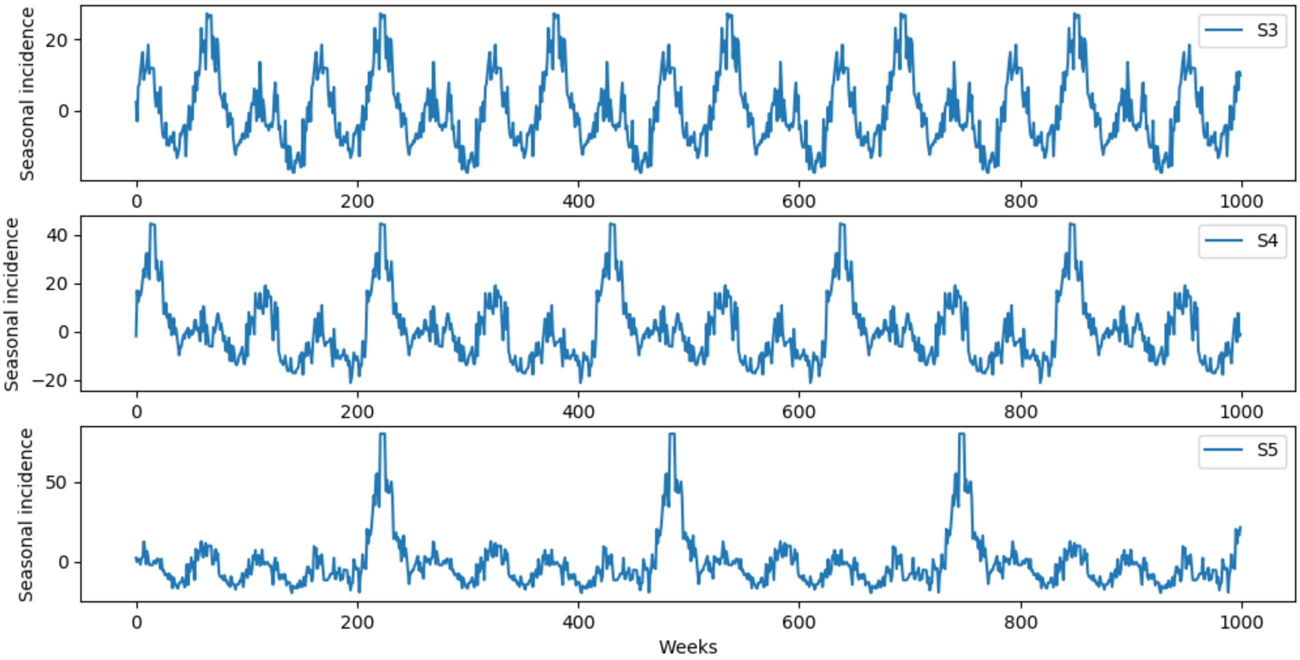


Figure 4.16 Seasonality patterns of 3 to 5 years.

The process of finding the optimal values for the SARIMAX terms is the same as for SARIMA. In this case it is a SARIMAX(0,1,1)(0,1,1,52), where the 52 represents a year (52 weeks). The data were divided into train and test sets. The latter is also 52 weeks long, as the forecasts were made yearly. The mean squared error (mse) is the method we chose to evaluate the results. Finally, the predictions were dynamic, which means that they used the forecast values for the next prediction (except for the first one).

The designed model showed good results, with a (train, test) mse of (65.71, 64.92). To avoid overfitting the model, it is important to have a similar or lower error in the test set with respect to the train set. Knowing that this model works well, we tested other combinations to make comparisons, as table 4.2 summarizes. For some reason, using 53 weeks instead of 52 had slightly better results, but they were negligible.

Table 4.2 Comparison of SARIMAX terms, evaluated with the mse.

Order	Seasonal	Train	Test
(0,1,1)	(0,1,1,52)	65.71	64.92
(0,1,0)	(1,1,1,53)	72.24	59.9
(0,1,4)	(0,1,1,53)	88.99	55.12
(1,1,4)	(1,1,1,53)	59.02	64.02

The chosen model was a SARIMAX(0,1,4)(0,1,1,53), which was very similar to the original one. This new model obtained the lowest mse for the test set. Choosing this way could be considered as filtering information from the test data, but the impact in this case would be minimal. Another important factor to look at is the significance level of each term in the model. Table 4.3 shows the SARIMAX model results after the training process. The table indicates that the variables were very significant, with the exception of MA(2,3,4). Nevertheless, these last ones still add value to the model by softening it, resulting in less sharp peaks.

To test the model, annual predictions were made for each year, starting from April, 2011. As it was said before, these are all dynamic. Figure 4.17 shows the resulting forecasts until 2024. These were slightly off in the first years, since the model captured the seasonality of the last ones. However, the forecast for the test year (color grey) was quite accurate and the future forecasts (yellow) make sense and capture the essence of the seasonality.

Table 4.3 SARIMAX model results.

```
                                 Statespace Model Results
===========================================================================================
Dep. Variable:                               Cases   No. Observations:                  365
Model:              SARIMAX(0, 1, 4)x(0, 1, 1, 53)   Log Likelihood                -652.339
Date:                             Fri, 10 May 2019   AIC                           1322.678
Time:                                     18:35:19   BIC                           1354.478
Sample:                                 04-24-2011   HQIC                          1335.472
                                      - 04-15-2018
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
s3             0.2409      0.053      4.581      0.000       0.138       0.344
s4             0.4889      0.051      9.654      0.000       0.390       0.588
s5             0.6702      0.050     13.482      0.000       0.573       0.768
ma.L1         -0.3590      0.046     -7.789      0.000      -0.449      -0.269
ma.L2         -0.0492      0.062     -0.794      0.427      -0.171       0.072
ma.L3          0.0877      0.062      1.421      0.155      -0.033       0.209
ma.L4          0.0661      0.056      1.190      0.234      -0.043       0.175
ma.S.L53      -0.5980      0.071     -8.394      0.000      -0.738      -0.458
sigma2         9.8830      0.752     13.148      0.000       8.410      11.356
====================================================================================
Ljung-Box (Q):                        48.17   Jarque-Bera (JB):                  30.75
Prob(Q):                               0.18   Prob(JB):                           0.00
Heteroskedasticity (H):                1.77   Skew:                               0.33
Prob(H) (two-sided):                   0.01   Kurtosis:                           4.57
====================================================================================
```

