UCLouvain
Institute for Information
and Communication Technologies,
Electronics and Applied Mathematics

Mathematical Modelling of Infectious Diseases: COVID-19 and Malaria

Ousmane DIAO

Thesis submitted in partial fulfillment
of the requirements for the degree of
Doctorat en sciences de l'ingénieur et technologie

Dissertation committee:

Prof. Pierre-Antoine Absil (UCLouvain, advisor)
Prof. Mouhamadou Diallo (UCAD, Dakar, advisor)
Prof. Emmanuel Hanert (UCLouvain, Belgium)
Prof. Ramsès Djidjiou Demasse (IRD, Montpellier, France)
Prof. Raphaël Forien (INRAE, Avignon, France)
Prof. Philippe Lefèvre (UCLouvain, Belgium, chair)

Version of August 7, 2023.

Mathematical Modelling of Infectious Diseases: COVID-19 and Malaria
by Ousmane Diao

Professional Email: ousmane.diao@uclouvain.be
Personal Email: diaoousmane1@gmail.com
ICTEAM
UCLouvain
1348 Louvain-la-Neuve
Belgium

This thesis is supported by a fellowship awarded by

- UCLouvain's Conseil de l'Action Internationale (CAI) South Partnership Doctoral grants
- ARES-CCD (Commission pour la coopération au développement) Doctoral mobility grants

0.0 Abstract

During the COVID-19 pandemic, authorities paid particular attention to the number of hospitalized individuals and had an interest in forecasting the evolution of the situation, especially at the hospital level. In this context, we propose a model that is a simplified version of the well-known SIR compartmental model of infectious diseases. On several occasions, with optimized parameters and initial conditions, this time-invariant two parameter two-dimensional model is able to fit COVID-19 hospitalization data over several months with high accuracy (e.g., the root relative squared error is below 10% for Belgium over the period from 2020-03-15 to 2020-07-15). We also investigate three methods such as stochastic uncertainty, parameter uncertainty and model uncertainty to produce confidence intervals for the SH model. Moreover, we observed that, when the model is trained on a suitable three-week period around the hospitalization peak for Belgium, it forecasts the subsequent two months with mean absolute percentage error (MAPE) under 4%. We repeated the experiment for each French department and found 14 of them where the MAPE was below 20%. However, when the model is trained in the increase phase, it is less successful at forecasting the subsequent evolution.

Malaria disease is endemic in Senegal and is the cause of most cases of hospital visits during the rainy season. We use generalized linear models (GLMs) with probability distributions such as Poisson, negative binomial and Gaussian to forecast malaria incidence taking into account climatic variables such as rainfall, average temperature and relative humidity, and the history of malaria incidence in Dakar, Fatick and Kedougou, three different endemic regions of Senegal. Forecasting algorithm methods are developed by considering protocols: (1) the meteorological explanatory variable X_j is taken at time $t - \ell_j$, where t represents the time of the forecasts, ℓ_j is the `lag` in X_j that maximizes its correlation with the malaria incidence; (2) the meteorological explanatory variables are assumed to be available at time t. We investigate the efficiency of the `scipy.optimize.fmin` solver as an alternative to the default method for GLMs, Newton-Raphson. Parameter uncertainty was studied in order to produce confidence intervals for forecasts. In addition, a saturation method is introduced on the rainfall variable to remedy some overestimations observed during the forecasts. As results, the lowest MASE_test's are observed with the protocol

(2), which gives 0.94 in Dakar, 0.92 in Fatick and 0.8 in Kedougou. These values indicate an excellent accuracy of the model with our datasets. The application of the saturated rainfall has reduced by 4% in the sense of MARE the over-estimation occurring at the end of 2015 only with protocol (1), in Dakar.

0.0 Research Publications

Journal articles

1- P.-A. Absil, O. Diao, and M. Diallo, "Assessment of COVID-19 hospitalization forecasts from a simplified SIR model", Letters in Biomathematics, vol. 8, no. 1, pp. 215–228, 2021, https://lettersinbiomath.journals.publicknowledgeproject.org/index.php/lib/article/view/403.
2- O. Diao, P.-A. Absil, and M. Diallo, "Generalized linear models to forecast malaria incidence in three endemic regions of Senegal", Int. J. Environ. Res. Public Health, Volume 20, Issue 13, 6303, 2023, https://www.mdpi.com/1660-4601/20/13/6303.

Conferences

- O. Diao, P.-A. Absil, and M. Diallo, "Optimizing generalized linear models for malaria incidence forecasting", Proceedings of the 42th Benelux meeting (in-person, talk), Elspeet, Netherlands, 2023.
- O. Diao, P.-A. Absil, and M. Diallo, "Modelling malaria incidence by a generalized linear model based on meteorological factors in three endemic regions, Senegal", Proceedings of the Ecology and Evolution of Infectious Diseases EEID (online, poster), Atlanta, USA, 2022.
- O. Diao, P.-A. Absil, and M. Diallo, "Saturated pluviometry as explanatory variable to forecast malaria incidence", Proceedings of the International Congress for Tropical Medicine and Malaria (ICTMM) (in-person, poster), Bangkok, Thailand, 2022.
- O. Diao, P.-A. Absil, and M. Diallo, "An SIR-type model for COVID-19 hospitalization forecasts", Proceedings of the Benelux workshop (in-person, talk), Rotterdam, Netherlands, 2021.
- O. Diao, P.-A. Absil, and M. Diallo, "Mathematical modelling of malaria transmission taking into account the influence of current prevention and treatment", Proceedings of the 39th Benelux meeting (in-person, talk), Elspeet, Netherlands, 2020.

0.0 Acknowledgements and dedications

This thesis marks the end of four years of research in the prestigious Belgium university that is UCLouvain, and thus the beginning of my future career as a doctor in engineering of mathematical modelling. This work consisted of several months of research, reflection, coding and writing. It has been a long job, and one that inevitably had its ups and downs. Then, achieving this thesis has allowed me to develop both from a professional/scientific and personal point of view.

I would like to thank the people who have contributed to this thesis and helped me along the way. Starting with a special thanks to my supervisors: Pierre-Antoine Absil (UCLouvain) and Mouhamadou Diallo (UCAD) without whom none of this would have been possible. They have both showed a lot of patience and gave me a great deal of guidance for this thesis. Their interest in my work, their availability and their precious advice have been of great help for me. I am very grateful for it.

I would also like to thank my thesis accompanying committee members Pr Emmanuel Hanert (ELI, UCLouvain, Louvain-la-Neuve, Belgium) and Pr Ramsès Djidjiou Demasse (IRD, Montpellier, France), who accepted to take the necessary time to read and evaluate my work.

I would also like to thank my colleagues at the Euler, specially to Loïc Van Hoorebeeck, Hazan Daglayan Sevim and Charles Monnoyer de Galland de Carnières.

I would also like to thank two staff members of “Institut Pasteur de Dakar, Sénégal" Mawlouth Diallo and Ibrahima Dia for accepting to meet me at their offices and share some data bases.

I would also like to thank my friends, from Sénégal, who have supported me for many years.

Of course, I dedicate this thesis to my wonderful family who always supported me in everything I do, which constitutes my main motivation source. In successes and failures, they have always been there to give me the strength to persevere and to get the best out of me.

Contents

⋆ Abbreviations and glossary

UCAD: “Université Cheikh Anta Diop”.
PNLP: “programme national de lutte contre le paludisme".
WHO: world heath organization.
GLM: generalized linear model.
SARIMAX: seasonal autoregressive integrated moving average with explanatory variables.
RDT: rapid diagnostic test.
Iid: independent and identically distributed.
PCR: polymerase chain reaction.

Important terms [VW10, WHOHM03].

Beta (β) The rate at which two specific individuals come into effective contact per unit time. Technically it is the *per capita* rate at which two specific individuals come into effective contact per unit time. It is also sometimes referred to using other names, such as the transmission coefficient, transmission rate, contact parameter, etc.

Compartmental model A model in which individuals in the population are subdivided into broad subgroups (compartments) and the model tracts individuals collectively.

Examples of compartmental model’s structures:

- SI: Susceptible–infectious.
- SIS: Susceptible–infectious–susceptible.
- SIR: Susceptible–infectious–recovered.
- SIRS: Susceptible–infectious–recovered–susceptible.
- SEIRS: Susceptible–pre-infectious–infectious–recovered–susceptible.

Deterministic model A model which describes what happens on average in a population and does not incorporate the effects of chance.

Endemic infection An infection that is present in a population at a similar level over a prolonged period.

Epidemic The increase and subsequent decrease in incidence following the (re) introduction of an infection in a population. Also refers to the occurrence of cases in a given locality at a frequency which greatly exceeds that expected.

Epidemiology This term is used to refer to the study of patterns of diseases (including those which are non-communicable) or infections in a population.
Fitting Fitting refers to the process of finding the best parameters for a model to solve a specific real problem with a high level of accuracy (accessed on 2023/07/17 at: `https://www.datarobot.com/wiki/fitting/`).
Forecast It is an estimate of what future observations will be if the underlying process continues as it has in the recent past [Bro04].
Hyperendemic areas Intense seasonal transmission but not sufficient enough for a very high proportion of the population to develop protective immunity.
Hypoendemic areas A very little transmission and low risk of infection to the population.
Holoendemic areas Year round transmission with a high degree of immunity among the population of all age groups, particularly adults.
Incidence (rate) The number of new events, such as infections or cases in the population at risk (usually susceptibles), per unit person per unit time. Note that the number of new events per person in the general population per unit time is also frequently referred to as the “incidence" in epidemiological and medical literature and surveillance reports. For rare events, this measure is identically to the incidence defined using the strict epidemiological definition.
Infection An infection may be defined as the invasion of an organism by a smaller (infecting) organism.
Mesoendemic areas Typically rural villages mainly in subtropical regions with varying intensity of transmission and often prone to malaria epidemics.
Pandemic An epidemic that occurs worldwide.
Susceptible An individual who has not yet been infected and is at risk of infection.
Time Series A time series is a sequential set of data points, measured typically over successive times. It is mathematically defined as a set of vectors $x(t), t = 0, 1, 2, \ldots$ where t represents the time elapsed [AA13]. The variable $x(t)$ is treated as a random variable. The measurements taken during an event in a time series are arranged in a proper chronological order.

⋆ Notations

$Y_o(t)$: malaria incidence at month t. Consequently, the history of malaria incidence is denoted by $Y_o(t-\delta), \delta \geq 1$.
$Rf_o(t)$: rainfall variable on month t.
$RH_o(t)$: relative humidity variable on month t.
$Tp_o(t)$: average temperature variable on month t.
$I_o(t)$: number of COVID-19 confirmed cases on day t.
$H_o(t)$: number of COVID-19 hospitalized patients on day t.
$E_o(t)$: number of COVID-19 patients entering the hospital (number of lab-confirmed hospital intakes) on day t.
$L_o(t)$: number of COVID-19 patients discharged from the hospital on day t.
$R_o(t)$: number of COVID-19 patients recovered on day t.
$D_o(t)$: number of COVID-19 patients deaths from the hospital on day t.
$B_o(t)$: insecticide-treated bed-nets (ITNs).
$A_o(t)$: artemisinin-based combination therapy (ACT).

The subscript $_o$ stands for "observed".

List of Figures

List of Tables

1

Introduction

1.1 Context

Modelling is to convert a real-world problem into a problem of a mathematical nature in order to enable us to understand and predict. Understanding the transmission characteristics of infectious diseases in communities, regions, and countries can lead to better approaches to decreasing the transmission of these diseases [Het00].

Indeed, mathematical models are used in comparing, planning, implementing, evaluating, optimizing various detection, prevention, therapy, and control programs. Epidemiology modeling can contribute to the design and analysis of epidemiological surveys, suggest crucial data that should be collected, identify trends, make general forecasts, and estimate the uncertainty in forecasts [Het89]. Examples of data can be the number of infected persons, the number of hospitalized, the recovered time, the mortality rate etc.

In this thesis, we are interested in two infectious diseases: COVID-19 pandemic in Belgium and malaria endemic in Senegal.

1.2 About COVID-19

1.2.1 Epidemiological situation

The coronavirus (SARS-CoV-2) is a pathogenic infectious agent causing severe acute respiratory syndrome. It was originally identified in Wuhan (China) [AWS+21] and was officially declared a pandemic by the WHO on March 11, 2020.

Its origin is not very sure, and has been intensely debated between the scientists. According to [VS22], SARS-CoV-2 would come from a form of coronavirus that exists on bats. Through zoonotic transmission, this virus has then evolved into a "human coronavirus", known today as COVID-19. Different animals could have been the host of the virus, transmitting it from the bats to the humans. However, the pangolin is the most suspected host, even if there still are some ongoing investigations on other possible host animals of this virus.

The SARS-CoV-2 virus is known to spread between people in several different ways [Who21].

- Current evidence suggests that the virus spreads mainly between people who are in close contact with each other, for example at a conversational distance. The virus can spread from an infected person's mouth or nose in small liquid particles when they cough, sneeze, speak, sing or breathe. Another person can then contract the virus when infectious particles that pass through the air are inhaled at short range (this is often called short-range aerosol or short-range airborne transmission) or if infectious particles come into direct contact with the eyes, nose, or mouth (droplet transmission).

- The virus can also spread in poorly ventilated and/or crowded indoor settings, where people tend to spend longer periods of time. This is because aerosols can remain suspended in the air or travel farther than conversational distance (this is often called long-range aerosol or long-range airborne transmission).

- People may also become infected when touching their eyes, nose or mouth after touching surfaces or objects that have been contaminated by the virus.

Laboratory data suggests that infected people appear to be most infectious just before they develop symptoms (namely 2 days before they develop symptoms) and early in their illness. People who develop severe disease can be infectious for longer.

The spread of the disease occurred so quickly that no one was prepared for it. A consequence of this has been the difficulty to manage the situation in the hospitals, and especially in the intensive care units. The hospital staff had to deal with an extremely large number of patients in need of intensive care, without either extra workforce nor extra material [WKJB20]. Hospital staff faced a lack of beds in the intensive care units, as well as a lack of oxygen masks and oxygen cylinders. The care units thus became rapidly overcrowded and the staff overwhelmed [GWT+20]. The situation generally became highly critical.

To relieve the situation, the governments of many countries put a number of measures in place with the most prominent ones being social distancing, quarantines, and testing with various tests, such as PCR and antigen tests. These measures have been considered short-term solutions. Soon, more durable solutions where introduced, among which five vaccines approved by the European Commission [Com22], which are BioNTech and Pfizer, Moderna, AstraZeneca, Janssen Pharmaceutica SA and Novavax. The main goal was to reduce the number of deaths caused by COVID-19 and to reduce the number of hospitalized patients. Therefore, lockdown was imposed which means that people weren't allowed to travel anymore and that almost every public place was closed: schools, restaurants, unessential shops, companies, cultural places.

Regarding this situation, it was judged very important to develop forecasting models in order to plan measures against the future evolution of this pandemic.

1.2.2 State of art

Many methods of modelling the transmission of the COVID-19 have been developed. These methods where based on compartmental models, and in particular the well-known SIR model [KMW27], used since the early 20th century to model infectious diseases; see [Het00] and references therein.

Certain methods are related to the prediction of the future of this pandemic. In [AWS+21], an SEIR model is proposed to predict hospitalization cases, based on real data, and to predict the resurgence of the COVID-19

epidemic taking into account the age structure of the population, and different government strategies (physical distancing, quarantine, lockdown, school closure, work at home and transport closure) to fight the pandemic. [Gab20] uses the common SIR model to show how the cumulative number of deaths can be regarded as a master variable to estimate some epidemiological parameters such as the reproduction number R_0, the size of the susceptible population and the infection rate. In this study, the data explored come from Spain, Italy, New York and Hubei Province.

In Belgium, some transmission models were developed during the pandemic of COVID-19.

Michel Rigo in “Modèles mathématiques et confinement" [Rig20] has studied the evolution of the epidemiological curves of S and R related to the COVID-19 spread through an SIR model.

Then, Niel Hens and al. [AWS+21] used a stochastic discrete time compartmental model to describe the transmission of COVID-19 in Belgium. In this study, the authors take into account the age-structure by integrating data on social contacts to (i) assess the impact of the lockdown as implemented on March 13, 2020 on the number of new hospitalizations in Belgium; (ii) conduct a scenario analysis estimating the impact of possible exit strategies on potential future COVID-19 waves. More specifically, their model is fitted to hospital admission data, data on the daily number of COVID-19 deaths and serial serological survey data informing the (sero)prevalence of the disease in the population while relying on a Bayesian MCMC approach. The Authors declare that despite an extensive exploration of various projections for the future course of the epidemic, based on the impact of adherence to measures of physical distancing and a potential increase in contacts as a result of the relaxation of the stringent lockdown measures, a lot of uncertainty remains about the evolution of the epidemic in the next months.

In [CLP+21], the authors developed a stochastic compartmental model (SEIR), data-informed, meta-population model that accounts for mixing and mobility of the age-structured population of Belgium. The model is calibrated to daily hospitalization data and is able to reproduce the outbreak at the national level. This model is able to successfully describe (fit) the initial wave of COVID-19 in Belgium and identifies interactions during leisure/other activities as pivotal in the exit strategy. According to the authors, scenarios predicting a second wave of hospitalizations were not observed, suggesting that the per-contact probability of infection has changed with respect to the pre-lockdown period.

The main difference between these studies and our work remains the approach related to the parameter estimation method, the data assimilation (for the generation of the initial conditions) defined in Section 3.2 and the machine learning (for the train-test approach) presented in Section 1.5.5.

1.3 About malaria

MALARIA is an infectious disease caused by a parasite called *Plasmodium* transmitted by a mosquito (female Anophele). It can also be passed to human by blood transfusion, sharing needle or congenitally [PJB14].

1.3.1 Life cycle of anopheline mosquitoes

Mosquitoes have four different stages in their life cycle: eggs, larva, pupa and adult [WHOHM03]. The time taken for the various stages to develop depends on temperature and nutritional factors in their environment. Development is shorter at higher temperatures.
Eggs: A female anopheline mosquito normally mates only once in her lifetime. She usually requires a blood meal after mating before the eggs can develop. Blood meals are generally taken every two to three days; before the next batch of eggs is laid. About 100 to 150 eggs are laid on the water surface during oviposition. Oviposition sites vary from small hoof prints and rain pools to streams, swamps, canals, rivers, ponds, lakes and rice fields. Each species of mosquito prefers different types of habitats to lay eggs. Under the best conditions in the tropics, the average lifespan of female anopheline mosquitoes is about three to four weeks. A female mosquito continues to lay eggs throughout her lifetime. Most females will lay between one and three batches of eggs during their life, though some may lay as many as seven batches.
Larva: A larva hatches from the egg after about one or two days and generally floats parallel under the water surface, since it needs to breathe air. It feeds by taking up food from the water. When disturbed, the larva quickly swims towards the bottom but soon needs to return to the surface to breathe. There are four larval stages or instars. The small larva emerging from the egg is called the first instar. After one or two days it sheds

its skin and becomes the second instar, followed by the third and fourth instars at further intervals of about two days each. The larva remains in the fourth instar stage for three or four more days before changing to a pupa. The total time spent in the larval stage is generally eight to ten days at normal tropical water temperatures. At lower temperatures, the aquatic stages take longer to develop.

Pupa: The pupa is the stage during which a major transformation takes place, from living in water to becoming a flying adult mosquito. The pupa is shaped like a comma. It stays under the surface and swims down when disturbed but does not feed. The pupal stage lasts for two to three days after which the skin of the pupa splits. Then the adult mosquito emerges and rests temporarily on the water's surface until it is able to fly.

Adult: Mating takes place soon after the adult emerges from the pupa. The female usually mates only once because she receives sufficient sperm from a single mating for all subsequent egg batches. Normally the female takes her first blood meal only after mating, but sometimes the first blood meal can be taken by young virgin females. The first batch of eggs develops after one or two blood meals (depending on the species), while successive batches usually require only one blood meal. The feeding and resting habits of mosquitoes are of great importance in control programmes and for this reason they must be well understood. Most anopheline mosquitoes bite at night. Some bite shortly after sunset while others bite later, around midnight or the early morning. Some mosquitoes enter houses to bite and are described as being endophagic; others bite mostly outside and are called exophagic. After the mosquito takes a blood meal she usually rests for a short period. Mosquitoes that enter a house usually rest on a wall, under furniture or on clothes hanging in the house after they bite and are said to be endophilic. Mosquitoes that bite outside usually rest on plants, in holes, in trees or on the ground or in other cool dark places and are called exophilic.

Authors in [GBMM16, OG17] have defined parameters related to mosquito development (growth) and activity as function of temperature and rainfall.

- Mosquito's biting rate (per day) defined as:

$$a(T,R) = \beta_J(T,R)/2.325. \tag{1.1}$$

- Birth rate of juveniles (per day) defined as:

$$\beta_J(T,R) = 10\beta_v(T,R). \tag{1.2}$$

- Adult mosquito birth rate (per day) defined as:

$$\beta_v(T,R) = n_E P_E(R) P_L(T,R) P_P(R) / (\tau_E + \tau_L(T) + \tau_P). \tag{1.3}$$

- Daily survival probability of eggs defined as:

$$P_E(R) = (4 \times 0.9 / R_L^2) R(R_L - R). \tag{1.4}$$

- Daily survival probability of larvae defined as:

$$P_L(T,R) = (4 \times 0.25 / R_L^2) R(R_L - R) \times e^{-0.00554T - 0.06737}. \tag{1.5}$$

- Daily survival probability of pupae defined as:

$$P_P(R) = (4 \times 0.75 / R_L^2) R(R_L - R). \tag{1.6}$$

- Duration of larvae development defined as:

$$\tau_L(T) = 1/(0.00554T - 0.06737) \text{ (days)}. \tag{1.7}$$

- Juvenile mosquito death defined as:

$$\mu_J(T) = 0.0025T^2 - 0.094T + 1.0257. \tag{1.8}$$

- Progression rate of mosquitoes death defined as:

$$\alpha_v(T) = e^{-1/(-0.03T^2 + 1.31T - 4.4)}. \tag{1.9}$$

- Adult mosquito death defined as:

$$\mu(T) = -\ln \alpha_v(T). \tag{1.10}$$

But the estimation of the parameters contained in these equations is possible only with adequate and reliable collected data.

1.3.2 Epidemiological situation

According to the 2020 World Health Organization (WHO) report, malaria caused 627,000 deaths, 95% of which were registered in the African Re-

gion. The number of malaria cases reduced from 2000 (238 millions reported cases) to 2019 (229 millions reported cases). In spite of this diminution, malaria is still endemic in many countries in the world, particularly in Africa.

In Senegal, the most of malaria cases occur during the rain season and in the entire country, from June to November. This disease constitutes a major public health problem according to the PNLP.

However, to reduce the vector density, the morbidity and the mortality due to malaria, several strategies have been developed in recent years, namely insecticide-treated bed-nets (ITNs), Vaccines, Anti-Malarial Drugs (mainly the artemisinin-based combination therapy (ACT)) and indoor residual spraying (IRS) [Mal15].

In Senegal, the following strategies are adopted for malaria control according to [dLclP16, FCG+18]:

- Long-lasting insecticide-treated nets (LLINs): the national strategy includes a goal of universal coverage of LLINs through mass campaigns and routine distribution during antenatal clinics and immunization campaigns.

- Indoor residual spraying (IRS): based on the malaria incidence and entomological data, spraying with the IRS is conducted by the selected district level administration. The districts are chosen in zones where the incidence was greater than 15 malaria cases per 1000.

- Intermittent preventive treatment for pregnant women (IPT): at least 80 % of pregnant women are protected by the IPT according to national guidelines. It is conducted at all health facilities offering antenatal care (ANC) nationwide. Moreover, all pregnant women should receive at least three doses of sulfadoxine-pyrimethamine (SP) starting in the second trimester, with at least 1 month between doses.

- Rapid diagnostic tests (RDTs): the malaria cases are confirmed on 100 % by RDT or “goutte épaisse", in accordance with national guidelines.

- Artemisinin combination therapy (ACT): any malaria case must be confirmed with sure and effective anti-malaria drugs accordance with national guidelines. The provision of drugs such as ACT, SP, Quinine, Arthésunate, Artemether, Rectocaps up to 99 % is taken care of by public services.

- “Prise en charge à domicile" (PECADOM): home-based management of malaria is implemented at the community level, where community health workers (“Dispensateur de soins à domicile-DSDOM") confirm malaria cases with RDTs, provide treatment with ACTs and refer complicated cases to health facilities.
- Focal test and treat (FTAT): it is a form of reactive case investigation (RCI) carried out in one northern district (Richard Toll).
- Seasonal malaria chemoprevention (SMC): at least 95 % of children between 3 to 120 months in the targeted zones are protected by the CPS. It is currently implemented in 16 high transmission districts meeting WHO SMC criteria.
- Case management: all patients receive RTDs and ACT free of charge at all levels of the health system. Testing with microscopy is available at health centers and hospital laboratories.

Some combinations named scale-up for impact are also developed for effective interventions in the form of targeted packages: (1) SUFI = bed nets, intermittent preventive treatment in pregnancy, rapid diagnostic tests and artemisinin combination therapy; (2) SUFI + reactive case investigation (focal test and treat); (3) SUFI + indoor residual spraying (IRS); (4) SUFI + seasonal malaria chemoprophylaxis (SMC) and (5) SUFI + SMC + IRS.

To meet the challenge of eradicating this infectious disease mathematical modelling can be a precious tool.

1.3.3 State of art

The first important contribution in mathematical modelling on malaria disease comes from the works of R. Ross (a medical doctor, a mathematician, and an epidemiologist) in 1911 [BN16]. He models the human (host) population divided into two compartments (S: susceptibles and I: infectious), and mosquito (vector) population divided into two compartments (S: susceptibles and I: infectious). Then, Kermack et McKendrick proposed an SIR model to explain the rise and the rapid decline in the number of infected patients observed during outbreaks such as plague (to London in 1665, in Bombay in 1906) or cholera (in London in 1865). It was brought in the forefront in the late 1970s by Anderson and May [Rig20]. Since then, several mathematical models have been used to provide an explicit framework for understanding disease transmission dynamics in human population.

In [ADB+13], an SI-SI model is proposed to show that the use of insecticide treated Nets (ITNs) would lead to the elimination of malaria if and only if, 75% of the population used them. For this, authors model how the bed-net limits the contact between vector (mosquito) and host (human), and also how it kills mosquitoes due to the insecticide applied on the bed-net. They define the transmission rate as $\beta(b) = \beta_{max} - b(\beta_{max} - \beta_{min})$, where b is the proportion of bed-net use ($0 \leq b \leq 1$), and β_{max} and β_{min}, are the maximum and the minimum transmission rates, respectively. Moreover, they define the death rate of the mosquitoes as $\mu_v(b) = \mu_{v_1} + b\mu_{max}$ where μ_{v_1} is the natural death rate for mosquitoes, and $b\mu_{max}$ is the death rate due to pesticide on treated bed-net. The usage of bed nets and indoor residual spraying is also studied in [OSJF13], to show the level impact of their combination into the community. Then, authors in [Res17], used an SEIR-SEI model to show how the increase of Vaccine or Anti-malarial Drugs effectiveness can reduce the number of female Anopheles. The disease disappears from the population within a certain time.

Some studies incorporate the climate change. In [GBMM16], the effect of the temperature and the rainfall is taken into account in a compartmental transmission model. Results reveal that malaria burden is likely to increase in the tropics, the highland regions, and East Africa and along the northern limit of falciparum malaria. Then, in [DDAAB20a], authors developed an SEIR-SEI mathematical model with seasonal mosquito life-history traits in Mpumalanga province in South Africa. A global analysis of the model was provided and it was concluded that there was an accordance to the simulation results with the seasonal variation of the reported cases.

But estimating parameters, with this kind of model, can be difficult, even relying as much as possible on data from the literature. That can justify a priori the use of simple statistical models, such as generalized linear models (GLMs) [PJ83, Lin97, GS93] or seasonal autoregressive moving average with exogenous variables (SARIMAX) [LAA17, KPT+10].

Examples of GLM includes the Poisson regression developed first in 1997 by Nelder and Wedderburn [Lee20]. But also, the negative binomial (NB) [AMA+19], the quasi-Poisson [Nak97], and the zero-inflated regression [AMA+19]. In general, the Poisson regression model is the most used discrete distribution for data fitting. But, the assumption of equality in observations mean and variance can limit its application on the over-dispersed data [Lee20, AMA+19, Nak97].

In [BVG+08], a study has been conducted to compare the performance

of three time series models such as the exponentially weighted moving average, the autoregressive integrated moving average (ARIMA) and the seasonal autoregressive integrated moving average (SARIMA), on monthly malaria cases at district level. In that study, the number of malaria cases in neighbouring districts and rainfall in the same district was used as a covariate to improve the predictions. The authors used the train-test approach and they defined the prediction criterion which is the mean absolute relative error (MARE) to evaluate the model's accuracy. As result in [BVG+08], the best model (without extrinsic explanatory variables) varied by district and forecasting horizon.

In another study [ALPP16], these model types are also used for malaria incidence fitting data in Afghanistan. In that work, Bayesian variable selection within a geostatistical model was employed to identify important environmental and climatic risk factors of parasitaemia. The fitting result of the models with the different values of the order parameters (p, d, q) was compared [ALPP16].

A GLM is applied in [GGK+12] where authors provide spatially explicit burden estimates of malaria in Senegal using the Senegal Malaria Indicator Survey (SMIS) data and Bayesian geostatistical Zero-Inflated Binomial models based on variable selection methods for spatial data.

A comparative work of existing models is made in [NPC21]. In these studies, a SARIMA model has been considered to work best with time series data that exhibited periodic or seasonal characteristics and was able to predict the seasonal trend of malaria. That model type is only suitable for a stationary or seasonal process. As for a NB regression model, it correctly identified the association between climate variables (taken as explanatory variables) and the rate of malaria transmission. That last model type can make good forecasts, but is not ideal for prediction in subsequent months or years.

The novelties of our work in relation to these studies are the use of the Machine Learning approach presented in Section 1.5.5, the saturation of the rainfall explanatory variable defined in Section 6.1.3 and the use of the history of the falciparum malaria incidence count as an explanatory variable defined in Section 6.4.1.

1.4 Thesis objectives

This thesis intervenes in the modelling of two infectious diseases: COVID-19 in Belgium and malaria in Senegal with the following objectives.

1. To make forecasts from COVID-19 data a few weeks or months ahead. We used data from Belgium due to the online open access of data by Sciensano and the reporting quality of these data. We also used data from France, the United Kingdom and Luxembourg for comparison analyses and model consistency analysis. All data sources are reported in Appendix A and Section A.2.
2. To propose models that produce h-days-ahead forecasts (with h a forecast horizon) and confidence intervals of the falciparum malaria incidence count per month in various locations in Senegal. To assess the accuracy of such forecasts.
3. To provide governments or public health officials with a model capable of forecasting the number of patients who would be infected at a given date with the aim of setting up an alert system for effective anticipatory decision-making, adequate and proportional to the evolution of an epidemic or pandemic.

It benefits from a collaboration between UCLouvain (Louvain-la-Neuve, Belgium) and UCAD (Dakar, Senegal).

In Section 1.5, we will present the different methods or models used in this thesis without a full description of their use.

1.5 Methodology

1.5.1 Transmission models

Let's consider a deterministic SIR model based on [KMW27, Het00], when population is divided on tree groups: $S(t)$ is the number of susceptible at time t, $I(t)$ is the number of infectives (all infected people are infectious), $R(t)$ is the number of recovered people with permanent immunity. If we do not take account the births, deaths and migration, the total population

size N is fixed. Thus, we can have a very simple model where the differential equations which describe the rate of change in the number of susceptible, infectious and immune individuals at time t are represented in the following system:

$$\begin{cases} \dot{S}(t) = -\frac{\beta}{N} S(t) I(t) \quad S(0) \geq 0 \\ \dot{I}(t) = \frac{\beta}{N} S(t) I(t) - \gamma I(t) \quad I(0) \geq 0 \\ \dot{R}(t) = \gamma I(t) \quad R(0) \geq 0 \end{cases} \tag{1.11}$$

where $N = S(t) + I(t) + R(t)$, β is the contact parameter and γ is the specific rate at which infected individuals recover from the disease. By comparison, we have the difference equations which describe the number of susceptible, infectious and immune individuals at time t [VW10]

$$\begin{cases} S(t+1) = S(t) - \frac{\beta}{N} S(t) I(t) \quad S(0) \geq 0 \\ I(t+1) = I(t) + \frac{\beta}{N} S(t) I(t) - \gamma I(t) \quad I(0) \geq 0 \, . \\ R(t+1) = R(t) + \gamma I(t) \quad R(0) \geq 0 \end{cases} \tag{1.12}$$

Because of the assumptions in the SIR model in equation (1.11), it is not possible to model an endemic situation (the rebounds or waves). To take into account the endemic situation of a disease we have to consider immigration, reproduction and reinfection, arriving to the so-called SIR endemic model [Het76, JKKPS21].

1.5.2 Generalized Linear Models

One of the most important developments in statistics in the last few years has been the introduction and widespread use of *generalized linear models* (GLMs), as formulated by Nelder and Wedderburn [Lee20], [GS93] (chapter 5) and [CT13] (Section 2.4).

A GLM respects the following steps:

- Random variable (R-v)

$$Y_t \sim f(Y_t; \mu_t), \tag{1.13}$$

  where Y_t is the explained variable with mean μ_t, f is the probability density or mass of an exponential family, and t is the observation time.

- Deterministic component (D-c)

$$\eta(X_t) = \sum_{j=1}^{k} \beta_j X_{tj}, \tag{1.14}$$

where η represents the linear predictor, X_t is the vector of explanatory variables at time t, with X_{tj} denoting the jth explanatory variable at time t, β_j is a regression coefficient and k represents the number of explanatory variables.

- Link function (L-f)

$$g(\mu_t) = \eta(X_t), \tag{1.15}$$

where g represents a transformer function and

$$\begin{aligned} \mu_t &= g^{-1}(X_t^T \beta) \\ &= g^{-1}(\beta_1 X_{t1} + \ldots + \beta_k X_{tk}). \end{aligned} \tag{1.16}$$

According to [Lin97] (Section 1.4.3), the link function g permits to relate the mean (μ) of the tth observation and its linear predictor (η). We let $X_{t1} = 1$, $t = 1, \ldots, n$ making β_1 the intercept. In the D-c step (equation (1.14)), the regression coefficients $\beta_1, \ldots, \beta_k$ are to be estimated on the train-set.

We now define the Poisson, negative binomial and Gaussian distributions, which we will consider for the density f in (1.13).

Poisson distribution

The Poisson distribution is probably the most used discrete distribution because of its simplicity [ADA20]. Often used to model the number of events occurring in a fixed period of time when the times at which events occur are independent.

Its conditional probability mass function is defined as in [ADA20] and also in [CT13, p. 71] by

$$\begin{aligned} p(Y_t; \mu_t) &= \frac{\mu_t^{Y_t} \exp(-\mu_t)}{Y_t!} \\ &= \exp[Y_t \log(\mu_t) - \mu_t - \log(Y_t!)]. \end{aligned} \tag{1.17}$$

Negative Binomial (NB) distribution

The negative binomial distribution describes a Poisson random variable whose rate parameter is gamma distributed. That's why it is called the

Poisson–Gamma mixture distribution [AMA+19].

If we have $x \sim \text{Poisson}(\theta)$ and $\theta \sim \text{gamma}(a, b)$ then their probability density functions are, respectively, defined by

$$p_1(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} \tag{1.18}$$

and

$$p_2(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}. \tag{1.19}$$

Therefore, the probability density function of this Poisson–Gamma mixture distribution is defined as follows

$$\begin{aligned}
p(x|\theta, a, b) &= \int_0^\infty p_1(x|\theta) p_2(\theta)\, \mathrm{d}\theta \\
&= \int_0^\infty \frac{\theta^x e^{-\theta}}{x!} \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}\, \mathrm{d}\theta \\
&= \frac{b^a}{x!\Gamma(a)} \int_0^\infty \theta^{x+a-1} e^{-\theta-b\theta}\, \mathrm{d}\theta \\
&= \frac{b^a \Gamma(x+a)}{x!\Gamma(a)(1+b)^{x+a}} \int_0^\infty \frac{(1+b)^{x+a}}{\Gamma(x+a)} \theta^{(x+a)-1} e^{-\theta(1+b)}\, \mathrm{d}\theta \\
&= \frac{b^a}{x!\Gamma(a)} \frac{\Gamma(x+a)}{(1+b)^{x+a}} \\
&= \frac{\Gamma(x+a)}{x!\Gamma(a)} \frac{b^a}{(1+b)^{x+a}} \\
&= \frac{\Gamma(x+a)}{x!\Gamma(a)} \frac{b^a}{(1+b)^a (1+b)^x} \qquad \text{with} \qquad \Gamma(n+1) = n! \\
&= \frac{\Gamma(x+a)}{\Gamma(x+1)\Gamma(a)} \left(\frac{b}{1+b}\right)^a \left(\frac{1}{1+b}\right)^x.
\end{aligned}$$

Let's denote by $Y_t = x$, $\alpha = 1/a$ and $\alpha\mu_t = b$. In this case, the probability density function of the negative binomial distribution is given as in [AMA+19, Lee20, Nak97] by

$$p(Y_t; \mu_t, \alpha) = \frac{\Gamma(Y_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(Y_t + 1)} \left(\frac{1}{1+\alpha\mu_t}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_t}{1+\alpha\mu_t}\right)^{Y_t}, \tag{1.20}$$

where Γ is the gamma function. Its mean is μ_t and its variance is $\mu_t + \alpha\mu_t^2$, where α is termed the dispersion parameter. Note that, if $\alpha \to 0$, the negative binomial converges to the Poisson distribution [CT13, p. 85].

Still in [CT13, p. 87], the GLM literature suggests choosing α so that the Pearson statistic or deviance statistic equals $n - k$. In [CT13, p. 90], authors suggest a advanced means to compute α using a technique they call auxiliary ordinary least squares (OLS) regression without a constant and described in capture 1.1. The principle is to fit the Poisson regression model (defined in Section 1.5.2) on the dataset to get the fitted values denoted by λ. Then, fitting the OLS regression model on the dataset with these λ values to get the estimate α.

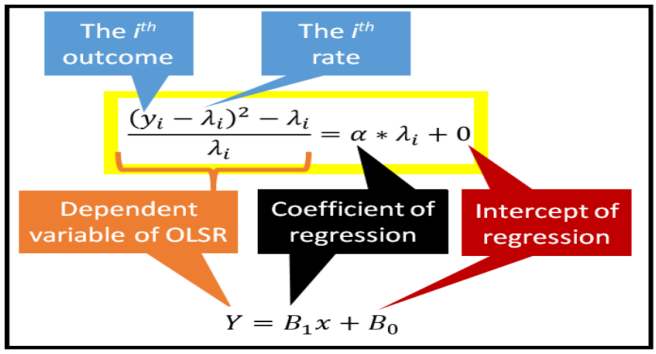


Fig. 1.1 Estimating the dispersion parameter α for negative binomial regression model by ordinary least squares (OLS) method [Dat20].

Normal distribution

Its probability density function is defined as in [Wik23] by

$$p(Y_t \mid \mu_t, \tau) = \sqrt{\frac{\tau}{2\pi}} \exp\left\{-\frac{\tau}{2}(Y_t - \mu_t)^2\right\}, \tag{1.21}$$

where μ defined in equation (1.16).

According to the `PyMC3` Python package, the normal distribution can be parameterized either in terms of precision (τ) or standard deviation (σ). The link between the two parametrizations is given by $\tau = \frac{1}{\sigma^2}$.

1.5.3 Optimization methods

Maximum likelihood estimator (MLE)

According to [RC04, p. 6], the method maximum likelihood estimation is quite a popular technique for deriving estimators. Starting from an iid sample $x = (x_1, \ldots, x_n)$ from a population with density $p(x|\theta_1, \ldots, \theta_k)$, the *likelihood* function is

$$\begin{aligned} L(\theta|x) &= L(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) \\ &= \prod_{i=1}^{n} p(x_i|\theta_1, \ldots, \theta_k). \end{aligned} \tag{1.22}$$

More generally, when the x_i's are not iid, the likelihood is defined as the joint density $p(x_1, \ldots, x_n|\theta)$ taken as a function of θ. The value of θ, denoted by $\hat{\theta}$, which is the parameter value at which $L(\theta|x)$ attains its maximum as a function of θ, with x held fixed, is known as a *maximum likelihood estimator*.

Newton-Raphson algorithm

When the goal is to solve an equation of the form $f(\theta) = 0$, a common approach is to use a *Newton-Raphson algorithm* [RC04, p. 19] presented in Algorithm 1, which produces a sequence x_n such that

$$\theta_{n+1} = \theta_n - (\frac{\partial f}{\partial \theta}|_{\theta=\theta_n})^{-1} f(\theta_n) \tag{1.23}$$

until it stabilizes around a solution of $f(\theta) = 0$. Optimization problems associated with smooth functions F are then based on this technique, using the equation $\nabla F(\theta) = 0$, where ∇F denotes the *gradient* (or *score function*) of F, that is, the vector of derivatives of F. The corresponding techniques are then *gradient methods*, where the sequence θ_n is such that

$$\theta_{n+1} = \theta_n - (\nabla \nabla^t F)^{-1}(\theta_n) \nabla F(\theta_n), \tag{1.24}$$

where $\nabla \nabla^t F$ denotes the matrix of second derivatives of F.

Least squares estimator (LSE)

According to [con22], the objective consists of adjusting the parameters of a model function to best fit a data set. A simple data set consists of n points (data pairs) (X_t, Y_t), $t = 1, \ldots, n$, where X_t is an independent variable and

Algorithm 1: Newton-Raphson [Cle19]

1 Choose initial parameter estimate $\theta_p = \theta_0$;
2 Calculate score $\nabla F(\theta) \mid_{\theta=\theta_p}$;
3 Calculate derivative of the function for which you want to calculate the roots;
4 Walk along first derivative until line (plane) of the derivative crosses zero;
5 Update the betas θ_{p+1};
6 Iterate from step 2 - 5 until convergence.

Y_t is a dependent variable whose value is found by observation. The model function has the form $f(X, \beta)$, where k adjustable parameters are held in the vector β. The goal is to find the parameter values for the model that "best" fits the data. The fit of a model to a data point is measured by its errors and residuals in statistics residual, defined as the difference between the observed value of the dependent variable and the value predicted by the model: $r_t = Y_t - f(X_t, \beta)$.

The least-squares method finds the optimal parameter values by minimizing the sum of squared residuals noted by S and defined by

$$S = \sum_{t=1}^{n} r_t^2. \tag{1.25}$$

1.5.4 Time series models: seasonal autoregressive integrated moving average with exogenous variables (SARIMAX)

The previous sections focus on examples of GLMs. This section presents the SARIMAX model. This model will be used as an alternative model to these GLMs for malaria incidence forecasting in chapter 7. The purpose of using a time series model is that it generates autocorrelated data sequences [CT13, p. 264]. Therefore, the history of malaria data will not be considered as an explanatory variable.

This model, whose concept was first developed by Box-Jenkins [LAA17], is obtained by the following process, according to [doc] and [Cha16] (chapter 3: Some Time Series Models):

$$(\mathrm{AR} + \mathrm{MA}) \rightarrow \mathrm{ARMA} \rightarrow \mathrm{ARIMA} \rightarrow \begin{cases} \mathrm{SARIMA} \\ \mathrm{ARIMAX} \end{cases} \rightarrow \mathrm{SARIMAX}.$$

For k exogenous variables defined at each time step t, denoted by $X_t^{[j]}$ for $j \leq k$, with coefficients β_j, the ARIMAX(p, d, q) model is defined by

$$\Theta(L)^p \Delta^d Y_t = \Phi(L)^q \Delta^d \epsilon_t + \sum_{j=1}^{k} \beta_j X_t^j, \tag{1.26}$$

and the SARIMAX$(p, d, q)(P, D, Q, s)$ model by

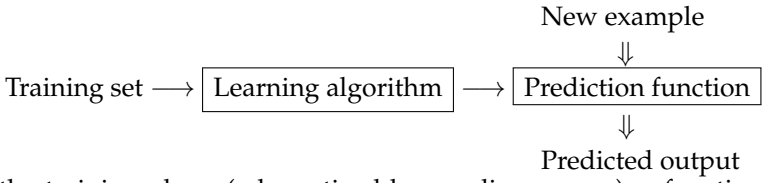
$$\Theta(L)^p \theta(L^s)^P \Delta^d \Delta_s^D Y_t = \Phi(L)^q \phi(L^s)^Q \Delta^d \Delta_s^D \epsilon_t + \sum_{j=1}^{k} \beta_j X_t^j. \tag{1.27}$$

It can be denoted by ARIMA$(p, d, q)(P, D, Q, s)$ [LAA17].

For non-seasonal data, p, d and q, respectively, represent the order of the AR term (i.e, the number of lag observations the model will use), the order of the MA term (i.e the number of times that the raw observations are differenced till stationarity) and the number of differencing required to make the time series stationary. While for seasonal data, we need to add also the terms P, D, Q and s, respectively, corresponding to the number of seasonal lag observations the model will use, the number of times that the seasonal observations are differenced till stationarity, the size of the seasonal moving average window and the seasonal order meaning the number of observations of one season.

1.5.5 Machine Learning principle

Machine Learning concerns the study and development of quantitative models that allow a computer to perform tasks without it being explicitly programmed to do them [MR20].



In the training phase (schematized by one-line arrows), a function minimizing empirical error on a training set is found among a class of predefined functions. In the test phase (schematized by two-line arrows), the outputs of new examples are predicted by the prediction function.

1.5.6 The relevant choice of models for each disease

For COVID-19 modelling we used an SIR model because of its adaptability to model the transmission of this disease and the data availability for each of the compartments S: Susceptible, I: Infectious and R: Recovered. We have not used GLMs with this disease as the external factors influencing its transmission are not clearly identified. For example, in contrast with malaria a climatic influence on the development of the virus (SARS-CoV-2) responsible for its transmission is not fully proved. There is a lack of well-identified explanatory variables for COVID-19, unlike malaria. The emergence of variants and vaccination campaigns could be explanatory variables, but their rather binary nature lends itself poorly to be taken into account in a GLM with an explanatory variable of the count type (number of infected or hospitalized persons).

As for malaria modelling, we did not use an SIR type model due to the need to model the population of the vector (mosquito) because these entomological data were not available. Instead, we opted for GLMs and the SARIMAX time series as alternative model. They allow for explanatory variables, a crucial asset since pluviometry is known to have a major influence on the malaria incidence. The meteorological explanatory variables such as the rainfall, the average temperature and the relative humidity are known to influence the mosquitoes (Anopheles) ecology by affecting its distribution, seasonality, and transmission intensity [LAA17, OG17]. Moreover, in contrast with an ordinary linear regression, GLMs allow the explained variable to follow an arbitrary distribution. In particular, GLMs subsume Poisson regression, which is a popular choice to model disease incidence expressed as count.