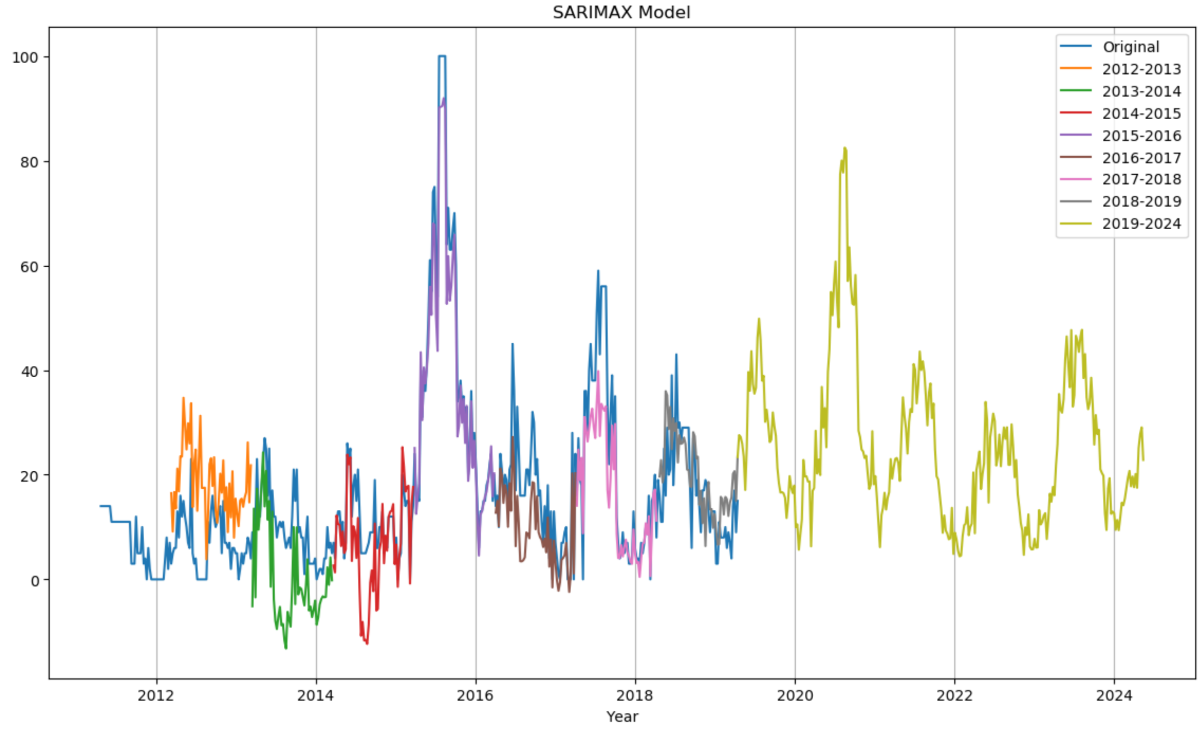


Figure 4.17 Annual forecast using the SARIMAX model.

The forecasts were accurate in the training data since they totally captured both seasonalities. The most interesting part is the forecast for the test data, shown in grey (2018-2019). Importantly, the model had no information at all about this year and the forecast was still quite accurate. Finally, the future forecasts, in yellow, indicate that the next epidemic may come between the years 2019 and 2020.

This new model was an improvement compared to the previous one. It also indicated that the next epidemic year is likely to be either this 2019 or the 2020, specifically by the end of the summer. Nevertheless, the forecast was not as extreme as the first. However, the model was far to be perfect. Using more data would certainly make it more robust to overfitting, especially because of the 5 year seasonality. In addition, there exist other exogenous variables that could add value to the model.

4.3.5.1 Trying to add strain data

Lastly we tried to add, without success, the tendencies of the alleles as predictor variables for the SARIMAX model. The first problem was that the allele data are annual and they were only available until 2015. Therefore, they had to be extrapolated to be able to add them to the model. Also, as it was seen in chapter 3, the alleles of the antigens prn2 and ptxP3 became dominant in recent years. In other words, the variance of the alleles almost vanished. The only exception was the fim3 antigen, which could be related to the 2011 and 2015 epidemics. Figure 4.18 shows the tendencies of the fim3 alleles, as well as that of the clades (there is a clear similarity between both). The observed fluctuation might be related to the cyclic epidemic pattern of Pertussis.

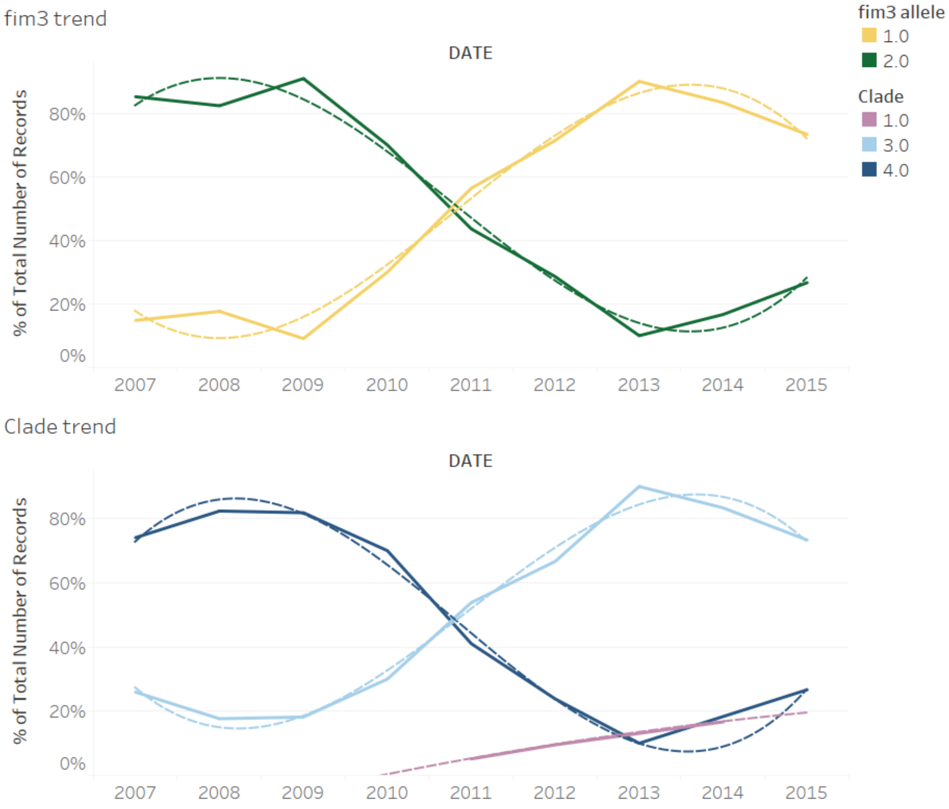


Figure 4.18 Fim3 and clade tendencies.

The dashed curves represent the tendencies, while the continuous curves refer to the original data. The clade tendencies were very similar to that of the fim3 alleles. The fluctuations coincide with the epidemic years. For instance, in 2011 there was a big change from fim3-2 to fim3-1. Also, in 2015 the fim3-2 allele was starting to regain terrain.

These tendencies were added to the model as an attempt to improve it. The p-value for the new variable is 0.731, thus indicating a very low significance. Moreover, it added noise to the model, worsening the predictions. Although the tendencies could be related to the cyclic epidemic pattern, there was not enough evidence within the data to prove it. A more extensive analysis (with more recent data) would be needed to extract appropriate conclusions. Regarding the model, these tendencies might be interfering with the seasonalities. Therefore, the new variable was removed from it.

Chapter 5

Comparisons and results

For the first section of this chapter a couple external studies were selected to make comparisons with the results of this project. The first one analyzes the global population and evolution of B.pertussis, using data about 343 strains isolated between 1920 and 2010 over the world [29]. The second analyzes 704 strains isolated from the Netherlands between 1949 and 2010 [39]. Both studies contain a lot of quality data, even though only the matching data (1986-2010) will be used for the comparison. It is left as a future line to deepen more in these studies.

The second section presents a web application aimed at making the results accessible. For that, a RESTful API was designed and deployed, which allows the interaction through the POST method. It does also provide a GET method to obtain html5 interfaces that facilitate this interaction.

5.1 Comparison with other studies

The data from the studies are older and they only cover until the years 2008 and 2010, respectively. That means that they do not encompass the recent resurgence of Pertussis, which started at 2011 (in Catalonia). Also, as it was seen in the state of the art, the PFGE technique does not allow inter-laboratory comparisons. Having that in mind, the idea was to compare the evolution of the alleles of the antigens, from the WCV period to the first years of the ACV, to see to what extent the results obtained with the data from the Vall d'Hebron Hospital differ from those obtained with other data sources.

The global evolution of the alleles is shown in figure 5.1. The overall evolution follows the same patterns found in chapter 3, even though there was more diversity. For instance, notice that ptxP3 does not became as dominant, since ptxP1 was still present in some places. This could be due to the differences in vaccination programs and vaccination coverage between countries.

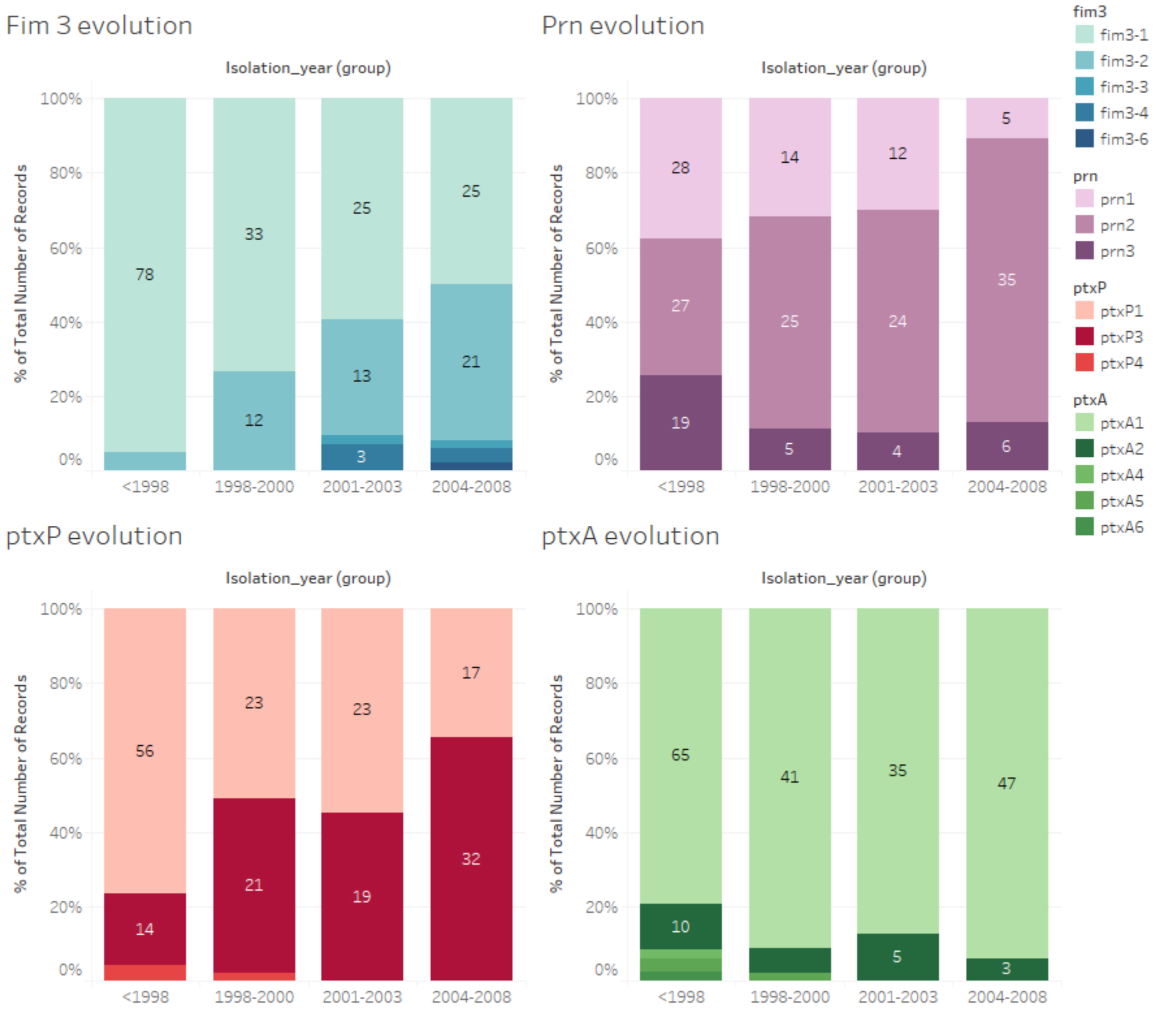


Figure 5.1 Global evolution of the antigen alleles. Data from [29].

The evolution of the alleles over the world was more diverse than what was seen from the data of the Vall d'Hebron Hospital. However, the main patterns were the same. The fim3 allele slowly changed from 1 to 3 while the prn2 became dominant. There were, however, slight differences regarding the ptx alleles. The ptxP3 was not as dominant and there were also other ptxA alleles in addition to ptxA1.

Furthermore, the study carried out in the Netherlands, shown in figure 5.2, obtained very similar results to those of this project.