1.6 Main contributions

We propose a method to learn the parameters and initial conditions of a basic two-compartment epidemiological model. We show that, on the Belgian data, when trained during a two-week period around the first peak, this basic two-compartment model is able to predict the subsequent decrease over several months with a remarkably high accuracy (Section 3.4.2). Unsurprisingly, predicting the start of the subsequent waves is much more challenging. Perhaps more surprisingly, the prediction quality is signifi-

cantly lower when the model is trained around the other peaks. We also investigate alternative methods such as fitting a polynomial or an exponential, which give less accurate forecasts than the proposed method.

As for malaria, we propose a method to forecast the falciparum malaria incidence count a few months ahead taking into account explanatory variables such as rainfall, average temperature, relative humidity and history of falciparum malaria incidence count per month, where each of them is taken from a well determined lag (Section 6.2.1), in Dakar, Fatick and Kedougou. We show that removing the history of falciparum malaria incidence count per month from explanatory variables (Section 6.4.4) has a strong adverse effect on the forecast accuracy. We develop algorithms (Sections 6.1.2, and 7.1) through a machine learning approach based on a separation of the data into a train-set to estimate the parameters and a test-set to assess the forecast accuracy. In order to remedy the overestimations observed during the forecasts, we develop a saturation method (Section 6.1.3) where an increasing of the forecast quality in Dakar and Kedougou is observed.

1.7 Contents of the thesis

Chapter 2: Data and indicators of measures

In this chapter, we present the datasets used in this thesis. We had the COVID-19 data in Belgium, but also in France, Luxembourg and the United Kingdom. We also had epidemiological data of malaria disease in three regions of Senegal and meteorological data such as rainfall, temperature and relative humidity in these regions. We present the discussion of these data.

Chapter 3: Assessment of COVID-19 hospitalization forecasts from a simplified SIR model

This chapter contains the formulation of the SH model and describes several methods to estimate the parameters and initial conditions. It also presents all experimental results.

Chapter 4: Alternative models

In this chapter, we use fatalities and patients discharged from the hospital data in Belgium to propose an estimation of parameters such as the proportion p, and γ, $\bar{N}$ and $\bar{\beta}$, defined in chapter 3. We also present alternative models to the SH model such as the polynomial model and the exponential model, and the results associated.

Chapter 5: Producing confidence intervals

In this chapter, we present three methods of producing confidence intervals for the SH model. We have a stochastic uncertainty method, a parameter uncertainty method and a model uncertainty method. We also present the results of each method in local and global test with the data from Belgium and France.

Chapter 6: Generalized linear models to forecast malaria incidence in three endemic regions of Senegal

In this chapter, we describe the models by presenting the estimation and the forecasting methods. We also present the saturation method. Finally, we present the experimental results and the discussion of these results.

Chapter 7: Alternative forecasting methods of malaria incidence

In this chapter, we investigate the SARIMAX model and a second forecasting method by considering the GLM with Poisson distribution. We also study a generalized parameter method for the three regions and we present their results associated.

Chapter 8: Producing confidence intervals for GLM

In this chapter, we use methods: the stochastic uncertainty and the parameter uncertainty, defined in chapter 5, and a GLM default method to produce confidence intervals for GLMs.

Chapter 9: Mosquito density forecasting based on weather factors

In this chapter, we present data collected by the Pasteur institute of Senegal such as the mosquito density, and the climatic variables: rainfall, temperature and relative humidity in four village in Senegal. We apply the GLM with Gaussian distribution and log as the link function to forecast the mosquito density based on this above cited climatic variables.

2

Data and indicators of measures

THIS chapter is organized as follows. We present the data such as the daily reported cases, the number of hospitalized patients at the moment of reporting (prevalence), the number of new hospital intakes in the last 24h (incidence), the number of new hospital discharges (alive) in the last 24h (incidence), the number of deaths, related to COVID-19 pandemic in Section 2.1 and we discuss them in Section 2.1.2. The data related to malaria epidemic (the monthly malaria cases, the ITNs distributed, the ACT distributed, the rainfall, the temperature, the humidity) are presented in Section 2.2 and discussed in Section 2.2.2. These data are needed to validate the proposed models. The accuracy measures are defined in Section 2.3.

2.1 COVID-19 data

2.1.1 Data origin

The COVID-19 dataset (Figure 2.1) for Belgium and the data related to COVID-19 death for Belgium are provided from t_s to t_e, where t_s is 2020-03-15 and t_e is 2020-07-15. The data related to the COVID-19 disease are

presented in Figure 2.3. We will also mention results obtained with the datasets (Figure 2.2) for France where t_s is 2020-03-18 and t_e is 2020-07-17, for Luxembourg where t_s is 2020-03-19 and t_e is 2020-06-26 and for the United Kingdom where t_s is 2020-03-27 and t_e is 2020-07-09. All datasets URL links are presented in Appendix A.

	DATE	PROVINCE	REGION	NR_REPOR	TOTAL_IN	TOTAL_IN	TOTAL_IN	TOTAL_IN	NEW_IN	NEW_OUT
2	3/15/2020	Antwerper	Flanders	14	50	9	4	0	8	8
3	3/15/2020	Brussels	Brussels	14	58	11	8	0	7	2
4	3/15/2020	Hainaut	Wallonia	15	56	13	11	1	26	1
5	3/15/2020	Limburg	Flanders	7	20	6	3	0	9	3
6	3/15/2020	LiÃ¨ge	Wallonia	12	22	2	1	0	4	1
7	3/15/2020	Luxembou	Wallonia	3	4	0	0	0	3	0
8	3/15/2020	Namur	Wallonia	6	2	1	1	0	0	0
9	3/15/2020	OostVlaan	Flanders	14	16	5	1	0	5	1
10	3/15/2020	VlaamsBra	Flanders	6	14	2	0	0	2	0
11	3/15/2020	BrabantW	Wallonia	2	5	2	2	0	1	0
12	3/15/2020	WestVlaar	Flanders	11	19	3	1	0	6	2

Fig. 2.1 Data base structure (Excel screenshot) for Belgium.

	dep	sexe	jour	hosp	rea	rad	dc
2	1	0	3/18/2020	2	0	1	0
3	1	1	3/18/2020	1	0	1	0
4	1	2	3/18/2020	1	0	0	0
5	2	0	3/18/2020	41	10	18	11
6	2	1	3/18/2020	19	4	11	6
7	2	2	3/18/2020	22	6	7	5
8	3	0	3/18/2020	4	0	1	0
9	3	1	3/18/2020	1	0	0	0
10	3	2	3/18/2020	3	0	1	0
11	4	0	3/18/2020	3	1	2	0
12	4	1	3/18/2020	3	1	0	0
13	4	2	3/18/2020	0	0	2	0
14	5	0	3/18/2020	8	1	9	0
15	5	1	3/18/2020	1	0	7	0
16	5	2	3/18/2020	7	1	2	0
17	6	0	3/18/2020	25	1	47	2
18	6	1	3/18/2020	15	1	19	0
19	6	2	3/18/2020	10	0	28	2

Fig. 2.2 Data base structure (Excel screenshot) for France.

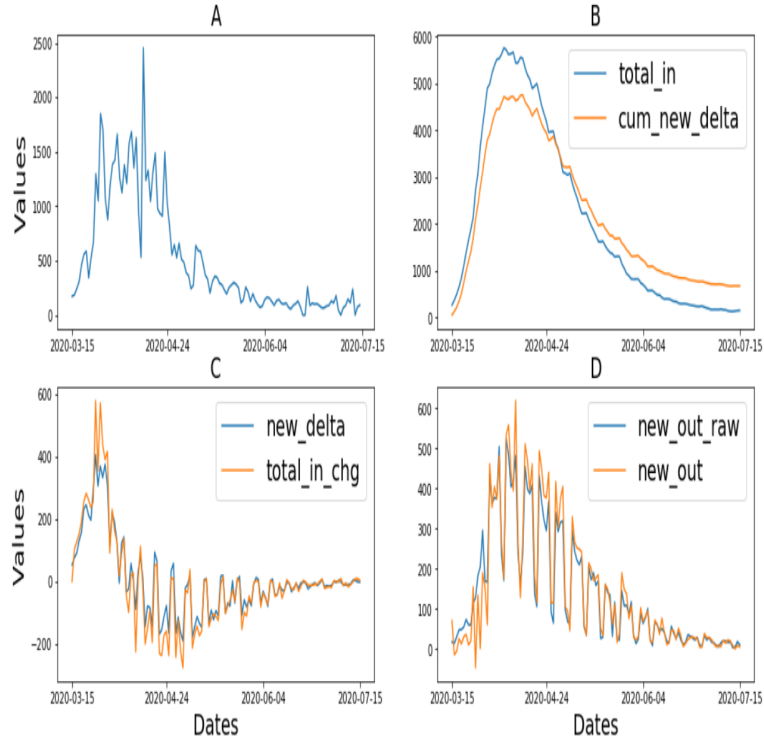


Fig. 2.3 Evolution of the COVID-19 epidemiological data from 2020-03-15 to 2020-07-15 in Belgium. We show the daily reported cases in (A), the number of hospitalized patients at the moment of reporting (prevalence) in (B), the number of new hospital intakes in the last 24h (incidence) in (C) and the number of new hospital discharges (alive) in the last 24h (incidence) in (D). We denote by cum_new_delta the cumulative of new_delta, total_in_chg the difference between the hospitalized a time t and the hospitalized at time $t-1$, new_out the difference between new_in and total_in_chg.

2.1.2 Discussion

In the data for Belgium (Figure 2.3), there is a mismatch between $H_o(t)$ and $H_o(t-1) + E_o(t) - L_o(t)$ for most t, and $H_o(t_s) + \sum_{t=t_s+1}^{t_e} E_o(t) - L_o(t)$ is significantly larger than $H_o(t_e)$. This can be due to the patients who get infected at the hospital (they would be counted in H_o without appearing

in E_o) and to the patients who die at the hospital (they would be removed from H_o without appearing in L_o). In order to remedy this mismatch, we redefine $L_o(t)$ by $L_o(t) := -H_o(t) + H_o(t-1) + E_o(t)$.

For the French data, we sum the "rad" (daily number of new home returns) and "dc" (daily number of newly deceased persons) columns to get $L_o(t)$. Since there is no column for E_o, we define $E_o(t) = H_o(t) - H_o(t-1) + L_o(t)$.

Several other COVID-19-related data are also available for Belgium and France. In particular, the daily number of infected individuals, $I_o(t)$, is also reported by health authorities. However, a visual inspection reveals that the graph of I_o (in Figure 2.3 A) looks less smooth than the graph of H_o (in Figure 2.3 B). Moreover, for the authorities, predicting H is more crucial than predicting I. Therefore, as in [Koz20], we focus on H.

2.2 Malaria data

2.2.1 Data origin

We conducted this study because Senegal is a large laboratory of data related to malaria disease. Malaria has been endemic there for many years with a high weather variability between regions due to the presence of the ocean in the West country part, but also the variation of vegetation between regions, from West (Dakar) and South-East (Kegoudou).

There are two principal malaria transmission zones in Senegal:

- *Faciès tropical*: corresponding the regions of Ziguinchor, Kolda, Tambacounda and Kedougou. In that zone, the raining season is the longest and most intensive in the country and covers 5 to 6 months. Malaria cases are observed between 4 to 6 months and the transmission is high (20 to 100 infected bites/ human/ year).

- *Faciès sahélien*: corresponding the regions such as Kaolack, Fatick, Diourbel, Dakar, Thies, Louga, Saint-Louis and Matam with a less intensive rainy season and covers 2 to 3 months. The transmission is very low in general (0 to 20 infected bites/ human/ year).

The historical data, such as the monthly malaria cases in the big sanitary districts in Dakar, Fatick and Kedougou (three regions of Senegal) and

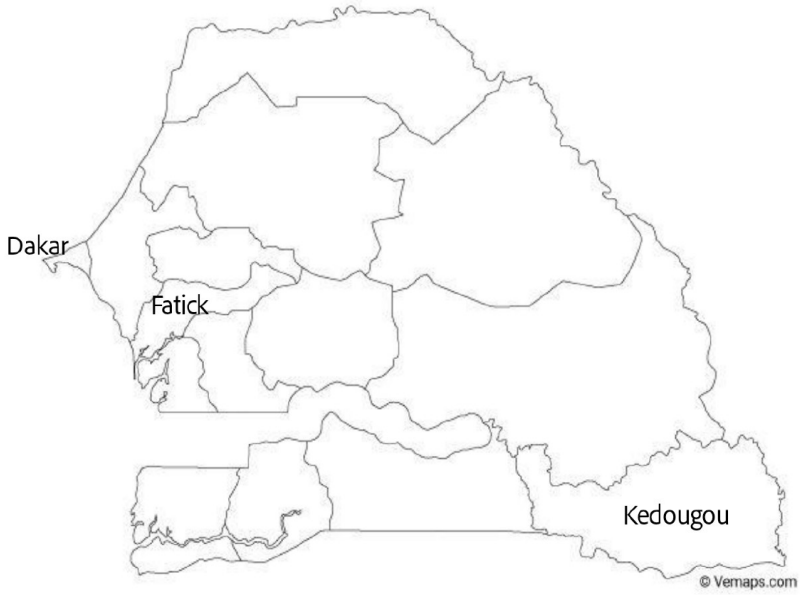


Fig. 2.4 Location of Dakar, Fatick and Kedougou, three regions in Senegal (western Africa). The choice of these three regions is motivated by data availability (notably the presence of villages where longitudinal studies have been conducted), but also by geographical differences: influence of the ocean in the Dakar peninsula, tropical climate in Kedougou, and savanna landscape in Fatick. These fundamental geographical differences allow us to test the applicability of GLMs under drastically different evolutions of the climate variables. Senegal is bordered by the Atlantic Ocean to the west, Mauritania to the north, Mali to the east, and Guinea and Guinea-Bissau to the south.

located in Figure 2.4, the (ITNs) distributed and the (ACT) distributed, between 2008 and 2016, come from the PNLP of Senegal (see Appendix A).

These data are registered in Excel folders where each of them contains the following informations: “fiche récapitulative de la situation des malades vus en consultation externe", “Fiche récapitulative des données du TPI", “Fiche récapitulative des MILDA distribuées", “Fiche récapitulative traitements ACT dispensés", “Fiche récapitulative de la situation des malades hospitalisés" and “Fiche récapitulative des patients décédés en hospitalisation".

We consider as malaria cases, the cumulative number of confirmed tests by RDTs during the month, in all individual groups [dLclP16, FCG+18]. For data by region, we grouped the data from all the different big sanitary districts.

The hourly meteorological data, such as the temperature, the relative humidity and the rainfall, between 2008 and 2016, come from meteoblue (see Appendix A). To obtain adequate meteorological data for our study (monthly time unit) we add up all the measured values of the month for rainfall, we calculate the mean of all the measured values of the month for temperature and humidity.

The districts in the three regions are

- Dakar: Centre, Ouest, Sud, Nord, Mbao and Pikine,
- Fatick: Foundiougne, Sokone, Passy, Diofior, Gossas and Fatick city,
- Kedougou: Kedougou city and Saraya.

Within each region, other districts different from those mentioned may exist, but we noticed some missing observations from 2008 to 2016.

Response variable

In this study, the response variable (or explained variable) is the falciparum malaria incidence count per month ($Y_o(t)$).

Independent variables

The available explanatory variables of this study are the malaria cases count per month in the past ($Y_o(t-\delta)$), the rainfall measured in millimetres ($Rf_o(t-\delta)$, mm), the average temperature measured in degree Celsius ($Tp_o(t-\delta)$, °C), the relative humidity measured in percentages ($RH_o(t-\delta)$, %), the number of insecticide treated bednets distributed per month ($B_o(t-\delta)$), and the number of antimalarial drugs distributed per month ($A_o(t-\delta)$) where we consider $\delta = h, h+1, \ldots, 6$, ($h$ represents a forecast horizon) and an artificial vector that we call intercept vector I equal to 1 for all t. These notations are explained in section called notations.

2.2.2 Discussion

We are interested in the meteorological explanatory variables such as the rainfall, the average temperature and the relative humidity because they are known to influence the mosquitoes (*Anopheles*) ecology by affecting its distribution, seasonality, and transmission intensity [LAA17, OG17]. For example, the temperature influences the sporogonic development duration of the parasite and many parameters related to the mosquitoes such as the biting rate, the egg deposition rate and the death rate of immature and adult [OG17]. The rainfall influences the availability and the quality of the larval breeding grounds [NHE+01] and the maturation of immature mosquitoes [OG17].

As for the bed-net and the drugs distributed, we took them because we suppose that they constitute the main factors fighting against malaria disease [FCG+18], reducing its morbidity and its mortality. Needless to say, the number of bed-nets actually used would be a more suitable explanatory variable, but this data is not available. Note that, in contrast with the physical explanatory variables (Rf_{o}, Tp_{o}, and RH_{o}), the human explanatory variables (B_{o} and A_{o}) may depend on the incidence in the past, and they may also be influenced by models used by the health authorities. For this reason, in most of our experiments, we only consider the physical explanatory variables.

Our experiments use the epidemiological data from PNLP and the meteorological data from meteoblue from 2008-01-31 to 2016-12-31, in Dakar, Fatick and Kedougou. In Figure 2.5–2.13, we present the plots of the variables in order to illustrate their annual distribution.

2.3 Accuracy measures

Since plotting all the results is out of the question, we have to return measures that give an account of the accuracy of the predictions. We denote by Z the variable to measure. Among the many possibilities, we will favor the measures defined as follows for an arbitrary period from t_{a} to t_{b}:

- Root mean square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=t_{\mathrm{a}}}^{t_{\mathrm{b}}} (Z_{\mathrm{o}}(t) - Z(t))^2}{t_{\mathrm{b}} - t_{\mathrm{a}} + 1}}. \tag{2.1}$$

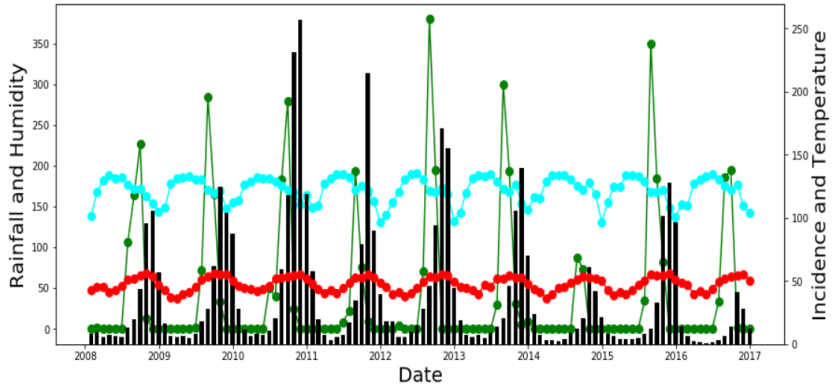


Fig. 2.5 Evolution of the monthly falciparum malaria incidence count (black), the rainfall (mm, green), the relative humidity (%, cyan) and the average temperature (°C, red) in Dakar. We make the following transformations in order to have a better display: $Y_{\mathrm{o}}(t)/\min\{Y_{\mathrm{o}}\}$, $Tp_{\mathrm{o}}(t) * 2$ and $RH_{\mathrm{o}}(t) * 2$. Hence be mindful that the curves for humidity, temperature and falciparum malaria incidence count are scaled.

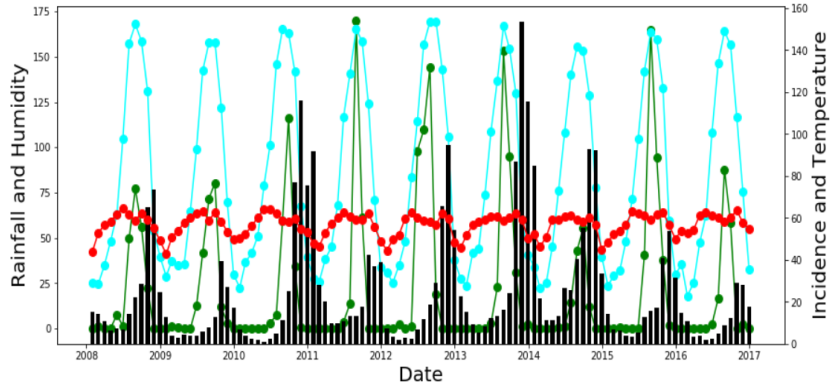


Fig. 2.6 Evolution of the monthly falciparum malaria incidence count (black) the rainfall (mm, green), the relative humidity (%, cyan) and the average temperature (°C, red) in Fatick. We make the following transformations in order to have a better display: $Y_{\mathrm{o}}(t)/\min\{Y_{\mathrm{o}}\}$, $Tp_{\mathrm{o}}(t) * 2$ and $RH_{\mathrm{o}}(t) * 2$. Hence be mindful that the curves for humidity, temperature and falciparum malaria incidence count are scaled.

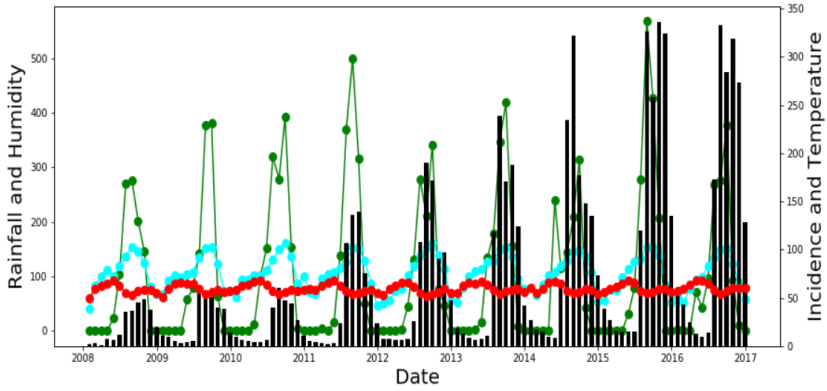


Fig. 2.7 Evolution of the monthly falciparum malaria incidence count (black) the rainfall (mm, green), the relative humidity (%, cyan) and the average temperature (°C, red) in Kedougou. We make the following transformations in order to have a better display: $Y_o(t)/\min\{Y_o\}$, $Tp_o(t) * 2$ and $RH_o(t) * 2$. Hence be mindful that the curves for humidity, temperature and falciparum malaria incidence count are scaled.

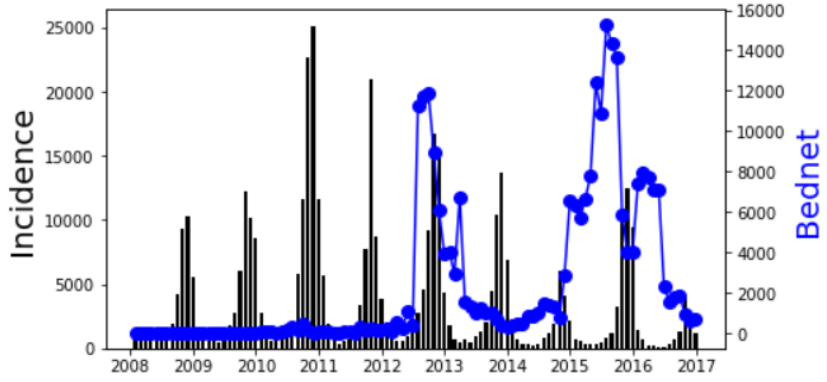


Fig. 2.8 Falciparum malaria incidence count per month and the number of insecticide treated bednets distributed per month in Dakar.

- Root relative squared error (RRSE):

$$\text{RRSE} = \sqrt{\frac{\sum_{t=t_a}^{t_b}(Z_o(t) - Z(t))^2}{\sum_{t=t_a}^{t_b}(Z_o(t) - \text{mean}(Z_o))^2}} \tag{2.2}$$

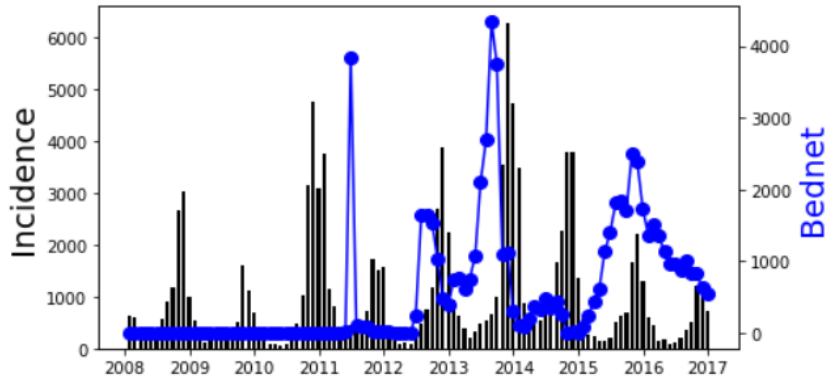


Fig. 2.9 Falciparum malaria incidence count per month and the number of insecticide treated bednets distributed per month in Fatick.

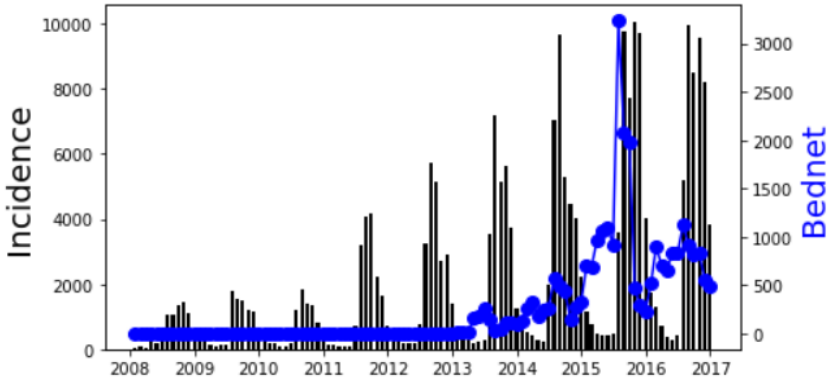


Fig. 2.10 Falciparum malaria incidence count per month and the number of insecticide treated bednets distributed per month in Kedougou.

where $\text{mean}(Z_o) = \frac{1}{t_b - t_a + 1} \sum_{t=t_a}^{t_b} Z_o(t)$. This measure can be interpreted as the ratio between the RMSE of Z and the RMSE of $\text{mean}(Z_o)$.

- Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{t_b - t_a + 1} \sum_{t=t_a}^{t_b} |Z_o(t) - Z(t)| . \tag{2.3}$$

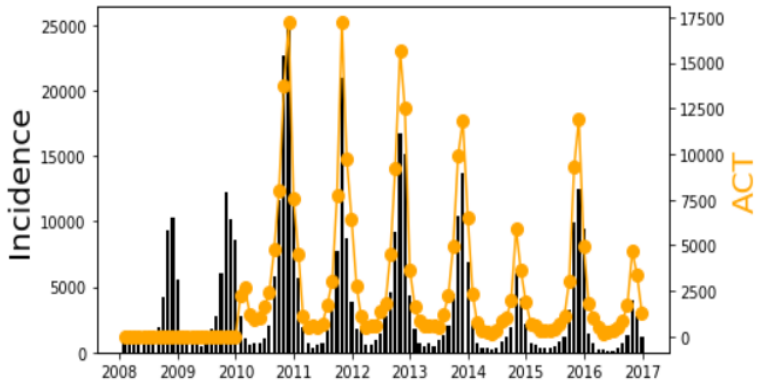


Fig. 2.11 Falciparum malaria incidence count per month and the number of ACT distributed per month in Dakar.

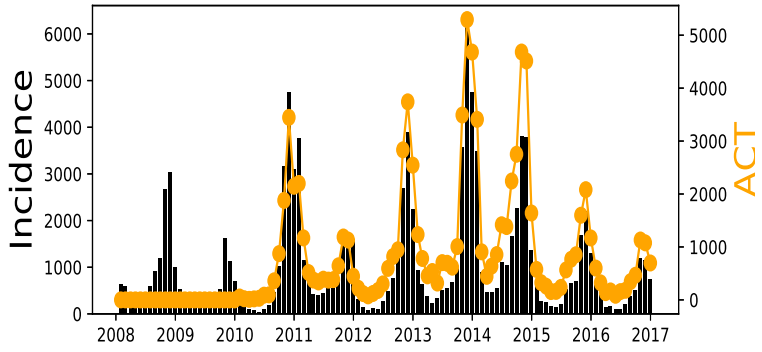


Fig. 2.12 Falciparum malaria incidence count per month and the number of ACT distributed per month in Fatick.

The MAE is considered to have a better interpretability than the RMSE. However, for comparisons between datasets, scale-invariant versions, given next, should be favored.

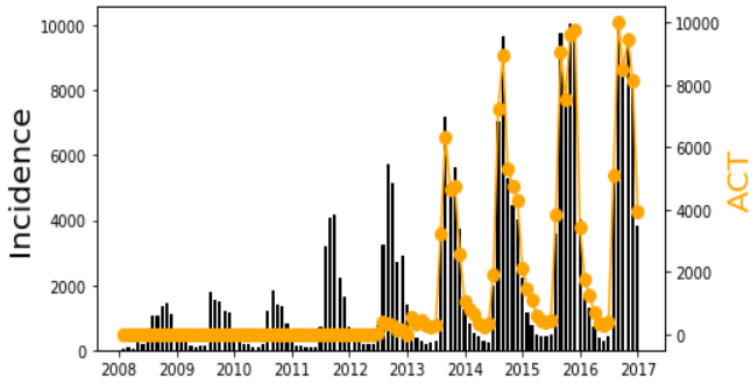


Fig. 2.13 Falciparum malaria incidence count per month and the number of ACT distributed per month in Kedougou.

- Mean absolute scaled error (MASE):

$$\mathrm{MASE} = \frac{\mathrm{MAE}}{\frac{1}{t_\mathrm{b}-t_\mathrm{a}}\sum_{t=t_\mathrm{a}}^{t_\mathrm{b}-1}|Z_\mathrm{o}(t+1)-Z_\mathrm{o}(t)|}. \tag{2.4}$$

  This measure is inspired from [HK06]. It consists of the ratio between the MAE and the mean monthly variation of the observed values. A MASE value around 1 or below indicates an excellent accuracy.

- Mean absolute percentage error (MAPE):

$$\mathrm{MAPE} = \frac{1}{t_\mathrm{b}-t_\mathrm{a}+1}\sum_{t=t_\mathrm{a}}^{t_\mathrm{b}}\frac{|Z_\mathrm{o}(t)-Z(t)|}{|Z_\mathrm{o}(t)|}. \tag{2.5}$$

  While easy to interpret, this measure may be considered to give an undue importance to errors on small values. Moreover, in the case, encountered in practice, where $Z_\mathrm{o}(t)=0$ for some t, it is undefined.

- Symmetric mean absolute percentage error (sMAPE):

$$\mathrm{sMAPE} = \frac{1}{t_\mathrm{b}-t_\mathrm{a}+1}\sum_{t=t_\mathrm{a}}^{t_\mathrm{b}}\frac{|Z_\mathrm{o}(t)-Z(t)|}{(|Z_\mathrm{o}(t)|+|Z(t)|)/2}. \tag{2.6}$$

- Scatter index (SI) [SMNM20]:

$$\mathrm{SI} = \frac{\sqrt{(1/n)\sum_{t=t_\mathrm{a}}^{t_\mathrm{b}}((Z_t - \bar{Z}) - (Z_\mathrm{o}(t) - \bar{Z}_\mathrm{o}))^2}}{\sqrt{(1/n)\sum_{t=t_\mathrm{a}}^{t_\mathrm{b}} Z_\mathrm{o}(t)}}. \tag{2.7}$$

- Reliability analysis [SMNM20]:
  It is a statistical method for measuring the overall consistency of a model by determining if this suggested model achieves a permissible level of performance.

$$\mathrm{RA} = (\frac{100\%}{t_\mathrm{b} - t_\mathrm{a} + 1}) \sum_{t=t_\mathrm{a}}^{t_\mathrm{b}} k(t), \tag{2.8}$$

  where k's are determined through two steps. First, the relative average error (RAE) is defined as a vector whose tth component is

$$\mathrm{RAE(t)} = |\frac{Z_\mathrm{o}(t) - Z(t)}{Z_\mathrm{o}(t)}|. \tag{2.9}$$

  Next, if $\mathrm{RAE(t)} \leq \Delta$, then $k(t) = 1$, otherwise $k(t) = 0$, where Δ is a threshold value that is 0.2 (20%) based on Chinese standards.

2.4 Partial conclusion

In this chapter, we presented all data related to COVID-19 and malaria diseases. We used Belgium data because of open access and the quality of reporting. As for malaria data, we considered the data in Dakar (hypoendemic zone), Fatick (endemic zone) and Kedougou (hyperendemic zone) because of their availability, the variability of epidemiological variables and climatic variables, and the diversity of malaria transmission zones between these regions.

PART I
COVID-19 modelling

3

Assessment of COVID-19 hospitalization forecasts from a simplified SIR model

THIS chapter is adapted from a published article [ADD21]. We further add experiments with data from Luxembourg and the United Kingdom.

The goal of this chapter is to use a simplified SH model for COVID-19 hospitalization forecasts.

The progression of an epidemic such as COVID-19 can be modeled by rate processes

$$\text{Susceptible} \xrightarrow{\frac{\beta SI}{N}} \text{Infectious} \xrightarrow{\gamma I} \text{Recovered}$$

Letting $S(t)$, $I(t)$, and $R(t)$ denote the number of susceptible, infectious and removed (or recovered) individuals at time t, and letting $\dot{S}(t)$, $\dot{I}(t)$, and $\dot{R}(t)$ denote their time derivatives, the SIR model consists in the following

three-dimensional continuous-time autonomous dynamical system

$$\dot{S}(t) = -\frac{\beta}{N} S(t) I(t) \tag{3.1a}$$

$$\dot{I}(t) = \frac{\beta}{N} S(t) I(t) - \gamma I(t) \tag{3.1b}$$

$$\dot{R}(t) = \gamma I(t), \tag{3.1c}$$

where $N = S(t) + I(t) + R(t)$ is the constant total population and β and γ are parameters. The SIR model, and several (sometimes deep) variations thereof, have been applied in several works to model the COVID-19 dynamics (see, e.g.,[LGWR20, Atk20, FP20, CFP20, Koz20, Nes20b, Nes20a]) with known limitations [RVHL20, BFG+20, WF20, CGG20]. In [BD20], an SIR-like model is used to make long-term forecasts. However, at the time of writing this paper, it appears that studies are still rare (see [SM20, GC20, GBAS20]) where the SIR-like model parameters and initial conditions are learned on a "train" part of the available data in order to predict a subsequent "test" part, making it possible to assess the forecast accuracy of the model. We adapt the SIR model to the situation where (i) S, I and R are hidden variables but $I(t)$ is observed through a "proxy" $H(t) = \alpha I(t)$, where α is unknown but constant, and (ii) not only β and γ but also the total concerned population N are unknown. In the context of the COVID-19 application, the proxy H will be the total number of lab-confirmed hospitalized patients; the equation $H(t) = \alpha I(t)$ thus posits that a constant fraction of the infected people is hospitalized. The proposed adapted SIR model, which we term *SH model*, is given in equation (3.8). It has two state variables ($\bar{S}$—a scaled version S—and H) and two parameters ($\bar{\beta}$—which lumps together the parameters β, N, and α—and γ).

We leverage the proposed SH model as follows in order to make hospitalization forecasts. Given observed values $(H_o(t))_{t=t_i,\dots,t_c}$, we estimate the parameters $\bar{\beta}$, γ, and the initial conditions $\bar{S}(t_i)$ and $H(t_i)$ of the SH model. Then we simulate the SH model in order to predict $(H(t))_{t=t_c+1,\dots,t_f}$ for a specified final prediction time t_f. This approach thus relates to the areas of parameter estimation (for obvious reasons), data assimilation (for the generation of the initial conditions) and machine learning (for the train-test approach).

The chapter is organized as follows. The SH model is derived in Section 3.1. Several methods to estimate the parameters and initial conditions are presented in Sections 3.2 and 3.3. Experimental results are reported in

Section 3.4 and conclusions are drawn in Section 3.5.

3.1 Models

3.1.1 Case hospitalization ratio

We assume that, for all t,

$$H(t) = \alpha I(t) \tag{3.2}$$

where α is unknown but constant over time. As already mentioned, equation (3.2) posits that a constant fraction of the infected people is hospitalized.

Equation (3.2) is reminiscent of [CFP20, (3)], where the number of dead individuals plays the role of H and α is time dependent.

The relation $H(t) = \alpha I(t - \Delta)$, where Δ is a delay, might be considered more realistic than equation (3.2). We point out anticipatively that, in this case, we can replace equation (3.6) by $\bar{S}(t) := \alpha S(t - \Delta)$, and we still get equation (3.8). It follows that the forecasts produced by the forthcoming model are independent of Δ, and we can thus restrict to $\Delta = 0$, i.e., equation (3.2), without loss of generality.

3.1.2 Observation models

We assume the following observation models with additive noise:

$$H_{\mathrm{o}}(t) = H(t) + \epsilon_H(t) \tag{3.3a}$$
$$E_{\mathrm{o}}(t) = E(t) + \epsilon_E(t) \tag{3.3b}$$
$$L_{\mathrm{o}}(t) = L(t) + \epsilon_L(t). \tag{3.3c}$$

The variables with the subscript $_{\mathrm{o}}$ stand for the values provided in the datasets; see chapter 2 and Section 2.1. The variables without the subscript stand for the values given by the forthcoming model. The ϵ variables account for the discrepancies between the former and the latter.

We do not make any explicit statistical assumption on the ϵ variables in this work. However, we point out that if we (questionably) assume them to be independent Gaussian centered random variables, then some subsequent estimators admit a maximum likelihood interpretation.

3.1.3 Proposed SH model

Multiplying equation (3.1a) and equation (3.1b) by α, and further multiplying the numerators and denominators by α, we obtain the equivalent system

$$\alpha\dot{S}(t) = -\frac{\beta}{N\alpha}\,\alpha S(t)\,\alpha I(t) \tag{3.4}$$

$$\alpha\dot{I}(t) = \frac{\beta}{N\alpha}\,\alpha S(t)\,\alpha I(t) - \gamma\alpha I(t). \tag{3.5}$$

Letting

$$\bar{S} := \alpha S \tag{3.6}$$

$$\bar{\beta} := \frac{\beta}{N\alpha} \tag{3.7}$$

and using equation (3.2), we obtain the simplified SIR model

$$\dot{\bar{S}}(t) = -\bar{\beta}\bar{S}(t)H(t) \tag{3.8a}$$

$$\dot{H}(t) = \bar{\beta}\bar{S}(t)H(t) - \gamma H(t) \tag{3.8b}$$

which we term the *SH model*. The "S" in this SH model can be interpreted as the number of individuals susceptible of being hospitalized. The SH model has only two parameters ($\bar{\beta}$ and γ), one hidden state variable ($\bar{S}$) and one observed state variable (H) with observation model equation (3.3a).

Note that, in the SH model equation (3.8), the number of patients entering the hospital by unit of time is

$$E(t) := \bar{\beta}\bar{S}(t)H(t) \tag{3.9}$$

and the number of patients leaving the hospital by unit of time is

$$L(t) := \gamma H(t). \tag{3.10}$$

3.2 Estimation and prediction method

The goal is now to leverage the SH model equation (3.8) in order to predict future values of H based on its past and current observations $(H_o(t))_{t=t_s,\ldots,t_c}$.

To this end, we have to estimate (or "learn") four variables, which we term *estimands*: the two parameters $\bar{\beta}$ and γ and the two initial values $\bar{S}(t_\mathrm{i})$ and $H(t_\mathrm{i})$, where t_i is the chosen initial time for the SH model equation (3.8). One possible approach is to minimize some error measure between the simulated values $(H(t))_{t=t_\mathrm{i},\ldots,t_\mathrm{c}}$ and the observed values $(H_\mathrm{o}(t))_{t=t_\mathrm{i},\ldots,t_\mathrm{c}}$ as a function of the four estimands. However, the error measure is not available as a closed-form expression of the four estimands, and this makes this four-variable optimization problem challenging. We show in this section that it is possible to estimate $H(t_\mathrm{i})$ and γ separately. This leaves us with an optimization problem in the two remaining estimands $\bar{\beta}$ and $\bar{S}(t_\mathrm{i})$, making it possible to visualize the objective function by means of a contour plot.

3.2.1 Train and test sets

To recap, we have $t_\mathrm{s} \leq t_\mathrm{i} < t_\mathrm{c} < t_\mathrm{e}$. The provided dataset goes from t_s to t_e. The *test set*, $(H_\mathrm{o}(t), E_\mathrm{o}(t), L_\mathrm{o}(t))_{t\in[t_\mathrm{c}+1,t_\mathrm{e}]}$, is not available to the forecasting algorithm; it is revealed only to compute the forecasting error (see chapter 2 and Section 2.3). The SH model is initialized at t_i, and we refer to the data $(H_\mathrm{o}(t), E_\mathrm{o}(t), L_\mathrm{o}(t))_{t\in[t_\mathrm{i},t_\mathrm{c}]}$ as the *train set*, though it is legitimate to widen it to $t \in [t_\mathrm{s}, t_\mathrm{c}]$.

3.2.2 Estimation of $H(t_\mathrm{i})$

We simply take

$$H(t_\mathrm{i}) := H_\mathrm{o}(t_\mathrm{i}).$$

In view of equation (3.3a), this amounts to $\epsilon_H(t_\mathrm{i}) = 0$.

3.2.3 Estimation of γ

We have $L(t) = \gamma H(t)$, see (3.10). In view of the observation model (3.3), we can estimate γ by a ratio of means:

$$\hat{\gamma}^{\mathrm{RM}} = \frac{\sum_{t=t_i}^{t_c} L_o(t)}{\sum_{t=t_i}^{t_c} H_o(t)}. \tag{3.11}$$

Several other estimators are possible, such as the least square estimator,

$$\hat{\gamma}^{\mathrm{LS}} = \frac{\sum_{t=t_i}^{t_c} H_o(t) L_o(t)}{\sum_{t=t_i}^{t_c} H_o(t) H_o(t)}.$$

Note that t_i in the expression of $\hat{\gamma}$ can legitimately be replaced by any time between t_s and t_c. Only data in the test set, i.e., occurring after t_c, are unavailable in the variable estimation phase.

3.2.4 Joint estimation of $\bar{\beta}$ and $\bar{S}(t_\mathrm{i})$

Now we have to estimate the two remaining estimands, namely $\bar{\beta}$ and $\bar{S}(t_\mathrm{i})$. We choose the following sum-of-squared-errors objective function

$$\begin{aligned}\phi(\bar{\beta}, \bar{S}(t_\mathrm{i})) = c_\mathrm{H} \sum_{t=t_\mathrm{i}}^{t_\mathrm{c}} (H(t) - H_\mathrm{o}(t))^2 + c_\mathrm{E} \sum_{t=t_\mathrm{i}}^{t_\mathrm{c}} (E(t) - E_\mathrm{o}(t))^2 \\ + c_\mathrm{L} \sum_{t=t_\mathrm{i}}^{t_\mathrm{c}} (L(t) - L_\mathrm{o}(t))^2,\end{aligned} \tag{3.12}$$

where the c coefficients are parameters, all set to 1 in our experiments unless otherwise stated. In equation (3.12), the values of $H(t)$, $E(t)$ (see equation (3.9)), and $L(t)$ (see equation (3.10)) are given by the (approximate) solution of the SH model equation (3.8) in which (i) $H(t_\mathrm{i})$ and γ take the values estimated as above, and (ii) $\bar{\beta}$ and $\bar{S}(t_\mathrm{i})$ take the values specified in the argument of ϕ.

Several methods exist to compute the required (approximate) solution of the SH model equation (3.8); see [BW20]. Since the method will have to be repeatedly called by the optimization solver, we opt for the simplicity of the explicit Euler method with a time step of one day, yielding, for $t = t_\mathrm{i}, \ldots, t_\mathrm{c} - 1$,

$$\bar{S}(t+1) = \bar{S}(t) - \bar{\beta} \bar{S}(t) H(t) \tag{3.13a}$$

$$H(t+1) = H(t) + \bar{\beta} \bar{S}(t) H(t) - \gamma H(t). \tag{3.13b}$$

Now that the objective function ϕ (also termed "cost function" or "loss function") is defined, we let the estimated $(\bar{\beta}, \bar{S}(t_\mathrm{i}))$ be the (approximate) minimizer of ϕ returned by some optimization solver. The initial guess that we give to the solver is the outcome of the successive estimation method

that we will present in Section 3.3.1. If this proposed initial guess is negative, a situation that we observed on rare occasions in the late part of the decrease phase, then we take its opposite.

3.2.5 Prediction of H

Recall that the time range between t_i and t_c is the train period and the time range between $t_\mathrm{c}+1$ and t_e is the test period. In order to predict (i.e., forecast) the values of H over the test period, we apply the above procedure to estimate the four estimand variables $\bar{\beta}$, γ, $\bar{S}(t_\mathrm{i})$, and $H(t_\mathrm{i})$, and we compute the solution $H(t)$ of equation (3.13) for t from t_i to t_e. The prediction is then $(H(t))_{t=t_\mathrm{c}+1,\ldots,t_\mathrm{e}}$. The discrepancy between $(H(t))_{t=t_\mathrm{c}+1,\ldots,t_\mathrm{e}}$ and $(H_\mathrm{o}(t))_{t=t_\mathrm{c}+1,\ldots,t_\mathrm{e}}$ reveals the accuracy of the prediction.

3.3 Alternative estimation and prediction methods

3.3.1 Successive estimation of $\bar{\beta}$ and $\bar{S}(t_\mathrm{i})$

As an alternative to Section 3.2.4, we now present a method to estimate $\bar{\beta}$ independently. We do not recommend this alternative, as we have observed that it usually yields less accurate forecasts. However, it provides a convenient initial guess for Section 3.2.4, and it also sheds light on the various forecast accuracies observed in Section 3.4. From equation (3.8a) and equation (3.9), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{E(t)}{H(t)} = -\bar{\beta}E(t).$$

Since $\frac{\mathrm{d}}{\mathrm{d}t}\frac{E}{H} = \frac{H\dot{E}-E\dot{H}}{H^2}$, this yields

$$\bar{\beta} = \frac{\dot{H}(t)}{(H(t))^2} - \frac{\dot{E}(t)}{E(t)H(t)}.$$

We have

$$\dot{X}(t) = \frac{X(t+\delta t) - X(t)}{\delta t}. \tag{3.14}$$

The finite-difference approximation with a time step of one day $\dot{X}(t) \approx$

$X(t+1) - X(t)$ yields

$$\bar{\beta} \approx \frac{H(t+1) - H(t)}{(H(t))^2} - \frac{E(t+1) - E(t)}{E(t)H(t)}, \tag{3.15}$$

and a ratio-of-mean approach leads to the estimation

$$\widehat{\bar{\beta}} = \frac{H_{\mathrm{o}}(t_{\mathrm{c}}) - H_{\mathrm{o}}(t_{\mathrm{i}})}{\sum_{t=t_{\mathrm{i}}}^{t_{\mathrm{c}}-1}(H_{\mathrm{o}}(t))^2} - \frac{E_{\mathrm{o}}(t_{\mathrm{c}}) - E_{\mathrm{o}}(t_{\mathrm{i}})}{\sum_{t=t_{\mathrm{i}}}^{t_{\mathrm{c}}-1} E_{\mathrm{o}}(t)H_{\mathrm{o}}(t)}. \tag{3.16}$$

From equation (3.9), a possible simple estimator for the remaining estimand is $\widehat{\bar{S}}(t_{\mathrm{i}}) = \frac{E(t_{\mathrm{i}})}{\widehat{\bar{\beta}}H(t_{\mathrm{i}})}$.

We can now investigate how the ϵ error terms in the observation model defined in equation (3.3) impact $\widehat{\bar{\beta}}$. We assume throughout that the errors in $H_{\mathrm{o}}(t+1) - H_{\mathrm{o}}(t)$ and $E_{\mathrm{o}}(t+1) - E_{\mathrm{o}}(t)$ are comparable. Except at the very beginning of the outbreak, $E_{\mathrm{o}}(t)H_{\mathrm{o}}(t)$ is considerably smaller than $(H_{\mathrm{o}}(t))^2$, and thus the second term of equation (3.15) drives the error.

Consequently, the estimation of $\bar{\beta}$ should be the most accurate when $E_{\mathrm{o}}(t)H_{\mathrm{o}}(t)$ is the largest. This occurs slightly before the peak of $H_{\mathrm{o}}(t)$. This means that the estimation of $\bar{\beta}$ should be the most accurate for a train period slightly before the peak. However, this does not mean that this position of the train period gives the most accurate forecasts, as we will see below.

Let us consider the situation where the train period is located *before* the peak. Then the estimation of $\bar{\beta}$ is less accurate, and this impacts $\widehat{\bar{S}}(t_{\mathrm{i}})$. At the initial time t_{i}, this does not impact the right-hand term of equation (3.8b) in view of the definition of $\widehat{\bar{S}}(t_{\mathrm{i}})$. However, an overestimation of $\bar{\beta}$ will induce an underestimation of $\bar{S}(t_i)$ and, in view of equation (3.8a), a subsequent even stronger underestimation of $\bar{S}(t)$. Hence the first term of equation (3.8b) will be underestimated. As a consequence, the peak in H will appear sooner and lower. The case of an underestimation of $\bar{\beta}$ leads to the opposite conclusion, namely a peak in H that appears later and higher. In summary, the further before the peak the train period is located, the more inaccurate the position and height of the peak is expected to be.

Finally, let us consider the situation where the train period is located *after* the peak. Then we can make the same observations as in the previous paragraph, except that predicting the peak is now irrelevant. Moreover, we are in the decrease phase, where the first term of equation (3.8b) (which involves $\bar{\beta}$ and $\bar{S}(t)$) is smaller than the second term (which does not in-

volve these quantities). Consequently, the possibly large estimation errors on $\bar{\beta}$ and $\bar{S}(t)$ will only slightly affect the forecast of $H(t)$.

3.3.2 Joint estimation of the four estimands

An alternative to Sections 3.2.2–3.2.4 is to reconsider equation (3.12) as a function of all four estimands: $K_1 = \sum_{t=t_i}^{t_c}(H(t) - H_o(t))^2$, $K_2 = \sum_{t=t_i}^{t_c}(E(t) - E_o(t))^2$, and $K_3 = \sum_{t=t_i}^{t_c}(L(t) - L_o(t))^2$

$$\tilde{\phi}(\bar{\beta}, \bar{S}(t_i), \gamma, H(t_i)) = c_H K_1 + c_E K_2 + c_L K_3. \tag{3.17}$$

In equation (3.17), variables $H(t)$, $E(t)$ (see equation (3.9)), and $L(t)$ (see equation (3.10)) are given by the solution of the discrete-time SH model equation (3.13) where the parameters $\bar{\beta}$ and γ and the initial conditions $\bar{S}(t_i)$ and $H(t_i)$ take the values specified in the argument of $\tilde{\phi}$. Minimizing $\tilde{\phi}$ is a more challenging problem than mimizing ϕ equation (3.12) in view of the larger number of optimization variables. It may be essential to give a good initial guess to the optimization solver, and a natural candidate for this is the values obtained by the procedure described in Sections 3.2.2–3.2.4.

In our preliminary experiments, we have found that this alternative does not present a clear advantage in terms of the prediction mean absolute percentage error (MAPE). The results reported in Section 3.4 are obtained with the prediction approach of Section 3.2, unless otherwise specified.

3.4 Results

We now apply the method of Section 3.2 (by default) or Section 3.3.2 (when specified) to the data of chapter 2 and Section 2.1.

3.4.1 Fitting experiment

We first check how well the SH model equation (3.8) can fit the available data. For this experiment, we use the method of Section 3.3.2 with $c_E = c_L = 0$ in order to get the best possible fit (in the least squares sense) to the H_o time series. The results are shown in Figure 3.1 for Belgium and France.

We also present the results for Luxembourg and the United Kingdom in Figure 3.2.

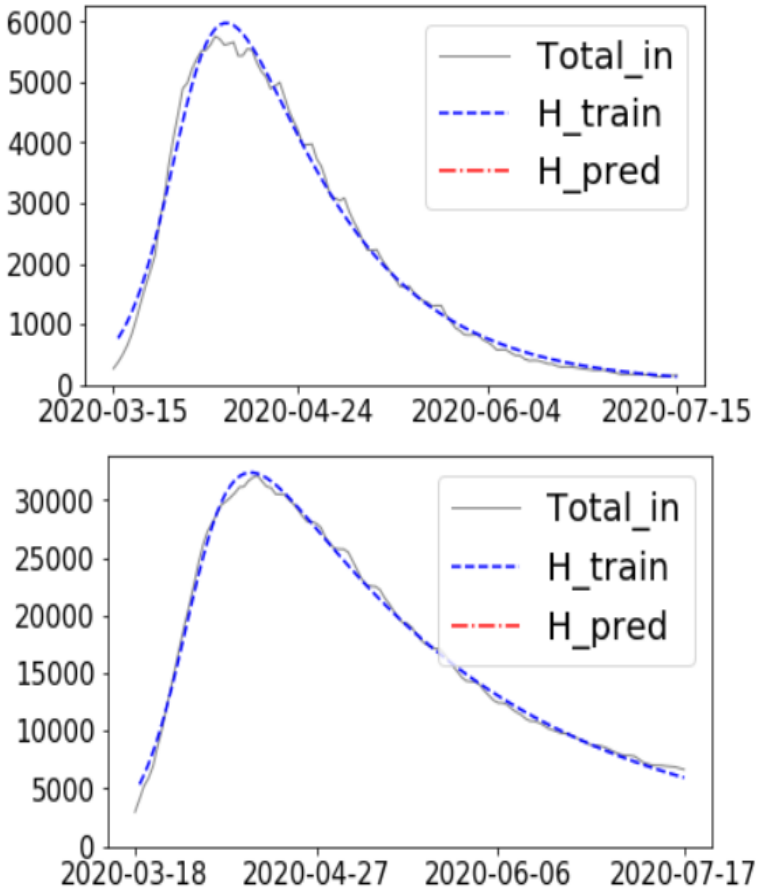


Fig. 3.1 Fitting the SH model to the H_o (Total_in: total hospitalized) curve. In this experiment, the train set is the whole dataset, hence there is no test (i.e., forecast) curve. Top: Belgium. Bottom: France.

For Belgium, the RRSE is 9% and the MAPE 14%. For France, the RRSE is 6% and the MAPE 3%. For Luxembourg, the RRSE is 13% and the MAPE 11%. For the United kingdom, the RRSE is 6% and the MAPE 3%. The comparatively large value of the MAPE for Belgium illustrates the comment made about MAPE in chapter 2 and Section 2.3.

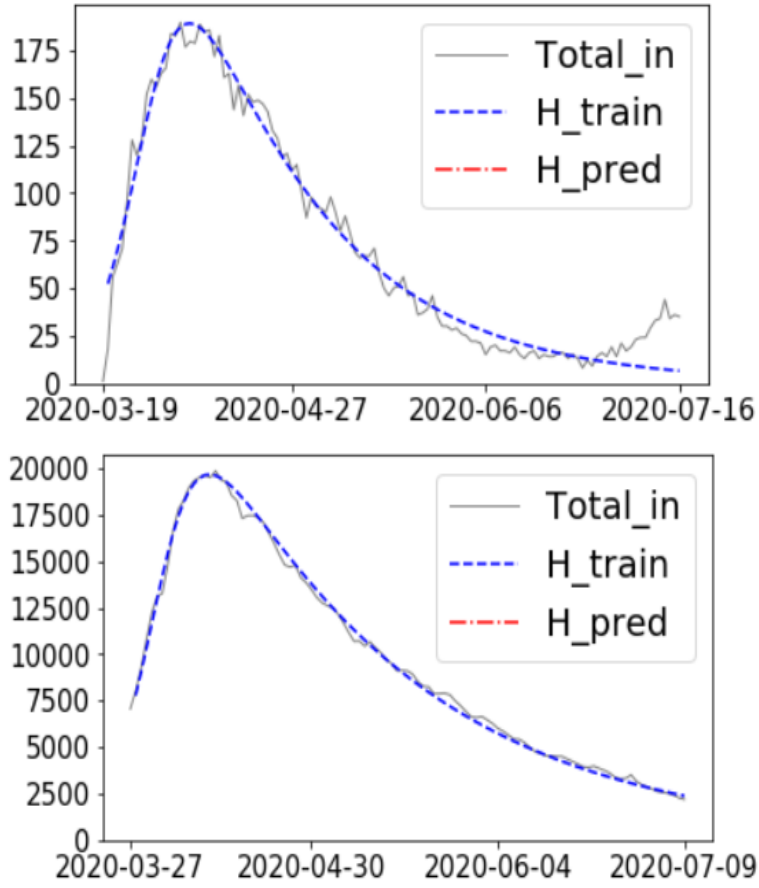


Fig. 3.2 Fitting the SH model to the H_o (Total_in: total hospitalized) curve. In this experiment, the train set is the whole dataset, hence there is no test (i.e., forecast) curve. Top: Luxembourg. Bottom: the United Kingdom.

The fitting error is remarkably small in view of the fact that there are some 120 data points for only 4 estimands. Recall indeed that the parameters of the SH model are constant with respect to time in our experiments. This contrasts with [Koz20] where there are two phases, and with [Nes20b] where the infection rate is piecewise constant with several pieces.

We stress that Figure (3.1 and 3.2) in itself does not imply that the model

leads to accurate *forecasts*. If the optimal fit over some period is bad, then forecasts over that period can only be bad. But if the fit is good (as it is the case here), forecasts can still be bad due to their sensitivity with respect to the data preceding the to-be-forecasted period. For example, a better fit (in the RMSE sense) than in Figure 3.1 (left) can be obtained with a polynomial of degree between 8 and 15; however, its forecast accuracy is abysmal.

In order to assess the forecast accuracy of the model, we have to learn the estimand variables over a *train period* that we make available to the algorithm, then use the learned estimands in order to predict H over a subsequent *test period* whose data is not available to the algorithm, and finally compare the prediction with the data on the test period. This is what we proceed to do in the rest of this Section 3.4.

3.4.2 Forecasts from a hand-picked train period chosen around the peak

We start with a forecast experiment where the train period is chosen around the peak. According to Section 3.3.1, this is a promising location.

Note that this proof-of-concept experiment is affected by a form of data leakage since we look at the whole data—including the to-become test part—in order to choose the train period. This flaw will be remedied in Section 3.4.3.

A contour plot of the objective function ϕ equation (3.12) is given in Figure 3.3. In order to make the minimizer easier to visualize, the plot shows equispaced-level curves of $\log(\phi - 0.90\,\phi_*)$, where ϕ_* is the approximation of the minimal value of ϕ provided by the optimization solver. The red circle shows the initial guess obtained from Section 3.3.1 and the red star shows the approximate minimizer returned by the optimization solver. In our experiments, we use the optimization solver scipy.optimize.fmin with its default parameters.

The bottom plot of Figure 3.3 shows $(H_o(t))_{t=t_s,\dots,t_e}$ (observed hospitalizations, gray solid line), $(H(t))_{t=t_i,\dots,t_c}$ (hospitalizations given by the model over the train period, blue dashed line), and $(H(t))_{t=t_c+1,\dots,t_e}$ (hospitalizations forecasted over the test period, in red dash-dot line). In order to give a sense of the sensitivity of the results, we superpose the curves obtained for three slightly different train periods. The test MASE values for the three curves are 3.16, 0.77, and 1.15. These values are compatible with the remark in chapter 2 and Section 2.3 that MASE values around 1 or lower indicate an excellent accuracy. The test MAPE values for the three

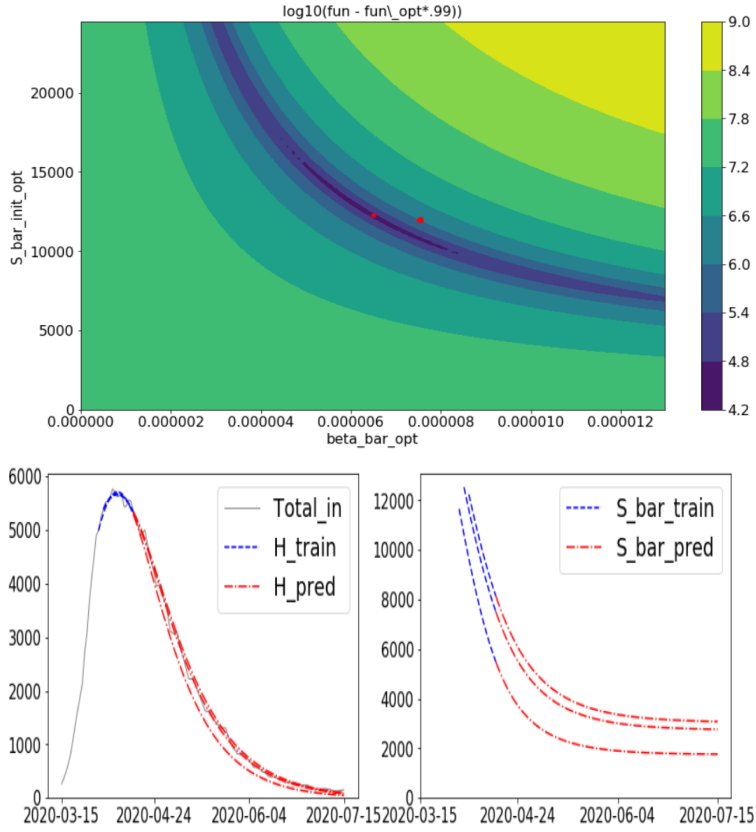


Fig. 3.3 Belgium, train period around the peak. Top: contour plot of ϕ equation (3.12). Bottom: fitting and predictions with the SH model.

curves are 27%, 7%, and 8%.

The right-hand plot of Figure 3.3 shows the evolution of $\bar{S}(t)$.

In summary, Figure 3.3 has shown that a high forecast accuracy can be achieved for the Belgian data with suitably chosen train periods. The next step is to investigate if this finding generalizes to other datasets (Section 3.4.3) and to other positions of the train period (Section 3.4.4).

3.4.3 Forecasts from an automatically chosen region around the peak

In order to conduct experiments on a large number of datasets, we have to automatize the selection of the train period around the peak. Moreover, the procedure has to be free of data leakage, i.e., it has to select the end time t_c of the train period without ever reading the data for $t > t_c$.

The proposed train-period selection procedure goes as follows. We first select a time span N; in our experiments, we choose $N = 7$. In order to robustify the procedure, we smooth out the H_o time series by compting its moving average (MA):

$$H_o^{MA}(t) = \text{mean}(H_o(t-N), \ldots, H_o(t+N)). \tag{3.18}$$

Then we compte $\hat{t} = \min\{t : \arg\max_{\tau \le t} H_o^{MA}(\tau) = t - N\}$. This $\hat{t}$ is computed by letting t increase until the argmax condition is satisfied. In view of equation (3.18), the data for $t > \hat{t} + N$ remains unseen, hence we choose $t_c := \hat{t} + N$ as the end time of the train period, thereby avoiding data leakage. The start of the train period is chosen as $t_i = \hat{t} - 2N$.

The test accuracies are measured over the 60-day period $[t_c + 1, \min(t_c + 60, t_e)]$.

The results obtained for each district (province in Belgium and department in France) are summarized in Table 3.1 (Belgium) and Table 3.2 (France). Moreover, for the whole of Belgium, we obtain the excellent MASE_test = 0.72, and for the whole of France, we get a considerably poorer MASE_test = 14.07. The tables reveal that none of the districts admits an MASE_test as low as the 0.72 obtained for the whole of Belgium. However, Section 3.4.2 has shown that an MASE_test around 3 is still appreciably low. In view of Table 3.1, more than half of the Belgian provinces have an MASE_test below 3. Table 3.2 indicates that an MASE_test below 3 occurs for at least 10% of the French departments.

3.4.4 Forecasts from various train periods

We now restrict to the total hospitalization numbers for Belgium, France, Luxembourg and the United Kingdom, and we investigate the impact of the position of the train period on the forecast accuracy. In Figure 3.4 we superpose the results obtained with various train periods of 14 days for Belgium. The figure corroborates the comments of Section 3.3.1. In particular, if the train period is fully located before the peak, then the forecasts

Percentiles	min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	max
RRSE_train	0.30	0.33	0.33	0.42	0.51	0.55	0.60
RRSE_test	0.13	0.17	0.18	0.29	0.36	0.38	0.75
Ratio_RRSE	0.24	0.33	0.54	0.71	0.90	1.15	1.25
MASE	0.61	0.79	1.10	1.31	1.75	2.60	6.69
MASE_train	0.65	0.75	0.76	0.85	0.94	0.98	1.27
MASE_test	1.51	1.81	2.20	2.84	3.30	4.73	11.18
Ratio_MASE	1.54	2.19	2.51	3.28	3.99	6.08	14.44
sMAPE_train	0.02	0.02	0.02	0.03	0.03	0.04	0.06
sMAPE_test	0.09	0.11	0.17	0.21	0.47	0.48	0.71
Ratio_sMAPE	4.36	5.20	6.05	7.34	10.25	20.01	35.55

Table 3.1 Belgian provinces. The ratio means that test/train values are reported.

are rather inaccurate. Placing the train period around the peak gives excellent forecast results. When the train period is fully located in the decrease phase, the estimation of $\bar{\beta}$ and $\bar{S}(t_i)$ is seen to be very sensitive, but this does not affect much the quality of the forecast of $H(t)$. Figure 3.5 is the equivalent of Figure 3.4 for France. Again, the experiments are compatible with the comments of Section 3.3.1. In experiments not reported here, we also considered some departments separately, with fairly similar results. In Figure 3.6, and 3.7, we observe the same results with data from Luxembourg and UK compared to various train periods. A disconcerting aspect is the evolution of the estimated γ as a function of the location of the train period. In Figure 3.4 and 3.6 (Belgium and Luxembourg), the estimation of γ is grouped around 0.08 for several train periods. However, in Figure 3.5 and 3.7 (France and UK), the estimation of γ keeps decreasing (not too for UK), indicating that the daily number of patients leaving the hospital is an increasingly small fraction of the number of patients at the hospital. In view of equation (3.11), this casts doubt on the L_o values (see chapter 2 and Section 2.1).

3.5 Partial conclusion

The experiments in Section 3.4 have shown that the proposed method has a remarkably good fitting accuracy over the whole available data. It also

Percentiles	min	P_{10}	P_{25}	P_{50}	P_{75}	P_{90}	max
RMSE_train	0.69	2.34	3.44	5.47	10.05	18.19	37.94
RMSE_test	0.78	3.80	7.60	21.51	41.03	93.60	369.17
Rati_RMSE	0.47	1.10	1.93	3.34	6.87	9.65	21.38
RRSE_train	0.17	0.35	0.40	0.52	0.67	0.86	1.27
RRSE_test	0.13	0.31	0.45	0.77	1.33	2.12	5.79
Rati_RRSE	0.33	0.51	0.78	1.37	2.94	4.52	11.85
MAE_train	0.59	1.84	2.67	4.34	8.09	15.29	30.21
MAE_test	0.66	3.01	6.95	17.94	37.76	87.36	351.34
Rati_MAE	0.47	1.11	2.12	3.81	8.10	11.45	26.11
MASE	0.37	1.27	2.10	4.16	7.63	11.36	22.95
MASE_train	0.50	0.75	0.87	1.00	1.20	1.41	1.89
MASE_test	1.88	2.96	4.42	8.42	14.05	18.75	90.10
Rati_MASE	2.05	2.84	4.07	7.51	16.05	21.11	68.08
RelMAE_train	0.10	0.18	0.28	0.37	0.50	0.67	0.98
RelMAE_test	0.04	0.10	0.17	0.33	0.70	1.15	4.95
Rati_RelMAE	0.13	0.22	0.41	0.90	2.15	4.07	21.22
nMAE_train	0.01	0.01	0.02	0.04	0.05	0.08	0.34
nMAE_test	0.05	0.14	0.20	0.30	0.51	0.66	1.00
Rati_nMAE	1.21	3.40	5.02	7.60	14.21	20.91	48.18
rnMAE_train	0.03	0.08	0.09	0.12	0.15	0.19	0.31
rnMAE_test	0.04	0.08	0.10	0.21	0.36	0.55	1.68
Rati_rnMAE	0.31	0.55	0.78	1.76	3.43	5.22	14.31
sMAPE_train	0.01	0.01	0.02	0.04	0.05	0.08	0.36
sMAPE_test	0.06	0.18	0.26	0.45	0.79	1.09	1.80
Rati_sMAPE	1.33	4.17	6.64	12.45	22.95	33.33	56.45

Table 3.2 French departments. The ratio means that test/train values are reported.

produces remarkably accurate forecasts over certain time ranges for some areas (Belgium, some Belgian provinces, and a few French departments). However, there are also time ranges and areas where the forecasts are very inaccurate. In particular, when it is trained before the peak, the model produces rather poor forecasts for the position and height of the peak and for the subsequent decrease. Because of how it is constructed, the model is also unable to produce multiple peaks in order to fit or forecast rebounds. The forecasts returned by the model should thus be taken with much caution.

Another source of caution is that we cannot rule out the situation where the considered objective function would be multimodal. The optimization

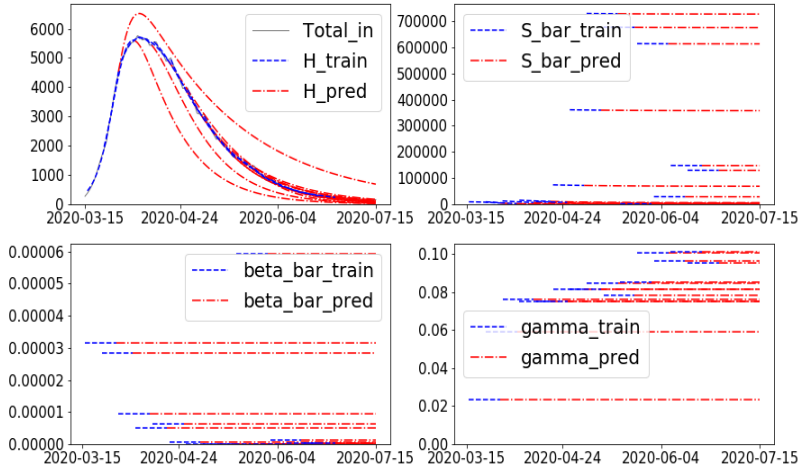


Fig. 3.4 Belgium, various train periods.

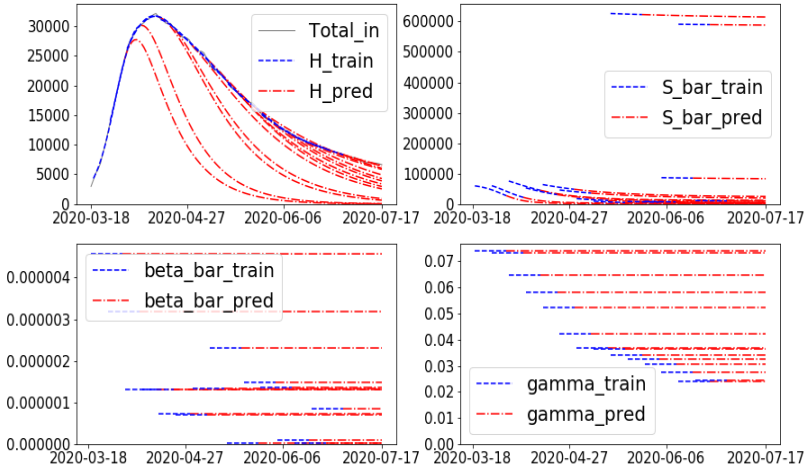


Fig. 3.5 France, various train periods.

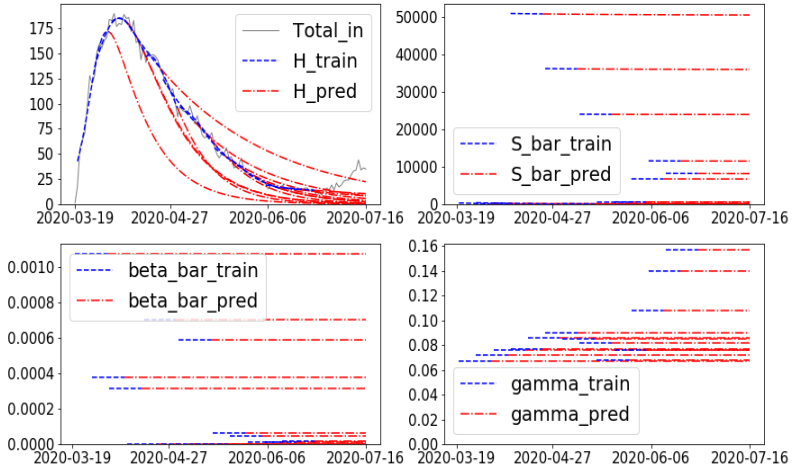


Fig. 3.6 Luxembourg, various train periods.

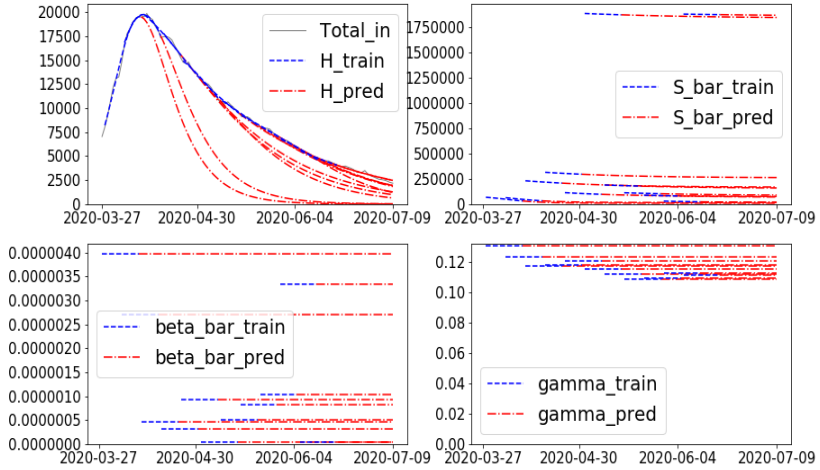


Fig. 3.7 UK, various train periods.

solver might thus get stuck in a local nonglobal minimum, yielding a suboptimal fit of the train data and possibly poorer forecasts than what the global minimum would achieve. Moreover, even if the objective function is unimodal, the stopping criterion of the solver may trigger before a suitably accurate approximation of the minimum is reached. If the proposed model is used to guide prevention policies, then further caveats are in order. We have seen that the estimation of $\bar{\beta}$ is very sensitive. Hence the proposed model can hardly help assess the impact of prevention measures on $\bar{\beta}$. Without knowing sufficiently accurately the impact of prevention measures on $\bar{\beta}$, we may not aptly use the model to predict their impact on the evolution of the hospitalizations. Yet another caveat is that it may be tempting to deduce from the excellent fit with a constant-parameter model (Figure 3.1) that the evolution of the prevention measures over the dataset period has had no impact on $\bar{\beta}$. But the deduction is flawed. Indeed, in view of the comments made in Section 3.3.1, the available data could also be very well explained with fairly large jumps in $\bar{\beta}$ during the decrease phase.

In spite of all these caveats, the hospitalization forecasts returned by the method, and also the evolution of $\bar{S}(t)$, might be of practical use in the context of various disease outbreaks, e.g., for resource planning. To this end, it will be important to understand which specific features of the COVID-19 outbreak in Belgium made it possible to forecast so accurately the hospitalization decrease several months ahead.

4

Alternative models

4.1 SH-D model

In this chapter, we present some alternative models to the SH model defined in the above chapter 3. First, we investigate a parameter estimation method by least squares with fatalities data denoted by $\bar{D}$ and patients discharged from the hospital alive (meaning the patients who leave the hospital after being recovered) denoted by $\bar{R}$ in order to predict hospitalizations cases. Second, we present polynomial and exponential decrease models.

The number of new deaths and the patients discharged alive from the hospital in Belgium are presented in Figure 4.1 where the weekly pattern can be explained by the fact that fewer patients were discharged on Saturdays and Sundays. The following sections 4.1.1 and 4.1.2 are inspired from [Gab20], and are related to the estimation of p (defined below), γ, $\bar{N}$ and $\bar{\beta}$ (remain the parameters defined in chapter 3) with fatalities and patients discharged from the hospital data.

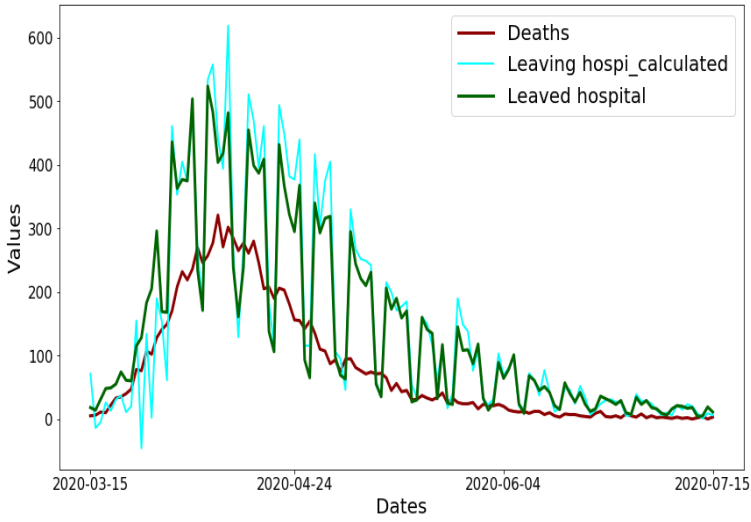


Fig. 4.1 Daily number of deaths and patients leaving alive from the hospital in Belgium.

4.1.1 Estimation of p, γ, $\bar{N}$ and $\bar{\beta}$ with fatalities and patients discharged from the hospital data.

It is assumed that the number of recovered and the number of deaths have a fix ratio

$$\bar{D}(t) = p\bar{R}(t), \tag{4.1}$$

where p can be determined like in estimation method of γ (ratio means) defined in chapter 3 and Section 3.2.3. Then, we can estimate p by a ratio of means:

$$\hat{p}^{\mathrm{RM}} = \frac{\sum_{t=t_{\mathrm{i}}}^{t_{\mathrm{c}}} \bar{D}_o(t)}{\sum_{t=t_{\mathrm{i}}}^{t_{\mathrm{c}}} \bar{R}_o(t)}. \tag{4.2}$$

By considering the two equations in system (3.8) defined in chapter 3 that are $\dot{\bar{S}}(t) = -\bar{\beta}\bar{S}(t)H(t)$ and $\dot{\bar{R}}(t) = \gamma H(t)$, we can write

$$\begin{aligned} \dot{\bar{S}}(t) &= -\bar{\beta}\bar{S}(t)H(t) \\ &= -\frac{\bar{\beta}}{\gamma}\dot{\bar{R}}(t)\bar{S}(t). \end{aligned} \tag{4.3}$$

By using the equation (4.1), we obtain

$$\dot{\bar{S}}(t) = -\frac{\bar{\beta}}{p\gamma}\dot{\bar{D}}(t)\bar{S}(t) \text{ or } \frac{1}{\bar{S}(t)}\frac{d\bar{S}(t)}{dt} = -\frac{\bar{\beta}}{p\gamma}\frac{d\bar{D}(t)}{dt}. \tag{4.4}$$

By integrating the equation (4.4), we obtain as in [Gab20]

$$\bar{S}(t) = \bar{N}\exp(-\frac{\bar{\beta}\bar{D}(t)}{p\gamma}), \tag{4.5}$$

where we assumed that at the beginning of the epidemic, the whole population was susceptible being hospitalized $\bar{S}(0) = \bar{N}$ and nobody died of yet $\bar{D} = 0$.
From equations $\bar{D}(t) = p\bar{R}(t)$ and $\frac{d\bar{R}(t)}{dt} = \gamma H(t)$ we can write

$$\begin{aligned}\frac{d\bar{D}(t)}{dt} &= p\frac{d\bar{R}(t)}{dt}\\ &= p\gamma H(t).\end{aligned} \tag{4.6}$$

Thus, by combining the equations (4.6) and (4.5), and $\frac{dH}{dt}(t) = \bar{\beta}\bar{S}(t)H(t) - \gamma H(t)$ (defined in system (3.8) in chapter 3), we write the following equation in terms of the cumulative number of fatalities

$$\begin{aligned}\frac{d^2\bar{D}}{dt^2}(t) &= p\gamma\frac{dH}{dt}(t)\\ &= p\gamma[\bar{\beta}\bar{S}(t)H(t) - \gamma H(t)]\\ &= [p\gamma\bar{\beta}\bar{N}\exp(-\frac{\bar{\beta}\bar{D}(t)}{p\gamma}) - p\gamma^2]H(t)\\ &= [p\gamma\bar{\beta}\bar{N}\exp(-\frac{\bar{\beta}\bar{D}(t)}{\gamma p}) - p\gamma^2] \times \frac{1}{\gamma}\frac{d\bar{R}(t)}{dt}\\ &= [p\bar{\beta}\bar{N}\exp(-\frac{\bar{\beta}\bar{D}(t)}{\gamma p}) - p\gamma] \times \frac{1}{p}\frac{d\bar{D}(t)}{dt}\\ &= [\bar{\beta}\bar{N}\exp(-\frac{\bar{\beta}\bar{D}(t)}{\gamma p}) - \gamma]\frac{d\bar{D}(t)}{dt}.\end{aligned} \tag{4.7}$$

The right hand side of this equation can be simplified to

$$\frac{d^2\bar{D}}{dt^2}(t) = \frac{dF(\bar{D}(t))}{dt}. \tag{4.8}$$

Integrating both sides we have

$$\frac{d\bar{D}}{dt}(t) = F(\bar{D}(t)) + c, \tag{4.9}$$

and setting the boundary condition $\frac{d\bar{D}}{dt} = 0$ when $\bar{D} = 0 \longrightarrow F(0) + c = 0 \longrightarrow c = 0$. Finally,

$$F(\bar{D}(t)) = p\gamma\bar{N}(1 - \exp(-\frac{\bar{\beta}\bar{D}(t)}{p\gamma})) - \gamma\bar{D}(t), \tag{4.10}$$

representing the daily number of deaths as in [Gab20].

We compute the parameters in equation (4.10) by performing the least squares estimation described in capture 4.2 and the statistics results are presented in Table 4.1, the fitting in Figure 4.3 and the forecast results in Figure 4.4. This γ estimate with Belgium data is compatible to the estimates in Spain and Italy if we compare the Table 4.1 and the Table 1 in [Gab20].

In Figure 4.4, in order to give a sense to the sensitivity of the results, we superpose the curves obtained for three slightly different train periods. With the discharged patients from the hospital data (in left) the MASE_test values for the three curves are 1.15, 1.21, and 1.28. These MASE values are around 1 indicating an excellent accuracy. With the number of deaths data (in right) the RMSE_test values are 15.81, 23.87 and 34.11. These RMSE values are low and indicate a good accuracy.

In Figure 4.5, these estimates give bad quality of forecasts around the peak and perform better in the decreasing phase.

p	γ	$\bar{N}$	$\bar{\beta}$
0.482	0.072	20417.588	$1.4017906318051884e^{-05}$

Table 4.1 Estimation of p, γ, $\bar{N}$ and $\bar{\beta}$ with data related to deaths and patients discharged from the hospital, Belgium.

```
def modell(x, u):
    return x[0]*x[1]*x[2]*(1-np.exp(-(x[3]*u)/(x[0]*x[1])))-x[1]*u
def funn(x, u, y):
    return modell(x, u) - y
y=death
u=np.cumsum(death)
x0=[p,gamma, N_bar,beta_bar_opt] #Initialization
res = least_squares(funn,x0, args=(u, y), verbose=1)
p_estimate=res.x[0]
gamma_estimate = res.x[1]
N_bar_estimate = res.x[2]
beta_bar_estimate = res.x[3]
```

Fig. 4.2 Pseudocode of computing of p, γ, $\bar{N}$ and $\bar{\beta}$ by `least squares` using `scipy.optimize` Python library, where $\bar{D}$ is represented by u and y represents the daily number of observed deaths.

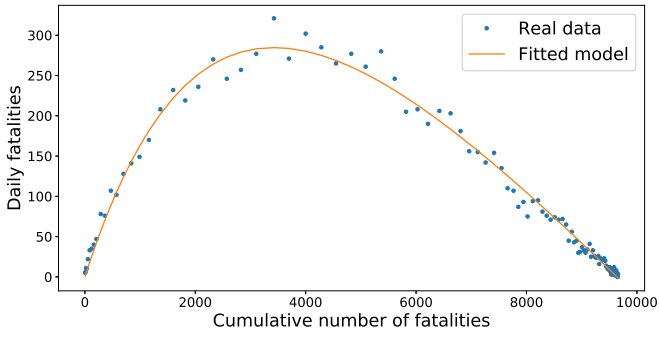


Fig. 4.3 Fitting result of the daily number of deaths as function of the cumulative number of deaths, with the estimates in Table 4.1, Belgium.

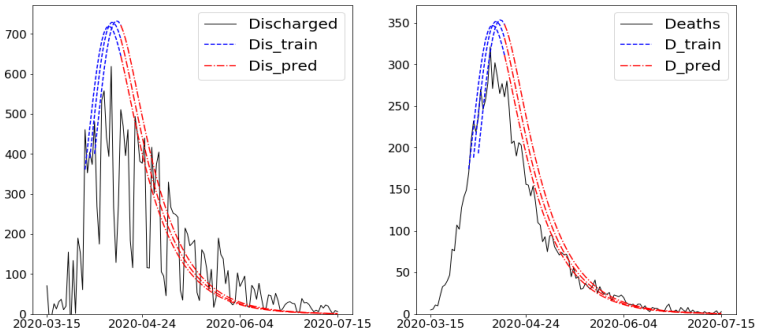


Fig. 4.4 Predictions and forecasts, with the estimates in Table 4.1, of the daily number of deaths and the daily number of patients discharged from the hospital, Belgium.

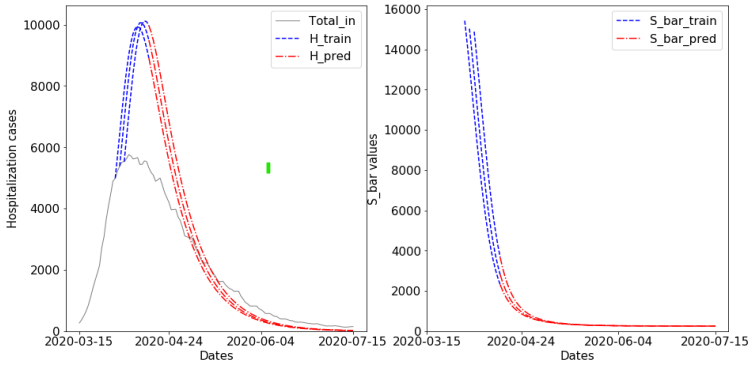


Fig. 4.5 Predictions and forecasts, with estimates in Table 4.1, of daily hospitalized of COVID-19 and $\bar{S}$ values, in Belgium.

4.1.2 Estimation of γ, $\bar{N}$ and $\bar{\beta}$ with discharged individuals from the hospital data.

By analogy with Section 4.1.1, we have

$$\dot{\bar{S}}(t) = -\frac{\bar{\beta}}{\gamma}\dot{\bar{R}}(t)\bar{S}(t) \quad \text{or} \quad \frac{1}{\bar{S}(t)}\frac{d\bar{S}(t)}{dt} = -\frac{\bar{\beta}}{\gamma}\frac{d\bar{R}(t)}{dt}. \tag{4.11}$$

By integrating the equation (4.11), we obtain as in [Gab20]

$$\bar{S}(t) = \bar{N}\exp(-\frac{\bar{\beta}\bar{R}(t)}{\gamma}), \tag{4.12}$$

where we assumed that at the beginning of the epidemic, the whole population was susceptible being hospitalized $\bar{S}(0) = \bar{N}$ and nobody discharges from the hospital yet $\bar{R} = 0$. For that, by deriving $\frac{d\bar{R}(t)}{dt} = \gamma H(t)$, we can have

$$\begin{aligned}\frac{d^2\bar{R}}{dt^2}(t) &= \gamma\frac{dH}{dt}(t)\\ &= \gamma[\bar{\beta}\bar{S}(t)H(t) - \gamma H(t)]\\ &= [\gamma\bar{\beta}\bar{N}\exp(-\frac{\bar{\beta}\bar{R}(t)}{\gamma}) - \gamma^2]\frac{1}{\gamma}\frac{d\bar{R}}{dt}(t)\\ &= [\bar{\beta}\bar{N}\exp(-\frac{\bar{\beta}\bar{R}(t)}{\gamma}) - \gamma]\frac{d\bar{R}(t)}{dt}.\end{aligned} \tag{4.13}$$

The right hand side of this equation can be simplified yo

$$\frac{d^2\bar{R}}{dt^2}(t) = \frac{dF(\bar{R}(t))}{dt}. \tag{4.14}$$

Integrating both sides

$$\frac{d\bar{R}}{dt}(t) = F(\bar{R}(t)) + c, \tag{4.15}$$

and setting the boundary condition $\frac{d\bar{R}}{dt}(t) = 0$ when $\bar{R} = 0 \longrightarrow F(0) + c = 0 \longrightarrow c = 0$. Finally,

$$F(\bar{R}(t)) = \gamma\bar{N}(1 - \exp(-\frac{\bar{\beta}\bar{R}(t)}{\gamma})) - \gamma\bar{R}(t), \tag{4.16}$$

representing the daily discharged individuals from the hospital.

We already use the optimization method (capture 4.2) to compute the parameters in equation (4.16). We present the results in Table 4.2, the fitting in Figure 4.6, and the forecast results in Figures 4.7 and 4.8. In Table 4.2, the values of γ and $\bar{N}$ are smaller than the values obtain in Table 4.1.

In Figure 4.7, in order to give a sense of the sensitivity of the results, we superpose the curves obtained for three slightly different train periods. With the discharged patients from the hospital data (in left) the MASE_test

values for the three curves are 0.91, 0.88, and 0.89 and the MAPE_test values are 104%, 111.6% and 121.5%. These MASE values are around 1 indicating an excellent accuracy with these estimates even if this forecast curve doesn't reflect the weekly variability of the real data.

γ	$\bar{N}$	$\bar{\beta}$
0.0387427	18525.15804	1.613671e-05

Table 4.2 Estimation of γ, $\bar{N}$ and $\bar{\beta}$ with the discharged individuals from the hospital in Belgium.

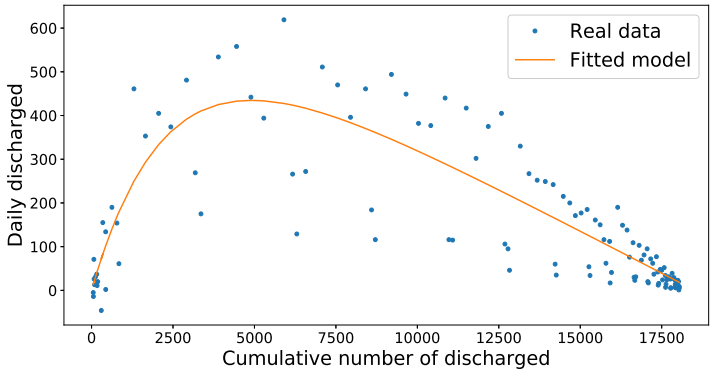


Fig. 4.6 Fitting result, with estimates in Table 4.2, of the daily number of patients discharged from the hospital as a function of the cumulative number of the patients discharged from the hospital, Belgium.

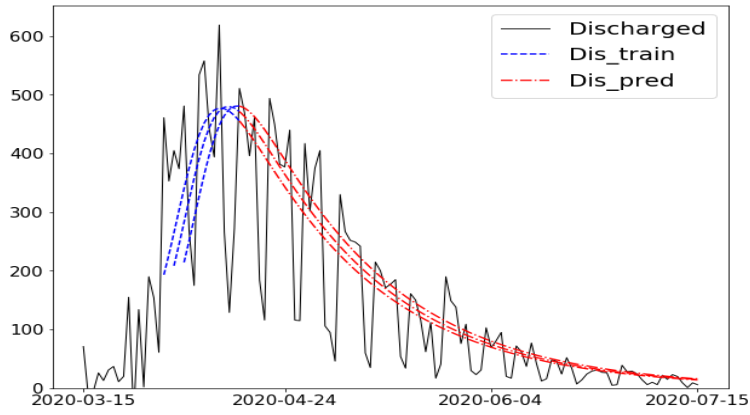


Fig. 4.7 Predictions and forecasts, with estimates in Table 4.2, of the daily number of patients discharged from the hospital in Belgium.

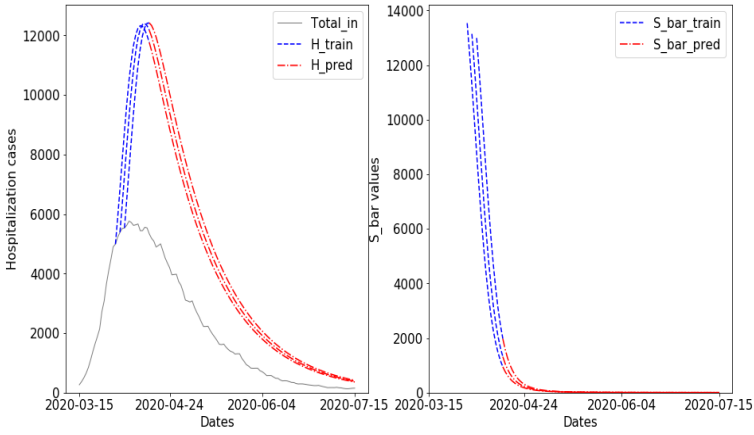


Fig. 4.8 Predictions and forecasts, with estimates in Table 4.2, of the daily hospitalizations cases and $\bar{S}$ values, in Belgium.