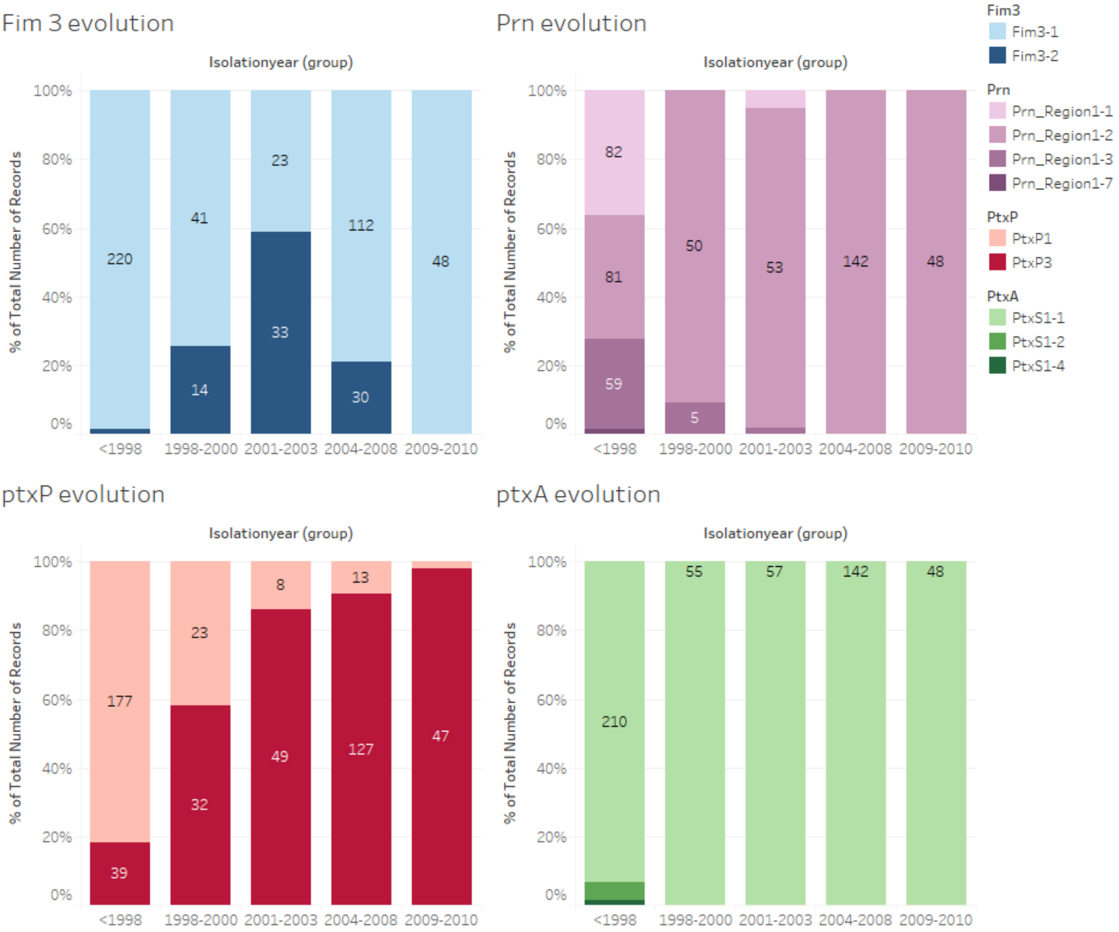


Figure 5.2 Evolution of the antigen alleles in the Netherlands. Data from [39].

The evolution of the alleles in the Netherlands was very similar to that of Catalonia. The ptxA1, ptxP3 and prn2 became dominant after the introduction of the ACV. However, in the Netherlands the fim3-1 dominated all along except for the years 2001 to 2003, while in Catalonia the dominating allele was the fim3-2 until 2011 (as seen in figure 3.6).

This makes sense since the Netherlands also have a high vaccination coverage and it is not that far away from Catalonia. There was, however, one big difference in the fim3 allele evolution time. It seems that in the Netherlands the fim3-2 dominance lasted only from 2001 to 2003 and in 2010 it was completely wiped out. This country is historically one of the most affected by Pertussis, despite its high vaccination coverage [41]. There, the resurgence started much earlier (back in 1996), coinciding with the spread of the ptxP3. Besides, the transition from WCV to ACV took place respectively later, in 2005.

In conclusion, the B.pertussis evolution followed similar patterns worldwide, but also differed in some aspects. Every country has and has had its own vaccination programs and it is a fact that the bacteria has been adapting to vaccination. Thus, the temporal or regional differences make sense, but the overall result was the global spread of the strains more resistant to the vaccine-induced immunity. Many (but not all [38]) of the Pertussis resurgences observed over the world seem to be linked to some changes in the antigen alleles. For instance, as the change from ptxP1 to ptxP3 caused the 1996 resurgence in the Netherlands, the changes in the fim3 alleles might have caused the last epidemics in Spain. It is important to emphasize that the effects of a given change in the B.pertussis bacteria may differ depending on the area. That is, there are many external factors to consider that are beyond the scope of this project.

5.1.1 Prediction models applied to the new data

To support the recent conclusions, the classifiers trained in chapter 4 were applied to the new data. Table 5.1 shows the accuracy of the resulting predictions. The first row corresponds to the accuracy achieved in the original data. Also, the base accuracy refers to the maximum class ratio of the Netherlands data. For the last row the new data were used to train a new model to apply it back to the original data.

Table 5.1 Accuracy of predictions with the Netherlands data.

Method/Accuracy	fim3	prn	ptxP
Original data	72%	91.3%	99%
Netherlands base acc.	84%	72%	58%
Netherlands data	62%	69%	84%
Netherlands prn model (back)	-	91.7%	-

The accuracy with respect to the base was lower for the fim3 (the PFGE profile variable was not used, since it was unknown in the Netherlands data), higher for the ptxP and pretty much the same for the prn. The first conclusion here is that the evolution of the ptxP allele in both countries was very similar. Furthermore, the prn results were good taking into account the differences in the class ratios for the two datasets. The prn model, as seen in chapter 4, could not classify correctly any class beyond the prn2, due to the imbalance of classes.

To solve that problem, a new prn model was trained using the Netherlands data (which have a better class ratio), and it was applied to the original data. The accuracy was equal or even higher than that obtained with the original model, thus indicating a better performance for the prn-1 and prn-3 classification. Also, this confirms that the evolution of the prn allele in both countries was also very similar. Finally, the low accuracy in the fim3 model indicates that there were differences in the evolution of this allele. In fact, within the Netherlands data a 84% of the observations have the fim3-1 allele.

To conclude this section, applying the same procedure to the global data resulted, as shown in table 5.2, in similar accuracy for the fim3 but lower for the other two. To summarize, these results support the hypothesis that the evolution tendency of B.pertussis was similar, in general terms, all over the world. However, it also shows some differences, either in timing or in allele persistence, depending on the region.

Table 5.2 Accuracy of predictions with the global data.