4.2 Polynomial method

We use the function *numpy.poly1d(numpy.polyfit(x_{train}, y_{train}, n))* in Python, where n is the degree of the polynomial function, x_{train} is the period of days, y_{train} is the hospitalized cases, to fit the observed hospitalization data. Consider a candidate list of integers for n representing polynomial degrees, we take the one that gives the best fit in the direction of MAPE. Thus, by visual inspection of Figure 4.9 (top), we observe a very nice fit curve with the whole dataset (MAPE=3.806) when $n = 15$. But the forecasts displayed in Figures 4.9–4.12 (in bottom, red curve) are extremely bad.

Train interval	MAPE_train	MAPE_test
20	0.71	2254082430652992389 12.00
40	0.99	1172675279413422.75
60	1.81	280792312941.52
80	2.32	564484412.66
100	3.63	2194481.87

Table 4.3 Statistics for polynomial model ($n = 15$) with Belgium data.

In Table 4.3, an anti-correlation between MAPE_train (with small values) and MAPE_test (with very high values) is noticed when the train set is increasing (i.e., the test set is decreasing). These observations clearly reveal the incapacity of the polynomial method to forecast.

In conclusion, we have observed that polynomials of sufficiently high degree are able to accurately fit the data. This was expected in view of the Weierstrass approximation theorem. However, polynomials produce very poor forecasts. This was also expected, as epidemiological models do not yield a polynomial evolution of the compartment populations with respect to time.

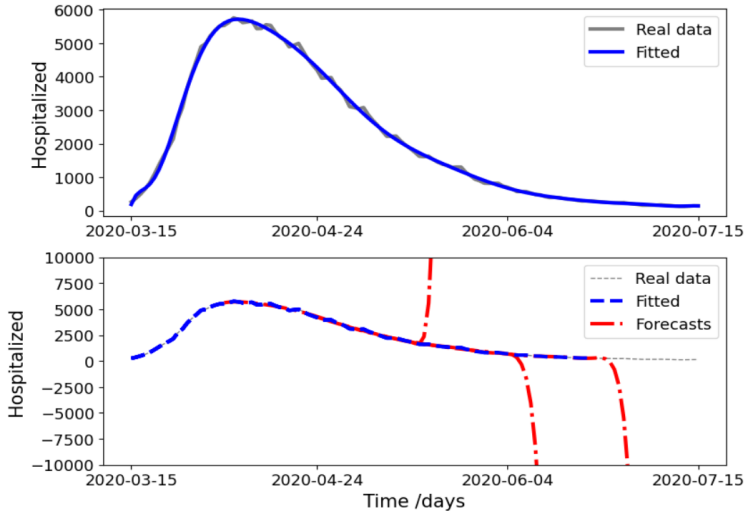


Fig. 4.9 Predictions and forecasts for polynomial method with various train periods, Belgium.

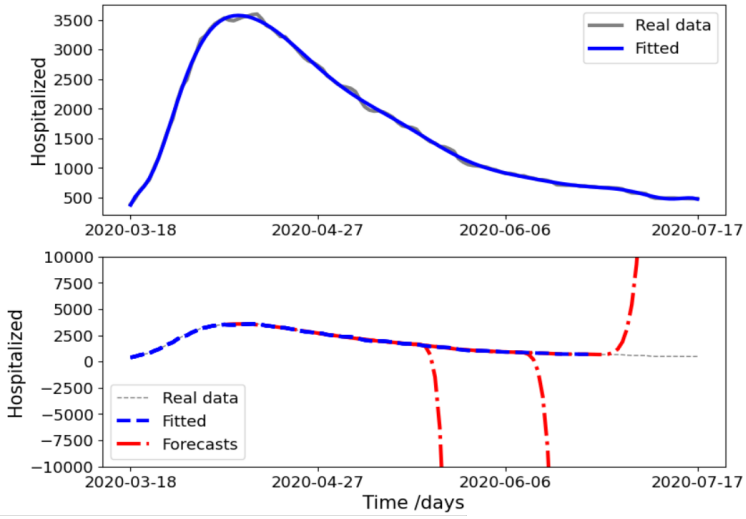


Fig. 4.10 Predictions and forecasts for polynomial method with various train periods, France.

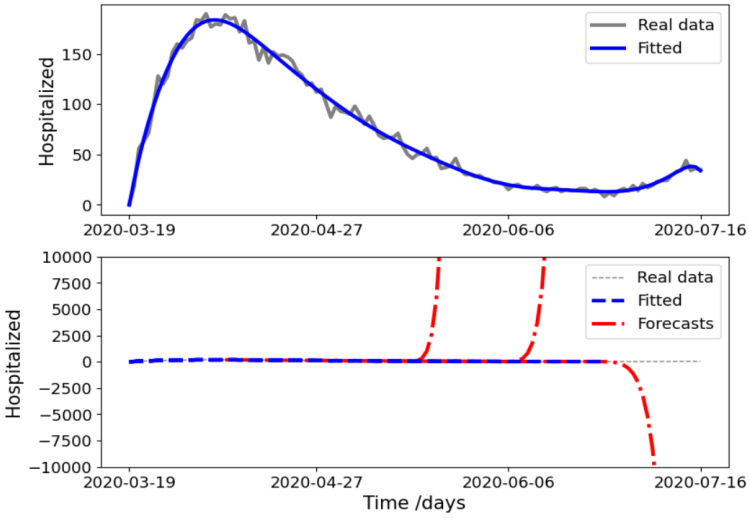


Fig. 4.11 Predictions and forecasts for polynomial method with various train periods, Luxembourg.

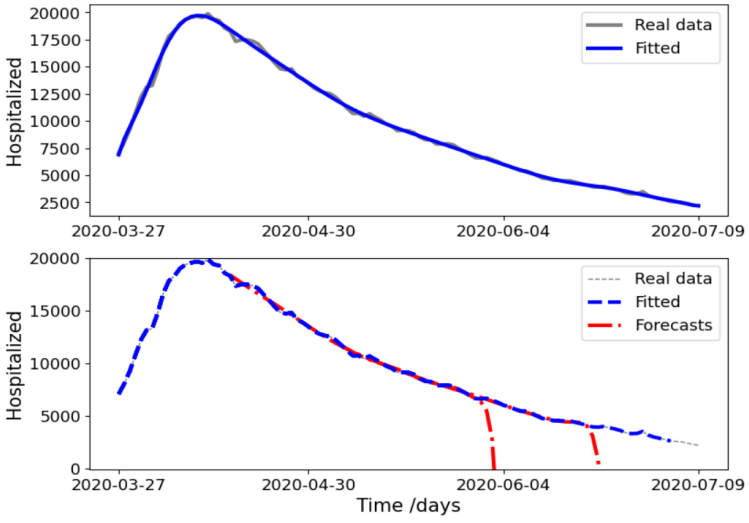


Fig. 4.12 Predictions and forecasts for polynomial method with various train periods, UK.

4.3 Exponential method

We consider the simple exponential (two parameters: initial value and decay rate) model for the decreasing exponential that can be defined as

$$H(t) = H(t_\mathrm{i}) \exp(-(t - t_\mathrm{i})/\tau). \tag{4.17}$$

The estimands are $H(t_\mathrm{i})$ and τ. In order to provide a first guess to the optimization solver, we can take the following initial parameters $H(t_\mathrm{i}) = H_\mathrm{o}(t_\mathrm{i})$ and $\tau = -(t_\mathrm{c} - t_\mathrm{i})/(\log(H_\mathrm{o}(t_c)) - \log(H(t_\mathrm{i}))$. The log has to be the natural logarithm (numpy.log, not the base-10 logarithm numpy.log10).

We compute the two estimands by minimizing

$$\phi(H(t_\mathrm{i}), \tau) = \sum_{t=t_\mathrm{i}}^{t_\mathrm{c}} (H(t) - H_\mathrm{o}(t))^2. \tag{4.18}$$

In order to give a sense of the sensitivity of the results, we superpose the curves obtained for different train periods where $t_\mathrm{c} = t_\mathrm{i} + 7$ and t_i is chosen arbitrary. We define the test periods in intervals form $[t_\mathrm{c}, t_\mathrm{e}]$ where t_e is the end time of the dataset.

Based on the forecast results in Figures 4.13–4.16, we can conclude that with an exponential model we get rather good results in the decreasing phase.

MAPE_test results in Table 4.4 show that in Belgium we have a fitting error (MAPE_test) which is remarkably small (about 12.8%) when the train period is taken just after three weeks (interval $[43, 50]$) of the peak (22^{th} to the beginning). For France in Table 4.5, we have a fitting error (MAPE_test) remarkably small (under 30%) when the beginning of the train period is taken after two weeks ($\geq$ 43) of the peak (27^{th} to the beginning). As for Luxembourg in Table 4.6, we have a fitting error (MAPE_test) remarkably small (about 14%) when the train period is taken after one week of the peak (16^{th} to the beginning) in the interval $[22, 29]$. Finally for UK in Table 4.7, we have all MAPE_test's remarkably small (under 17.2%). We observe that this method makes very good predictions on the decreasing phase in all graphs of three countries (Belgium, France, and UK) while the predictions are not good in all decreasing phase from Luxembourg because of the curve irregularity (or the small population in this country).

In Figures 4.13–4.16, we see that the forecasts are rather good in the

ascending and descending phases, but not at the peak level. Some explanations of these results can be based on the SH model. Let's recall the system (3.1) defined in chapter 3 where we replace I by H

$$\dot{S}(t) = -\frac{\beta}{N} S(t)H(t) \tag{4.19a}$$

$$\dot{H}(t) = \frac{\beta}{N} S(t)H(t) - \gamma H(t) \tag{4.19b}$$

$$\dot{R}(t) = \gamma H(t), \tag{4.19c}$$

where β and γ are parameters. Equation (4.19b) can be re-written as follows

$$\begin{aligned} \dot{H}(t) &= \beta S(t)H(t) - \gamma H(t) \\ &= (\beta S(t) - \gamma)H(t) \\ &= \lambda_t H(t) \text{ where } \lambda_t = \beta S(t) - \gamma. \end{aligned} \tag{4.20}$$

If $\lambda_t = \lambda$ constant, then $\dot{H}(t) = \lambda H(t)$. Thus,

$$H(t) = \exp(\lambda t)H(0). \tag{4.21}$$

There are two periods where S varies little, leading to λ almost constant.

- At the beginning of a wave, few people are hospitalized ($H \approx 0$), so from equation (4.19a) $\dot{S}$ is small then S varies little.
- At the end of a wave, a very small number of patients is hospitalized ($H \approx 0$), so from equation (4.19a) $\dot{S}$ is small then S varies little.

4.4 Partial conclusion

We investigated the polynomial model and the exponential decrease model as alternative models to the SH model defined in chapter 3. Results show that the SH model gives more accurate forecasts than the two other models especially when the train period is chosen around the peak.

We also used the fatalities and the patients discharged from the hospital data in Belgium to compute parameters included in the SH model in the sense of least squares estimation. Due to the weekly variability of the data the least squares method is not very accurate, then it gives less good

Train interval	MAPE_test (%)
[22, 29]	501.8
[29, 36]	166.8
[36, 43]	14.5
[43, 50]	12.8
[50, 57]	35.07
[57, 64]	28.9
[64, 71]	32.5
[71, 78]	56.09
[78, 85]	35.9
[85, 92]	35.5

Table 4.4 Statistics for exponential method, Belgium (120 days).

Train interval	MAPE_test (%)
[22, 29]	278.4
[29, 36]	61.03
[36, 43]	35.8
[43, 50]	29.1
[50, 57]	20.5
[57, 64]	4.2
[64, 71]	4.9
[71, 78]	4.1
[78, 85]	1.8
[85, 92]	1.7

Table 4.5 Statistics for exponential method, France (122 days).

forecasts than the parameter estimation method defined in chapter 3 and Section 3.2 with the SH model, compare Figures 4.5, 4.8 and 3.3 (in chapter 3).

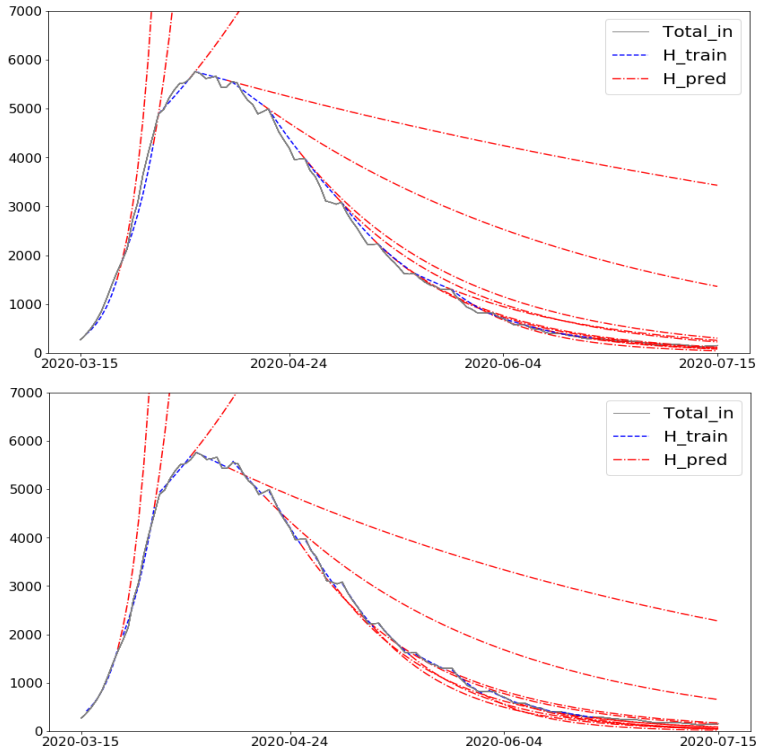


Fig. 4.13 Fitting and forecasting result of H_o with no optimized (in top) and optimized parameters (in bottom), Belgium total hospitalized.

Train interval	MAPE_test (%)
[22, 29]	14
[29, 36]	213.5
[36, 43]	32.6
[43, 50]	90.7
[50, 57]	16.84
[57, 64]	90.8
[64, 71]	15.4

Table 4.6 Statistics for exponential method, Luxembourg (99 days).

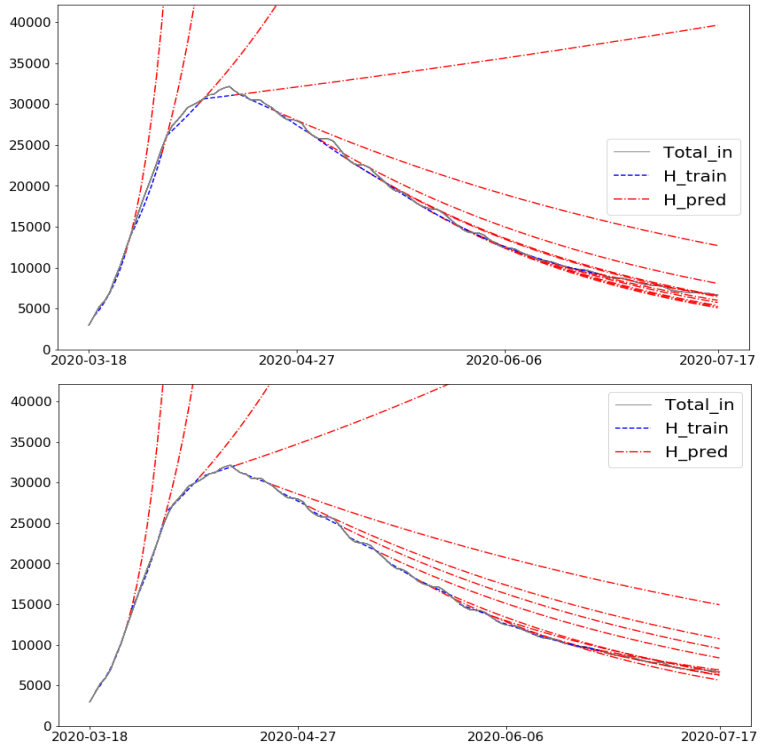


Fig. 4.14 Fitting and forecasting result of H_o with no optimized (in top) and optimized parameters (in bottom), French total hospitalized.

Train interval	MAPE_test (%)
[15, 22]	15.8
[22, 29]	10.7
[29, 36]	9.11
[36, 43]	17.2
[43, 50]	16.1
[50, 57]	15.3
[57, 64]	12.7
[64, 71]	10.8
[71, 78]	12.2

Table 4.7 Statistics for exponential method, UK (105 days).

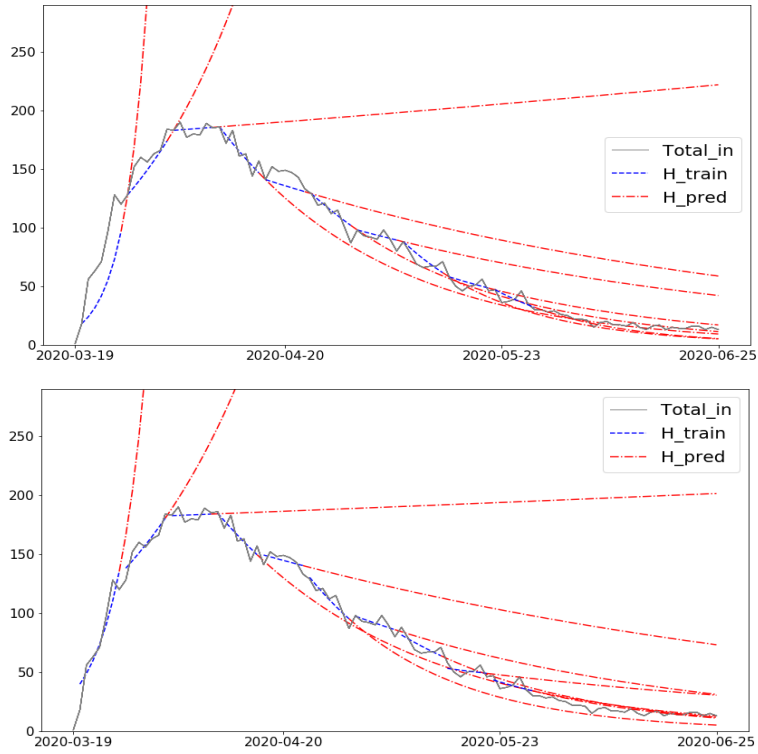


Fig. 4.15 Fitting and forecasting result of H_o with no optimized (in top) and optimized parameters (in bottom), Luxembourg total hospitalized.

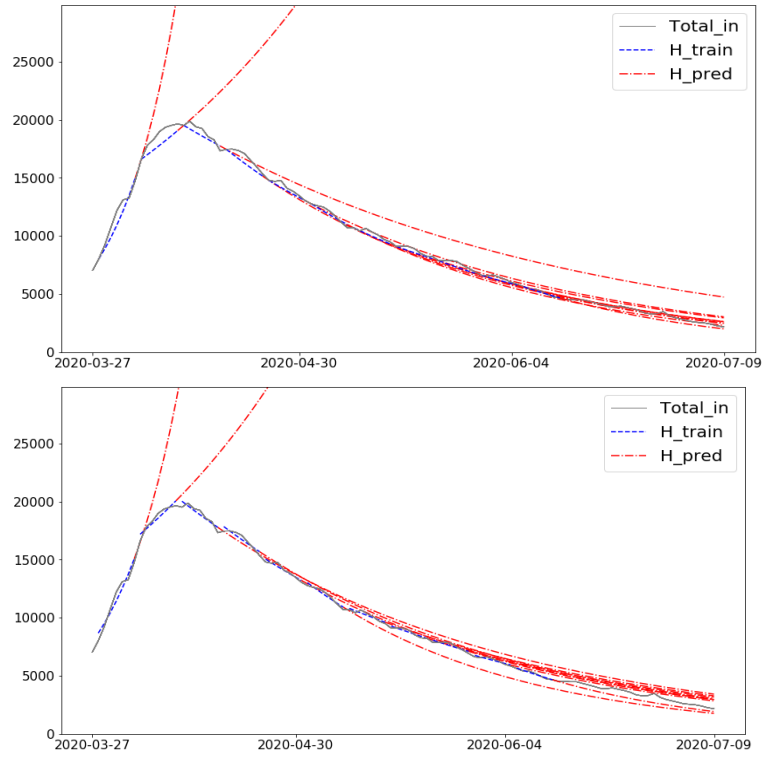


Fig. 4.16 Fitting and forecasting results of H_o with no optimized (in top) and optimized parameters (in bottom), the United Kingdom total hospitalized data.

5

Producing confidence intervals

THIS work is the outcome of a collaboration with Sylvie Van Schendel in the context of her master's thesis [VS22]. When forecasting an infectious disease such as COVID-19, we are facing uncertainty. Let us explain three possible sources of this uncertainty, according to [ZREG21]:

- Stochastic uncertainty: stochastic uncertainty arises from the uncertainty that an event may occur. In particular, the event that an infectious individual will contaminate a susceptible individual if they are in contact. The event that an infectious individual will be hospitalized. Or, the event that an infectious individual will recover.

- Parameter uncertainty: parameter uncertainty arises from the estimation of the two parameters of the SH model: $\bar{\beta}$ and γ.

- Model uncertainty: model uncertainty arises from the assumptions that have been made to build a model. For instance, to build the SH model, we have assumed that a constant fraction of the infected individuals will be hospitalized (see Section 3.1.1 in chapter 3). And, this is actually uncertain.

Therefore, while forecasting the hospitalizations due to COVID-19 with the SH model, it is very interesting to consider those different sources of uncertainty. To do so, we will produce confidence intervals, also called uncertainty intervals or credible intervals. When running the SH model several times, each time modifying the available data or the parameters that are given to the forecasting algorithm, the confidence interval corresponds to a range of values in which we expect the forecasts to fall. We will use different methods to produce those confidence intervals, according to the three sources of uncertainty. In the scientific literature, it is quite complicated to find out how the confidence intervals of the forecasts are produced, while using an SIR model or a derivative of it. Most of the time, no particular method is used but the statistical formula of a confidence interval, which is the following:

$$\mathrm{CI} = \bar{x} \pm z \frac{s}{\sqrt{n}} \tag{5.1}$$

where $\bar{x}$ is the sample mean, s is the standard deviation, n is the sample size and z has to be chosen according to the confidence level (e.g., $z = 1.96$ for a 95 % confidence interval).

However, in a paper about traffic forecasts [dJDP+07], some interesting information about the study of uncertainty is given. It is explained that different types of uncertainty exist, coming from either the model or the model inputs. Therefore, we will investigate the stochastic, the parameter and the model uncertainty. Moreover, to study the uncertainty coming from the model inputs, the most used method is based on a repeated model simulation. For each simulation of the model, the inputs are modified and give thus different output forecasts. Most of the time, this method ignores the correlation that exists between the inputs. And, even though this was about traffic forecasts and about a model that is totally different from the SH model, the repeated model simulation is the method we will use to measure uncertainty in the next two sections, to study stochastic and parameter uncertainty.

Finally, for the third method of this chapter, we will study the model uncertainty. This method consists to consider $\bar{\beta}$ non-constant. In a paper written by G. M. de Athayde and R. M. Damião [dAD20], we could find something similar. They have developed a derivative of the SIR model by considering deaths and by using time-varying parameters. Unfortunately, they didn't explain how the 95% confidence intervals are computed.

5.1 Stochastic uncertainty - method 1: a noised initial data set

Let us start by considering the stochastic uncertainty. This means considering the uncertainty that events may occur or not. To translate this uncertainty source into our SH model, we have to slightly modify the initial data given to the forecasting algorithm.

So, the first method we will use to produce confidence intervals is the following. Choosing the train period in the initial data set, we will do a first forecast of the COVID-19 hospitalizations. Then, we will choose new data points. Each new point will be chosen randomly, under a uniform distribution, in the interval build around the initial data point plus and minus ε. Then, we will produce a new forecasting curve, using a new train period based on those new noised data points. And finally, we will repeat this several times, using the same value for ε. Each time, the noise is randomly chosen to noisy the initial data points, and we will thus obtain a different forecasting curve. By doing this, we will see a confidence interval appear which will contain all those forecasting curves.

Now, the question is "how do we choose the value of ε?". The idea is to think about of how much we could slightly modify each initial data point, so that we still have data that follow the initial shape of the curve representing the initial data set. Therefore, let us choose the average gap between two adjacent data points:

$$\varepsilon := \frac{1}{n-1} \sum_{k=t_s}^{t_e-1} |H_o(k+1) - H_o(k)|,$$

where n is the total number of initial data points.

Using this idea, the obtained values of ε are: 69.161 for Belgium and 272.966 for France. The noised data points will thus be randomly chosen, under a uniform distribution, in the interval $[H_o(t) - \varepsilon, H_o(t) + \varepsilon] \quad \forall\, t \in [t_i, t_c]$ and with a value for ε that is different according to the country.

5.1.1 Local tests

From [ADD21], we know that the SH model gives better forecasts when the train period is located around a peak of the pandemic. Therefore, let us first try to obtain confidence intervals for train periods located around peaks.

For both countries, Belgium and France, we will use the data summed up over all districts, and choose a train period around the first peak of the pandemic, and then around the second to last peak. As a reference, let us consider the forecasting curves produced by the algorithm without modifying the initial data at all. Those are represented on Figures 5.1 and 5.2. We observe that the forecasts are slightly better for Belgium than for France. However, for both countries, let us see if we obtain confidence intervals that are preferably tight around those initial forecasting curves.

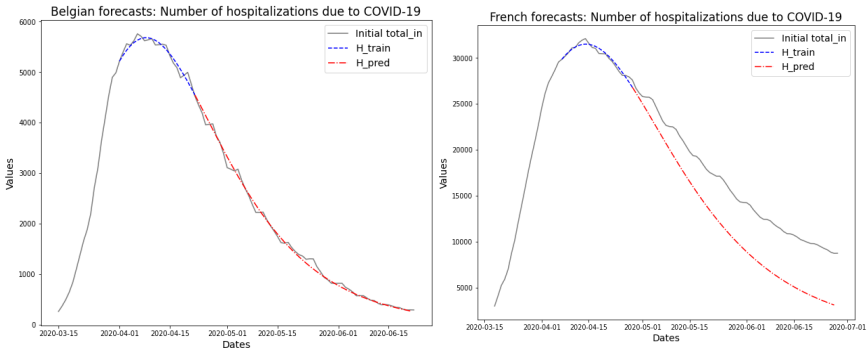


Fig. 5.1 Belgian and French COVID-19 hospitalization forecasts. Train period around the first peak.

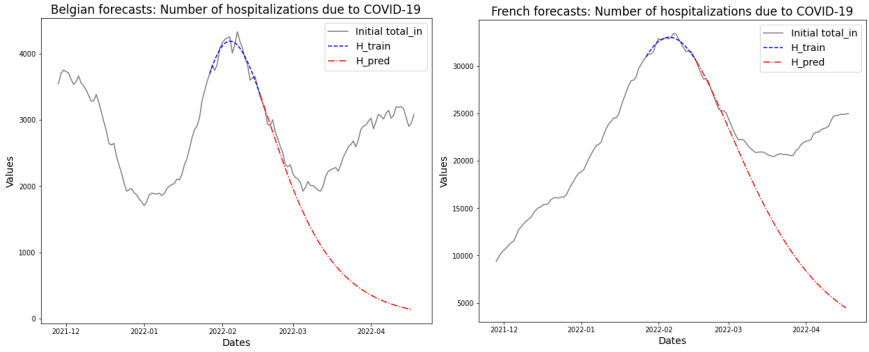


Fig. 5.2 Belgian and French COVID-19 hospitalization forecasts. Train period around the second to last peak.

On Figure 5.3 and 5.4, we can observe the results for Belgium and for France, choosing a train period around the first peak of the pandemic.

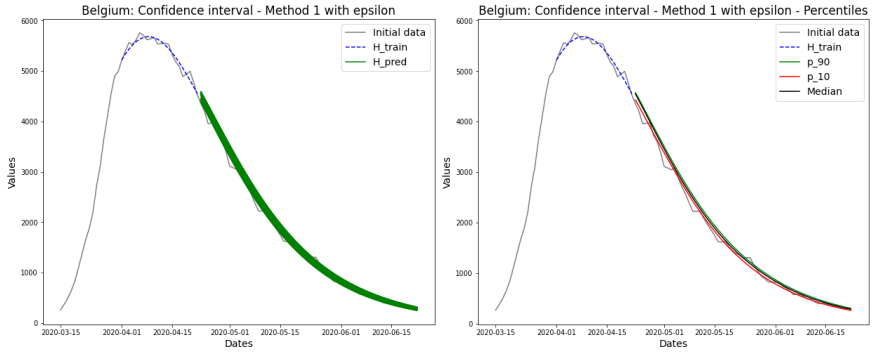


Fig. 5.3 Confidence interval and percentiles for the Belgian data, using method 1 with ε. Train period around the first peak.

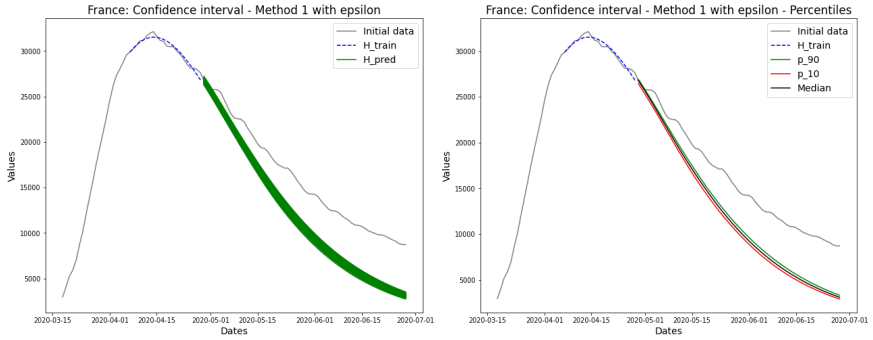


Fig. 5.4 Confidence interval and percentiles for the French data, using method 1 with ε. Train period around the first peak.

And, on Figures 5.5 and 5.6, we can observe the results for Belgium and for France, choosing a train period around the second to last peak of the pandemic.

The values of ε, equal to 69.161 for Belgium and to 272.966 for France, are small compared nation-wide which means that the initial data points are only slightly noised. As a result, the forecasting curves are close to each other as we can see on the left-side figures. The confidence interval is thus tight, as we had hoped for. Moreover, on the right-side figures, we can see the percentiles 10, 50 and 90. Note that a percentile is a value below which a percentage of data falls. For instance, the green curve p_{90} on the right-side figures is the value below which 90% of the forecasts fall. The percentiles

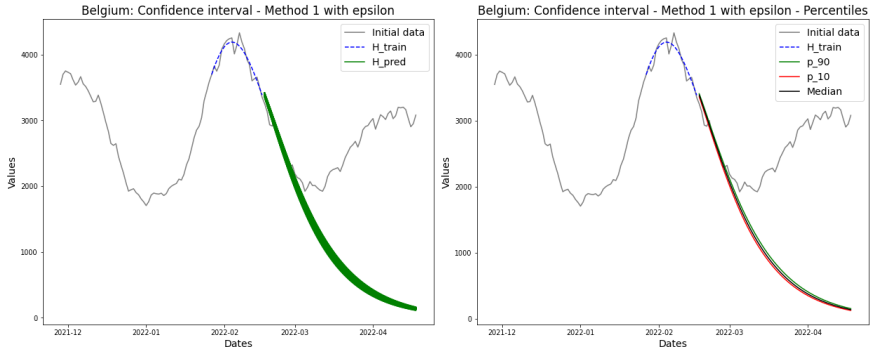


Fig. 5.5 Confidence interval and percentiles for the Belgian data, using method 1 with ε. Train period around the second to last peak.

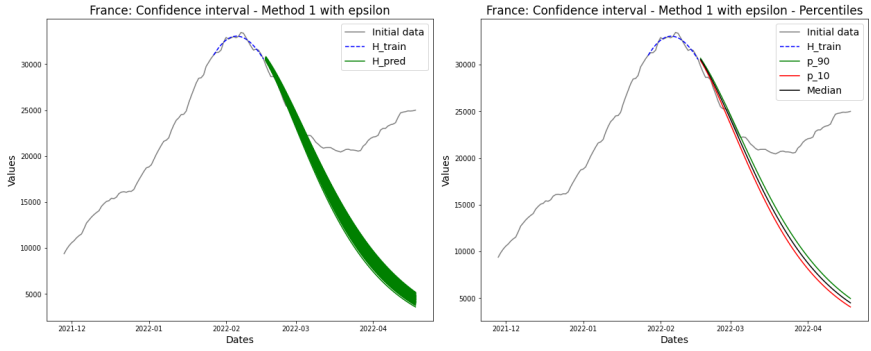


Fig. 5.6 Confidence interval and percentiles for the French data, using method 1 with ε. Train period around the second to last peak.

10, 50 and 90 are also very close to each other since the confidence interval is tight.

5.1.2 Global results

Now that we have done some tests on particular peaks of the pandemic, let us have a look at the entire data set in order to have a more global view of the SH model and its confidence intervals. We would like to know how the forecasts given by our SH model behave when considering the stochastic uncertainty, and even when choosing train periods that are not located around peaks.

So, the idea is the following. We will consider a train period of 14 days long, and a test period of 60 days. Then, we will have a particular look at a point that is 15 days after the end of the train period, and thus located in the test period. This 15 days period is called the horizon. Based on the chosen train period, we will produce several forecasting curves with the method explained previously. For our particular point located at a horizon from the train period, we will thus obtain several forecasts and will thus be able to produce the percentiles of those results. This will be done for the entire data set, and gives us a recapitulating graph about the confidence intervals of the forecasts produced by the SH model. For execution time reasons, the forecasts and their percentiles have not been computed for every single day of the data set but for one day a week.

The obtained results for Belgium are represented on Figures 5.7, 5.8 and 5.9.

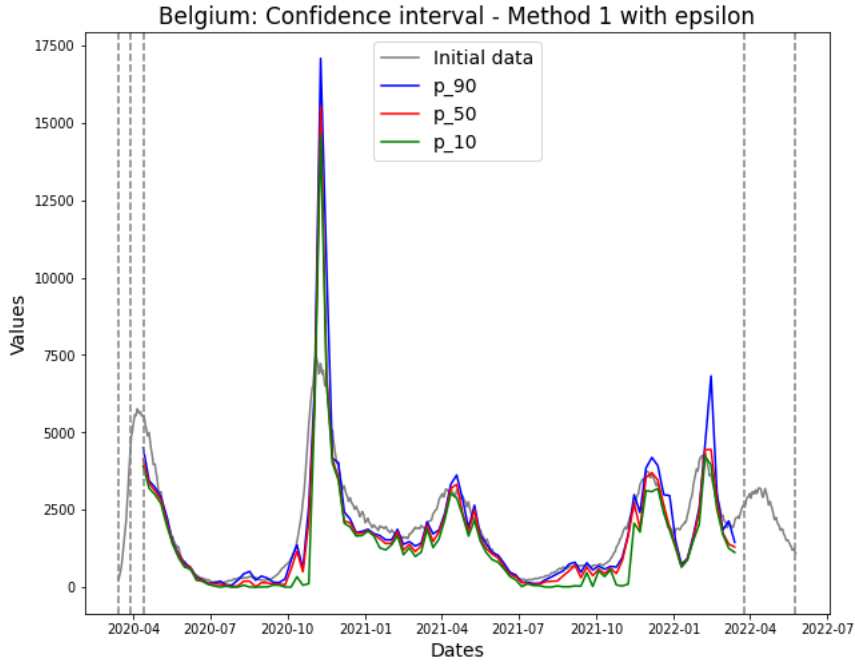


Fig. 5.7 Confidence intervals for the Belgian data, using method 1 with ε. Percentiles 10, 50 and 90.

On Figure 5.7, the percentiles 10, 50 and 90 of the hospitalization fore-

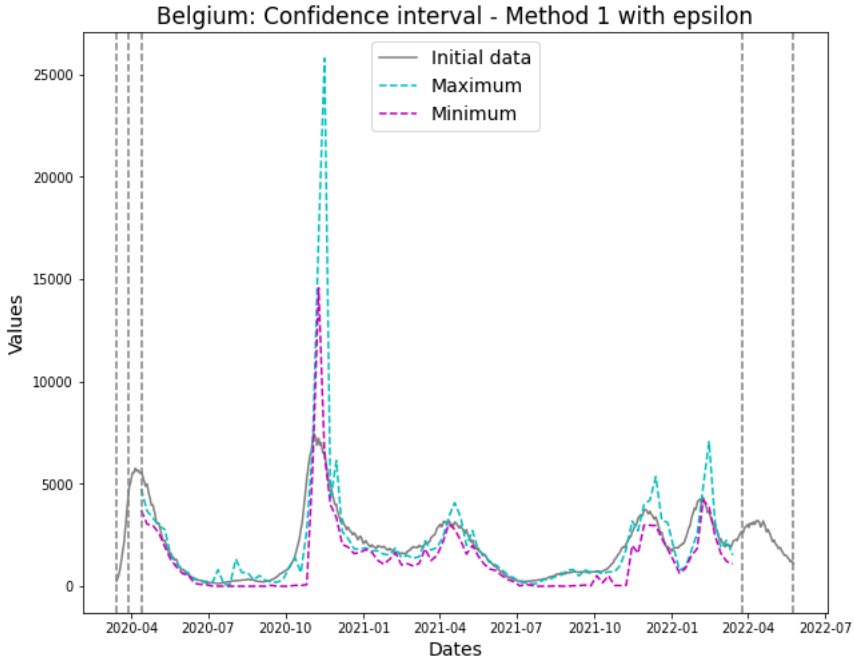


Fig. 5.8 Confidence intervals for the Belgian data, using method 1 with ε. Minimum and maximum.

casts are represented, and on Figure 5.8, the minimum and maximum forecasts. The results are quite good except for the second peak of the pandemic where the forecasts are much higher than what they should be. This is actually not really surprising since those forecasts are based on train periods located in an increasing phase of the number of hospitalizations. And, from [ADD21], we know that those train periods don't give accurate forecasts. Note that 30.69% of the real data are under p_{90}, 26.73% of the real data are between p_{10} and p_{90}, and 34.65% of the real data are between the minimum and the maximum curves. And finally, on Figure 5.9, we can observe the standard deviation between the initial data set and the median of the forecasts produced by the SH model, which corresponds to the percentile 50 represented on Figure 5.7. This last figure confirms what has been observed earlier. We have small values for the standard deviation, which means that the forecasts are close to the real data, except for the peaks of the pandemic in particular for the second peak which is very

high.

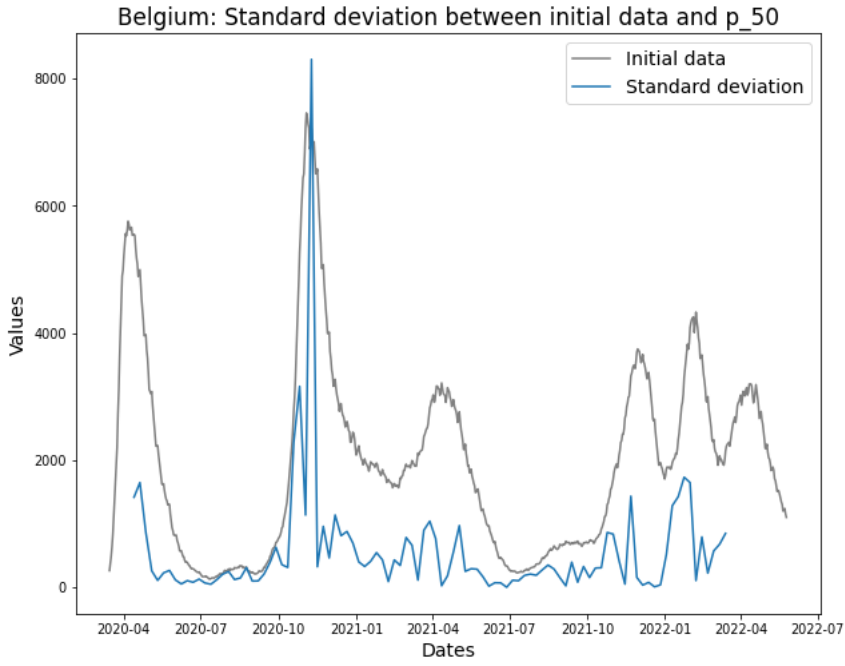


Fig. 5.9 Confidence intervals for the Belgian data, using method 1 with ε. Standard deviation between the median forecasts and the initial data set.

Now, let us observe the obtained results for France, represented on Figures 5.10, 5.11 and 5.12.

On Figure 5.10, the percentiles 10, 50 and 90 of the hospitalization forecasts are represented, and on Figure 5.11, the minimum and maximum forecasts. Similarly to Belgium, the results are quite good except for the peaks of the pandemic where the forecasts are higher than what they should be. Note that 34.65% of the real data are under p_{90}, 16.83% of the real data are between p_{10} and p_{90}, and 25.74% of the real data are between the minimum and the maximum curves. And finally, on Figure 5.12, we can observe the standard deviation between the initial data set and the median of the forecasts. Once again, this last figure confirms what has been observed earlier. We have higher values for the standard deviation at the different peaks of the pandemic, meaning that the gap between the median forecasts and the real data is larger when using a train period in an increasing phase.

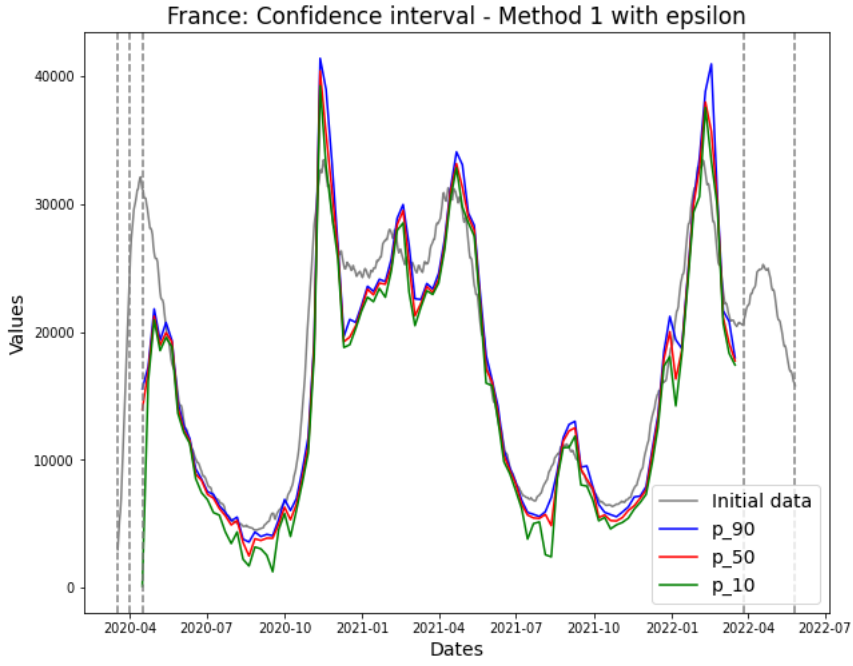


Fig. 5.10 Confidence intervals for the French data, using method 1 with ε. Percentiles 10, 50 and 90.

5.2 Parameter uncertainty - method 2: a sliding train period

Now, let us consider another source of uncertainty: the parameter uncertainty, which arises from the estimation of the two parameters of the SH model, γ and $\bar{\beta}$.

For this second method, let us have a closer look on the impact that the parameter estimation has on the COVID-19 hospitalization forecasts. Therefore, we will use a method that works with a sliding train period. The idea is the following. We consider a first train period, on which we estimate the model parameters γ and $\bar{\beta}$, and we observe the first forecasting curve produced by our algorithm. Then, we shift the train period by one day. This new train period will give us a new set of estimated parameters $(\gamma, \bar{\beta})$ and thus a new forecasting curve. Repeating this several times, each time

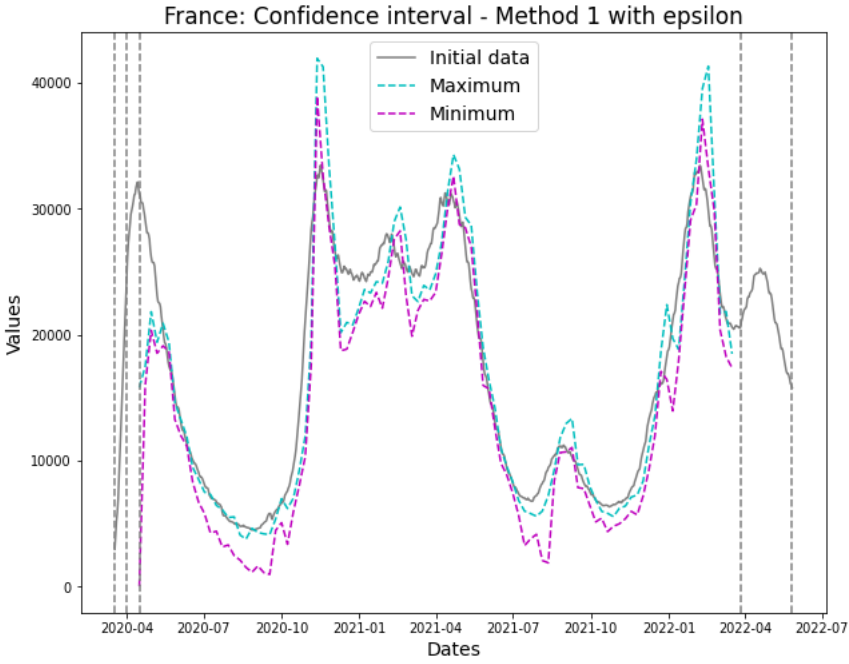


Fig. 5.11 Confidence intervals for the French data, using method 1 with ε. Minimum and maximum.

shifting the train period by one day, we will obtain an interval of values for the estimated parameters γ and $\bar{\beta}$ of the SH model, and a confidence interval for the COVID-19 hospitalization forecasts on the test set.

5.2.1 Local tests

As we did for the first method, let us first try to obtain confidence intervals with a train period located around the first peak of the pandemic. And for both countries, Belgium and France, we will use the data summed up over all districts.

The obtained results for Belgium are represented on Figure 5.13. On the left-side figure, we can observe the forecasting curves. Each curve corresponds to one train period, and thus to one set $(\gamma, \bar{\beta})$ of parameters. The train period is sliding. And, on the right-side figure, we observe the per-

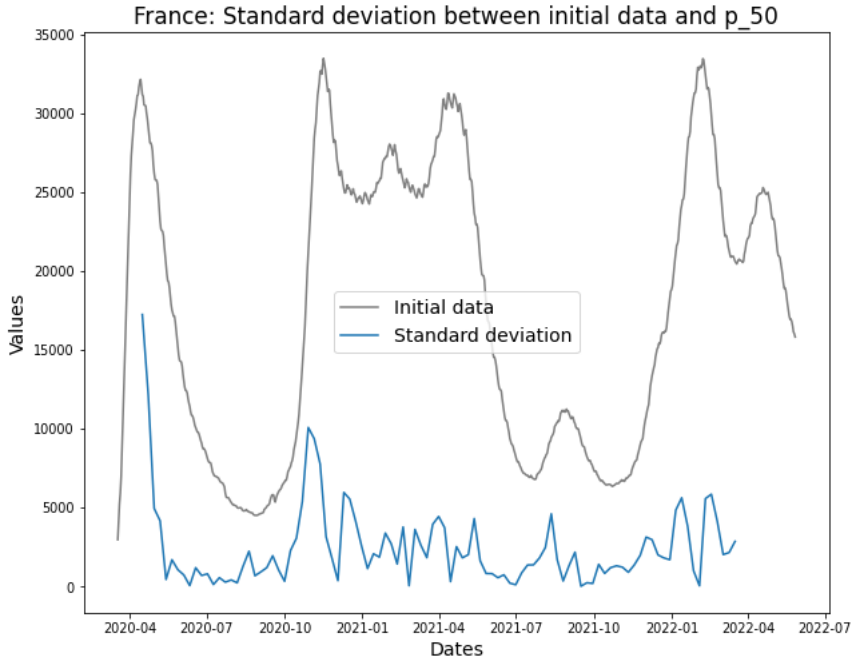


Fig. 5.12 Confidence intervals for the French data, using method 1 with ε. Standard deviation between the median forecasts and the initial data set.

centiles 10, 50 and 90 of the forecasting curves. Those percentiles are represented only on the time period that is common to all the forecasting curves. This time period is shorter than the two-months test set since we use a sliding train period. However, it is sufficient to observe that the confidence interval containing the forecasting curves is really tight.

Now, let us see how the model parameters behave with this sliding train period. The values of γ are in between 0.0713 and 0.0783, and the values of $\bar{\beta}$ are in between $4.769 \cdot 10^{-6}$ and $9.201 \cdot 10^{-6}$. The values of the parameters don't vary a lot. They are not very impacted by the exact location of the train period. And so, the forecasting curves are neither impacted. Moreover, the MAE on the test set, between forecasts and real data, has its value in between 52.363 and 136.087. In other words, the value of the MAE remains low compared nation-wide, and thus the confidence interval containing the forecasting curves is very tight and close to the real data.

Afterwards, let us observe the results obtained for France that are rep-

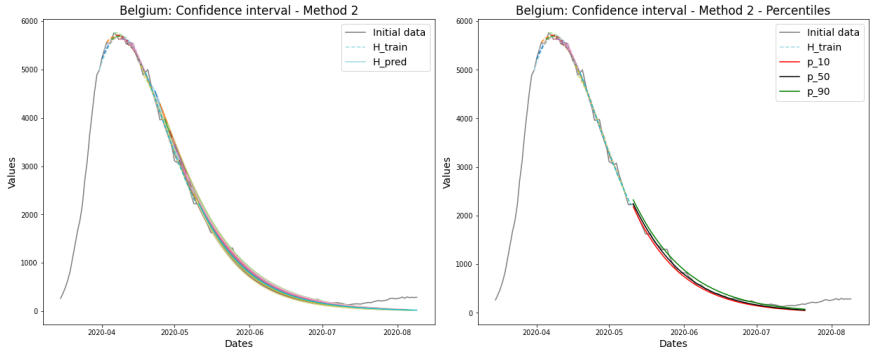


Fig. 5.13 Confidence interval and percentiles for the Belgian data, using method 2. Train period around the first peak.

resented on Figure 5.14.

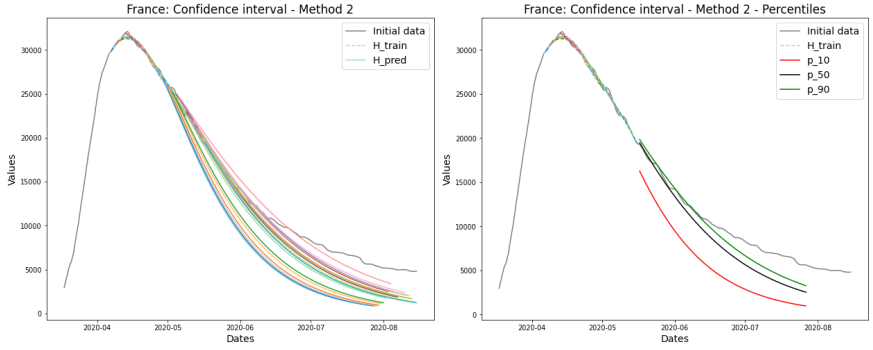


Fig. 5.14 Confidence interval and percentiles for the French data, using method 2. Train period around the first peak.

Once again, we observe the forecasting curves on the left-side figure, and their percentiles on the right-side figure. Contrary to Belgium, we can observe that the confidence interval containing the COVID-19 hospitalization forecasts is much larger. This comes from the estimated parameters of the SH model which are more sensitive to the sliding train period. The value of γ varies in between 0.0399 and 0.0569, and $\bar{\beta}$ varies between $2.866 \cdot 10^{-7}$ and $1.212 \cdot 10^{-6}$. This variation has an impact on the forecasting curves, and has for consequence a larger confidence interval.

5.2.2 Global results

Now, let us have a more global view of the SH model and its confidence intervals by looking at the entire data set. As for the first method, we would like to know how the forecasts behave when considering the uncertainty source, here caused by parameter estimation, and even when choosing train periods that are not located around peaks.

So, once again, we will consider a train set of 14 days long, and a test set of 60 days. Then, we will have a particular look at a point that is at a horizon of 15 days after the end of the train period. Based on the chosen train period that will slide by one day each time, we will produce several forecasting curves. For our particular point located at a horizon from the train period, we will thus obtain several forecasts and will thus be able to produce the percentiles of those results. This will be done for the entire data set, and gives us a recapitulating graph about the confidence intervals of the forecasts produced by the SH model. And for the same reasons as those for the first method, the forecasts and their percentiles have not been computed for every single day of the data set but for one day a week.

The obtained results for Belgium are represented on Figures 5.15, 5.16 and 5.17.

Note that 40.59% of the real Belgian data are under the percentile 90, and 37.62% are between the percentiles 10 and 90, represented on Figure 5.15. 49.50% of the real data are between the minimum and maximum forecasting curves, represented on Figure 5.16. And, the average gap between the real data and the percentile 50 is equal to 498.957. Once again, on Figure 5.17, we can observe a low standard deviation between the real data and the percentile 50, except for the peaks of the pandemic, especially the second peak which is very high.

Now, for France, the results are represented on Figures 5.18, 5.19 and 5.20.

For France, 38.61% of the real data are under the percentile 90, and 32.67% are between the percentiles 10 and 90, represented on Figure 5.18. 42.57% of the real data are between the minimum and maximum forecasting curves, represented on Figure 5.19. And, the average gap between the real data and the percentile 50 is equal to 2165.970. Moreover, on Figure 5.20, we can notice that the standard deviation between the real data and the percentile 50 remains low, even for the second peak of the pandemic. This actually confirms what we observe on Figure 5.18: the median of the hospitalization forecasts remains close to the real data.

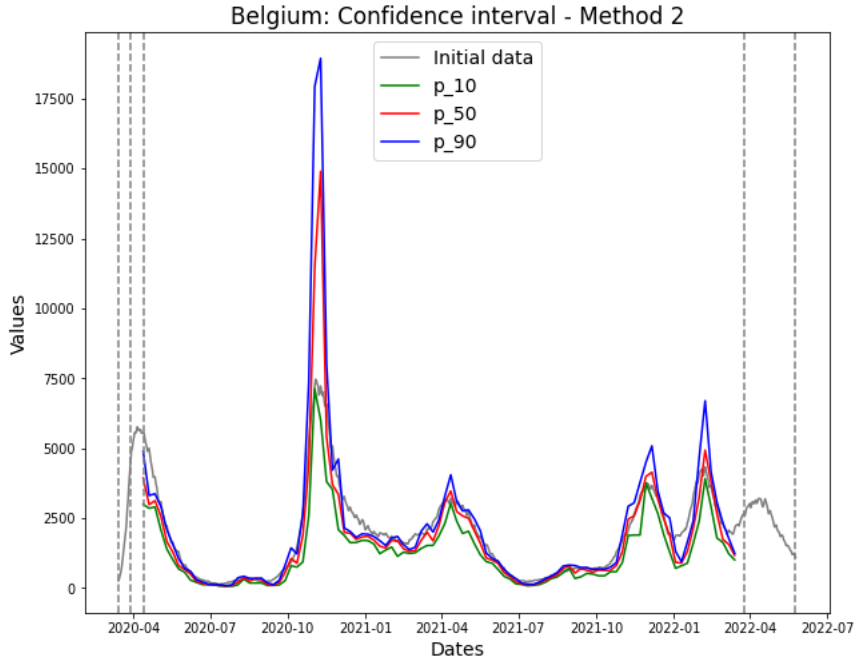


Fig. 5.15 Confidence intervals for the Belgian data, using method 2. Percentiles 10, 50 and 90.

5.3 Model uncertainty - method 3: a non-constant value for $\bar{\beta}$

Now, let us take into account the model uncertainty which is probably the most difficult uncertainty source to handle. Indeed, it could question every assumption that has been done to build our SH model. In this section, we conduct proof-of-concept experiments where $\bar{\beta}$ is allowed to vary linearly with t. Further work could consider other laws for the variation of $\bar{\beta}$, variations of other parameters, and even other models than the SH model, such as those mentioned in Section 1.2.2.

For this third method, the first idea would be to change the following assumption defined in Section 3.1.1 (chapter 3):

$$H(t) = \alpha I(t),\ \alpha = cst,\ \forall t,$$

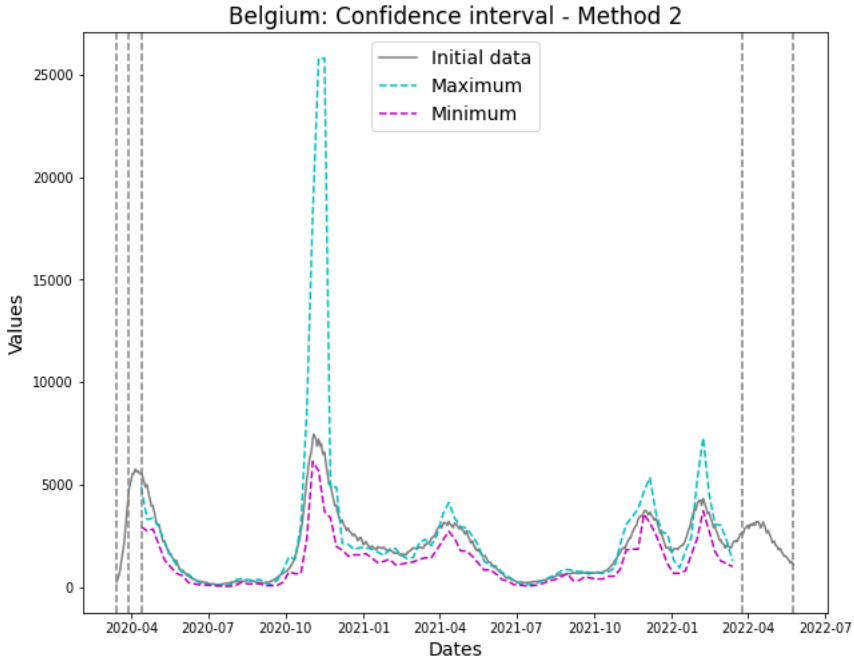


Fig. 5.16 Confidence intervals for the Belgian data, using method 2. Minimum and maximum.

meaning that a constant fraction of the infected individuals will be hospitalized. To do so, we could choose a non-constant value for α by considering α as a linear time-dependent value:

$$\alpha(t) = \alpha_0 + \alpha_1 t \tag{5.2}$$

which means that the number of infected individuals that will be hospitalized depends on time.

However, this isn't so simple since α is involved in the definition of the parameter $\bar{\beta}$ with N which is unknown to the SH model:

$$\bar{\beta} := \frac{\beta}{N\alpha}.$$

So, instead of considering a non-constant value for α, let us try to choose a non-constant value for $\bar{\beta}$ by considering a linear-time dependent value

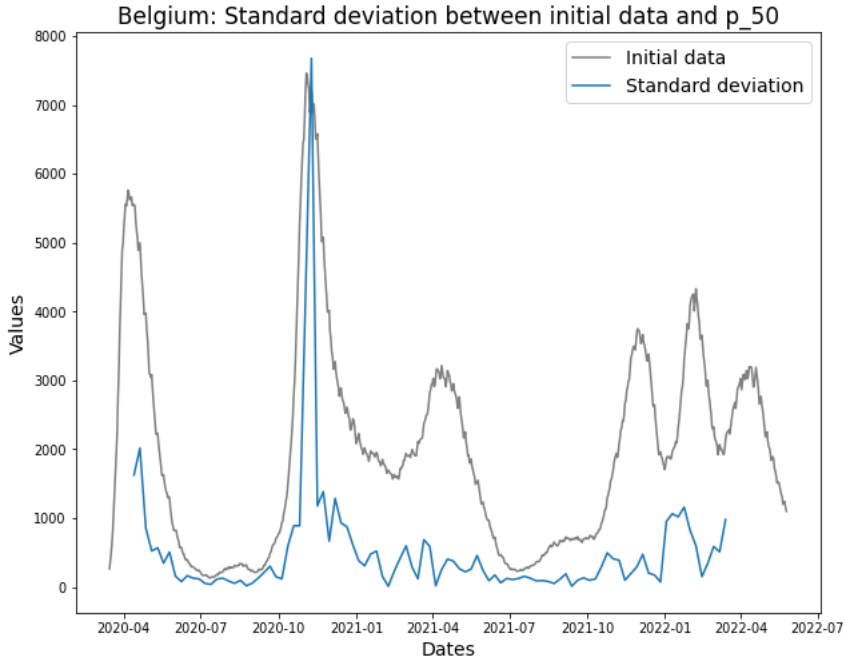


Fig. 5.17 Confidence intervals for the Belgian data, using method 2. Standard deviation between the median forecasts and the initial data set.

for it:

$$\bar{\beta}(t) = \bar{\beta}_0 + \bar{\beta}_1 t, \ \forall t. \tag{5.3}$$

Taking this new assumption into account, the SH model can be rewritten as:

$$\dot{\bar{S}}(t) = -\bar{\beta}(t)\bar{S}(t)H(t) \tag{5.4}$$

$$\dot{H}(t) = \bar{\beta}(t)\bar{S}(t)H(t) - \gamma H(t) \tag{5.5}$$

with $\bar{\beta}(t) = \bar{\beta}_0 + \bar{\beta}_1 t \quad \forall t$.

The initial state $H(t_i)$ and the parameter γ can both be estimated as previously (see Sections 3.2.2 and 3.2.3, in chapter 3). However, for $\bar{S}(t_i)$ and $\bar{\beta}$, we will have to estimate $\bar{S}(t_i)$, $\bar{\beta}_0$ and $\bar{\beta}_1$, giving us to minimize ϕ with those 3 estimands as arguments of the function. The optimization solver that will minimize ϕ needs convenient initial guesses for those 3 ar-

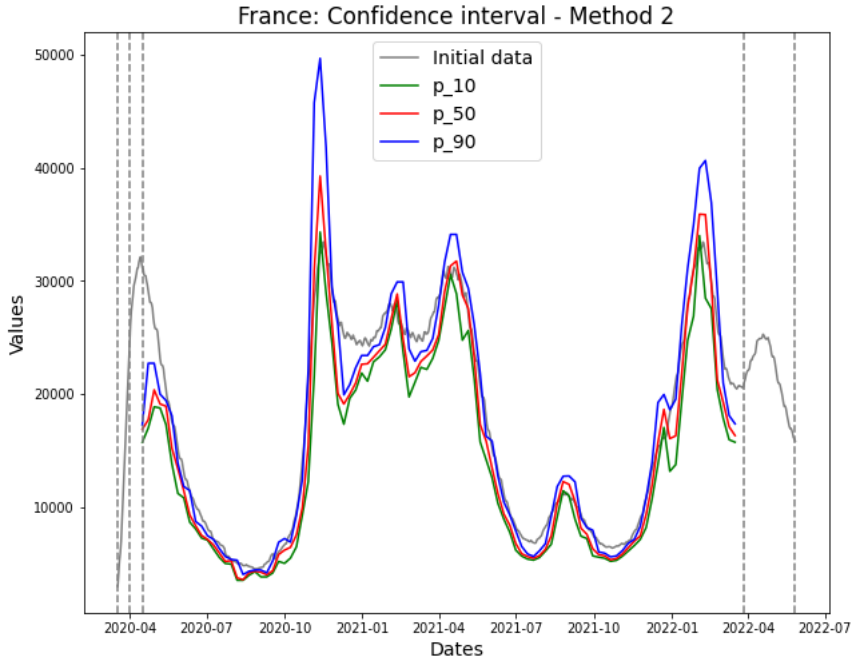


Fig. 5.18 Confidence intervals for the French data, using method 2. Percentiles 10, 50 and 90.

guments. To do so, we will use the values of $\bar{\beta}$ considered in the initial SH model with a constant $\bar{\beta}$; its initial guess computed as explained in Section 3.3.1 (chapter 3) and its optimal value computed by the optimization solver. So, for $\bar{\beta}_0$, we have chosen the initial guess of $\bar{\beta}$:

$$\hat{\bar{\beta}}_0 = \bar{\beta}_{\text{initial guess}}. \tag{5.6}$$

And, using the optimal value of the constant $\bar{\beta}$ and using (5.3), we have chosen the initial guess of $\bar{\beta}_1$ as:

$$\hat{\bar{\beta}}_1 = \frac{\bar{\beta}_{\text{optimal}} - \hat{\bar{\beta}}_0}{t_i}. \tag{5.7}$$

Finally, using those initial guesses of $\bar{\beta}_0$ and $\bar{\beta}_1$, we could obtain the initial

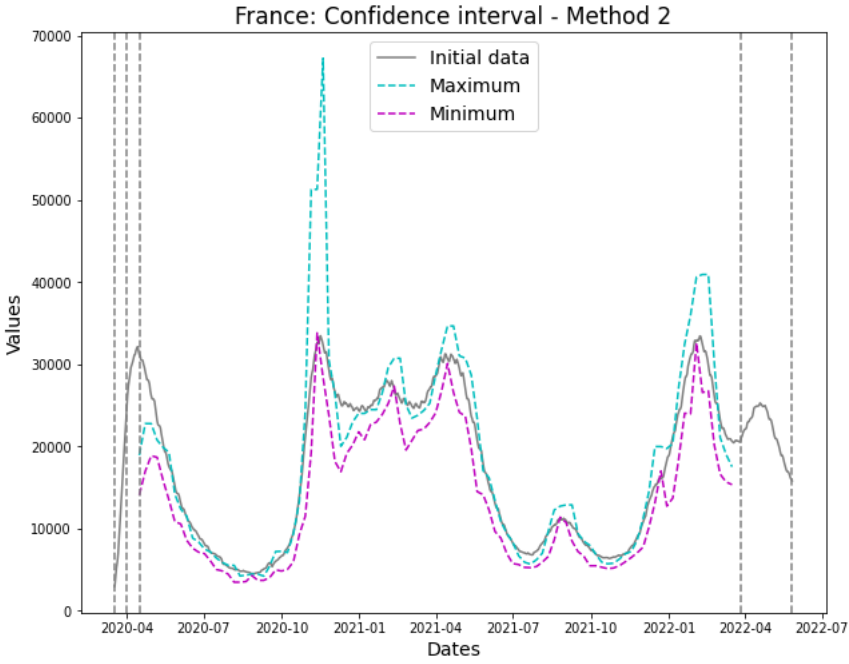


Fig. 5.19 Confidence intervals for the French data, using method 2. Minimum and maximum.

guess of $\bar{S}(t_i)$:

$$\hat{\bar{S}}(t_i) = \frac{E_o(t_i)}{(\hat{\bar{\beta}}_0 + \hat{\bar{\beta}}_1 t_i) H_o(t_i)}. \tag{5.8}$$

5.3.1 Local tests

Now, let us try to obtain confidence intervals with a train period located around the first peak of the pandemic. For both countries, Belgium and France, we will use the data summed up over all districts.

So, as explained in the method, the first step is to estimate $H(t_i)$ and γ; then, to find initial guesses for $\bar{\beta}_0$, $\bar{\beta}_1$ and $\bar{S}(t_i)$; and finally to obtain the optimal values of $\bar{\beta}_0$, $\bar{\beta}_1$ and $\bar{S}(t_i)$ by minimizing the function ϕ with an optimization solver. After that, we are able to produce a confidence interval. Indeed, knowing $\bar{\beta}_0$ and $\bar{\beta}_1$, we obtain a different value of $\bar{\beta}(t)$ for

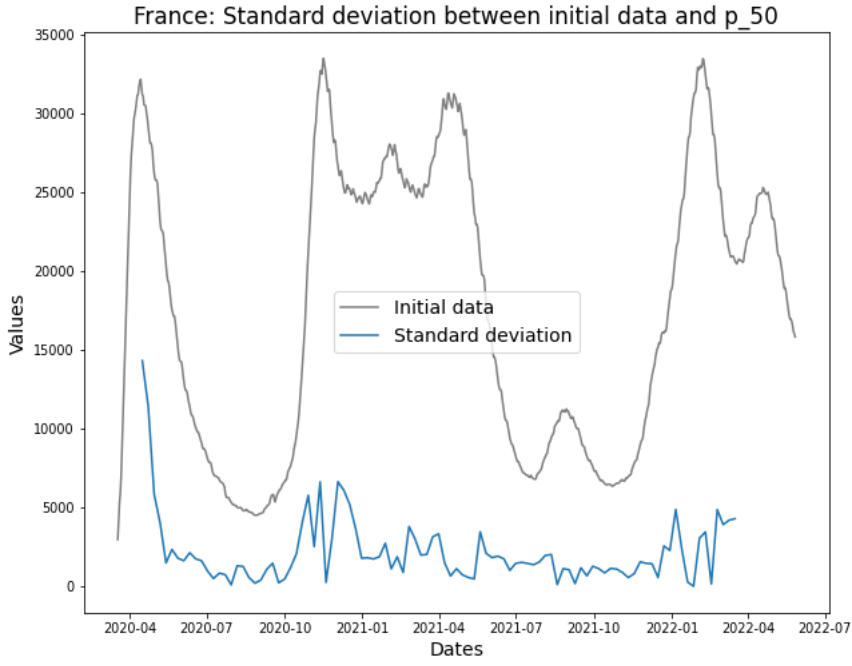


Fig. 5.20 Confidence intervals for the French data, using method 2. Standard deviation between the median forecasts and the initial data set.

each different t ($t \in [t_i, t_c]$). So, each value of $\bar{\beta}(t)$ will produce a different forecasting curve and the confidence interval will contain all those curves.

The obtained results for Belgium are represented on Figure 5.21. On the left-side figure, we can observe the forecasting curves. And, on the right-side figure, we observe the percentiles 10, 50 and 90 of those forecasting curves.

For Belgium, the optimal values of $\bar{\beta}_0$ and $\bar{\beta}_1$ are:

$$\bar{\beta}_0 = 6.996 \cdot 10^{-6},$$
$$\bar{\beta}_1 = 1.646 \cdot 10^{-19}.$$

Because the value of $\bar{\beta}_1$ is so small, the value of $\bar{\beta}(t)$ don't vary a lot and is close to the value of $\bar{\beta}_0$. Indeed, $\bar{\beta}(t)$ varies between:

$$6.995957786616053 \cdot 10^{-6} \quad \leq \quad \bar{\beta}(t) \quad \leq \quad 6.995957786619511 \cdot 10^{-6}.$$

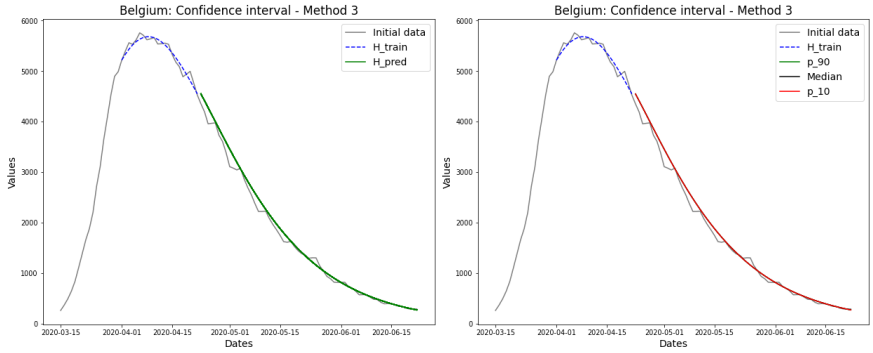


Fig. 5.21 Confidence interval and percentiles for the Belgian data, using method 3. Train period around the first peak.

This explains what we can observe on Figure 5.21. The forecasting curves are almost identical, and we cannot distinguish the percentiles 10, 50 and 90 from each other.

For France, we obtain similar results. Those are represented on Figure 5.22, with the forecasting curves on the left-side figure and its percentiles on the right-side figure.

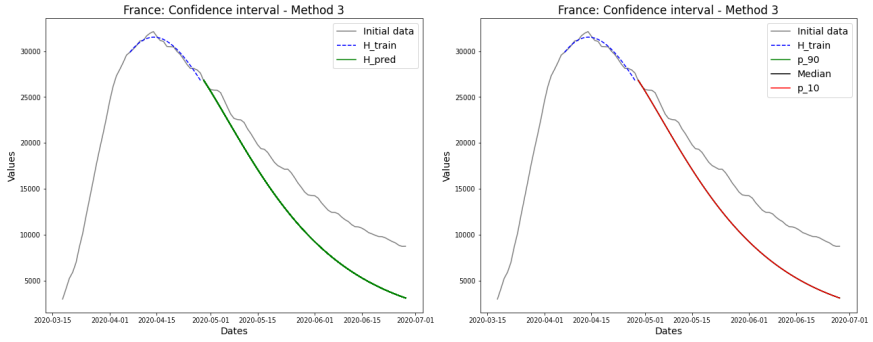


Fig. 5.22 Confidence interval and percentiles for the French data, using method 3. Train period around the first peak.

For France, the optimal values of $\bar{\beta}_0$ and $\bar{\beta}_1$ are:

$$\bar{\beta}_0 = 1.103 \cdot 10^{-6},$$
$$\bar{\beta}_1 = 3.590 \cdot 10^{-22}.$$

Once again, the value of $\bar{\beta}_1$ is close to zero which has as consequence that $\bar{\beta}(t)$ don't vary a lot. Here, $\bar{\beta}(t)$ varies between:

$$1.1033424629152106 \cdot 10^{-6} \quad \leq \quad \bar{\beta}(t) \quad \leq \quad 1.1033424629152182 \cdot 10^{-6}.$$

This is why the forecasting curves are almost identical on Figure 5.22. The confidence interval is thus so tight that it looks like one single forecasting curve.

5.3.2 Global results

Now, let us have a more global view of the SH model and its confidence intervals by looking at the entire data set. We would like to know how the forecasts behave when considering the model uncertainty, even when choosing train periods that are not located around peaks. We would like to see if there is a better distinction between the forecasting curves when $\bar{\beta}(t)$ varies on other train periods.

As for methods 1 and 2, we will consider a train set of 14 days long, and a test set of 60 days. Then, we will have a particular look at a point that is at a horizon of 15 days after the end of the train period. Based on the chosen train period, we will produce several forecasting curves according to the value of $\bar{\beta}(t)$ which is different for each t in the train set. For our particular point located at a horizon from the train period, we will thus obtain several forecasts and will thus be able to produce the percentiles of those results. This will be done for the entire data set, and gives us a recapitulating graph about the confidence intervals of the forecasts produced by the SH model. And once again, the forecasts and their percentiles have not been computed for every single day of the data set but for one day a week.

The results obtained for Belgium are represented on Figures 5.23, 5.24 and 5.25.

For Belgium, 25.74% of the real data are under the percentile 90, and 4.95% are between the percentiles 10 and 90, represented on Figure 5.23. 7.92% of the real data are between the minimum and the maximum forecasting curves, represented on Figure 5.24. Those percentages are quite low compared to the ones we have obtained with methods 1 and 2. This confirms our observations from the local test (see section 5.3.1). For most of the train periods, $\bar{\beta}(t)$ don't vary a lot because $\bar{\beta}_1$ is close to zero. The forecasting curves are thus very close to each other, and the confidence in-

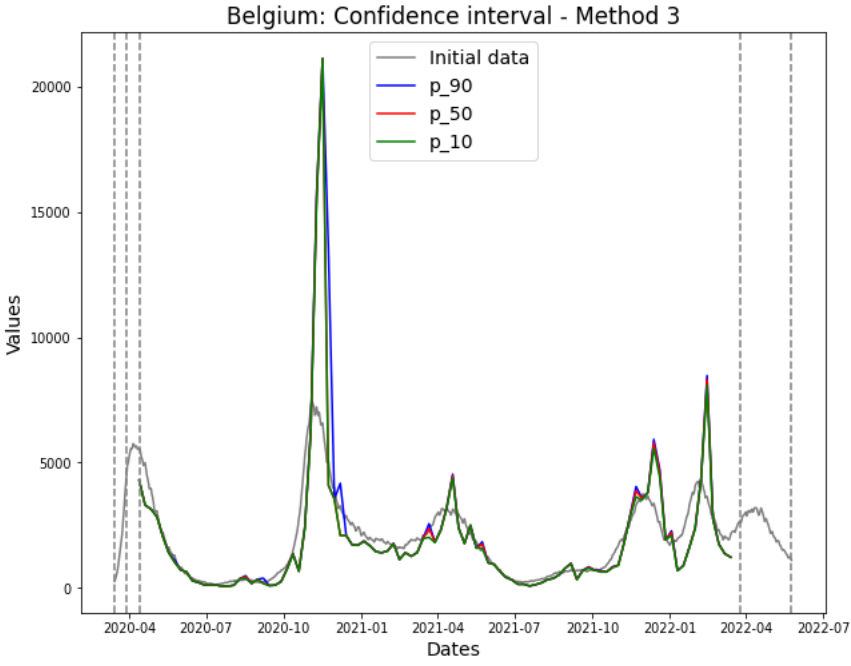


Fig. 5.23 Confidence interval for the Belgian data, using method 3. Percentiles 10, 50 and 90.

terval is really tight. And, on Figure 5.25, we can observe a low standard deviation between the real data and the percentile 50, except for the peaks of the pandemic, especially the second peak which is very high. This is similar to what we have obtained with the previous methods.

Now, for France, the results are represented on Figures 5.26, 5.27 and 5.28.

For France, the values of $\bar{\beta}(t)$ vary even less. 28.71% of the real data are under the percentile 90, and 0% are between the percentiles 10 and 90, represented on Figure 5.26. 0% of the real data are between the minimum and the maximum forecasting curves, represented on Figure 5.27. The only periods where we can distinguish the forecasting curves from each other are at the peaks of the pandemic. But, since the forecasts are much higher than the real data at the peaks, 0% of the data are located in the confidence interval. And on Figure 5.28, we can observe the standard deviation between the real data and the percentile 50 which is similar to previous results.

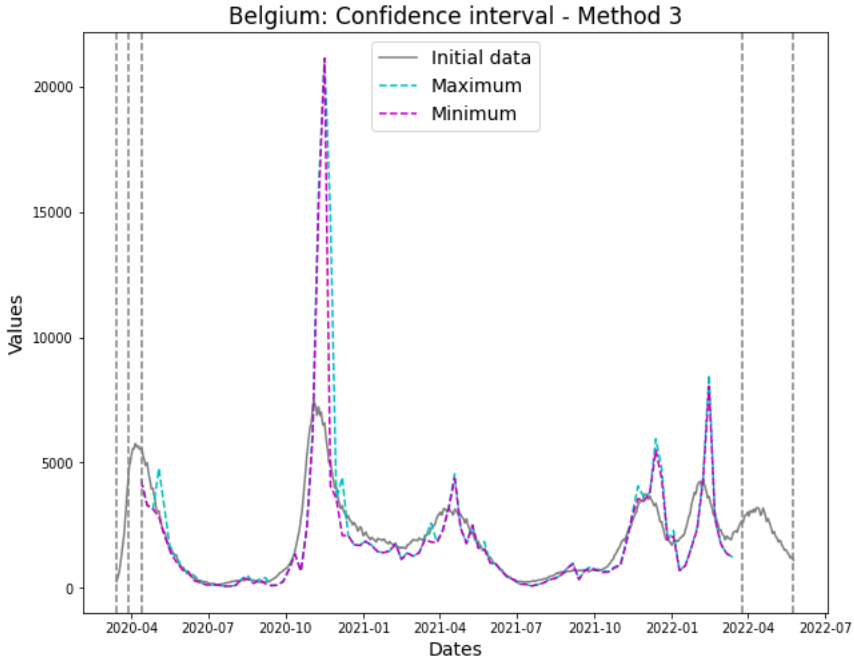


Fig. 5.24 Confidence interval for the Belgian data, using method 3. Minimum and maximum.

5.4 Conclusion and discussion

To conclude this chapter about the confidence intervals of the SH model, let us sum up the results we have obtained and compare the three methods that have been developed to take into account the different sources of uncertainty. The results obtained from the recapitulating graphs, concerning the entire data set, are presented in Table 5.1 for Belgium and in Table 5.2 for France. For the three methods, and for each of the two countries, you will find in those tables the percentage of real data located between the minimum and maximum forecasts, the percentage of real data located between the percentiles 10 and 90 of the forecasts and the percentage of real data located under the percentile 90 of the forecasts.

First of all, with method 1, the obtained confidence intervals are tight

	Method 1	Method 2	Method 3
Between min and max	34.65%	49.50%	7.92%
Between p_{10} and p_{90}	26.73%	37.62%	4.95%
Under p_{90}	30.69%	40.59%	25.74%

Table 5.1 Percentage of the real data located between the minimum and maximum forecasts, between the percentiles 10 and 90, and under the percentile 90, using the 3 different methods that are producing confidence intervals for the SH model, with the Belgian data set.

	Method 1	Method 2	Method 3
Between min and max	25.74%	42.57%	0%
Between p_{10} and p_{90}	16.83%	32.67%	0%
Under p_{90}	34.65%	38.61%	28.71%

Table 5.2 Percentage of the real data located between the minimum and maximum forecasts, between the percentiles 10 and 90, and under the percentile 90, using the 3 different methods that are producing confidence intervals for the SH model, with the French data set.

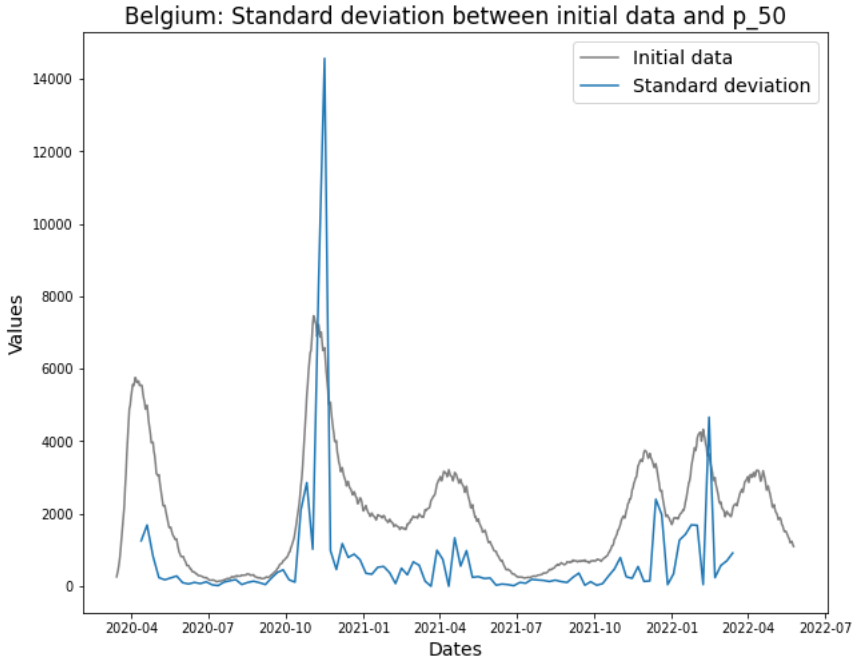


Fig. 5.25 Confidence interval for the Belgian data, using method 3. Standard deviation between the median forecasts and the initial data set.

around the initial forecasting curves, for both Belgium and France. As said earlier, this is probably because the values of ε are relatively small compared nation-wide. The initial data sets were thus only slightly noised to take account of the stochastic uncertainty. Moreover, we can notice that about 35% of the real data are located in between the minimum and the maximum forecasts for Belgium, and about 26% for France. These percentages can be understood by observing the recapitulating graphs on Figures 5.8 and 5.11. At the peaks of the pandemic, the SH model overestimates the number of hospitalizations due to COVID-19. And, on the contrary, in the troughs of the pandemic, it underestimates them a little. So, globally, the confidence intervals containing the forecasting curves remain close to the real data sets but do not follow them exactly. And, we observe once again that the better forecasts are given in the decreasing phases of the pandemic, corresponding to train periods that are located around peaks.

Now, let us have a look at the results obtained for the confidence in-

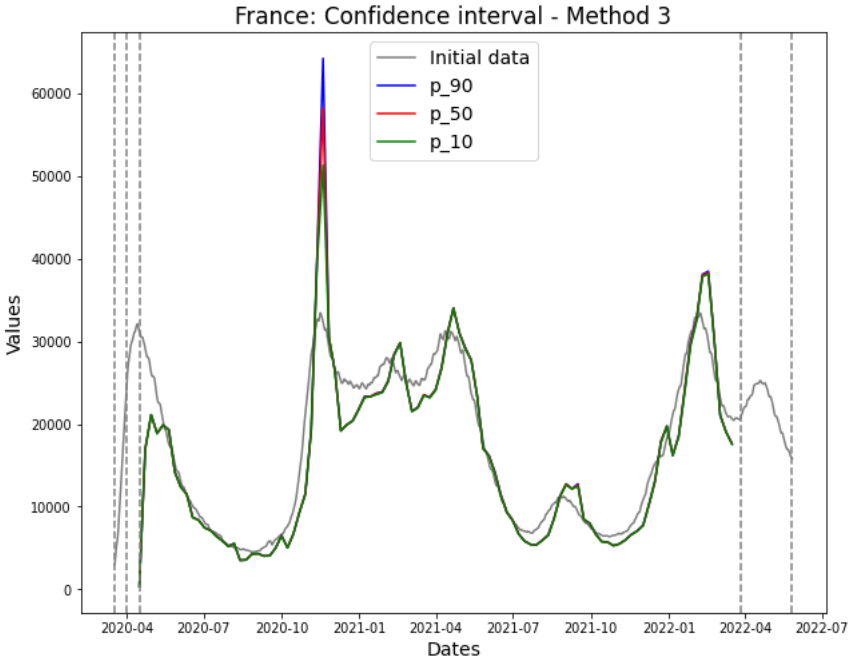


Fig. 5.26 Confidence interval for the French data, using method 3. Percentiles 10, 50 and 90.

tervals of the SH model produced by using the second method. The main difference is that the confidence intervals are a bit wider. A direct consequence is that a greater percentage of the real data are located in the confidence intervals of the SH model. For Belgium, almost 50% of the real data are located between the minimum and the maximum forecasts; and for France, it is about 43%. This means that the parameter uncertainty is an interesting uncertainty source to study. Taking it into account in our forecasts produces a confidence interval which is a little bit wider but one that is even closer to the real propagation of the pandemic, and thus even more relevant.

Finally, the third and last method taking into account the model uncertainty by considering a non-constant value for $\bar{\beta}$ has produced very narrow confidence intervals. Indeed, we have observed that, when using $\bar{\beta}(t) = \bar{\beta}_0 + \bar{\beta}_1 t$, the value of $\bar{\beta}_1$ is very close to zero, and thus $\bar{\beta}(t)$ is almost constant and equal to $\bar{\beta}_0$. The confidence intervals that are produced are

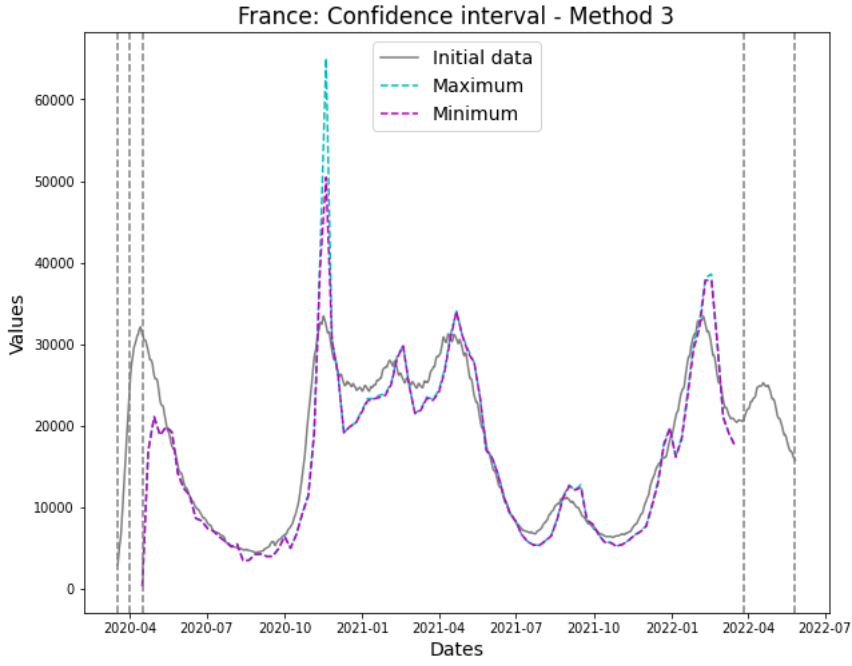


Fig. 5.27 Confidence interval for the French data, using method 3. Minimum and maximum.

so tight that very few real data are in it. For Belgium, about 8% of the real data are located between the minimum and the maximum forecasts, and for France, it is 0%. The value of $\bar{\beta}(t)$ is even less varying for the French data set.

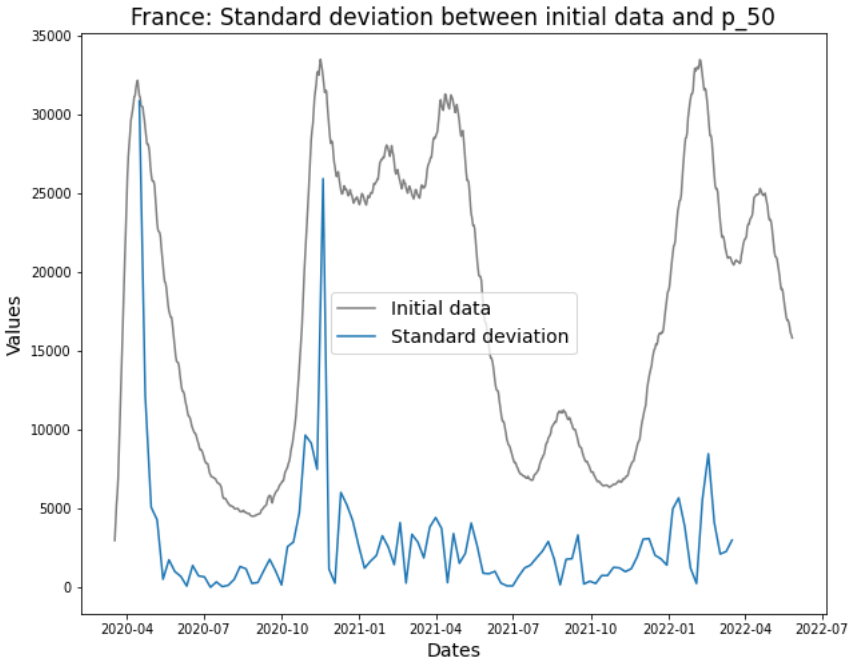


Fig. 5.28 Confidence interval for the French data, using method 3. Standard deviation between the median forecasts and the initial data set.

PART II
Malaria modelling

6

Generalized linear models to forecast malaria incidence in three endemic regions of Senegal

THIS chapter is adapted from a published paper [DAD23]. In this study, our goal is to forecast the monthly malaria incidence a few months ahead in the three endemic regions of Senegal (Dakar, Fatick, Kedougou) taking into account meteorological explanatory variables (rainfall, temperature, humidity) and history of malaria incidence. To this end, we assess the quality of the incidence forecasts obtained by GLMs with Poisson and negative binomial distributions. We use the statistical Vuong test and the reliability analysis in order to investigate their adequacy and to compare their performance. Then, we introduce a saturation on the rainfall variable to remedy some overestimations observed during the forecasts. Finally, we develop a machine learning approach based on a separation of the data into a train-set to estimate the parameters and a test-set to assess the forecast accuracy.

6.1 Materials and methods

The data used in this study are presented and discussed in the chapter 2 and Section 2.2.

6.1.1 Models

We have n data points $(X_{t1}, X_{t2}, ..., X_{tk}, Y_t) \in \mathbb{R}^{k+1}$ for $t = 1, ..., n$ where k is the number of explanatory variables (including the intercept vector) and n is the number of observations (months). These data are organized as in Table 6.1.

Y	X_1	$\cdots$	X_k
Y_1	$X_{1,1}$	$\cdots$	$X_{1,k}$
Y_2	$X_{2,1}$	$\cdots$	$X_{2,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_n	$X_{n,1}$	$\cdots$	$X_{n,k}$

Table 6.1 Data organization

We want to apply a GLM of the response vector Y using the k explanatory variables $X_1, ..., X_k$.

According to the studies in [Lin97, Mas19, YMMA20], candidate distributions for viable modeling include the Poisson and negative binomial distributions. These two probability distributions and the normal distribution are defined in the chapter 1 and Section 1.5.

The goal of fitting these GLMs is to find estimates for $\beta_1, ..., \beta_k$ and also α (with NB distribution), which in turn give us estimates for $g(\mu_t)$ by MLE (defined in Section 1.5.3) by using Algorithm 1 (defined in Section 1.5.3. In Appendix B, we describe how the regression coefficients are computed.

6.1.2 Estimation and forecasting methods

Train and Test Sets

We have $t_s < t_i < t_c < t_e$, where t_s is the initial time of the data, t_i is the initial time of the observed malaria incidence, t_c is the end time of the train

set, t_e is the end time of the test set.

Parameter estimation and principles of forecasting

We train the model by taking the observed malaria incidence ($Y_o(t)$) in $[t_i, t_c]$ and the explanatory variables ($X_{tj}, j = 1, \ldots, k$) in $[t_i - \delta, t_c - \delta]$. We made that in order to have the regression coefficients β's and the dispersion parameter α (only with NB regression model).