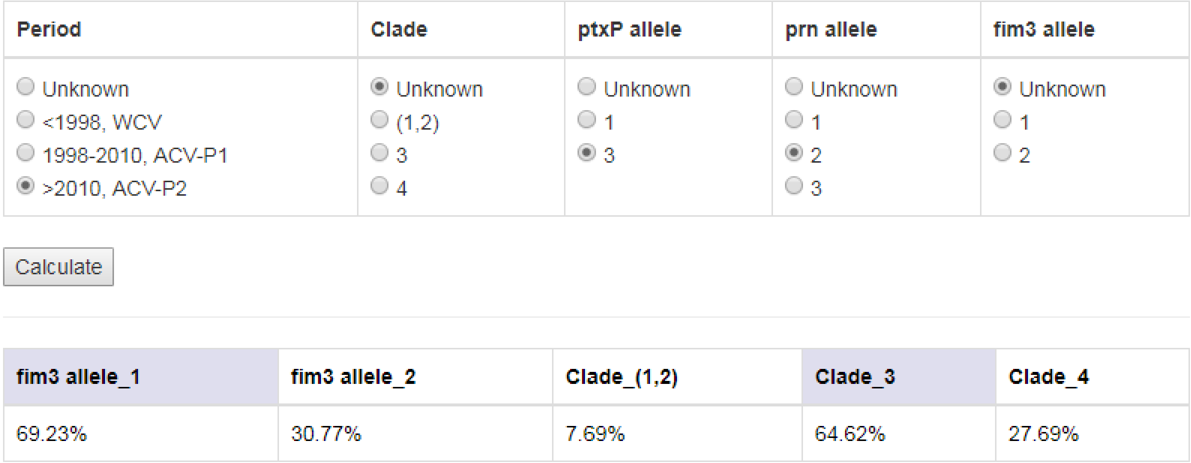
Method/Accuracy	fim3	prn	ptxP
Original data	72%	91.3%	99%
Global base acc.	74%	57%	56%
Global data	65%	54%	65%

5.2 Web application

A project in data science encompasses many areas like data acquisition, data cleaning, statistics, data visualization or machine learning. All this result in models, tables, graphs, new data, etc. It is common to translate all these into documents or presentations, but ideally the final users should be able to use the trained models or to interact with the graphics. Instead of having a code working in a notebook, it is much better to have an easy-to-use interface available anywhere. That is what this section is about. The chosen method for that was to build an API with a website as its interface.

This API, written with the Python package Flask and deployed to Heroku [12, 54], currently contains three tabs each with its own interface, even though it is possible to use them directly through code. The first one allows the use of the Bayesian network trained in chapter 4. For instance, figure 5.3 shows the conditional probabilities for the clade and the fim3 allele of a strain isolated after the year 2010, with the ptxP3 and prn2 alleles.

The second tab shows the evolution of the alleles in the form of interactive graphs. These were made with Tableau, based on the data from the studies of the previous section. Although the visualizations are simple, these allow to quickly see and compare the overall evolution with



Period	Clade	ptxP allele	prn allele	fim3 allele
○ Unknown ○ <1998, WCV ○ 1998-2010, ACV-P1 ● >2010, ACV-P2	● Unknown ○ (1,2) ○ 3 ○ 4	○ Unknown ○ 1 ● 3	○ Unknown ○ 1 ● 2 ○ 3	● Unknown ○ 1 ○ 2

fim3 allele_1	fim3 allele_2	Clade_(1,2)	Clade_3	Clade_4
69.23%	30.77%	7.69%	64.62%	27.69%

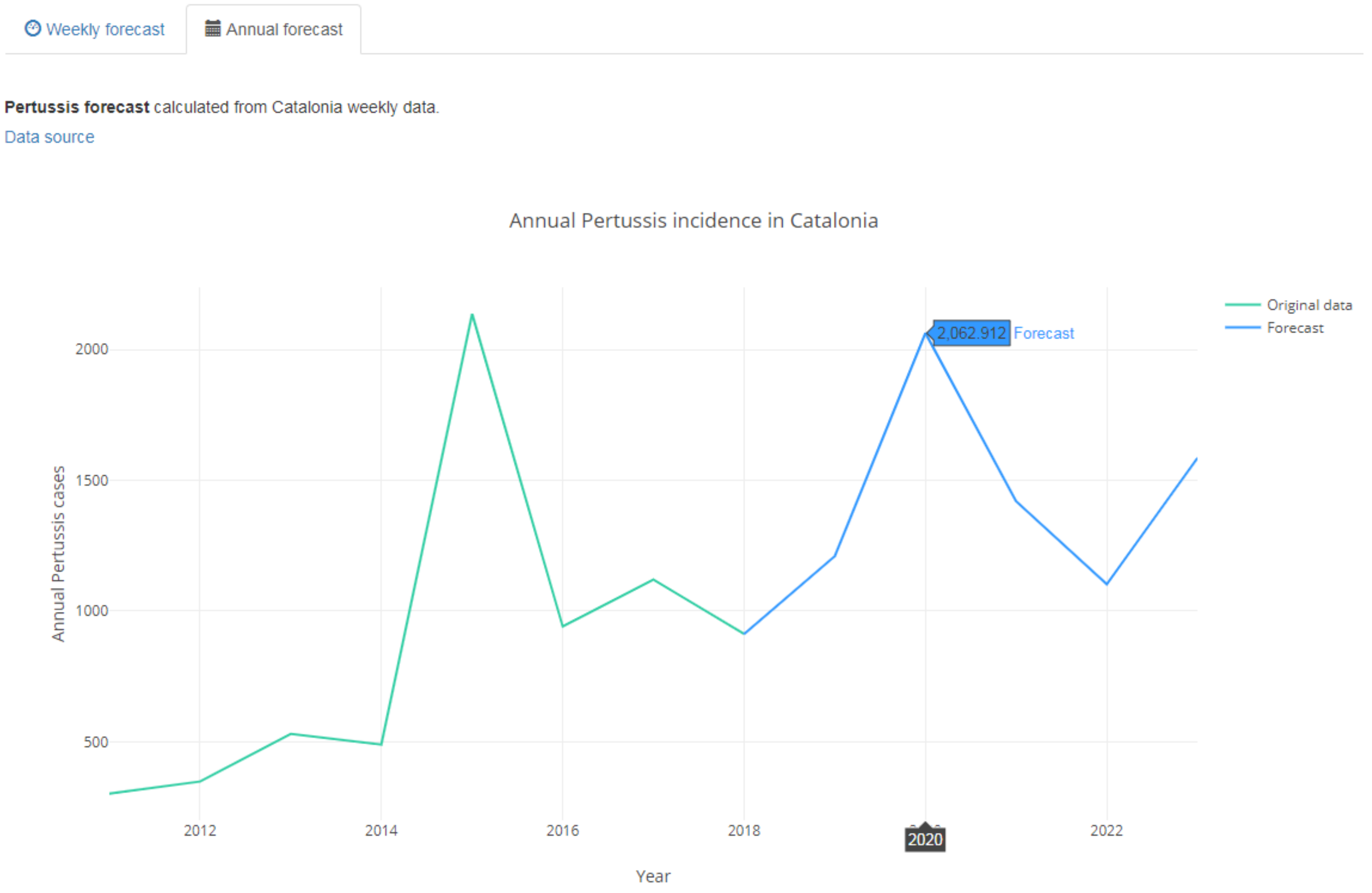
Figure 5.3 Bayesian network web interface.

a more specific one such as that of the Netherlands.

Finally, the last tab contains an interactive version of the SARIMAX model, which was made with the Python package Plotly. As shown in figure 5.4, it is possible to select between weekly or annual forecast, as well as to see the values on hover. Besides, the visualizations adapt to the size of the screen.

The API is available at:

https://pmajortfm.herokuapp.com/bayesian



Pertussis forecast calculated from Catalonia weekly data.

Data source

Figure 5.4 SARIMAX forecast model web interface.

Chapter 6

Conclusions

In this final chapter the project conclusions, as well as its future lines, will be discussed. Although the objectives of this project were successfully achieved according to the planning, the redefinition periods were useful to reconsider the best approaches and to mark new goals. As the project progressed, the initial results were taken into consideration to make better decisions in the next steps. Accordingly, not only all the questions posed in the project objectives were answered, but also some more.

6.1 Discussion of results and methodologies

Before 1998 (WCV period) the main antigen alleles were the ptxP1, the fim3-1 and the prn(1,2,3). During the transition to the ACV (1998-2003), the fim3-2 appeared and the combination ptxP3 and prn2 became dominant. Thus, we concluded that the bacteria actually adapted to the ACV-induced immunity. Beyond the transition period, a fluctuation between the fim3-1 and fim3-2 alleles was observed. Moreover, the fluctuation coincides with the 2011 and 2015 Pertussis epidemics. In fact, the fim3 is the only antigen that changed between the years 2003 and 2015.

Along the years, B.pertussis strains became more specific and the variances of the antigen alleles and the epidemiological profiles almost vanished. During the transition period and beyond, it went from more than 10 different MLVA types to only 4, of which the MLVA-27 became dominant (82% of the strains). Regarding the PFGE profiles, the VH19 and the VH20 dominated the first ACV period (2003-2010) along with the fim3-2 allele. After that, the VH2 and the VH26 replaced them, bringing back the fim3-1 allele. Notice that the VH2 profile, which reappeared in 2011, is the only survivor from the WCV period.

The prediction models were accurate even if using only the date (year of isolation) as input, thus confirming the specification of the strains. It is important to emphasize that the year of isolation was the only variable that could be controlled. One can decide when to isolate a strain, but not the results obtained. Therefore, the Bayesian network was designed through date clusters based on the most relevant periods. Then, the clusters were randomly reduced to ensure that all of them had the same size. That is, ensuring fair probabilities using the same amount of examples for each period. The resulting network was a fast tool that allowed to validate any relationship between the dates, the clades and the evolution of the antigen alleles.

Having explained the evolution of the B.pertussis strains, the next idea was to forecast the Pertussis incidence using a SARIMAX model. Unfortunately, adding strain information as exogenous variables had poor results, due to the lack of variance of the strains. Instead, different seasonalities were added to a model to forecast the Pertussis incidence in Catalonia. We saw that the season of B.pertussis is between summer and fall. Also, there was a 3 to 5 years seasonality, which is likely related to the evolution of the bacteria. The weekly forecasts for the test year were good and, according to the model, the next epidemic year will be the 2020. Nevertheless, there is always the possibility that an external factor (which the model ignores) interfere with this forecast. Therefore, this result should serve as a recommendation to take the appropriate preventive measures.

Since all these conclusions were specifically for Catalonia, comparisons with a couple of other studies were made. While the evolution of B.pertussis was generally similar all over the world, there were some differences either in timing or in allele persistence depending on the region. Applying our prediction models to external data from other developed countries, such as the Netherlands, resulted in an acceptable accuracy. Therefore, we concluded that a deeper analysis between countries, comparing the timings of Pertussis resurgence, the vaccination programs and other relevant factors could help to find key patterns.

Finally, mention that no correlations were found regarding the patients age, their gender, the number of vaccine doses administered and the vaccine type (WCV, ACV). One explanation could be that the vaccine effects vanish over time, thus allowing the infection regardless of the vaccine. Besides, there was no information about the decision criterion between isolating a strain from one patient or another. Therefore, we considered that no further conclusions could be made about this aspect.

From the data science point of view, this project encompassed, iteratively, data acquisition, cleaning, visualization, modelling and validation, finally presenting the results as a web service. A variety of techniques were studied and implemented, such as the Bayesian network, the SARIMAX model or the Flask application. Each technique was carefully chosen for a specific task, which allowed to achieve better results for the project objectives. To summarize, the chosen approaches and planning successfully led the project to succeed.

6.2 Future lines of work

As the project progressed, there were some aspects that were left as possible future lines of work, either because they were out of the scope of the project or because there were not enough data. The following list summarizes these future lines:

- The Bayesian network could be improved by adding more balanced data. Notice that in this project, clustering and data reduction techniques were used to achieve this balance. However, the ideal would be to have the same amount of examples per year, thus avoiding having to cluster or reduce the data. This could be achieved either by isolating more strains (which might be too expensive) or by using external sources of data.

- Using the seasonalities as exogenous variables for the SARIMAX model resulted in a black box model, where the results might be accurate but we cannot explain why. That is, it would be interesting to know exactly why these seasonalities exist. Adding more recent strain data to the model would help. Also, there might be other relevant characteristics that were not available in our data.

- Following the same lines, deeper comparisons with other similar studies could help to find key patterns of Pertussis resurgence. For instance, in the Netherlands the resurgence took place much earlier. Also, differences (with respect to Catalonia) in the evolution of the fim3 antigen were found, which leads to the hypothesis of the relationship between this antigen and the Pertussis resurgence.

- The specialization of the epidemiological profiles, especially that of the MLVA-27, should be studied more deeply. As seen in the state of the art, genomic data could be analysed to understand the evolution of the B.pertussis bacteria. Deep learning techniques, such as recurrent neural networks (used to deal with large sequential data), could be used to find the key differences between strains with very similar profiles. For instance, knowing what differentiates the MLVA-27 strains from the others could lead to important discoveries and better preventive measures for Pertussis.

Chapter 7

Glossary

ACF: Autocorrelation function.

ACV: Acellular Vaccine.

API: Application programming interface.

Allele: variant form of a given gene.

Antigen: substance that triggers the formation of antibodies. It can also cause an immune response.

Bordetella: genus of a gram-negative coccobacilli of the phylum Proteobacteria. Two Bordetella species (pertussis and parapertussis) can cause Whooping cough.

Clade: group of organisms with a common ancestor and its lineal descendants.

Deep Learning: subfield of machine learning that models high level abstractions.

Filamentous hemagglutinin: filamentous protein that serves as a dominant attachment factor for adherence to host that B.pertussis uses as a virulence factor.

Fimbria: appendage that can be found on many Gram-negative bacteria, such as the Bordetella pertussis.

Gene: lineal sequence of nucleotides in RNA or DNA, coding a molecule that has a function.

Genus: biological taxonomic rank, between species and family.

Insertion Sequence: simplest transposon that contains a transposase gene flanked.

Machine Learning: subfield of artificial intelligence, that uses models to learn from data.

MLVA: Multiple-Locus VNTR Analysis is a technique used to perform molecular typing.

PACF: Partial Autocorrelation function.

Pertactin: highly immunological protein that participates in the virulent process caused by the Bordetella pertussis.

PFGE: Pulsed-Field Gel Electrophoresis is technique used to fragment nucleic acids and perform strain typing or DNA fingerprint.