In [CT13, p. 90], the authors suggest to calculate α using a technique that they call auxiliary ordinary least squares (OLS) regression without a constant. In the negative binomial case, a first estimation of the β's is obtained by the procedure of the Poisson case. Then, α is computed by performing the OLS regression. Finally, the β's are re-estimated by maximizing the log likelihood (equation (B.9) defined in chapter B) wrt β.

Then, we make the forecasts, according to Algorithm 2, with the coefficients found in the train period. We assess the model accuracy in the test period $[t_c + 1, t_e]$ by comparing the theoretical (mean μ_t) and the observed ($Y_o(t)$) incidences.

Algorithm 2 describes the train-test procedure. For the link function g, the choices identity, log and sqrt are available in the following Python library `statsmodels.genmod.families.links`. The forecasts are obtained with the formula

$$\mu_t = g^{-1}(\beta_1 X_1(t) + \sum_{j=2}^{k} \beta_j X_j(t - \delta)), \tag{6.1}$$

where g is the link function introduced in Section 1.5.2.

6.1.3 Saturation method

We would like to test a saturation in the explanatory variables, in particular rainfall. The motivation is that additional rainfall should have less impact on the malaria incidence in a wet period than in a dry period. The saturation can be simply a hyperbolic tangent function. We posit that, instead of being linear, the contribution of rainfall to $\eta(X)$ (defined in Section 1.5.2) is affected by a saturation, which we model by

$$Rf_{sat} = \gamma \tanh(Rf/\gamma). \tag{6.2}$$

Algorithm 2: Forecasting Algorithm

Input: t_s, t_i, t_c, $t_e \leftarrow$ times of Section 6.1.2;
$\ell \leftarrow$ vector (Section 6.2.1, equation (6.3));
$h \leftarrow$ forecast horizon ($h \geq 1$);
$Y_o \leftarrow$ observed malaria incidence (dependant variable);
$X = \{X_1, X_2, \ldots, X_k\} \leftarrow$ Set of explanatory variables;
$f \leftarrow$ distribution (Poisson or NB or Normal);
$g \leftarrow$ link (identity, log, or sqrt);

Output: $\hat{Y} \leftarrow$ the forecasted vector of malaria incidence;

1 $Y_{\text{train}} = Y_o(t_i : t_c)$;
2 $X_{\text{train}} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_k(t_i - \ell_k : t_c - \ell_k)\}$;
3 Fit the GLM with distribution f and link g in the train period using the Python library `statsmodels.genmod.families`;
4 Get the regression coefficients β_j, $j = 1, \ldots, k$;
5 **for** $t \in [t_c + 1, t_e]$ **do**
6 $\quad \hat{Y}(t) = g^{-1}\left(\beta_1 X_1(t) + \sum_{j=2}^{k} \beta_j X_j(t - \ell_j)\right)$.

We estimate the parameter γ by making a research in many initial values of γ in order to find the more adapted value which gives the low RMSE_train after fitting the GLM. This procedure is entirely described in Algorithm 3.

6.2 Results and discussion

In this section, we first present and discuss the results of the correlation between the explained variable and each explanatory variable. Then, we present the forecast results from the Algorithm 2. Finally, we present the results from other methods such as addition study, ablation study and saturation.

6.2.1 Determination of lags

For Dakar data in Figure 2.5 (in chapter 2), we observe every year, an increase of malaria cases in the rainy season (from May or June to October or November), reaching a peak around the month of October or November, and a decrease to become stationary along the dry season. This situation is also observed in Fatick (Figure 2.6) and Kedougou (Figure 2.7) (both these

Algorithm 3: Forecasting Algorithm with Saturation

Input: $t_s, t_i, t_c, t_e \leftarrow$ times of Section 6.1.2;
$\ell \leftarrow$ vector (Section 6.2.1, equation (6.3));
$h \leftarrow$ forecast horizon ($h \geq 1$);
$Y_o \leftarrow$ observed malaria incidence (dependant variable);
$X = \{X_1, X_2, \ldots, X_{k-1}, Rf\} \leftarrow$ Set of explanatory variables;
$f \leftarrow$ distribution (Poisson or NB or Normal);
$g \leftarrow$ link (identity, log, or sqrt);

Output: $\hat{Y} \leftarrow$ the forecasted vector of malaria incidence;

$Y_{\text{train}} = Y_o(t_i : t_c)$;
$Rf_{\text{train}} = Rf(t_i - \ell_R f : t_c - \ell_R f)$;
$X_{\text{train}} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_{k-1}(t_i - \ell_{k-1} : t_c - \ell_{k-1}), Rf_{\text{train}}\}$;
for $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ **do**
  $Rf_{\text{sat}} = \gamma \tanh(Rf(t_i - \ell_R f : t_c - \ell_R f)/\gamma)$ // saturated rainfall;
  $X_{\text{train-sat}} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_{k-1}(t_i - \ell_{k-1} : t_c - \ell_{k-1}), Rf_{\text{sat}}\}$;
  Fit the Poisson distribution and link g with Y_{train} and $X_{\text{train-sat}}$ in order to obtain the predicted mean (μ);
  $\text{RMSE}(\gamma) = \sqrt{\frac{\sum_{t=t_i}^{t_c} (Y_o(t) - \mu_t)^2}{t_c - t_i + 1}}$ // RMSE for every value of γ
$\gamma_{\text{opt}} = \arg\min \text{RMSE}$ // Get the γ_{opt};
$Rf_{\text{sat-opt}} = \gamma_{\text{opt}} \tanh(Rf(t_i - \ell_R f : t_c - \ell_R f)/\gamma_{\text{opt}})$ // re-calculate the saturated rainfall with γ_{opt};
$X_{\text{train-opt}} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_{k-1}(t_i - \ell_{k-1} : t_c - \ell_{k-1}), Rf_{\text{sat-opt}}\}$;
Fit the GLM with distribution f and link g with Y_{train} and $X_{\text{train-opt}}$ using the Python library `statsmodels.genmod.families`;
Get the regression coefficients β_j, $j = 1, \ldots, k$;
for $t \in [t_c + 1, t_e]$ **do**
  $\hat{Y}(t) = g^{-1}\left(\beta_1 X_1(t) + \sum_{j=2}^{k-1} \beta_j X_j(t - \ell_j) + \gamma_{\text{opt}} \tanh(Rf(t - \ell_R f)/\gamma_{\text{opt}})\right)$.

two figures are reported in chapter 2), and proves the seasonality of the malaria cases from these three regions.

The plots of malaria cases and the explanatory variables (e.g., Figure 2.5–2.7, in chapter 2) reveal that there is a delay between the maximum of malaria cases and the maximum of the explanatory variable) in the three data sets. This delay is called `lag` and represented by ℓ in the formulae. We set

$$\ell_j := \underset{\delta \in \{h, h+1, \ldots, 6\}}{\arg\max} \left| r(Y_o(t_i : t_e), X_j(t_i - \delta : t_e - \delta)) \right|, \tag{6.3}$$

where r denotes the sample Pearson correlation coefficient. The statistical results are presented in Table 6.2 and the evolution of correlations as function of `lags` is illustrated in Figure 6.1–6.3.

	Dakar	Fatick	Kedougou
$Y_{o\{t\}}\&R_{t-\ell}$	0.79 $\ell = 2$ $4.48e^{-23}$	0.62 $\ell = 3$ $2.14e^{-12}$	0.61 $\ell = 1$ $9.93e^{-12}$
$Y_{o\{t\}}\&T_{t-\ell}$	0.65 $\ell = 1$ $8.73e^{-14}$	0.43 $\ell = 4$ $5.04e^{-06}$	0.56 $\ell = 5$ $4.86e^{-10}$
$Y_{o\{t\}}\&H_{t-\ell}$	0.56 $\ell = 5$ $7.43e^{-10}$	0.69 $\ell = 3$ $5.88e^{-16}$	0.58 $\ell = 1$ $1.50e^{-10}$
$Y_{o\{t\}}\&B_{t-\ell}$	0.15 $\ell = 6$ 0.129	0.38 $\ell = 3$ $5.99e^{-05}$	0.60 $\ell = 3$ $3.16e^{-11}$
$Y_{o\{t\}}\&A_{t-\ell}$	0.85 $\ell = 0$ $8.61e^{-30}$	0.88 $\ell = 0$ $5.84e^{-35}$	0.92 $\ell = 0$ $6.48e^{-43}$
$Y_{o\{t\}}\&Y_{o\{t-\ell\}}$	0.72 $\ell = 1$ $9.55e^{-18}$	0.75 $\ell = 1$ $1.31e^{-19}$	0.80 $\ell = 1$ $4.28e^{-24}$

Table 6.2 Sample lagged Correlations between the falciparum malaria incidence count per month and the explanatory variables in Dakar, Fatick and Kedougou, from 2008 to 2016. The table shows the maximal lagged correlation and the lag at which this maximal value is reached. The statistical significance of these correlations is tested by calculating the p-value associated with the Pearson correlation coefficient by using the `Scipy pearsonr()` function, which returns the Pearson correlation coefficient along with the two-tailed p-value. Correlation, lag and p-values are reported.

Since these p-values are less than 0.05, we conclude that there is a statistically significant correlation between the malaria incidence and the explanatory variable at the delay considered. The only exception is the bednet distributed in Dakar where the p-value is 0.129>0.05.

The rainfall is highly and positively correlated, meaning that if the rainfall increases, the malaria incidence will increase. These results are also found in [ADA20], revealing that rainfall was a strong positive predictor of increasing the annual incidence rate. In Senegal, it is known that the mosquito population increases with rainfall. Indeed, based on [GBMM16], the variation of the mosquito's birth rate, the daily survival probability of eggs, the daily survival probability of larvae, the daily survival probability of pupae in function of rainfall is positive.

The bed-net is weakly correlated in Dakar. Then, it is positively correlated in Fatick (even if the value is low) and in Kedougou with a `lag` of three months. Indeed, the situation in Fatick and Kedougou can be explained by the fact that the authorities anticipate the bed-net distribution three months before the beginning of the rainy season. On the other hand, in Figure 2.10 (in chapter 2), we observe that the bed-net distribution is not regular in Kedougou because its behavior is very different in the last two years of the dataset. For all these reasons, we do not use the bed-net distribution (denoted by B) as an explanatory variable.

The optimal `lag` between the malaria cases and the drugs (denoted by A) is equal to 0 in all the regions. This result means that the drugs is not a real predictor because it is usually taken after appearing some malaria symptoms from a patient. That reason leads us to do not consider the drug as an explanatory variable. But it can help to cure some sick people and to participate in the reduction of future malaria cases.

In Figure 6.1–6.3, the correlation between malaria cases at time t and the malaria in the past $(t - \delta)$ decreases for all values of δ meaning that the malaria in the past, as an explanatory variable, becomes less and less important when δ becomes more and more big.

The explanatory variables for $Y(t)$ are thus finally $Y_{\text{o}}(t - \ell_{Y_{\text{o}}}), Rf(t - \ell_{\text{Rf}}), Tp(t - \ell_{\text{Tp}}), RH(t - \ell_{\text{RH}})$, and 1 for the intercept.

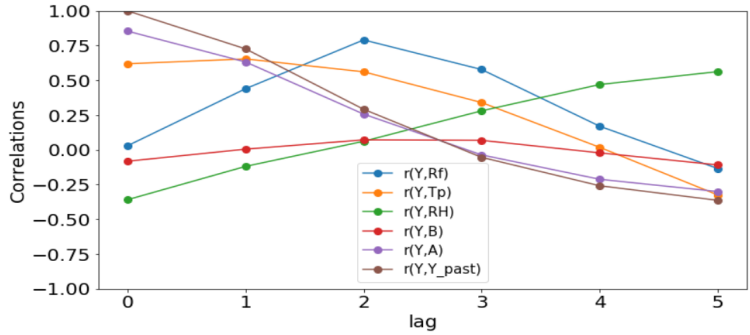


Fig. 6.1 Correlations vs lag (count) in Dakar.

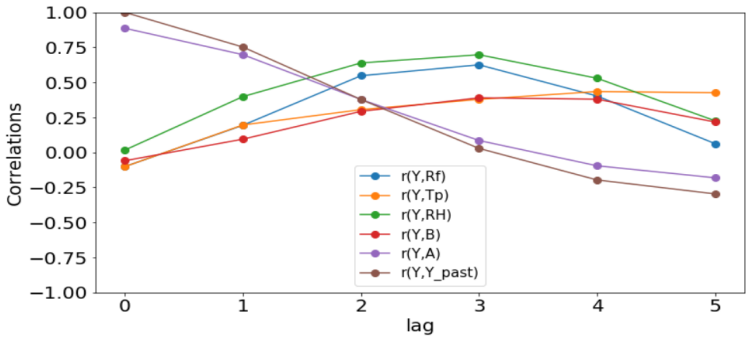


Fig. 6.2 Correlations vs lag (count) in Fatick.

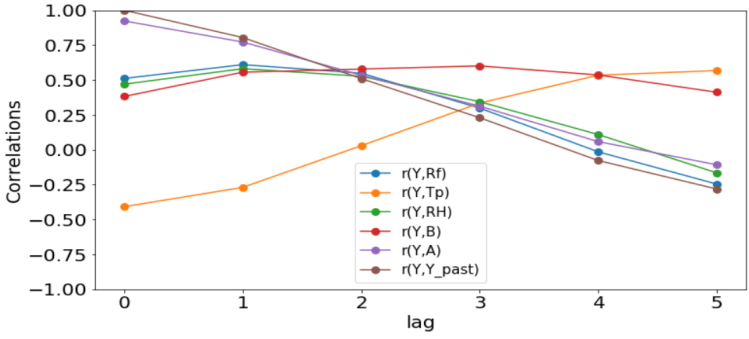


Fig. 6.3 Correlations vs lag (count) in Kedougou.

6.3 Model selection and result comparison

6.3.1 Model selection by using the Vuong test

In order to assess the adequacy of the distribution, we apply the Vuong statistical test as in [GRD19, HDHS21]. The Vuong test is defined as follows

$$V = \frac{\sqrt{n}\frac{1}{n}\sum_t^n m_t}{\sqrt{\frac{1}{n}\sum_t^n (m_t - \bar{m})^2}}, \tag{6.4}$$

where $m_t = \log(f_1(Y_t|X_t)/f_2(Y_t|X_t))$ in which $f_1(Y_t|X_t)$ is the first probability mass function and $f_2(Y_t|X_t)$ is the second probability mass function. If $V > 1.96$, then the first model is preferred. If $V < -1.96$, then the second one is preferred. If $-1.96 < V < 1.96$, none of the models are preferred. In our case, we let f_1 be the Poisson distribution and f_2 the NB distribution. This statistical test permits to choose the most adequate between the two regression models in order to fit the data.

In Table 6.3, the results of the Vuong test reveal that Poisson model is preferred to negative binomial model in Dakar, Fatick and Kedougou because $V > 1.96$. However, none one is preferred in Kedougou with link equal to log because the result is $-1.96 < V = 0.0 < 1.96$.

Regions	Link	V
Dakar	identity	5.09
Dakar	log	3.22
Dakar	sqrt	4.36
Fatick	identity	3.29
Fatick	log	4.59
Fatick	sqrt	2.5
Kedougou	identity	2.31
Kedougou	log	0.0
Kedougou	sqrt	2.55

Table 6.3 Results of the Vuong test.

6.3.2 Results comparison by using metrics

In this part, we made a comparative analysis between the three models that are Gaussian (identity, log), Poisson (identity, log, sqrt) and negative binomial (identity, log, sqrt) in Table 6.4. Experiments are made with Algorithm 2 without saturation. We include the minimum values of the predicted mean obtained after fitting the model. A negative sign indicates that the model is not adequate for our count data. All these metric values permit to validate and to compare the performance of the models. Additionally, we present the RA results in Table 6.5.

Results in Table 6.4 display that the most of the MASE's and MARE's in the train and test periods are around 1 in the three regions. Then, we also have the R^2_{COR}'s indicator whose values are high (>50 %) indicating a good contribution of the explanatory variables in the forecasts. In Table 6.5, the predictions provided by Poisson (identity) are globally more reliable in the train period when compared to other models due to the highest values of RA, in the three regions. They also are very reliable in the test period where the RA values are the highest in 2/3 of regions. The forecast accuracy of GLMs with various distributions (Poisson, negative binomial, and Gaussian) and link functions (identity, log, and sqrt) is compared in terms of several model performance metrics. Whereas the best distribution-link combination varies according to the endemic region of interest and the performance metric, the experiments lead to the conclusion that the Poisson distribution with the identity link is overall the most suitable combination.

We can conclude that Poisson (identity) model can be nicely used to fit our data for parameter estimation and to make the forecasts with these found parameters in the three regions. All the additional studies will base on this model.

6.4 Forecast results by various sets of explanatory variables

6.4.1 Forecast results using history of malaria incidence only

For transmissible diseases, the incidence at a given time $t - \delta$ is very important to predict the expected incidence at a given time t. That reason leads us to first consider a set of explanatory variables composed of two variables that are the history of malaria incidence ($Y_o(t - \delta)$) and the intercept

	RMSE	MASE	MARE	R^2_{COR}	min
G(identity)	2197.29 / 2384.26	0.52 / 1.01	0.68 / 1.54	0.84 / 0.79	-517.03
G(log)	2466.29 / 2352.81	0.6 / 1.38	0.85 / 2.66	0.79 / 0.78	511.81
P(identity)	2245.27 / 2689.74	0.52 / 1.02	0.54 / 1.28	0.83 / 0.78	282.6
P(log)	2523.01 / 2297.45	0.57 / 1.23	0.65 / 2.05	0.79 / 0.77	341.22
P(sqrt)	2354.97 / 2303.14	0.54 / 1.08	0.62 / 1.77	0.81 / 0.8	279.49
NB(identity)	2558.87 / 3555.94	0.58 / 1.28	0.5 / 1.24	0.82 / 0.76	460.63
NB(log)	3736.91 / 2424.02	0.74 / 1.1	0.61 / 1.89	0.73 / 0.79	471.42
NB(sqrt)	3409.99 / 3234.53	0.73 / 1.23	0.55 / 1.56	0.79 / 0.79	482.42
G(identity)	768.52 / 578.88	0.83 / 1.33	0.66 / 0.59	0.67 / 0.75	87.78
G(log)	721.66 / 484.07	0.81 / 1.31	0.73 / 0.7	0.71 / 0.83	83.29
P(identity)	772.32 / 543.56	0.82 / 1.26	0.65 / 0.6	0.67 / 0.72	99.12
P(log)	741.92 / 491.42	0.83 / 1.37	0.85 / 0.88	0.69 / 0.78	215.44
P(sqrt)	763.78 / 526.94	0.84 / 1.32	0.79 / 0.74	0.68 / 0.73	215.13
NB(identity)	839.53 / 527.32	0.87 / 1.22	0.61 / 0.59	0.64 / 0.67	69.22
NB(log)	861.98 / 466.03	0.91 / 1.21	0.84 / 0.79	0.63 / 0.7	268.19
NB(sqrt)	942.27 / 504.91	0.98 / 1.13	0.75 / 0.59	0.62 / 0.64	180.39
G(identity)	1230.61 / 2467.27	1.01 / 0.92	1.19 / 0.63	0.62 / 0.62	29.85
(log)	1409.28 / 2720.94	1.27 / 1.01	2.18 / 0.73	0.51 / 0.56	872.92
P(identity)	1241.89 / 2431.91	0.99 / 0.87	0.9 / 0.47	0.61 / 0.63	126.76
P(log)	1488.67 / 3009.49	1.15 / 1.1	1.47 / 0.52	0.46 / 0.59	510.81
P(sqrt)	1352.9 / 2523.28	1.03 / 0.89	1.07 / 0.42	0.56 / 0.62	324.21
NB(identity)	1287.33 / 2346.07	1.07 / 0.83	0.9 / 0.44	0.61 / 0.65	-52.81
NB(log)	2160.53 / 7024.86	1.42 / 2.47	1.08 / 0.73	0.4 / 0.58	275.72
NB(sqrt)	1567.23 / 3037.79	1.19 / 1.14	0.91 / 0.44	0.54 / 0.63	132.16

Table 6.4 Results of accuracy measures with Algorithm 2 and the explanatory variables $Y_o(t-h)$, $R(t-\ell_R)$, $T(t-\ell_T)$, $H(t-\ell_H)$, and 1 for the intercept. We have considered $t_s = 0$, $t_i = 5$, $t_c = 84$ and $t_e = 108$ and $h = 1$. Refer to Table 6.2 for ℓ_R, ℓ_T and ℓ_H values. Train/test values are reported and $\min_{t\in\{t_i,\ldots,t_c\}} \mu(t)$. We denote by G: Gaussian, P: Poisson, NB: negative binomial. The first bloc represents for Dakar, the second for Fatick and the last for Kedougou.

vector $(I(t))$. We defined this set as follows

$$\text{start\_set} \leftarrow \{Y_o(t-\delta), I(t)\}. \tag{6.5}$$

Thus, our linear predictor becomes like a simple Markov model with the Gaussian distribution that is called first order autoregressive AR(1) [Lin97] (Section 5.2.1) and defined by

$$\mu_{t|t-\delta} = g^{-1}(\beta_1 I(t) + \beta_2 Y_o(t-\delta)). \tag{6.6}$$

In Figure 6.4–6.6, we present the forecast results (noted by A) and the curves of $\beta_j X_j$ (noted by B). These results are obtained from equation (6.6)

Dist	Dakar			Fatick			Kedougou		
Train	id	log	sqrt	id	log	sqrt	id	log	sqrt
G	28.75	22.5		27.5	25		10	13.75	
P	32.5	27.5	30	31.25	25	26.25	16.25	21.25	18.75
NB	27.5	28.75	28.75	30.0	21.25	22.5	16.25	16.25	20
Test									
G	20	4		24	16		16	20	
P	16	4	12	24	8	20	24	16	12
NB	16	8	16	20	20	28	28	8	16

Table 6.5 Comparison of the performance (reliability) results for GLM with Gaussian (G), Poisson (P) and Negative Binomial (NB) distributions and identity (id), log and sqrt as link functions, in Dakar, Fatick and Kedougou.

with Poisson (id) model when we take $h = 1$. The figures noted by B permit to show which variable is highly ($\beta_j X_j(t) \gg 0$) or weakly ($\beta_j X_j(t) \sim 0$) used during the forecasting process. In all the three regions, we observe that the sole use of the history of malaria incidence at time $t - 1$ gives good results. That can be explained by the fact that it is highly used based on figures (B) compared to the intercept whose values are closed to 0. This situation is biologically true because with the transmissible diseases like malaria the history of cases is very important to estimate the future cases.

In addition, in Table 6.6, we remark that the model is less accurate when the values of h become large meaning that the history of malaria incidence is weakly used when we take a high forecast horizon ($h > 1$).

The conclusion of these results is that malaria in the past was a very good explanatory variable when $h = 1$.

6.4.2 Forecasts results by using all explanatory variables

In this part, we use the whole explanatory variables and the results from Algorithm 2 and equation (6.1) are presented in Figures 6.7–6.9.

Figure 6.7 (B) of Dakar data shows with some peaks in the rainfall distribution. While in Figure 6.8 (B) about Fatick data, the rainfall variable is weakly used and this variable is negatively used in Figure 6.9 of Kedougou data. The peaks observed in Dakar and the situation in Kedougou lead us to develop the method of saturation in Section 6.1.3. Then, in Figure 6.7 (B)

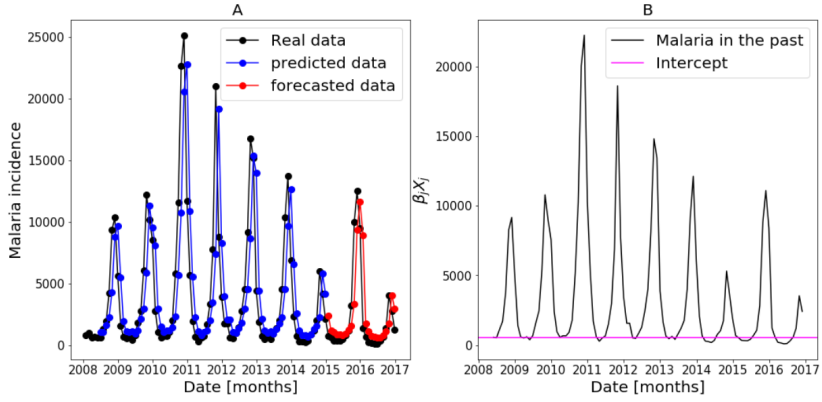


Fig. 6.4 Forecasting results (A) and the curves of $\beta_j X_j$ (B) in Dakar: Algorithm 2 with Poisson for f and identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$. Malaria incidence means the falciparum malaria incidence count per month. The train/test accuracy measures are RMSE: 3886.43 / 2367.55, MASE: 0.95 / 1.06, MARE: 0.85 / 1.59 and R^2_{COR}: 0.51 / 0.54.

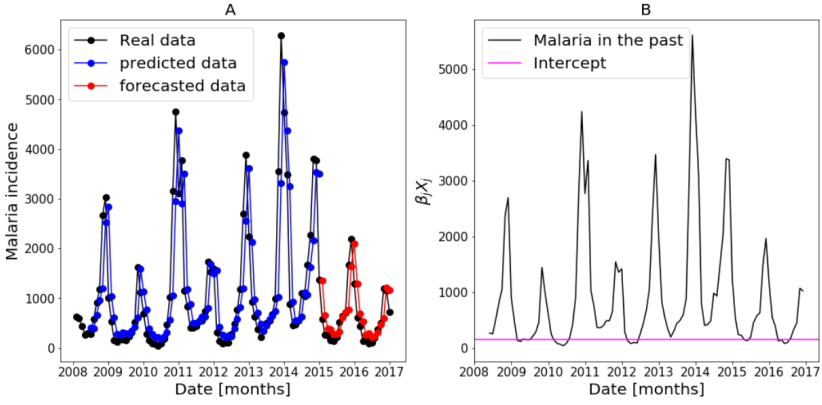


Fig. 6.5 Forecasting results (A) and the curves of $\beta_j X_j$ (B) in Fatick: Algorithm 2 with Poisson for f and identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$. Malaria incidence means the falciparum malaria incidence count per month. The train/test accuracy measures are RMSE: 916.35 / 408.29, MASE: 0.96 / 1.11, MARE: 0.76 / 0.75 and R^2_{COR}: 0.55 / 0.5.

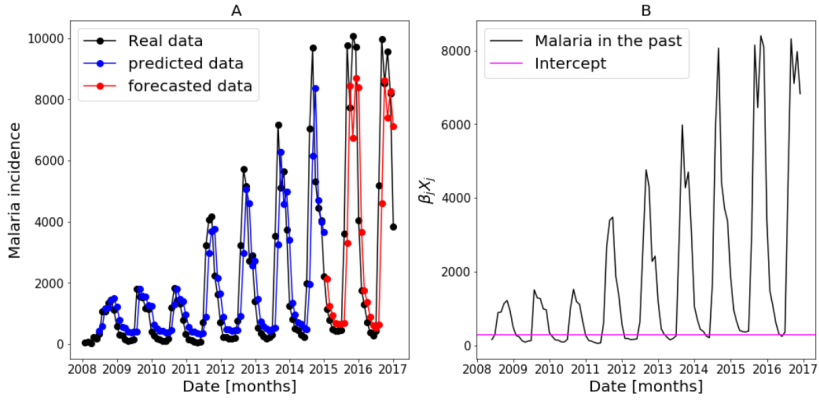


Fig. 6.6 Forecasting results (A) and the curves of $\beta_j X_j$ (B) in Kedougou: Algorithm 2 with Poisson for f and identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$. Malaria incidence means the falciparum malaria incidence count per month. The train/test accuracy measures are RMSE: 1250.53 / 2523.74, MASE: 1.02 / 0.93, MARE: 1.06 / 0.61 and R^2_{COR}: 0.61 / 0.59.

	h	RMSE	MASE	MARE	R^2_{COR}
Dk	1	3886.43 / 2367.55	0.95 / 1.06	0.85 / 1.59	0.51 / 0.54
	2	5265.41 / 3565.7	1.47 / 2.15	2.4 / 5.97	0.08 / 0.06
	3	5399.89 / 4075.83	1.55 / 2.79	3.15 / 9.38	0.01 / 0.01
Ft	1	916.35 / 408.29	0.96 / 1.11	0.76 / 0.75	0.55 / 0.5
	2	1250.88 / 741.17	1.47 / 2.26	1.99 / 2.21	0.14 / 0.03
	3	1339.06 / 807.71	1.67 / 2.71	2.85 / 3.32	0.0 / 0.01
Kd	1	1250.53 / 2523.74	1.02 / 0.93	1.06 / 0.61	0.61 / 0.59
	2	1790.88 / 3770.72	1.58 / 1.59	2.67 / 1.23	0.19 / 0.18
	3	1972.37 / 4427.78	1.87 / 1.84	3.65 / 1.43	0.02 / 0.01

Table 6.6 Accuracy measures in the three regions: Algorithm 2 with Poisson for f and identity for g where the set of explanatory variables is equation (6.5) and $h = 1, 2, 3$. Train/test values are reported. We denote by Dk: Dakar, Ft: Fatick and Kd: Kedougou.

of Dakar data, it is also shown that the humidity is weakly used to mean that this variable has a little contribution in the forecasts. Contrary in Figure 6.8 (B) about Fatick data, the humidity is high used, so it has a big participation in the forecasts. As for Kedougou, in Figure 6.9 (B), this vari-

able (humidity) is moderately used. As for temperature, its contribution changes depending on the region. For example, in Dakar and Fatick, this variable (temperature) is negatively used as it is shown in Figure 6.7 and 6.8. However, in Kedougou, it is positive and highly used as it is shown in Figure 6.9.

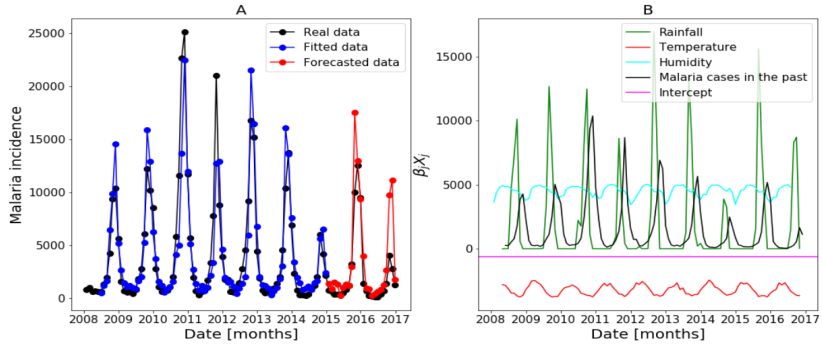


Fig. 6.7 Forecast results (A) and the curves of $\beta_j X_j$ (B) in Dakar: Algorithm 2 with Poisson for f and identity for g, no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$.

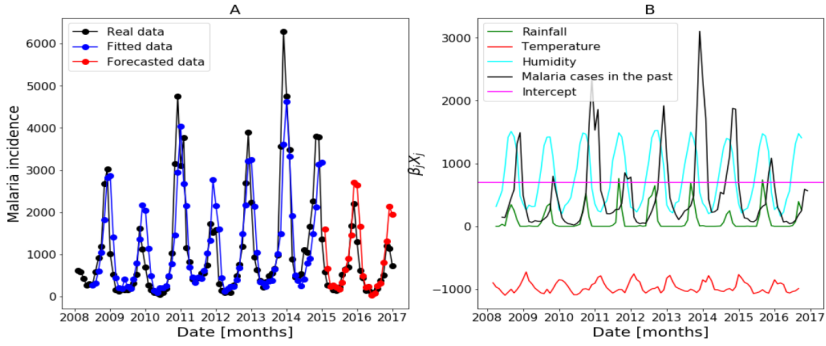


Fig. 6.8 Forecast results (A) and the curves of $\beta_j X_j$ (B) in Fatick: Algorithm 2 with Poisson for f and identity for g, no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$.

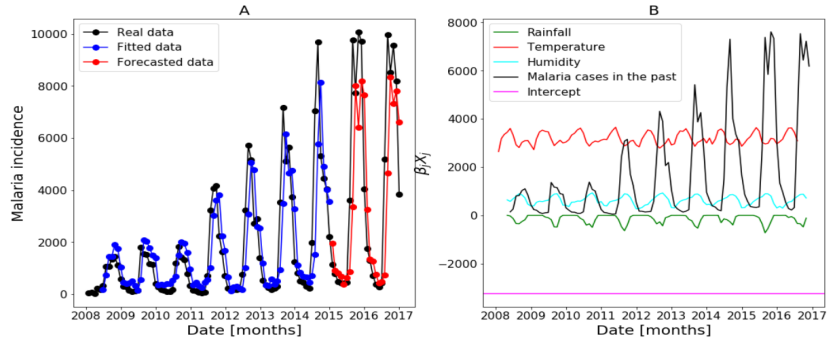


Fig. 6.9 Forecast results (A) and the curves of $\beta_j X_j$ (B) in Kedougou: Algorithm 2 with Poisson for f and identity for g, no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$.

We can summarize here that each explanatory variable doesn't use the same in the three regions. Thus, these analyses and observations lead the development of the methods in Sections 6.4.3 and 6.4.4.

6.4.3 Addition study

In this section, we experiment with different combinations of sets from equation (6.5) and we define them as follows

$$\{\text{start\_set} \cup X_j, j = 3, \ldots, k\}.$$

The purpose is to investigate the influence of each additional variable compared to the observation in Section 6.4.1.

The Table 6.7 reveals that, in Kedougou, the forecast results are improved by adding to start_set explanatory variables such as temperature or humidity. In contrast, the addition of the rainfall variable gives some less good results in the sense of the RMSE, MASE and MARE, in Dakar and Fatick, even if the higher values of R^2_{COR} are observed there.

Var	RMSE w/wo	MASE w/wo	MARE w/wo	R^2_{COR} w/wo
Rf	1.18	1.02	0.99	1.43
Tp	0.99	1.05	1.15	1.04
RH	0.94	1.04	1.06	1.11
Rf	1.24	1.22	1.04	1.34
Tp	1	1.06	1.08	1.1
RH	1.18	1.05	0.77	1.34
Rf	0.99	0.97	0.92	1.02
Tp	0.96	0.95	0.88	1.05
RH	0.98	0.93	0.7	1.03

Table 6.7 Results of the addition study in the three regions: Algorithm 2 with Poisson for f and identity for g, $t_{\text{s}} = 0$, $t_{\text{i}} = 5$, $t_{\text{c}} = 84$, $t_{\text{e}} = 108$ and $h = 1$. We denote by w: situation in the test period with the added variable and wo: situation in the test period without the indicated variable. Ratios of w/wo are reported. Note that, for the metrics: RMSE, MASE and MARE, a ratio lower than 1 implies an improvement of the model when the variable is added as an explanatory variable. A ratio of R^2_{COR} higher than 1 indicates that adding the variable improves the forecasts. The first three lines concern for Dakar, the second for Fatick and the last for Kedougou.

6.4.4 Ablation study

We now present ablation studies where we start with all the available explanatory variables and investigate the effect of removing any one of them. The sets of explanatory variables after ablation are thus

$$\{X_1, X_2, \ldots, X_k\} \setminus X_j, j = 2, \ldots, k.$$

By doing that, we can know which variable is more responsible of the over-forecasting (or under-forecasting) observed.

The results in Table 6.8 show a low accuracy in terms of RMSE, MASE and MARE in the whole three regions data sets when the malaria incidence in the past was deleted. This situation was expected when we refer to the forecast results in Section 6.4.1. It is also shown there that the deletion of one of the others variables such as rainfall, temperature and humidity generally reduces a bit the performance of the models in the sense of RMSE, MASE and MARE.

Var	RMSE wo/w	MASE wo/w	MARE wo/w	R^2_{COR} wo/w
$Y_{o(t-1)}$	1.5	1.74	2.04	0.82
Rf	0.83	1.06	1.27	0.75
Tp	1.05	1.07	1.11	0.98
RH	1.04	1.05	1.25	0.98
$Y_{o(t-1)}$	1.48	1.63	1.58	1.11
Rf	0.94	0.95	1.06	0.93
Tp	0.96	0.98	0.96	1.01
RH	0.92	1.05	1.3	0.98
$Y_{o(t-1)}$	1.5	1.64	1.25	1.21
Rf	1	0.99	0.92	1.01
Tp	1.02	1.02	0.92	0.97
RH	1.01	1.03	1.14	0.98

Table 6.8 Results of the ablation study in the three regions: Algorithm 2 with Poisson for f and identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$ and $h = 1$. We denote by wo: situation in the test period without the indicated variable and w: situation in the test period with all the explanatory variables. Ratios of wo/w are reported. Note that, for the metrics: RMSE, MASE and MARE, a ratio lower than 1 implies that the result is good without the indicated variable. A ratio of R^2_{COR} higher than 1 indicates adding the variable improves the forecasts. The first four lines concern for Dakar, the second for Fatick and the last for Kedougou.

6.4.5 Forecast results using saturation

In this section, we made a main modification in the rainfall variable. We call this method by saturation and the procedure is entirely detailed in Section 6.1.3 and Algorithm 3. Statistical results of this method are presented in Table 6.9. They permit to distinguish the performance given by this novelty compared to Table 6.4. The forecasts are displayed in Figure 6.10 and 6.11.

As a result, an improvement of the forecasts is observed based on Table 6.9. For Dakar, we observed an improvement of the results after applying the saturation according to all the metrics. It is interesting to see that, in Dakar, the introduction of the saturated rainfall has reduced the over-estimation occurring at the end of 2015; compare Figure 6.7 and Figure 6.10. Then, in Fatick, all the values are 1, meaning that the saturation method does not improve the results. That can be explained by the fact that the rainfall is less used in this region, according to Figure 6.8 (B). As

for Kedougou, we have a MARE smaller than its value before applying the method due to the ratio under 1 (ratio = 0.87) and a R^2_{COR} higher than the value before applying the method (ratio = 1.04). There is no improvement in terms of ratio of the RMSE and MASE. That may be explained by the fact that there was not a peak as that is illustrated in Figure 6.9 or the forecasts have been already good.

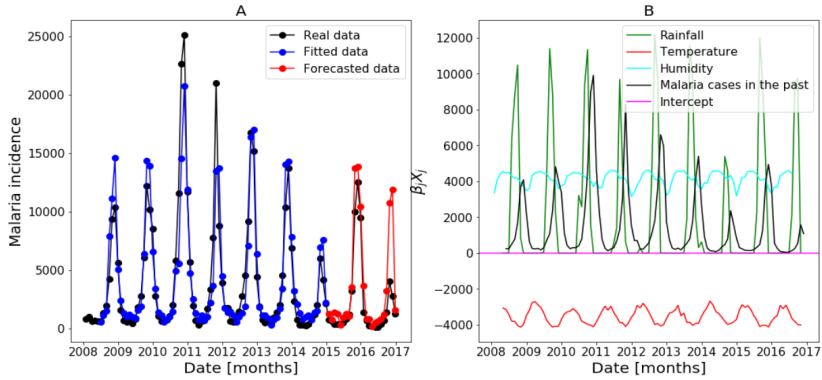


Fig. 6.10 Forecast results (A) and the curves of $\beta_j X_j$ (B) of the saturation method in Dakar: Algorithm 3 with Poisson for f and identity for g. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_{\text{s}} = 0$, $t_{\text{i}} = 5$, $t_{\text{c}} = 84$, $t_{\text{e}} = 108$, $h = 1$.

	RMSE w/wo	MASE w/wo	MARE w/wo	R^2_{COR} w/wo
Dk	0.95	0.97	0.96	1.01
Ft	1	1	1	1
Kd	1.01	1.02	0.87	1.04

Table 6.9 Comparison results after and before the saturation: Algorithm 3 with Poisson for f and identity for g, $t_{\text{s}} = 0$, $t_{\text{i}} = 5$, $t_{\text{c}} = 84$, $t_{\text{e}} = 108$ and $h = 1$. We denote by w: the metric with saturation and wo: the metric without saturation, in the test period. Ratios of w/wo are reported. For the metrics: RMSE, MASE and MARE, a ratio lower than 1 implies an improvement of the model to make good forecasts with saturation. A ratio of R^2_{COR} higher than 1 indicates that applying the saturation improves the forecasts. We denote by Dk: Dakar, Ft: Fatick and Kd: Kedougou.

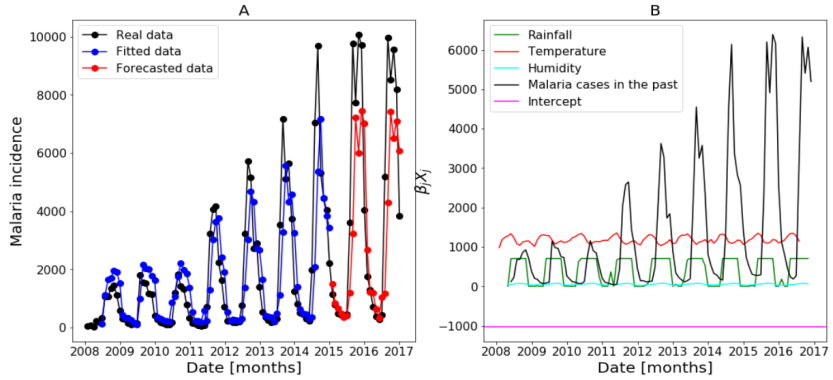


Fig. 6.11 Forecast results (A) and the curves of $\beta_j X_j$ (B) of the saturation method in Kedougou: Algorithm 3 with Poisson for f and identity for g. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$.

Another favorable effect of the saturation is that, in Kedougou, the non-saturated rainfall is negatively used (Figure 6.9 B) while the saturated rainfall is positively used (Figure 6.11 B).

6.5 Partial conclusion

For three endemic regions of Senegal, we have investigated the accuracy, in the sense of the metrics such as RMSE, MASE and MARE, of the falciparum malaria incidence count per month forecasts obtained with GLM's by using meteorological data and history of falciparum malaria incidence count per month as explanatory variables. Using the Vuong test, we compared the adequacy of Poisson-based and NB-based GLM's. And the GLM with Poisson and identity as a link function is in practice the more adequate regression model to make forecasts of falciparum malaria incidence count per month based on meteorological factors. We have observed that the choice of the GLM's link function and the use of adequate `lags` in the explanatory variables may have a considerable impact on the forecast accuracy.

We also have observed that the application of saturation in the rainfall

variable increases the quality of the forecasts in Dakar and Kedougou.

An ablation study shows that removing the history of malaria cases from the explanatory variables has a strong adverse effect on the forecast accuracy.

7

Alternative forecasting methods of malaria incidence

THIS chapter is organized as follows. A forecasting algorithm different to the Algorithm 2 defined in chapter 6, the SARIMAX model algorithm and comparative results to these three algorithms are presented in Section 7.1. The `scipy.optimize.fmin` optimization method is described in Section 7.2. For further experiments, we investigate the forecast from various train period study in Section 7.3, the forecast from districts level study in Section 7.4 and parameter generalization study in Section 7.5.

7.1 Alternative forecasting methods

The data used in this study are presented and discussed in the chapter 2 and Section 2.2.

7.1.1 Forecasting method without `lag` in the climatic variables

This method corresponds to where we train the model by taking the explained variable $Y_o(t)$, and the explanatory variables $X_{tj}, j = 2, \ldots, k$ in $[t_i, t_c]$, while the history of malaria incidence $Y_o(t-\delta)$ (represented by X_p) is taken in $[t_i - \ell_p, t_c - \ell_p]$. The initial times t_i and t_c are the same like in Section 6.1.2 and ℓ is also defined in Section 6.2.1 (equation (6.3)). Then, the forecasts are obtained as follows

$$\mu_t = g^{-1}(\beta_p X_p(t-\ell_p) + \sum_{j=1, j\neq p}^{k} \beta_j X_j(t)), \tag{7.1}$$

where g represents the link function introduced in Section 1.5.2. This method is identical to the method defined in equation (6.1), with the exception that the meteorological explanatory variables for $Y(t)$ are assumed to be available at the forecasts time t. This allows us to investigate the forecast improvement that perfectly accurate weather predictions might bring. The train-test process is presented in Algorithm 4. Experimental results can be found in Section 7.1.3.

Algorithm 4: Alternative forecasting algorithm

Input: t_s, t_i, t_c, $t_e \leftarrow$ times of Section 6.1.2, $\ell \leftarrow$ vector (Section 6.2.1, equation 6.3), $h \leftarrow$ forecast horizon, Y_o and $X = \{X_1, X_2, \ldots, X_k\}$: set of explanatory variables;

Output: $\hat{Y} \leftarrow$ the forecasted incidence;

1 $Y_{\text{train}} = Y_o(t_i : t_c)$;

2 $X_{\text{train}} = \{X_1(t_i : t_c), X_2(t_i : t_c), \ldots, X_{k-1}(t_i : t_c), X_k(t_i - \ell_k : t_c - \ell_k)\} \leftarrow k$ indicates the index of $Y_o(t-\delta)$;

3 Fit the GLM with Poisson for f and identity for g in the train period using the Python library `statsmodels.genmod.families`;

4 Get the regression coefficients $\beta_j, j = 1, \ldots, k$;

5 **for** $t \in [t_c + 1, t_e]$ **do**

6 $\quad \hat{Y}_t = g^{-1}(\sum_{j=1}^{k-1} \beta_j X_j(t) + \beta_k X_k(t - \ell_k))$.

7.1.2 Forecasting method with SARIMAX model

In the past few years, a class of time series models named SARIMAX (seasonal autoregressive integrated moving average with exogenous variables) appear to be very popular in short-term load forecasting [TA17]. The most of the time, the order and seasonal order values are separately calculated based on Box-Jenkins approach [Gou21].

Algorithm 5: SARIAMAX forecasting algorithm

Data: t_s, t_i, t_c, $t_e \leftarrow$ times of Section 6.1.2; Y_o; $X = \{X_1, X_2, \ldots, X_k\}$; $s = 12$;
Result: $Y_{\text{forecasts}} \leftarrow$ the forecasted malaria incidence;

$Y_{\text{train}} = Y_o(t_i : t_c)$;
$X_{\text{train}} = \{X_1(t_i : t_c), X_2(t_i : t_c), \ldots, X_k(t_i : t_c)\}$;
$X_{\text{test}} = \{X_1(t_c : t_e), X_2(t_c : t_e), \ldots, X_k(t_c : t_e)\}$;
$i = 0$, $p_{\max} = d_{\max} = q_{\max} = m$ where m is an initial guess;
for $p, d, q \in [0, m]$ **do**
    for $P, D, Q, s \in \{[0, m], s\}$ **do**
        Fit SARIMAX model using the Python library `statsmodels.tsa.statespace.SARIMAX`;
        Call the `predict()` with X_{train} in order to get μ_t;
        order$[i] = [p, d, q]$;
        seasonal_order$[i] = [P, D, Q, s]$;
        RMSE$[i] = \sqrt{\frac{\sum_{t=t_i}^{t_c} (Y(t) - \mu_t)^2}{t_c - t_i + 1}}$;
        $i = i + 1$;
[order][seasonal_order]_opt = arg min RMSE;
Fit SARIMAX model with these optimal parameters;
Call the `predict()` function to get μ_t;
$interval = t_e - t_c \leftarrow$ define the forecast horizon;
Call the `forecast()` with this horizon to get Y_{forecast}.