PTX: pertussis toxin. PtxA stands for the active subunit while ptxP covers the alleles of the promoter region.

RESTful: web service that implements the architecture Representational State Transfer.

SARIMAX: a Seasonal Auto Regressive Integrated Moving Average model with exogenous variables.

Strain: genetic variant or subtype of a microorganism.

Strain typing: is used to identify an specific strain.

Tandem repeat: happens when a certain pattern of nucleotides is repeated consecutively in DNA.

Transposase: enzyme that binds to the end of a transposon and catalyses its movement to another part of the genome.

Transposon: DNA sequence that can change its position within a genome.

VNTR: Variable number tandem repeat is a location in a genome where a tandem repeat occurs. It is used as a DNA fingerprint.

WCV: Whole-Cell Vaccine.

Web scraping: technique used for extracting data from websites.

Bibliography

[1] Genome Sequences of 28 Bordetella pertussis U.S. Outbreak Strains Dating from 2010 to 2012, . URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3868863/.

[2] Recombinación del DNA, . URL http://fbio.uh.cu/sites/genmol/confs/conf5/p04.htm.

[3] Multiple Locus Variable-number Tandem Repeat Analysis (MLVA) | PulseNet Methods| PulseNet | CDC, October 2017. URL https://www.cdc.gov/pulsenet/pathogens/mlva.html.

[4] Learn About Pertussis, 2018. URL https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/pertussis/learn-about-pertussis.html.

[5] Pinkbook | Pertussis | Epidemiology of Vaccine Preventable Diseases | CDC, July 2018. URL https://www.cdc.gov/vaccines/pubs/pinkbook/pert.html.

[6] Pulsed-field Gel Electrophoresis (PFGE) | PulseNet Methods| PulseNet | CDC, March 2018. URL https://www.cdc.gov/pulsenet/pathogens/pfge.html.

[7] Bordetella pertussis Serotypes in the United States, 2019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC378047/.

[8] Detectan cambios evolutivos en el agente causal de la tos ferina en Barcelona - Biotech Spain, February 2019. URL http://biotech-spain.com/es/articles/detectan-cambios-evolutivos-en-el-agente-causal-de-la-tos-ferina-en-barcelona/.

[9] Export dataset - Bordetella PubMLST, 2019. URL https://pubmlst.org/bigsdb?page=plugin&name=Export&db=pubmlst_bordetella_isolates.

[10] GanttProject, 2019. URL https://www.ganttproject.biz.

[11] Global Population Structure and Evolution of Bordetella pertussis and Their Relationship with Vaccination, March 2019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3994516/.

[12] Heroku, 2019. URL https://www.heroku.com.

[13] Immunization schedules by antigens, 2019. URL http://apps.who.int/immunization_monitoring/globalsummary/schedules.

[14] MLVA, 2019. URL https://www.mlva.net/bpertussis/default.asp.

[15] Pertussis (Whooping Cough) Questions and Answers Information About the Disease and Vaccines. *Questions and Answers*, page 4, 2019. URL http://www.immunize.org/catg.d/p4212.pdf.

[16] Pertussis - Museum of health care - Kingston, 2019. URL http://www.museumofhealthcare.ca/explore/exhibits/vaccinations/pertussis.html.

[17] Rules for identifying ARIMA models, 2019. URL http://people.duke.edu/~rnau/arimrule.htm.

[18] What Is Systems Biology · Institute for Systems Biology, 2019. URL https://systemsbiology.org/about/what-is-systems-biology/.

[19] WHO | Data, statistics and graphics, 2019. URL http://www.who.int/immunization/monitoring_surveillance/data/en/.

[20] Zotero, February 2019. URL https://www.zotero.org/.

[21] BART MJ . Global population structure and evolution of Bordetella pertussis and their relationship with vaccination. - PubMed - NCBI, May 2014. URL https://www.ncbi.nlm.nih.gov/pubmed/24757216.

[22] Gorringe AR and Vaughan TE. Expert Rev. Vaccines 2014 . Bordetella pertussis fimbriae (Fim): relevance for vaccines | Vacunas / Asociación Española de Vacunología, March 2019. URL https://www.vacunas.org/bordetella-pertussis-fimbriae-fim-relevance-for-vaccines/.

[23] James D. Cherry . Pertussis: Challenges Today and for the Future, July 2013. URL https://www.researchgate.net/publication/255736534_Pertussis_Challenges_Today_and_for_the_Future.

[24] María de Viarce Torres de Mier and Josefa Masa Calles Noemí López-Perea. Situación de la Tos ferina en España, 1998-2016, 2016. URL http://www.isciii.es/ISCIII/es/contenidos/fd-servicios-cientifico-tecnicos/fd-vigilancias-alertas/fd-enfermedades/fd-enfermedades-prevenibles-vacunacion/pdf_2018/Situacion_de_la_Tos_ferina_en_Espana_1998-2016.pdf.

[25] María Isabel Fernández Cano . Thesis - Comparision of notified incidence of whooping cough in Spain., 2016. URL https://ddd.uab.cat/pub/tesis/2016/hdl_10803_399576/mifc1de1.pdf.

[26] The Sanger Institute UK . Comparative analysis of the genome sequences of Bordetella, 2003. URL https://www.ncbi.nlm.nih.gov/pubmed/12910271.

[27] A. Darwiche. Handbook of Knowledge Representation - Bayesian networks, 2008. URL http://dai.fmph.uniba.sk/~sefranek/kri/handbook/chapter11.pdf.

[28] Michael Barber. Data science concepts you need to know!, January 2018. URL https://towardsdatascience.com/introduction-to-statistics-e9d72d818745.

[29] Marieke J. Bart. Global Population Structure and Evolution of Bordetella pertussis and Their Relationship with Vaccination. *mBio*, 5(2), April 2014. ISSN 2150-7511. doi: 10.1128/mBio.01074-14. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3994516/.

[30] Thomas Belcher and Andrew Preston. Bordetella pertussis evolution in the (functional) genomics era. *Pathog Dis*, 73(8), 2019. ISSN 2049-632X. doi: 10.1093/femspd/ftv064. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4626590/.

[31] Daniel Bourke. Bioinformatics: Where code meets biology, February 2019. URL https://towardsdatascience.com/bioinformatics-where-code-meets-biology-faa2b99cdfcb.

[32] Jason Brownlee. What is Deep Learning?, August 2016. URL https://machinelearningmastery.com/what-is-deep-learning/.

[33] Jason Brownlee. Autoregression Models for Time Series Forecasting With Python, January 2017. URL https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/.

[34] Jason Brownlee. A Gentle Introduction to SARIMA for Time Series Forecasting in Python, August 2018. URL https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/.

[35] Rotem Lapidot Christopher J. Gill. The Pertussis resurgence: putting together the pieces of the puzzle. 2016. URL https://www.researchgate.net/publication/311642076_The_Pertussis_resurgence_Putting_together_the_pieces_of_the_puzzle.