In [VYKK+22], a Bayesian structural time series (BSTS) and seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) models were implemented to predict the malaria incidence in the northeast states of India. In that study, authors apply the auto ARIMA method from `pmdarima` package of Python to automatically identify and estimate the best fitting model among the grid of hyper-parameters. SARIMAX model was also used in [MSH+12] to quantify the relationship between malaria cases and remotely sensed environmental variables, includ-

ing rainfall, land-surface temperature (LST), vegetation indices (NDVI and EVI), and actual evapotranspiration (ETa) with lags ranging from one to three months in the highlands of Ethiopia. In that study, the results reveal that the SARIMA models that included these environmental covariates had better fits and more accurate predictions, as evidenced by lower Akaike Information Criterion (AIC) and RMSE values, compared to models without environmental covariates.

In our case, we develop the SARIMAX model in order to compare its performance to the performance of GLMs with the datasets from Dakar, Fatick and Kedougou. Concerning the seasonal parameter s, we use $s = 12$ corresponding to annual seasonality. The choice of annual seasonality is backed by our previous findings about the periodic structure of the data; see Figures 2.5, 2.6 and 2.7 in chapter 2. We perform grid-search for the optimal ARIMA order (p, d, q) and the seasonal order (P, D, Q) based on the lower RMSE, as was also done by [VYKK+22]. The procedure is described in Algorithm 5. We consider neither the history of malaria incidence in the explanatory variables because the model itself incorporates the autoregression, nor an intercept vector.

7.1.3 Results

In Figures 7.1–7.3, we display the forecast results of the method in Algorithm 4. For comparison, we calculate a ratio (R) of RMSE_test's of the two methods considered. After that, we calculate a percentage by $p = 100(1 - R)$. When calculating the ratio between Algorithm 4 and Algorithm 2, the results reveal a reduction about 22% in Dakar, about 44% in Fatick and about 9% in Kedougou. The conclusion is that Algorithm 4 performs better than Algorithm 2.

We compare the performance between Algorithms 2, 4 and 5. The accuracy measures are presented in Table 7.1. We observe low MASE_test with Algorithm 4: 0.94, 0.92 and 0.8, respectively, in Dakar, Fatick and Kedougou. These values are better than the values otained with Algorithms 2 and 5. We can notice that the SARIMAX model (Algorithm 5) produces the best results in the sense of the RMSE and MASE measures in the training period, compared to GLMs in the three regions. This result confirms the findings in [NPC21] revealing that the SARIMAX model generally gives a good fitting. But it produces less good quality of forecasts. This poor forecast quality can be explained by the large number of parameters to be

estimated included in this model. This may still be due to a large forecast horizon (two years in this case).

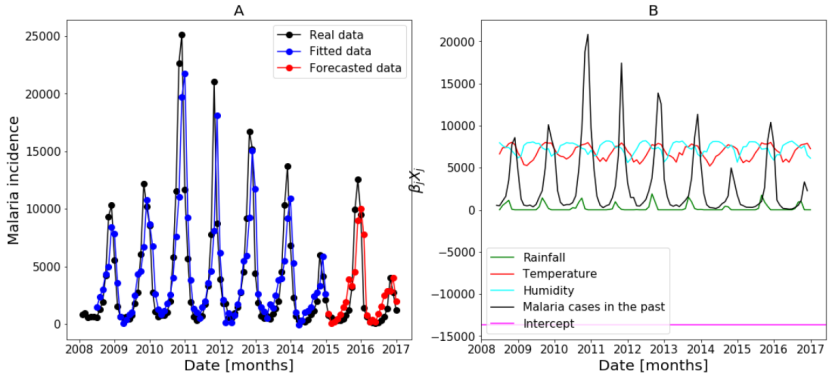


Fig. 7.1 Forecasting results (A) and the curves of $\beta_j X_j$ (B) in Dakar: Algorithm 4 with Poisson for f and identity for g, no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$.

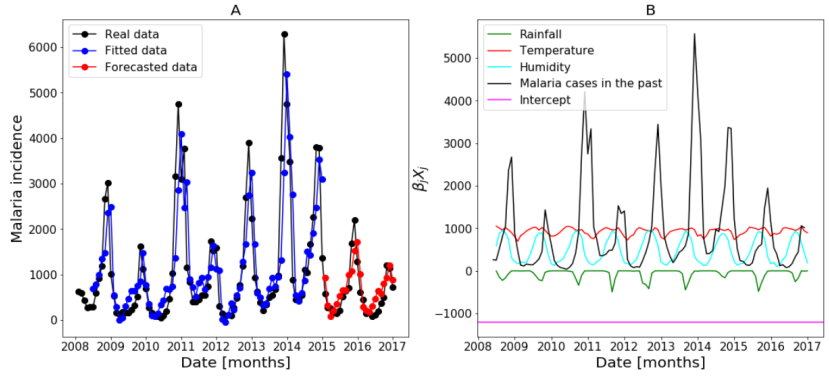


Fig. 7.2 Forecasting results (A) and the curves of $\beta_j X_j$ (B) in Fatick: Algorithm 4 with Poisson for f and identity for g, no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$.

We decide to show in Table 7.2 the performance of the Algorithms 2 (representing the protocol P1) and the 4 (representing the protocol P2)

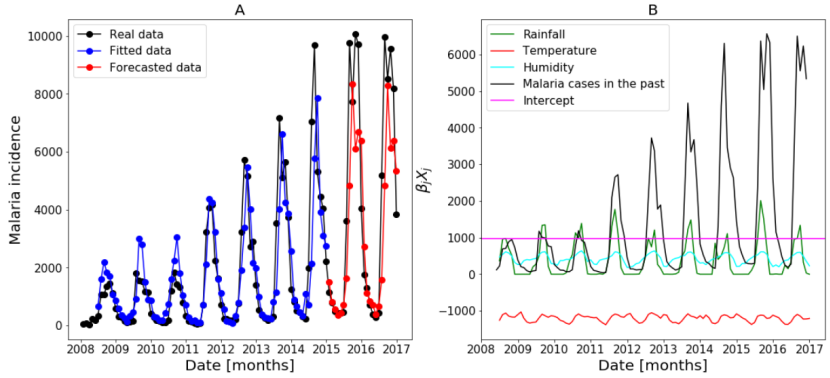


Fig. 7.3 Forecasting results (A) and the curves of $\beta_j X_j$ (B) in Kedougou: Algorithm 4 with Poisson for f and identity for g, no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$.

Region	Model	RMSE	MASE
Dk	Algo 2	2245.27 / 2689.74	0.52 / 1.02
	Algo 4	3338.28 / 2084.9	0.77 / 0.94
	SARIMAX [1,0,0][1,0,1,12]	2106.02 / 3602.89	0.52 / 1.9
Ft	Algo 2	772.32 / 543.56	0.82 / 1.26
	Algo 4	787.26 / 303.2	0.82 / 0.92
	SARIMAX [1 0 1][1 0 1 12]	632.89 / 767.43	0.72 / 1.96
Kd	Algo 2	1241.89 / 2431.91	0.99 / 0.87
	Algo 4	1095.23 / 2211.8	0.82 / 0.8
	SARIMAX [1 0 0][1 0 1 12]	763.25 / 2520.71	0.67 / 0.99

Table 7.1 Comparison results between the GLM with Poisson distribution (Algorithm 2, 4), and SARIMAX model (Algorithm 5), in the three regions where $t_s = 0$, $t_i = 5$, $t_c = 84$ and $t_e = 108$ and $h = 1$. Train/test values are reported. We denote by Dk: Dakar, Ft: Fatick and Kd: Kedougou.

when the test period is one year. These results reveal a clear improvement of the forecast quality in term of MASE, compared to the results in Table 7.1. A reduction of up to 44% is observed with the protocol P1 when the saturation method, defined in chapter 6 and Section 6.1.3, is applied in Dakar. In Fatick, a reduction of 11% is noted with the protocol P1 without

saturation.

Regions			RMSE	MASE
Dakar	P1	No_saturation	2245.27 / 2253.28	0.52 / 0.73
	P1	Saturation	2087.76 / 1269.15	0.49 / 0.58
	P2	No_saturation	3396.35 / 2110.79	0.78 / 0.93
	P2	Saturation	3232.91 / 2046.5	0.76 / 0.88
Fatick	P1	No_saturation	772.32 / 532.61	0.82 / 1.11
	P1	Saturation	772.94 / 530.94	0.82 / 1.1
	P2	No_saturation	788.0 / 342.97	0.82 / 0.89
	P2	Saturation	793.23 / 349.88	0.82 / 0.92
Kedougou	P1	No_saturation	1241.89 / 2553.82	0.99 / 0.9
	P1	Saturation	1208.47 / 2573.59	0.96 / 0.87
	P2	No_saturation	1074.69 / 2505.0	0.83 / 0.86
	P2	Saturation	1081.49 / 2521.66	0.84 / 0.87

Table 7.2 Performance measures of Algorithms 2 and 4 with saturation or no saturation when $t_s = 0$, $t_i = 5$, $t_c = 84$ and $t_e = 96$ and $h = 1$, in the three regions. Train/test values are reported.

7.2 Alternative optimization method

An alternative estimation method consists to compute the estimates $\hat{\beta}$ and α (for NB only) by applying the `scipy.optimize.fmin` optimization solver. This method, that we denote by method B, uses the *downhill simplex algorithm* with the log-likelihood as the objective function. In application, with the Poisson distribution, we solve

$$\arg\min_{\beta} -L(\beta, Y, X), \tag{7.2}$$

where L is the log likelihood function defined in equation (B.2) (defined in Appendix B). As for the NB distribution, we solve

$$\arg\min_{\beta,\alpha} -L(\beta, \alpha, Y, X), \tag{7.3}$$

where L is the log likelihood function defined in equation (B.8).

The purpose is to know if an optimization solver like method B can give better forecast results than the default and commonly used method for GLM: Newton-Raphson, that we denote here by method A. First, the solver requires an initial guess parameters, which is difficult with this method. We conduct experiments to investigate the importance of the choice of the initial guess for method B. We observe that method B stays fixed when we initialize it with the outcome of method A and that it gives worse RMSE scores when initialized at others points; see Table 7.3. This suggests that the log-likelihood function equation (B.2) is a multimodal function with several local nonglobal minimizers. The various initial guest for method B are defined as follows: β, 0, 2β, β^2, 10β, $\beta/10$ where β represents here the outcome of method A. The same conclusions with this method B are obtained in Fatick and Kedougou. That's why we decide to only show the table of results for Dakar.

	β_R	β_T	β_H	$\beta_{Y_{o\{t-1\}}}$	β_I	RMSE
A	44.61	-134.89	52.75	0.41	-598.59	2689.7
B	44.61	-134.89	52.75	0.41	-598.59	
B	**44.611**	**-134.89**	**52.75**	**0.41**	**-598.59**	**2689.7**
B	0	0	0	0	0	
B	**45.79**	**-11.56**	**12.86**	**0.385**	**-52.64**	**2812.6**
B	2×44.61	2×-134.89	2×52.75	2×0.41	2×-598.59	
B	**44.61**	**-134.89**	**52.75**	**0.412**	**-598.59**	**2689.7**
B	44.61^2	-134.89^2	52.75^2	0.41^2	-598.592^2	
B	**13.24**	**445.76**	**178.18**	**0.63**	**-21179.73**	**5189.8**
B	10×44.61	10×-134.89	10×52.75	10×0.41	10×-598.59	
B	**36.86**	**-1616.72**	**136.07**	**1.22**	**26872.21**	**3419.4**

Table 7.3 Results of the two methods (A and B) in Dakar. RMSE_test is reported. In method B, we report the initial guess and the estimates (in dark).

7.3 Forecast from various train periods

In this section, we want to chose the training period in a reasonable way. For that, we investigate the impact of the position of the train period on the forecast accuracy. The idea is to consider a fix test period and to slide

the train period by increasing the size by one year, in the past. To this end, we fix $t_c = 84$ and $t_e = 108$ (the end of the dataset), yielding the test period equal to $[t_c, t_e]$. Then, we choose t_i in the span values [72, 60, 48, 36, 24, 5]. We stop in 5 because the use of the lags in the explanatory variables. Recall that the reference train period was [5, 84] in the previous experiments. We calculate the accuracy measures such as RMSE, MASE and MARE in each step and the results are presented in Table 7.4. We also present the evolution of the RMSE in these various train periods in Figure 7.4.

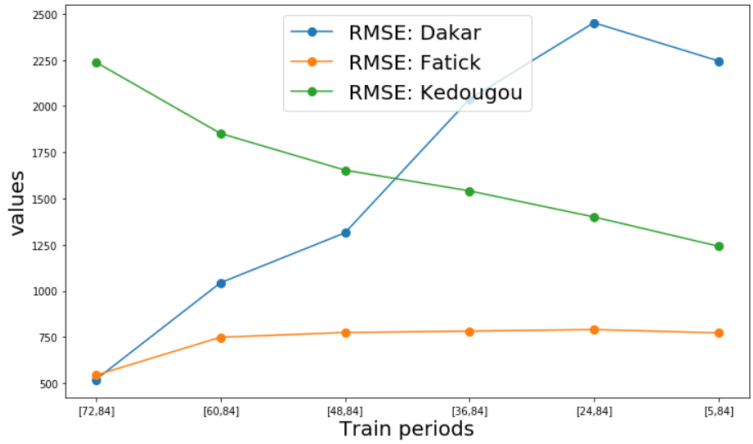


Fig. 7.4 Evolution of the RMSE values when the train period changes by increasing, in Dakar, Fatick and Kedougou, with Algorithm 2.

Results in Figure 7.4 show that RMSE and the train period are negatively correlated in Kedougou, and the minimum value is reached with the reference period [5,84] (RMSE = 1241.89). As for Fatick, the RMSE's present a stationary evolution close to the reference period [5,84] (RMSE = 772.32). And for Dakar, we observe an increasing of the RMSE's when the training period increases, but the maximum value is also reached around the reference period [5,84] (RMSE=2245.27).

We can conclude that the choice of a fix train period [5,84] is very credible with these data sets. Choosing a local train period for each region could be significant for the forecasts because we can't find a period where the RMSE is the minimum in the three regions due to variables are not identically distributed.

Regions	Period	RMSE	MASE	MARE
Dakar	[72,84]	519.99 / 2357.24	0.39 / 0.97	0.45 / 0.97
	[60,84]	1043.82 / 2051.23	0.44 / 0.87	0.46 / 0.99
	[48,84]	1314.17 / 2121.01	0.46 / 0.91	0.61 / 1.3
	[36,84]	2037.61 / 2723.05	0.48 / 1.05	0.63 / 1.41
	[24,84]	2451.89 / 3327.34	0.52 / 1.22	0.59 / 1.4
	[5,84]	2245.27 / 2689.74	0.52 / 1.02	0.54 / 1.28
Fatick	[72,84]	543.46 / 1340.96	0.53 / 3.59	0.31 / 2.21
	[60,84]	748.92 / 1480.54	0.65 / 3.71	0.46 / 1.79
	[48,84]	774.23 / 814.35	0.71 / 1.84	0.65 / 0.82
	[36,84]	781.87 / 657.38	0.75 / 1.52	0.56 / 0.72
	[24,84]	790.49 / 653.73	0.78 / 1.51	0.68 / 0.72
	[5,84]	772.32 / 543.56	0.82 / 1.26	0.65 / 0.6
Kedougou	[72,84]	2237.8 / 1980.91	0.94 / 0.73	0.59 / 0.74
	[60,84]	1852.45 / 2113.42	0.92 / 0.67	0.75 / 0.48
	[48,84]	1653.63 / 2199.98	0.91 / 0.75	0.81 / 0.5
	[36,84]	1541.43 / 2292.84	0.94 / 0.81	1.1 / 0.48
	[24,84]	1400.21 / 2402.79	0.97 / 0.85	1.0 / 0.48
	[5,84]	1241.89 / 2431.91	0.99 / 0.87	0.9 / 0.47

Table 7.4 Accuracy measures results from various train periods, with Algorithm 2, in Dakar, Fatick and Kedougou. Train/test values are reported.

7.4 Forecast from districts level

In this section, we study the performance of the three algorithms: 2, 4 and 5 in the different districts of Dakar, Fatick and Kedougou by displaying their MASE_test values in Figure 7.5.

The Figure 7.5 reveals that the Poisson regression model with Algorithm 4 gives the lowest MASE_test's (around 1) in the whole districts. Some peaks are observed with Algorithms 2 and 5 in Pikine, and a large peak is observed in Passy with only Algorithm 5.

In conclusion, we demonstrate that Poisson regression model through the Algorithm 4 performs better than the Poisson regression model through the Algorithm 2 and the SARIMAX model (Algorithm 5), using these datasets.

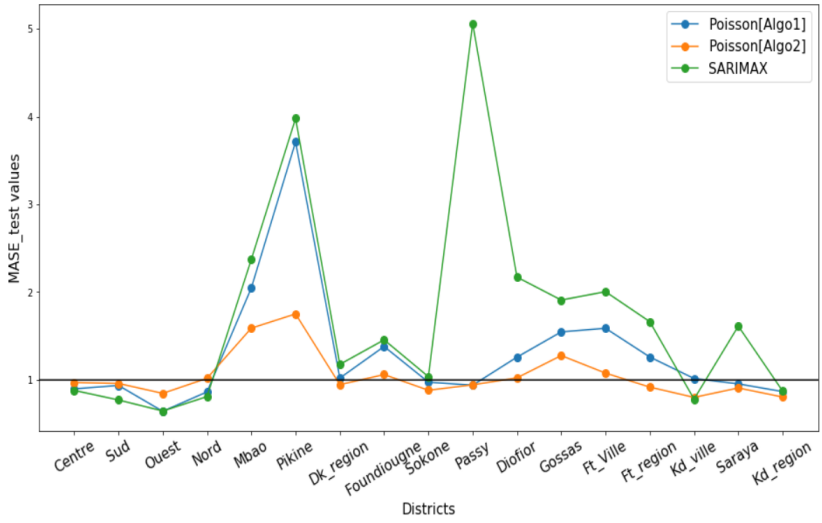


Fig. 7.5 MASE_test's as function of districts with the three algorithms: 2, 4 and 5. Dk_region represents the sum up off all districts of the said region including Dakar. Ft_ville means Fatick city and Ft_region represents the sum up off all districts of the said region including Fatick. Kd_ville means Kedougou city and Kd_region represents the sum up off all districts of the said region including Kedougou. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$ and $h = 1$.

7.5 Parameter generalization method

In this section, we decide to study how the parameters are distributed by regions in order to reflect on a generalized method of forecasting across regions. More clearly, we want to know if the parameters estimated with data in one region can be used to make forecasting in the other regions. The results of these experiments are presented in Table 7.5. We also show the distribution of each parameter into each region in Figure 7.6.

In Figure 7.6, we can observe that the parameter distribution is not homogeneous. For example, the variable intercept is distributed as follows: a median around -250 in Dakar, a median around -100 in Fatick and, a median between -2000 and -2500 in Kedougou. We also have the temperature variable with a median higher than the median values in Dakar and Fat-

Model	Dakar	Fatick	Kedougou
Fitted in Dakar	2689.74/1.02	3031.24/10.34	6374.86/2.77
Fitted in Fatick	2144.66/1.23	543.55/1.26	2478.32/0.88
Fitted in Kedougou	2343.30/0.98	549.01/1.65	2431.9/0.87

Table 7.5 Accuracy measures results by using the same parameter in the three regions, with Algorithm 2: Poisson for f, identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$, in Dakar, Fatick and Kedougou. RMSE_test/MASE_test are reported.

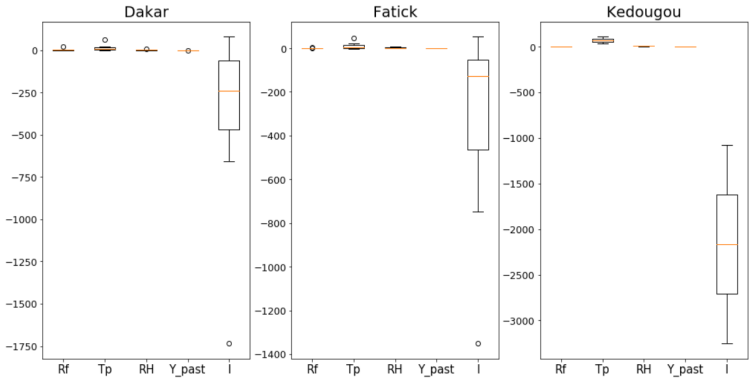


Fig. 7.6 Distribution of the regression parameters for each variable in the three regions: Algorithm 2 with Poisson for f and identity for g. The initial values are $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, $h = 1$. Recall: Rf for rainfall, Tp for average temperature, RH for relative humidity, Y_{past} for falciparum malaria count in the past and I for the intercept vector.

ick. The Table 7.5 reveals high RMSE_test's and MASE_test's in Fatick and Kedougou by using the estimates computed with Dakar data.

We can conclude that using the same parameter values in the three regions yields much poorer forecast results when the model is trained with Dakar data in terms of RMSE and MASE. Therefore, we suggest using this model specifically by locality to avoid overestimations or underestimations that may occur due to the difference in the size of malaria cases between localities or the climatic zone.

7.6 Partial conclusion

We made a comparison of the GLM with Poisson distribution and identity as the link function (implemented in Algorithms 2 and 4) and the SARIMAX model (implemented in Algorithm 5) with the historical monthly malaria cases and meteorological data from the three endemic regions of Senegal presented in chapter 2 and Section 2.2.

Algorithm 2 (the meteorological explanatory variable X_j is taken at time $t - \ell_j$, where t represents the time of the forecasts, ℓ_j is the `lag` in X_j that maximizes its correlation with the malaria incidence) seems to be very consistent on malaria incidence forecasting and corresponds more to the reality, even if Algorithm 4 (the meteorological explanatory variables are assumed to be available at time t) can improve the results. The results reveal a reduction in the sense of the RMSE_test about 22% in Dakar, about 44% in Fatick and about 9% in Kedougou, using Algorithm 4, compared to Algorithm 2 with the two years test period. We also tested the Poisson regression model in the one year test period and we compared its result to the results in Table 7.1 (corresponding to the two years test period). The results reveal a reduction of up to 44%, in terms of MASE, with the Algorithm 2 when the saturation method is applied in Dakar, and a reduction of 11% in term of the MASE with Algorithm 2 and without saturation in Fatick. This leads us to say that the GLM with Poisson distribution and identity as the link function can make good forecasts, but is not ideal for forecasting in subsequent years.

Furthermore, the comparative study of the two optimization methods: the Newton-Raphson and the `scipy.optimize.fmin` solver reveals that the Newton-Raphson is the more accurate method. Then, we tested the performance of these three algorithms on the district level and the Algorithm 4 seems to be better than the Algorithm 2 and the Algorithm 5. Finally, the parameter generalization method results show that the use of the same parameters leads to worse results in the three regions, both on the train and test sets and should not be considered with these datasets.

8

Producing confidence intervals for GLM

After producing the forecasts in chapters 6 and 7, now we would like to produce confidence intervals. Indeed, the major goal of the uncertainty analysis is to restrict the expected range in which the true value of the outcome of an experiment lies. This estimated range is in the form of an interval and is called the uncertainty interval [SMNM20].

Therefore, we present some methods for parameter sampling. First, we have the parameter uncertainty computed when the model is fitted. Second, we apply the parameter uncertainty method and the stochastic uncertainty method, which are defined in chapter 5. In these three methods, we fit the GLM with Poisson distribution (f) and id as link function (g) in the train period using the Python library `statsmodels.genmod.families`.

8.1 GLM default uncertainty method

Fitting a GLM with Poisson distribution also gives lower and upper bounds of the 95% confidence intervals for parameters. These values are computed with the covariance-variance matrix where the standard deviation is obtained by calculating the square root of the covariance-variance diagonal

matrix. Therefore, using these lower and upper bounds of parameters, we calculate the lower and upper bounds of the forecasts, which constitutes the confidence interval for forecasts. We present in capture 8.1, the statistics table when the model is fitted, only in Dakar in order to show the model outcome. The forecasts are displayed in Figures 8.2–8.4.

```
                Results: Generalized linear model
================================================================
Model:              GLM              AIC:            53009.3345
Link Function:      identity         BIC:            51924.5375
Dependent Variable: y                Log-Likelihood: -26500.
Date:               2023-03-27 12:35 LL-Null:        -2.0802e+05
No. Observations:   79               Deviance:       52248.
Df Model:           4                Pearson chi2:   5.70e+04
Df Residuals:       74               Scale:          1.0000
Method:             IRLS
----------------------------------------------------------------
            Coef.   Std.Err.     z     P>|z|    [0.025    0.975]
----------------------------------------------------------------
x1          44.6109   0.1684 264.9779 0.0000   44.2810   44.9409
x2        -134.8941   3.2066 -42.0670 0.0000 -141.1790 -128.6092
x3          52.7499   0.7139  73.8908 0.0000   51.3507   54.1491
x4           0.4125   0.0025 163.5850 0.0000    0.4075    0.4174
const     -598.5928  60.1922  -9.9447 0.0000 -716.5672 -480.6183
================================================================
```

Fig. 8.1 Summary statistics for GLM with Poisson distribution with Algorithm 2 in Dakar.

8.2 Application of the parameter uncertainty method

This method consists to slide the train period in order to get a sample of parameters. Letting $\hat{\beta}$, $\hat{\beta}^{\text{lower}}$ and $\hat{\beta}^{\text{upper}}$, respectively, the median of estimates vector, the mean of the lower bound estimates vector and the mean of the upper bound estimates vector. Then, we compute the forecasts by $\hat{\beta}X_{\text{test}}$ and the confidence interval of the forecasts by $[\hat{\beta}^{\text{lower}}X_{\text{test}}, \hat{\beta}^{\text{upper}}X_{\text{test}}]$. The results are displayed in Figures 8.5–8.7.

8.3 Application of the stochastic uncertainty method

In this section, we have to slightly modify the initial data (i.e, the trained data) given to the parameter estimation algorithm (see details in the chapter 5 and Section 5.1). We investigate two situations. First, we add noise only in the response variable (Y_o), and the other variables are not modified (Stochastic_1). Second, we add noise in all variables where a ϵ is calculated related to each of them (Stochastic_2). Results are displayed in Figures 8.8–8.13.

8.4 Discussion

As results, we observe from a visual inspection that these three methods defined in Sections 8.1, 8.2 and 8.3 give tight confidence intervals with the two forecasting algorithms, in the three regions. This can be explained by the fact that the parameter estimation method is very accurate, then produces small range of parameters between lower values (0.025 lower confidence limit) and upper values (0.975 upper confidence limit).

To assess the quality of these confidence intervals, we evaluate the validity and the *interquartile range (IQR)* [HL05]. Validity consists to evaluate the number of real data located into the confidence intervals. Optimality consists to evaluate the range of the interval by calculating the mean of the difference between the upper (97.5 percentile) and lower (2.5 percentile) of the forecasts.

Table 8.1 shows high percentages of real data located between confidence intervals of forecasts with parameter uncertainty method (section 8.2) where very closed results are observed between the two algorithms, in the three regions. This result is due to the higher values of IQR obtained with this method, compared to the other methods. In Dakar and Fatick, the Algorithm 2 gives less large confidence intervals than Algorithm 4, which is not the case in Kedougou. Additionally, we remark that when we only add the noise to the response variable, the outcome changes a bit from the default method (compare GLM default and Stochastic_1). The stochastic_2 method gives the worst values in the whole three regions. These bad results in this situation 2 can be caused by a non convenient determination method of ϵ in certain variables. In particular, in rainfall where high val-

ues are registered in rainy season and about 0 in dry season. But also, it can be explained by the fact that the model is less robust when we make a disturbance.

A possible investigation with the parameter uncertainty method can be to reduce the train period length because data are not uniformly distributed year by year and this variability may seriously affect the parameter estimation.

	GLM default		Parameter		Stochastic_1		Stochastic_2	
Validity								
Algo	2	4	2	4	2	4	2	4
Dakar	37.5	33.33	58.33	50.0	37.5	33.33	4.16	25.0
Fatick	62.5	45.83	45.83	70.83	54.16	45.83	25.0	16.66
Ked	41.67	33.33	50.0	54.16	41.66	29.16	12.5	33.33
IQR								
Dakar	824.99	970.49	1254.00	1374.02	796.49	878.74	719.39	838.36
Fatick	387.04	433.50	689.14	917.83	355.64	376.16	296.81	325.19
Ked	622.06	491.75	1182.80	1172.43	601.09	505.90	496.86	427.45

Table 8.1 Validity and IQR results, using the 3 different methods with Algorithm 2 and Algorithm 4, in Dakar, Fatick and Kedougou.

8.5 Partial conclusion

We investigated three methods to produce confidence intervals for GLMs: the GLM default uncertainty method (Section 8.1), the parameter uncertainty method (Section 8.2) and the stochastic uncertainty method (Section 8.3) using Algorithms 2 and 4.

Results show very tight confidence intervals with these methods. The parameter uncertainty method generally performs better with higher percentages of real data located into the confidence intervals with the two algorithms in the three regions.

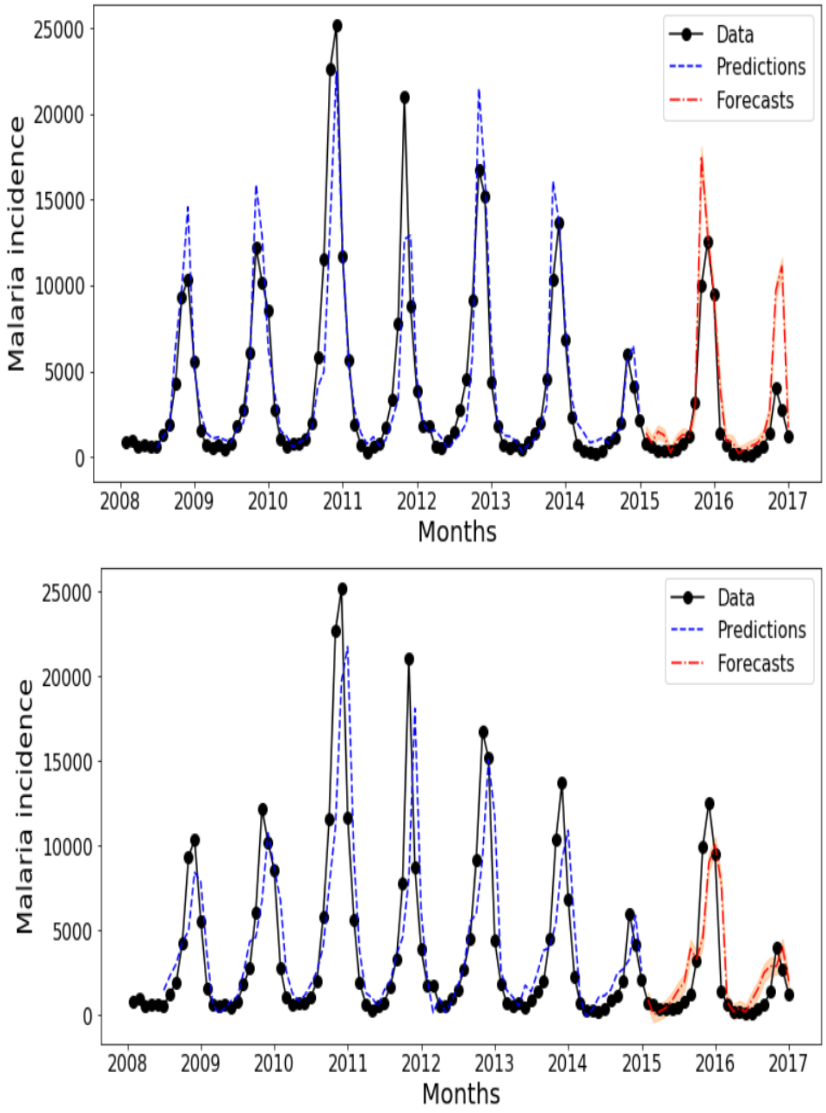


Fig. 8.2 Confidence interval result (orange) for Dakar data, using GLM default uncertainty method defined in Section 8.1 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month. The initial time values are $t_i = 5$, $t_c = 84$, $t_e = 108$.

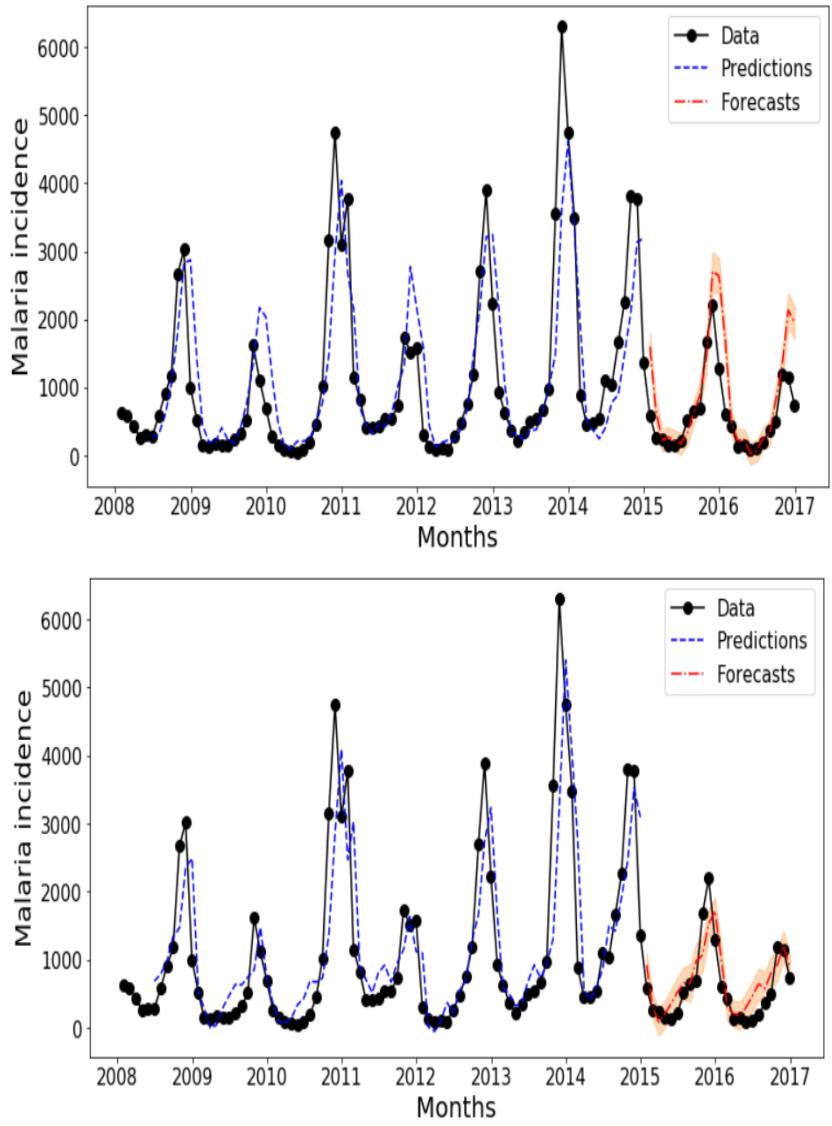


Fig. 8.3 Confidence interval result (orange) for Fatick data, using GLM default uncertainty method defined in Section 8.1 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month. The initial time values are $t_{\mathrm{i}} = 5$, $t_{\mathrm{c}} = 84$, $t_{\mathrm{e}} = 108$.

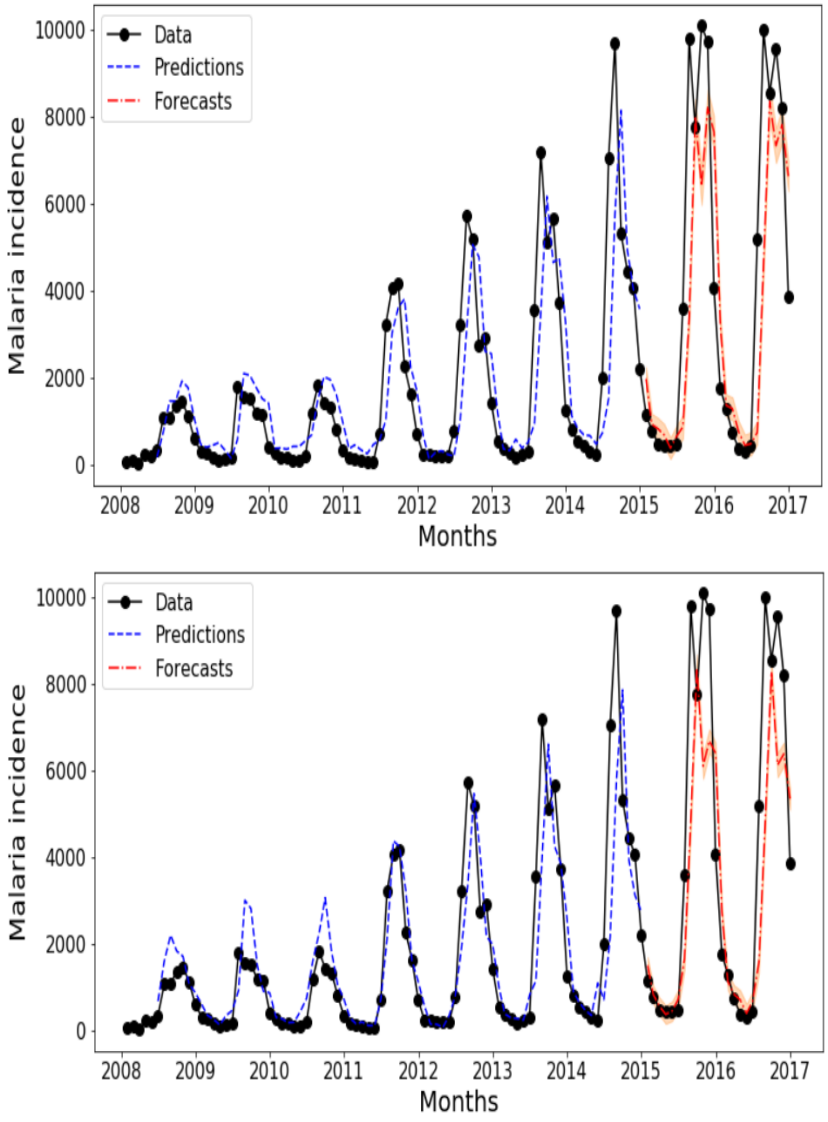


Fig. 8.4 Confidence interval result (orange) for Kedougou data, using GLM default uncertainty method defined in Section 8.1 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month. The initial time values are $t_i = 5$, $t_c = 84$, $t_e = 108$.

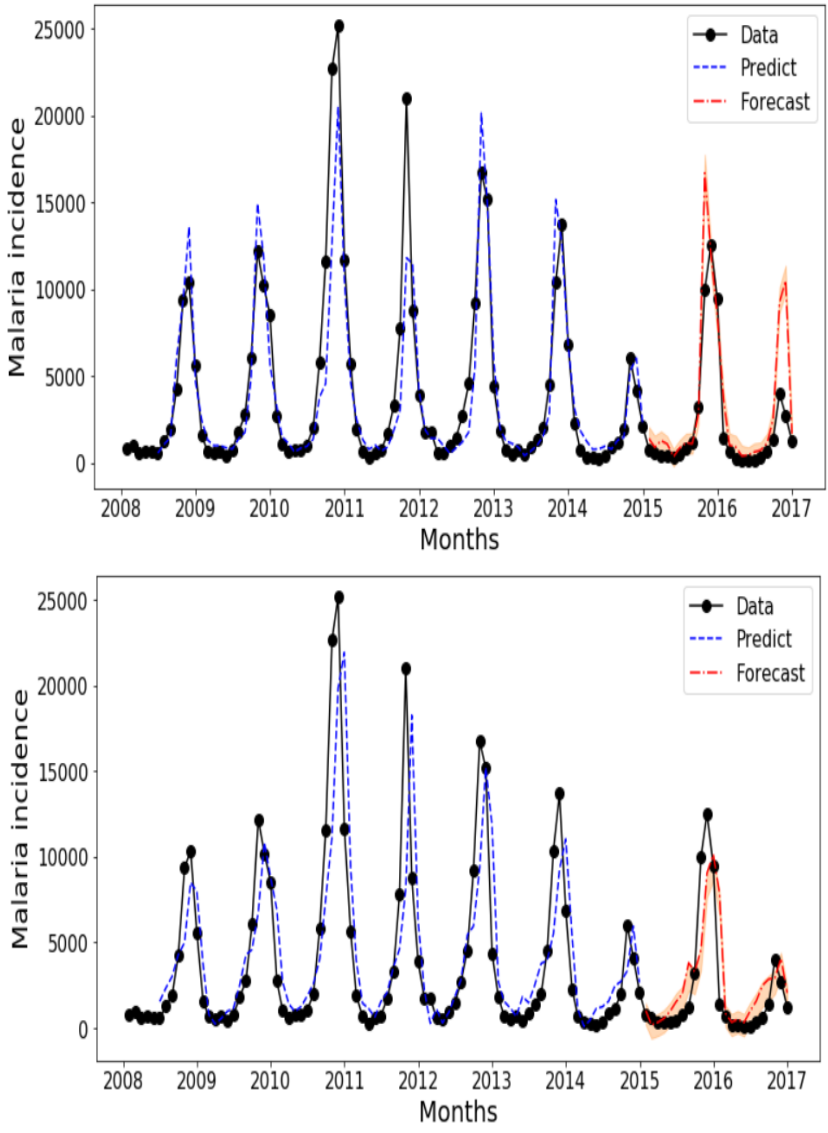


Fig. 8.5 Confidence interval for Dakar data, using parameter uncertainty method defined in Section 8.2 where $t_i \in [72, 60, 48, 36, 24, 5]$, $t_c = 84$, $t_e = 108$, with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

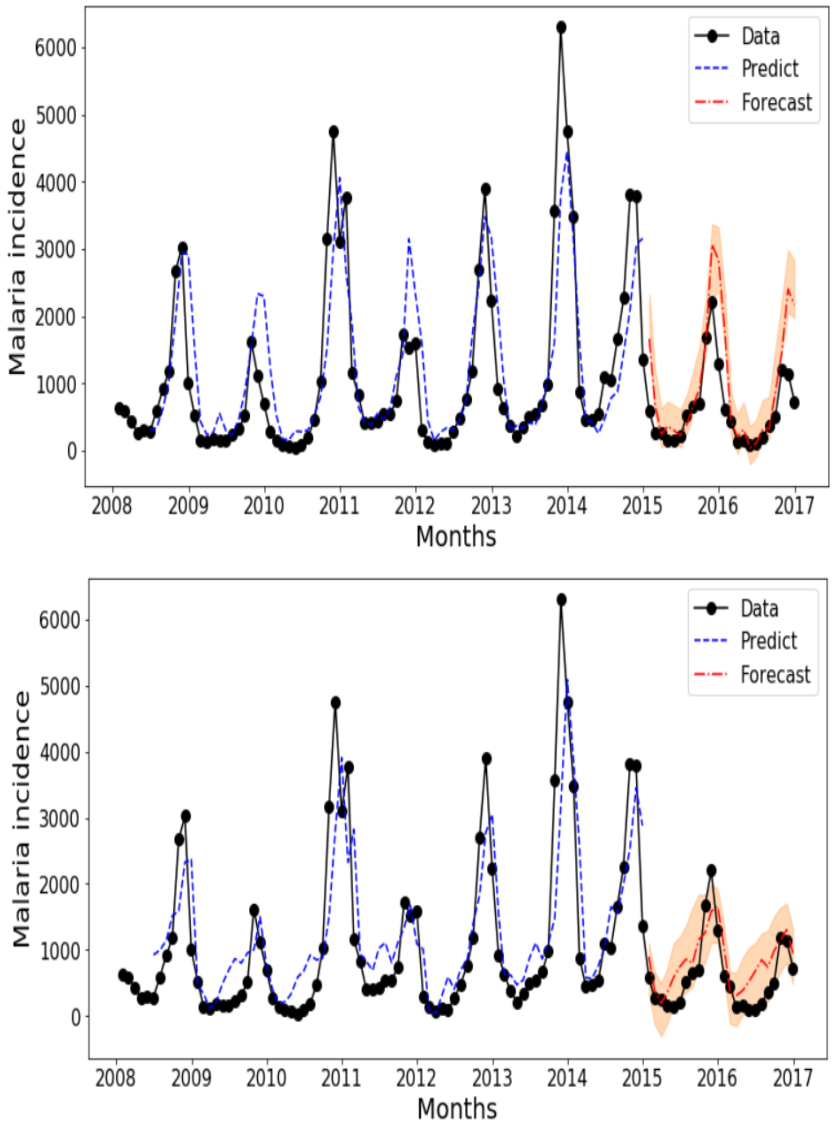


Fig. 8.6 Confidence interval for Fatick data, using parameter uncertainty method defined in Section 8.2 where $t_i \in [72, 60, 48, 36, 24, 5]$, $t_c = 84$, $t_e = 108$, with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