[36] Silvia Díaz. Las tasas de vacunación en España son excelentes: así se distribuyen las coberturas vacunales por CC.AA, July 2018. URL https://www.bebesymas.com/salud-infantil/tasas-vacunacion-espana-excelentes-asi-se-distribuyen-coberturas-vacunales-cc-aa.

[37] M. Domenech de Cellès, A. A. King, and P. Rohani. Resolving pertussis resurgence and vaccine immunity using mathematical transmission models. *Human Vaccines & Immunotherapeutics*, pages 1–4, November 2018. ISSN 2164-5515, 2164-554X. doi: 10.1080/21645515.2018.1549432. URL https://www.tandfonline.com/doi/full/10.1080/21645515.2018.1549432.

[38] Matthieu Domenech de Cellès, Felicia M. G. Magpantay, Aaron A. King, and Pejman Rohani. The pertussis enigma: reconciling epidemiology, immunology and evolution. *Proc. R. Soc. B*, 283(1822):20152309, January 2016. ISSN 0962-8452, 1471-2954. doi: 10.1098/rspb.2015.2309. URL http://rspb.royalsocietypublishing.org/lookup/doi/10.1098/rspb.2015.2309.

[39] Marjolein van Gent and Frits R. Mooi. Small Mutations in Bordetella pertussis Are Associated with Selective Sweeps. *PLOS ONE*, (9):e46407, September 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0046407. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0046407.

[40] Kwang-Il Goh and Albert-László Barabási. The human disease network. *PNAS*, 104(21): 8685–8690, May 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0701361104. URL https://www.pnas.org/content/104/21/8685.

[41] Sabine C. de Greeff and Guy A. M. Berbers. Seroprevalence of Pertussis in the Netherlands: Evidence for Increased Circulation of Bordetella pertussis. *PLOS ONE*, 5(12):e14183, 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0014183. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014183.

[42] Institut de Salut Juan Carlos. Boletín Epidemiológico Semanal en Red, 2019. URL http://www.cat.isciii.es/ISCIII/es/contenidos/fd-servicios-cientifico-tecnicos/fd-vigilancias-alertas/fd-boletines/boletin-epidemiologico-semanal-red.shtml.

[43] Josef Perktold,, Skipper Seabold, and Jonathan Taylor. endog, exog, what's that? — statsmodels 0.9.0 documentation, 2019. URL https://www.statsmodels.org/stable/endog_exog.html.

[44] Wenjun Li and Pierre-Edouard Fournier. Bacterial strain typing in the genomic era. *FEMS Microbiol Rev*, 33(5):892–916, September 2009. ISSN 0168-6445. doi: 10.1111/j.1574-6976.2009.00182.x. URL https://academic.oup.com/femsre/article/33/5/892/562963.

[45] Machine Learning Plus. ARIMA Model - Complete Guide to Time Series Forecasting in Python | ML+, February 2019. URL https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/.

[46] Natalia Malachowa and Waleria Hryniewicz. Comparison of Multiple-Locus Variable-Number Tandem-Repeat Analysis with Pulsed-Field Gel Electrophoresis, spa Typing, and Multilocus Sequence Typing for Clonal Characterization of Staphylococcus aureus Isolates. *Journal of Clinical Microbiology*, 43(7):3095, July 2005. doi: 10.1128/JCM.43.7.3095-3100.2005. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1169160/.

[47] Marianna Sockrider MD and Chantal Spencer MD . Que es la Pertussis (tos ferina)., 2015. URL https://www.thoracic.org/patients/patient-resources/resources/spanish/pertusis.pdf.

[48] Chris Mattmann. tika: Apache Tika Python library, 2019. URL http://github.com/chrismattmann/tika-python.

[49] Madhav Mishra. Unboxing ARIMA Models, June 2018. URL https://towardsdatascience.com/unboxing-arima-models-1dc09d2746f8.

[50] David Moreno-Pérez. Calendario de vacunaciones de la Asociación Española de Pediatría: recomendaciones 2019. *Anales de Pediatría*, 90(1):56.e1–56.e9, January 2019. ISSN 16954033. doi: 10.1016/j.anpedi.2018.10.006. URL https://linkinghub.elsevier.com/retrieve/pii/S169540331830376X.

[51] pdfkungfoo. Using ‘tabula-extractor‘ to liberate tables from their PDF imprisonment, 2016. URL https://asciinema.org/a/22761.

[52] Pol Major i Munich. pmajor TFM - API, 2019. URL https://pmajortfm.herokuapp.com/bayesian.

[53] Anne Marie Queenan and David J. Dowling. Increasing FIM2/3 antigen-content improves efficacy of Bordetella pertussis vaccines in mice in vivo without altering vaccine-induced human reactogenicity biomarkers in vitro. *Vaccine*, 37(1):80–89, January 2019. ISSN 0264-410X. doi: 10.1016/j.vaccine.2018.11.028. URL http://www.sciencedirect.com/science/article/pii/S0264410X18315512.

[54] Rachael Tatman. CareerCon: Intro to APIs, 2019. URL https://kaggle.com/rtatman/careercon-intro-to-apis.

[55] René H. M. Raeven, Elly van Riet, and Hugo D. Meiring. Systems vaccinology and big data in the vaccine development chain. *Immunology*, 156(1):33–46, 2019. ISSN 1365-2567. doi: 10.1111/imm.13012. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/imm.13012.

[56] T. F. Rice and B. Kampmann. Antibody responses to Bordetella pertussis and other childhood vaccines in infants born to mothers who received pertussis vaccine in pregnancy. *Clinical & Experimental Immunology*, 0(0), 2019. ISSN 1365-2249. doi: 10.1111/cei.13275. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cei.13275.

[57] Sangarshanan. Time series Forecasting — ARIMA models, October 2018. URL https://towardsdatascience.com/time-series-forecasting-arima-models-7f221e9eee06.

[58] Sean Abu. Seasonal ARIMA with Python, 2016. URL https://www.seanabu.com/2016/03/22/time-series-seasonal-ARIMA-model-in-python/.

[59] Harshdeep Singh. Understanding Gradient Boosting Machines, November 2018. URL https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab.

[60] Devin Soni. Introduction to Bayesian Networks, June 2018. URL https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e.

[61] Stephanie. Autoregressive Model: Definition & The AR Process, August 2015. URL https://www.statisticshowto.datasciencecentral.com/autoregressive-model/.

[62] Joan Torres and Carme Trilla. Brote de tos ferina con elevada tasa de ataque en niños y adolescentes bien vacunados. *Enferm Infecc Microbiol Clin*, 29(8):564–567, October 2011. ISSN 0213-005X. doi: 10.1016/j.eimc.2011.04.005. URL http://www.elsevier.es/es-revista-enfermedades-infecciosas-microbiologia-clinica-28-articulo-brote-tos-ferina-con-elevada-S0213005X11001613.

[63] Andrea Trevino. Introduction to K-means Clustering, 2016. URL https://www.datascience.com/blog/k-means-clustering.