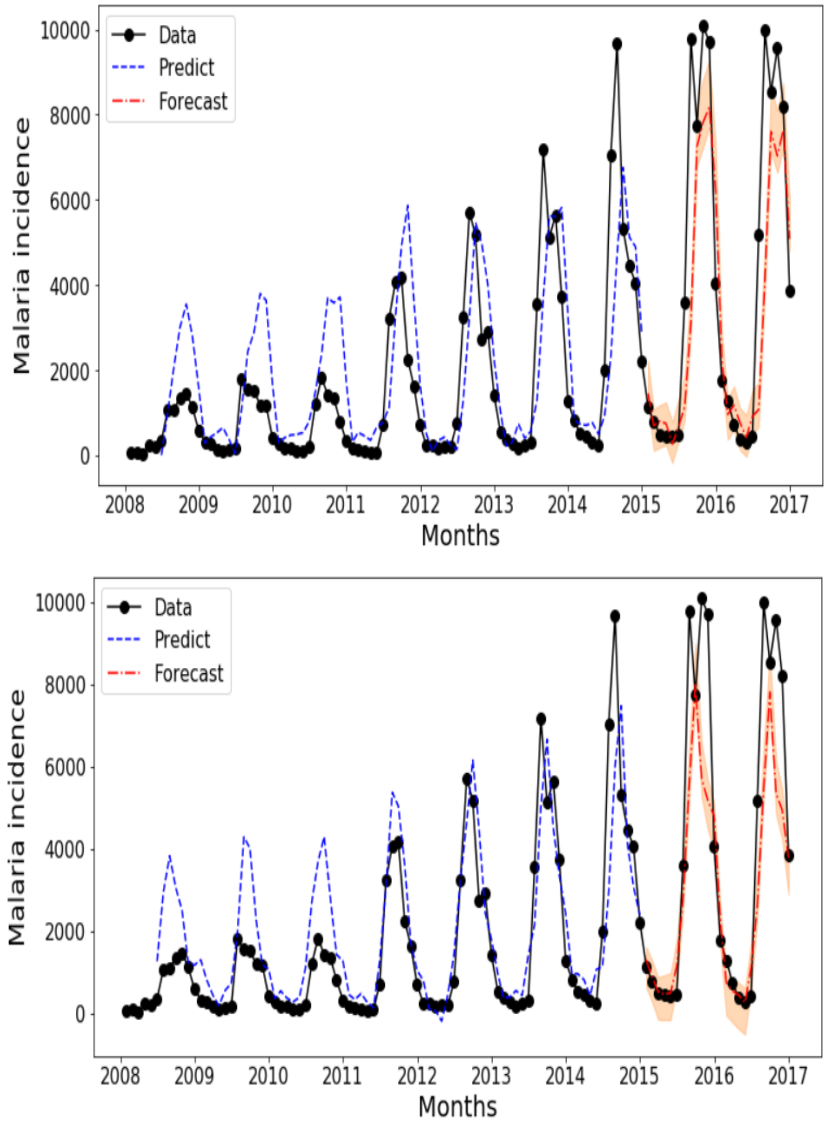


Fig. 8.7 Confidence interval for Kedougou data, using parameter uncertainty method defined in Section 8.2 where $t_i \in [72, 60, 48, 36, 24, 5]$, $t_c = 84$, $t_e = 108$, with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

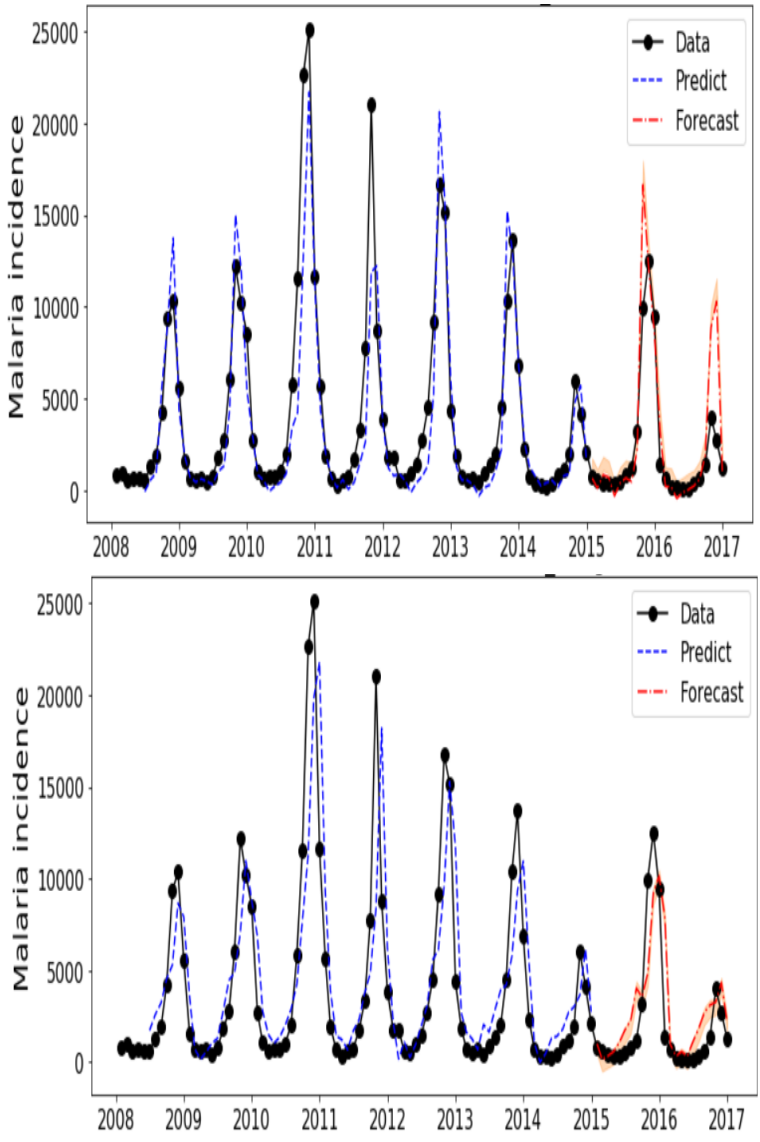


Fig. 8.8 Confidence interval for Dakar data in situation 1, using stochastic uncertainty method defined in Section 8.3 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

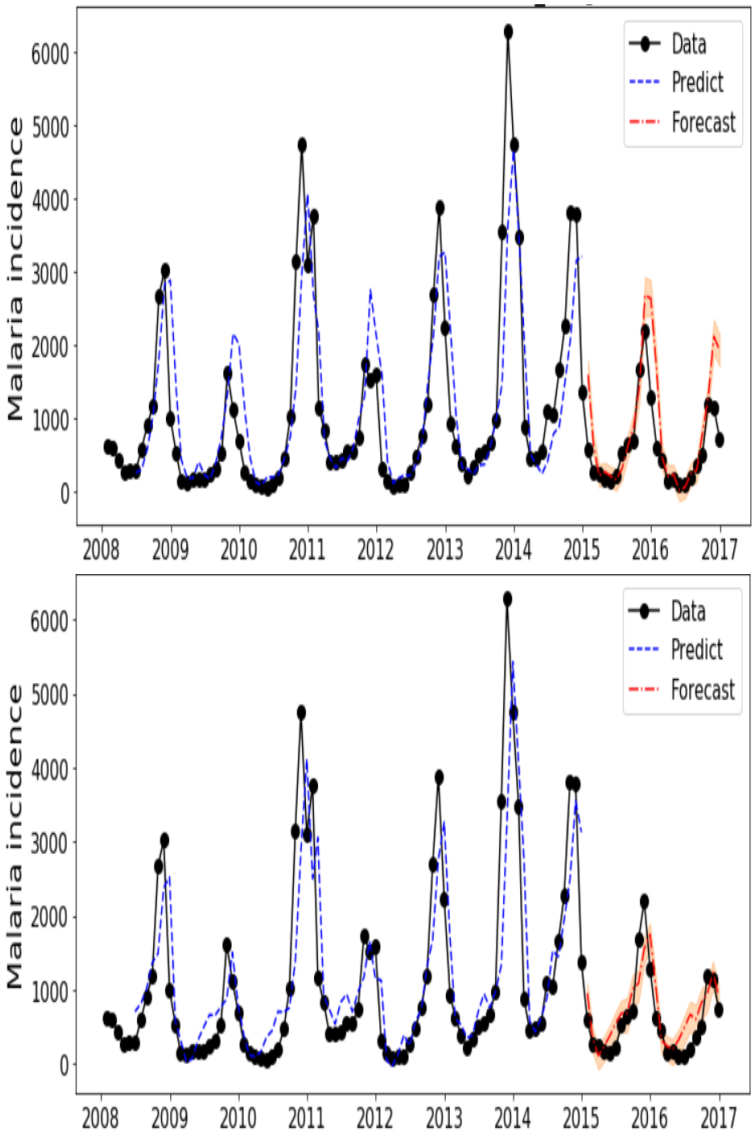


Fig. 8.9 Confidence interval for Fatick data in situation 1, using stochastic uncertainty method defined in Section 8.3 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

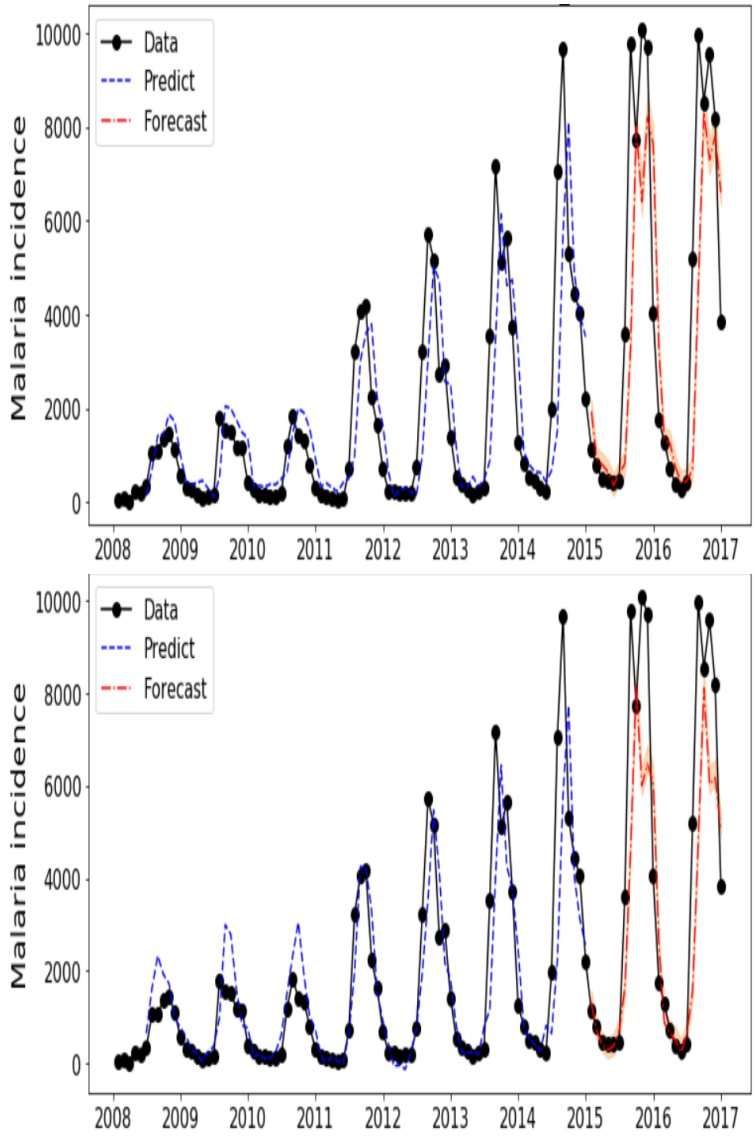


Fig. 8.10 Confidence interval for Kedougou data in situation 1, using stochastic uncertainty method defined in Section 8.3 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

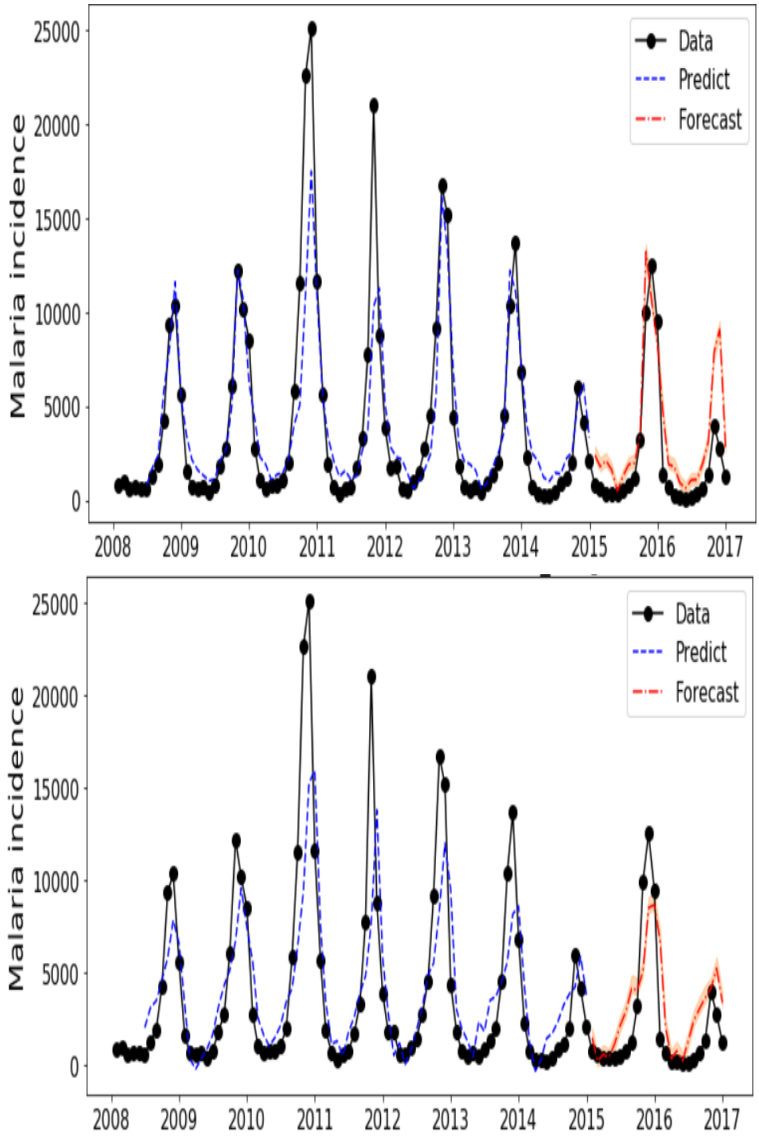


Fig. 8.11 Confidence interval for Dakar data in situation 2, using stochastic uncertainty method defined in Section 8.3 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

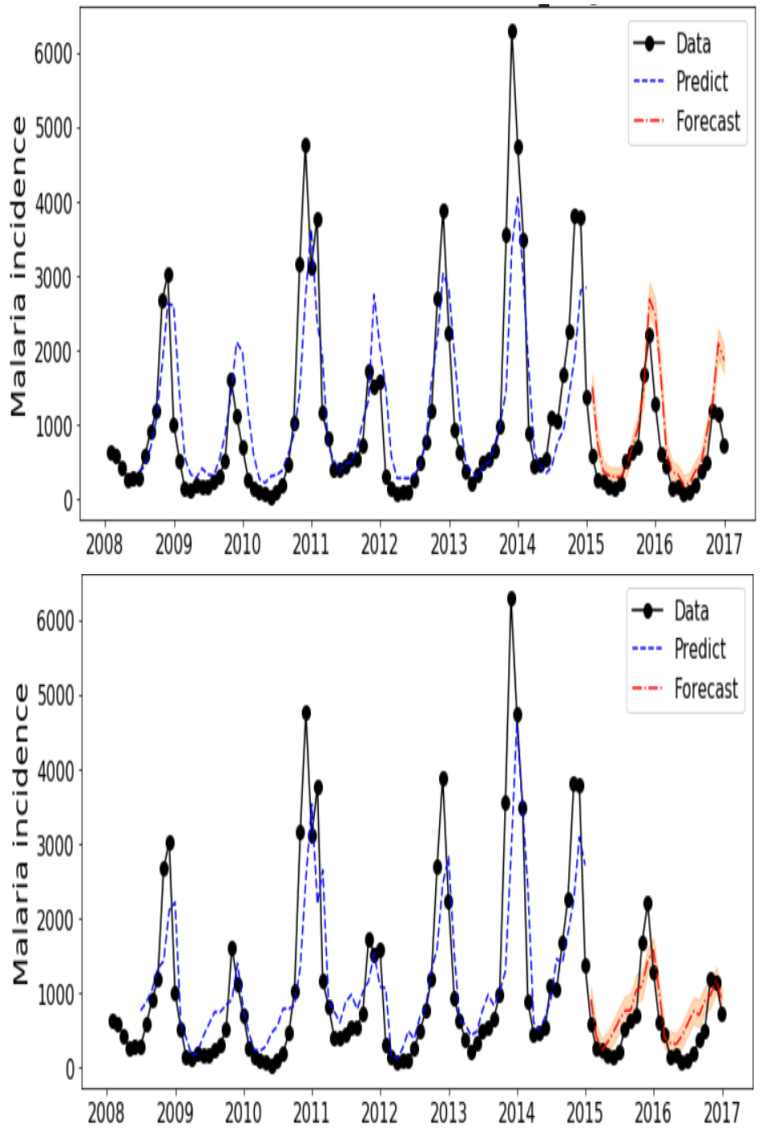


Fig. 8.12 Confidence interval for Fatick data in situation 2, using stochastic uncertainty method defined in Section 8.3 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

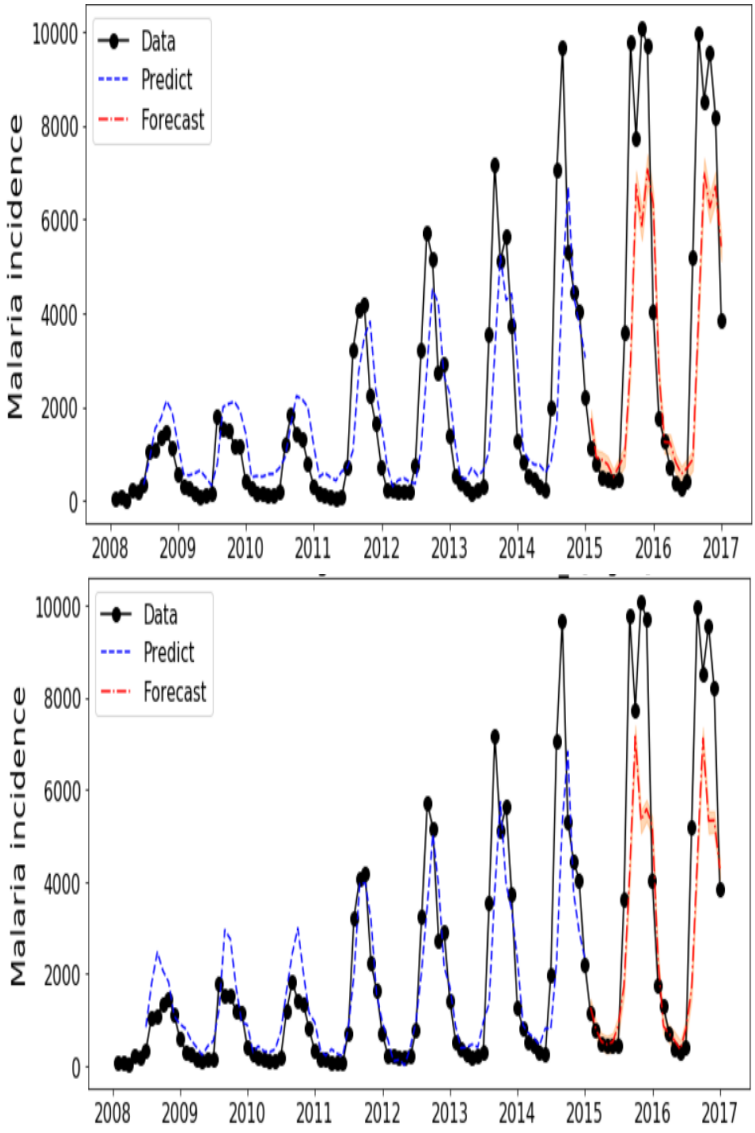


Fig. 8.13 Confidence interval for Kedougou data in situation 2, using stochastic uncertainty method defined in Section 8.3 with Algorithm 2 in top and Algorithm 4 in bottom. Malaria incidence means the falciparum malaria incidence count per month.

9

Mosquito density forecasting based on weather factors

METEOROLOGICAL factors such as rainfall, temperature and relative humidity are known to play a big influence in the vector development and the malaria incidence occurring.

Several studies have been constructed based on the climatic variables and the environment area of the vector life, that are considered to be susceptible to affect the parasite vectors (mosquitoes) development. For example in [ADH+06], a study of the seasonality and the dynamics of infectious diseases mentioned that the temperature affects the mosquito vector abundance, the biting rates and parasite development within the vector. In [GBMM16], the effect of the temperature and the rainfall was incorporated into a compartmental transmission model. The results in [GBMM16] reveal that malaria burden is likely to increase in the tropics, the highland regions, and east Africa and along the northern limit of falciparum malaria. In [NNB+13], a spatiotemporal analysis is made in order to identify the host preferences and feeding patterns of malaria vectors in the sylvo-pastoral area of Senegal. Then, the authors also show the impact of landscape classes. Results in [NNB+13] reveal that the differences in the feeding patterns of malaria vectors are merely linked to the specific localization of villages and are not influenced by landscape class distribution. In another study [NFT+14], the influence of landscape and climatic pa-

rameters on `Anopheles arabiensis` seasonal densities and infection rates is shown in a sahelian area of Senegal. As findings in [NFT+14], the mean densities of `Anopheles arabiensis` displayed a high seasonal differences with peaks observed in August or September. Then, a positive and non significant correlation was observed between `Anopheles arabiensis` densities for rainfall and humidity whereas a negative non-significant correlation was reported for temperature.

To quantify the effects of the climatic variables in the vector density, mathematical or statistical models are developed. In most of the time, generalized linear models (GLMs) have been used in the literature. GLMs are composed by exponential family distributions such as numerous distributions: Poisson and Negative Binomial for count data; Binomial, Bernoulli, and Geometric for discrete data; Gamma, Normal, Inverse Gaussian, Beta, and Exponential for the study of continuous response dataset [YMMA20]. Therefore, if the data follows a normal distribution so the log-normal regression model could be appropriated [ADA20].

In this study, we apply a GLM with log-normal regression model for mosquito densities (representing the response variable (Y_o)) forecasting based on the rainfall, the temperature and the relative humidity, in four villages of Senegal.

9.1 Models and methods

We supposed that the data follow a log-normal distribution defined in Section 1.5.2. We use Algorithm 2, defined in chapter 6, for forecasting.

9.2 Data origin

Data was kindly provided by Pasteur institute of Dakar entomological team. They were collected from July to November in four villages of Senegal:

- Keur_Aliou: Longitudes (15°15’86"N), Latitudes (14°50’30"W) and Landscape (Bare soils),

- Keur_Alpha: Longitudes (15°14′42"N), Latitudes (14°47′57"W) and Landscape (Shrubby savanna),

- Keur_Diallo: Longitudes (15°19′10"N), Latitudes (14°50′24"W) and Landscape (Steppe),

- Barkedji: Longitudes (15°17′0"N), Latitudes (14°53′0"W) and Landscape (Wooded savanna).

The density represents the number of mosquitoes per person (mos/per, Y_o). The relative humidity (RH), the temperature (Tp) and the rainfall (Rf) are measured directly in the villages because the weather stations were installed. Data are presented in Tables 9.1–9.4.

Months	Density (mos/per)	Rainfall (mm)	Temperature (°C)	Humidity (%)
July	0.0	72.4	28.31	79.61
August	5.4000	199.4	28.38	81.79
September	7.636364	7.0	28.71	84.66
October	2.000000	3.0	29.39	63.90
November	4.285714	0.0	28.88	28.20

Table 9.1 Data base in Keur_Aliou.

Months	Density (mos/per)	Rainfall (mm)	Temperature (°C)	Humidity (%)
July	0.285714	31.2	27.037363	83.602929
August	2.2000	125.3	27.111398	86.200000
September	24.25000	85.0	27.645486	87.895069
October	4.666667	21.0	28.311207	67.112555
November	0.333333	12.3	27.453319	31.401042

Table 9.2 Data base in Keur_Alpha.

Months	Density (mos/per)	Rainfall (mm)	Temperature (°C)	Humidity (%)
July	1.333333	118.2	29.5	67.8
August	2.666667	146.5	27.5	76.0
September	6.00000	121.2	29.8	71.3
October	3.666667	13.1	30.9	52.2
November	1.00000	0.0	30.2	29.4

Table 9.3 Data base in Keur_Diallo.

Months	Density (mos/per)	Rainfall (mm)	Temperature (°C)	Humidity (%)
July	0.125000	33.1	28.31	79.61
August	0.908587	118.3	28.38	81.79
September	3.210366	192.0	28.71	84.66
October	3.376855	137.3	29.39	63.90
November	3.149856	14.8	28.88	28.20

Table 9.4 Data base in Barkedji.

9.3 Results

We present the correlations results in Table 9.5, the accuracy measures in Table 9.6 and the forecasts in Figure 9.1–9.4. Figures noted by B have the same explanations as in chapter 6, Section 6.4.1. When we refer to Algorithm 2 (defined in chapter 6), we only apply a lag to the rainfall variable. Thus, because after applying the equation (6.3) (defined in Section 6.2.1) with these data, we find that the maximum correlation between rainfall and density is reached at lag 1, and 0 with other variables. Results in Table 9.6 show that the model fits accurately the data, but the forecasts are generally less good. For example, a very good forecasts is only observed in Keur_Alpha (RMSE_test=0.09) and Keur_Diallo (RMSE_test=0.78). That can be explained by the fact that there are not enough data to test, and/or the optimization method makes fewer iterations to find the adequate regression coefficients, due to the small data base.

Villages	Y_o&Rf	Y_o&Tp	Y_o&RH
Keur_Aliou	0.047	-0.03	0.108
Keur_Alpha	0.357	0.286	0.422
Keur_Diallo	0.279	0.109	0.466
Barkedji	0.379	0.839	-0.474

Table 9.5 Pearson correlations between mosquito density and explanatory variables. A positive value of the r implies correlation between density and a variable while negative value proves anti-correlation. Any lag is considered here.

Village	RMSE_train	RMSE_test
Keur Aliou	0.18	3.67
Keur Alpha	1.15	0.09
Keur Diallo	0.07	0.78
Barkedji	0.1	2.65

Table 9.6 Accuracy measures in the four villages.

9.4 Partial conclusion

A GLM can be nicely used on small database to fit and forecast the vector density taking into account climatic variables such as rainfall, temperature, humidity. This study gives acceptable results that can be taken with high consideration despite the small dimensions of the dataset. However, the choice of the probability distribution and link function can be very important to get good quality of forecasts.

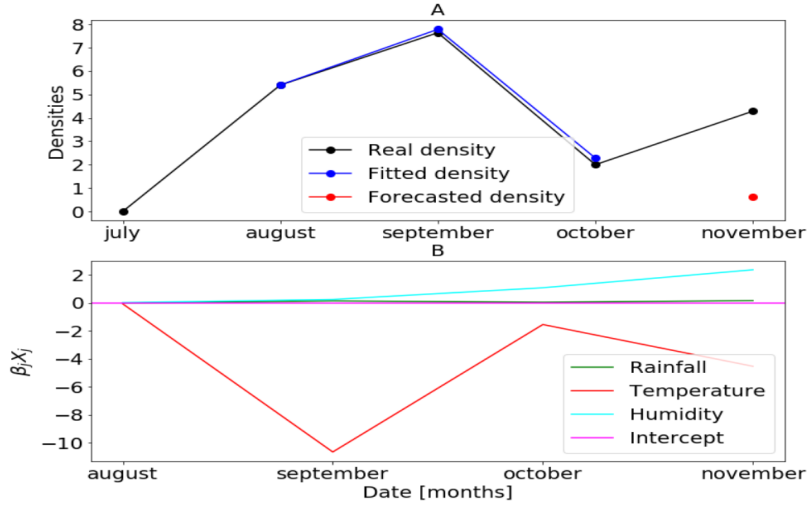


Fig. 9.1 Fitting and forecasting results in Keur_Aliou (A) and $\beta_j X_j$ curves (B), $t_i = 1$, $t_c = 3$, $t_e = 4$, $h = 1$ and GLM with log-normal distribution.

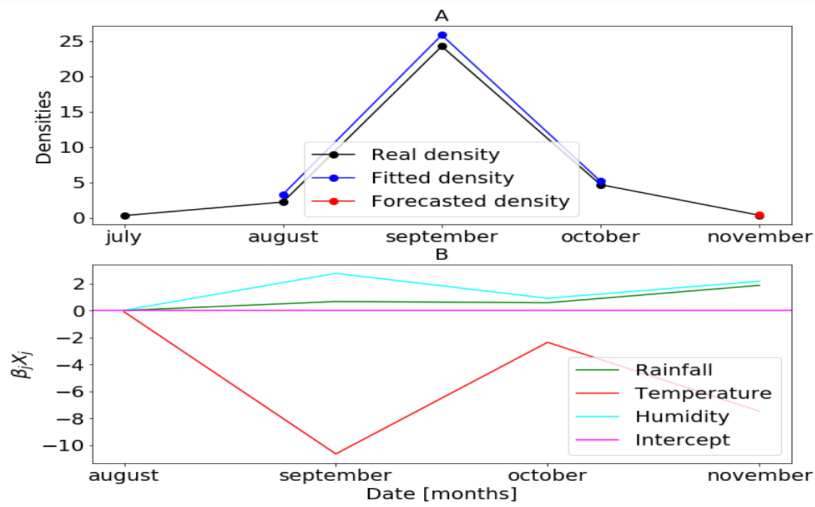


Fig. 9.2 Fitting and forecasting results in Keur_Alpha (A) and $\beta_j X_j$ curves (B), $t_i = 1$, $t_c = 3$, $t_e = 4$, $h = 1$ and GLM with log-normal distribution.

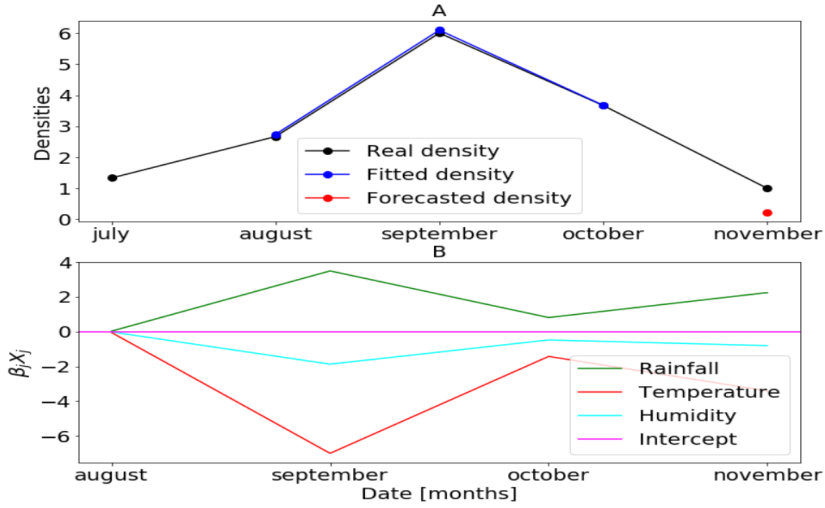


Fig. 9.3 Fitting and forecasting results in Keur_Diallo (A) and $\beta_j X_j$ curves (B), $t_i = 1$, $t_c = 3$, $t_e = 4$, $h = 1$ and GLM with log-normal distribution.

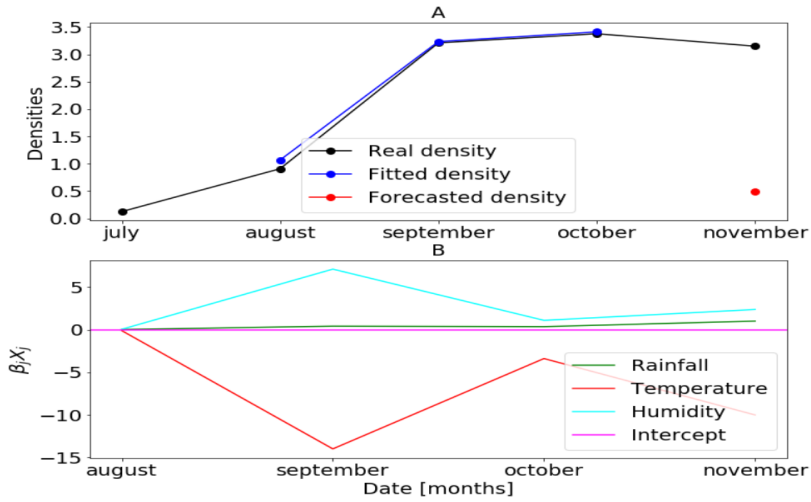


Fig. 9.4 Fitting and forecasting results in Barkedji (A) and $\beta_j X_j$ curves (B), $t_i = 1$, $t_c = 3$, $t_e = 4$, $h = 1$ and GLM with log-normal distribution.

PART III
Conclusion

10

Conclusion

In this thesis, we used datasets related to COVID-19 pandemic in Belgium, France, Luxembourg and the United Kingdom, and malaria epidemic in three endemic regions of Senegal: Dakar (hypoendemic zone), Fatick (endemic zone), and Kedougou (hyperendemic zone). We also used the mosquito densities data collected from July to November in Keur Aliou, Keur Alpha, Keur Diallo and Barkedji, four endemic villages of Senegal where weather stations were installed.

10.1 Main findings, scientific contributions

10.1.1 In COVID-19 modelling

The experiments from the SH model have shown that the proposed method has a remarkably good fitting accuracy over the whole available COVID-19 data. It also produces remarkably accurate forecasts over certain time ranges for some areas (Belgium, some Belgian provinces, and a few French departments), also for Luxembourg and the United Kingdom. We have investigated the polynomial model and the exponential decrease model as alternative models to this SH model. Results have shown that the SH model gives more accurate forecasts than the two alternative models especially when the train period is chosen around the peak. These results have

demonstrated that the parameter estimation method proposed in chapter 3 is more reliable than the least squares method proposed in chapter 4.

The proposed method to forecast hospitalized cases with data related to deaths and patients discharged from the hospital in Belgium has shown less good quality of forecasting than when the data are entirely based on the observed hospitalized cases. The weekly variability data reduce the performance of the least squares estimation method, compare Figures 4.5, 4.8 (chapter 4) and 3.3 (chapter 3).

The uncertainty results of the SH model has shown that the method 2 (a sliding train period) gives better results than the method 1 (a noised initial data set) and method 3 (a non-constant value for the hospitalization rate $\bar{\beta}$) with almost 50% of the real data located between the minimum and the maximum forecasts in Belgium, and 43% in France.

10.1.2 In malaria modelling

We have investigated the accuracy, in the sense of the RMSE, MASE and MARE, of the malaria incidence forecasts obtained with GLM's by taking into account the meteorological data and history of malaria incidence as explanatory variables, in Dakar, Fatick and Kedougou. A machine learning approach is developed, based on a separation of the data into a train-set to estimate the parameters by maximum likelihood and a test-set to assess the forecast accuracy. The forecast accuracy of GLMs with various distributions (Poisson, negative binomial, and Gaussian) and link functions (identity, log, and sqrt) is compared in terms of several model performance metrics. Whereas the best distribution-link combination varies according to the endemic region of interest and the performance metric, the experiments lead to the conclusion that the Poisson distribution with the identity link is overall the most suitable combination.

Algorithm 2 defined in chapter 6 (the meteorological explanatory variable X_j is taken at time $t - \ell_j$, where t represents the time of the forecasts, ℓ_j is the `lag` in X_j that maximizes its correlation with the malaria incidence) seems to be very consistent on malaria incidence forecasting and corresponds more to the reality, even if Algorithm 4 defined in chapter 7 (the meteorological explanatory variables are assumed to be available at time t) can improve the results. The results reveal a reduction in terms of the RMSE_test about 22% in Dakar, about 44% in Fatick and about 9% in Kedougou, using Algorithm 4, compared to Algorithm 2 with the two years

test period.

In addition, a saturation method was introduced on the rainfall variable to remedy some overestimations observed during the forecasts. This method has reduced by 4% in the sense of MARE the over-estimation occurring at the end of 2015, in Dakar. Addition and ablation studies were developed to show the influence of each explanatory variable in the forecasts. Results have shown that removing the history of malaria cases from the explanatory variables has a strong adverse effect on the forecast accuracy.

We have observed that the SARIMAX model (Algorithm 5) produces the best results in the sense of the RMSE and MASE in the train period, compared to GLMs (Algorithm 2 and 4) in the three regions. This result confirms the findings in [NPC21] revealing that the SARIMAX model generally gives a good fitting. But it produces less good quality of forecasts.

The results of the parameter generalization method have shown that the use of the same parameter values in the three regions leads some poorer results, both on the train and test sets and should not be considered with these datasets.

We applied three methods producing confidence intervals for GLM: GLM default method, parameter uncertainty method and stochastic uncertainty method that give very tight confidence intervals. Among these methods, the parameter uncertainty method seems to be more reliable in the sense of validity and the GLM default method gives better results in the sense of IQR.

We have also shown that the GLMs can be applied in small datasets if an adequate probability distribution and link function is well chosen, in the chapter 9. But these results should be considered with caution as they are not very indicative due to the small size of the dataset used.

10.2 Limitations and perspectives for future research

10.2.1 In COVID-19 modelling

If the proposed SH model is used to guide prevention policies for COVID-19 pandemic, then further caveats are in order. The model is unable to produce multiple peaks in order to fit or forecast rebounds due to the way it is constructed. We have seen that the estimation of $\bar{\beta}$ is very sensitive.

Hence the proposed model can hardly help assess the impact of prevention measures on $\bar{\beta}$ (see Figure 10.1 presented below). Without knowing sufficiently the impact of prevention measures on $\bar{\beta}$, we may not aptly use the model to predict their impact on the evolution of the hospitalizations. Yet another caveat is that it may be tempting to deduce from the excellent fit with a constant-parameter model (Figure 3.1 in chapter 3) that the evolution of the prevention measures over the dataset period has had no impact on $\bar{\beta}$. But the deduction is flawed. Indeed, in view of the comments made in Section 3.3.1 (chapter 3), the available data could also be very well explained with fairly large jumps in $\bar{\beta}$ during the decrease phase.

Furthermore, in order to address data variability (weekly variability) and to robustify the least squares procedure, we can smooth out the time series by compting its moving average (MA) in a time span of seven days as the method defined in chapter 3 and Section 3.4.3.

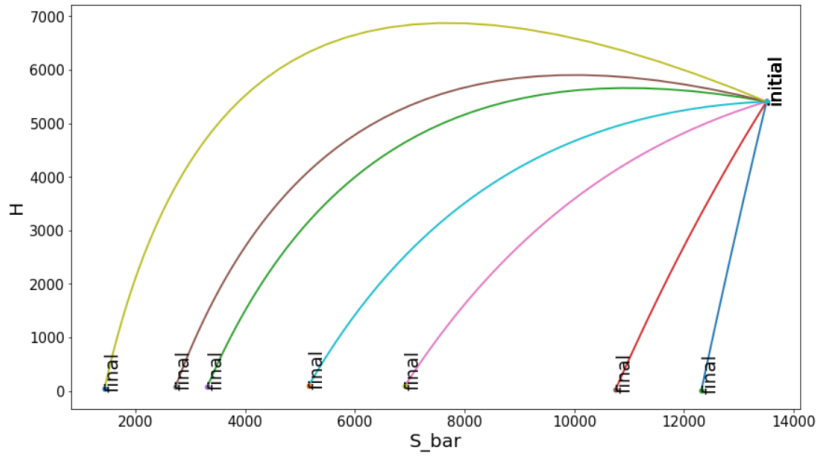


Fig. 10.1 The phase portrait trajectories of the hospitalized individuals (H) and the susceptible to be hospitalized individuals (S_bar) from various values of $\bar{\beta}$ (corresponding to the different steps of modification of the prevention measures).

It could be interesting to insert a delay in the SH model as in [Nes20b, Koz20]. Indeed, with a delay, it may be wise to consider right away the discrete-time version (like (3.13)) and skip continuous-time considerations (like (3.8)) defined in chapter 3. I'm thinking that equation (3.2) (defined in chapter 3) could be replaced by something like $E(t+\Delta) = \alpha\beta I(t)S(t)$,

where E is the newly hospitalized people, βIS is the newly infected people (in the classical SIR model), Δ is the delay between infection and hospitalization, and α is the proportion of infected people who get hospitalized.

10.2.2 In malaria modelling

This proposed GLM model gives some peaks in the forecasts (end of 2016 in Dakar in Figure 6.7 and Fatick in Figure 6.8). These peaks can be caused by an inadequacy of this model due to its linear character even if a non-linearity is then applied for rainfall. This may also be due to a variable ratio of infected people being sufficiently ill to go to the hospital and be confirmed, thus causing a bias in the collected data. Another and important explanation to the over-forecasting (or under-forecasting) observed can be the unavailability of some explanatory variables: the distance to water bodies, the normalized differenced vegetation index (NDVI), the night and day LST (land surface temperature), the ownership and use of insecticide treated nets (ITNs) and the intermittent preventive treatment distributed for pregnant women (IPTp) that are considered to fit malaria incidence [GGK+12]. Providing the enhanced vegetation index (EVI), and actual evapotranspiration (ETa) could help to improve the malaria incidence fitting model [MSH+12]. In our available data, there was the ITNs distributed because it constitutes the main factor fighting against malaria according to [FCG+18] but its distribution was not very regular (see Figures 2.8–2.10 in chapter 2). Needless to say, the ownership and use of ITNs would be a more suitable explanatory variable. As for ACT distributed which also was available, we have concluded that it is not a real predictor because it is usually taken after the appearance of some malaria symptoms from a patient (see Figures 2.11–2.13 in chapter 2).

The study is led with monthly incidence due to the unavailability of the daily malaria incidence reported that could have helped us to better understand the influence of the climatic data.

It could be interesting to develop a malaria transmission model like an SEIR (for humans)-SEI (for mosquitoes) type in these three regions including the effects of meteorological factors and some preventive measures as in [DDAAB20b, JHX20].

10.3 Implications for policy

The models developed in this thesis and used for the modelling of COVID-19 and malaria represent a precious tool in comparing the disease outbreak in different localities and evaluating the distribution of preventive measures. Thus, the results will be useful at various levels of decision making, for example, in setting up an early warning and sustainable strategies for climate change and adaptation for malaria vectors control programs in these regions. The algorithms developed can be applied to other regions outside Senegal.

For COVID-19, in spite of all the mentioned caveats, the hospitalization forecasts returned by the method, and also the evolution of $\bar{S}(t)$, might be of practical use in the context of various disease outbreaks, e.g., for resource planning. To this end, it will be important to understand which specific features of the COVID-19 outbreak made it possible to forecast so accurately the hospitalization decrease several months ahead.

As for malaria, the proposed model helps to quantify in advance the malaria incidence in order to optimize the shipment of necessary materials (medical or equipment) to the regions and especially the districts. In addition, it could allow the PNLP to properly manage the distribution of nets in a timely manner.

A

Implementation and data bases

A.1 Codes

The code used for this thesis is available on my GitHub
`https://github.com/OD1992/Python-programs.git`.
All simulations and data analysis are done with Jupyter Notebook (Anaconda 3) and Spyder (Python 3.7).

A.2 Data bases URL link

We present all the data bases used for experiments in this thesis.

- COVID-19 hospitalizations dataset for Belgium:
  `https://epistat.sciensano.be/Data/COVID19BE_HOSP.csv`
  come from `https://epistat.wiv-isp.be/covid/`.
- Daily reported cases for Belgium: `https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv`.

- COVID-19 deaths for Belgium:
  https://ourworldindata.org/coronavirus-source-data.
- COVID-19 epidemiological dataset for France:
  donnees-hospitalieres-covid19-2020-07-17-19h00.csv file comes from
  https://www.data.gouv.fr/en/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/.
- COVID-19 epidemiological dataset for Luxembourg:
  datapublic-covid19.csv is obtained from
  https://data.public.lu/fr/datasets/donnees-covid19.
- COVID-19 epidemiological dataset for United Kingdom:
  the Patients-admitted and Patients-in-hospital are obtained from
  https://ourworldindata.org/coronavirus-source-data.
- Malaria epidemiological data for Senegal: data come from PNLP
  https://www.dropbox.com/s/0p4uc2dihfhr9cb/Dakar.csv?dl=0.
- Climatic data: obtained from meteoblue
  https://www.meteoblue.com/historyplus.

B

Appendix

B.1 Maximum likelihood estimation method for Poisson

The likelihood function is defined as follows

$$\begin{aligned}\ell(\beta) &= \prod_{t=1}^{n} p(Y_t|X_t;\beta) \\ &= \prod_{t=1}^{n} \exp[Y_t \log(\mu_t) - \mu_t - \log(Y_t!)],\end{aligned} \tag{B.1}$$

where the right-hand side depends on β through equation (1.16) (defined in chapter 1 and Section 1.5).

The Log-likelihood is defined as follows

$$\begin{aligned}\log \ell(\beta, Y, X) &= \log \prod_{t=1}^{n} \exp[Y_t \log(\mu_t) - \mu_t - \log(Y_t!)] \\ &= \sum_{t=1}^{n} \{Y_t \log(\mu_t) - \mu_t - \log(Y_t!)\},\end{aligned} \tag{B.2}$$

where μ_t depends on X and β and is defined by equation (1.16). Observe that the term $\log(Y_t!)$ does not depend on the parameters to be estimated.

The first derivative of the log-likelihood is named the gradient and de-

noted by ∇.
If $g = \log$, we have $\log(\mu_t) = \eta_t = X_t^T\beta$. So this function is defined as follows

$$\begin{aligned}\log \ell(\beta) &= \sum_{t=1}^{n} \{Y_t \log(\mu_t) - \mu_t - \log(Y_t!)\} \\ &= \sum_{t=1}^{n} \{Y_t\eta_t - e^{\eta_t} - \log(Y_t!)\}.\end{aligned}$$

Then, the gradient at state j is defined as follows

$$\begin{aligned}\nabla L(\beta_j) = & \frac{\partial \log \ell(\beta_j)}{\partial \beta_j} \\ = & \sum_{t=1}^{n} \frac{\partial \left\{Y_t\eta_{tj} - e^{\eta_{tj}} - \log(Y_t!)\right\}}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\ = & \sum_{t=1}^{n} \{Y_t - \mu_t\} X_{tj}\end{aligned} \tag{B.3}$$

Then, if g=identity, we have $\mu_t = \eta_t = X_t^T\beta$. So

$$\log \ell(\beta) = \sum_{t=1}^{n} \{Y_t \log(\eta_t) - \eta_t - \log(Y_t!)\}.$$

Then, the gradient at state j is defined as follows

$$\begin{aligned}\nabla L(\beta_j) = & \sum_{t=1}^{n} \frac{\partial \left\{Y_t \log(\eta_{tj}) - \eta_{tj} - \log(Y_t!)\right\}}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\ = & \sum_{t=1}^{n} \{\frac{Y_t - \mu_t}{\mu_t}\} X_{tj}.\end{aligned} \tag{B.4}$$

Finally, if $g = \sqrt{}$, we have $\sqrt{\mu_t} = \eta_t = X_t^T\beta$. So

$$\log \ell(\beta) = \sum_{t=1}^{n} \left\{Y_t \log(\eta_t^2) - \eta_t^2 - \log(Y_t!)\right\}.$$

Then, the gradient at state j is defined as follows

$$\begin{aligned} \nabla L(\beta_j) = & \sum_{t=1}^{n} \frac{\partial \left\{ 2Y_t \log(\eta_{tj}) - \eta_{tj}^2 - \log(Y_t!) \right\}}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\ = & \quad 2 \sum_{t=1}^{n} \{ \frac{Y_t - \mu_t}{\sqrt{\mu_t}} \} X_{tj}. \end{aligned} \tag{B.5}$$

We summarize the gradient expressions depending on the link function:

$$\begin{cases} \nabla L(\beta_j) = \sum\limits_{t=1}^{n} \{Y_t - \mu_t\} X_{tj} \text{ if } g = \log \\ \nabla L(\beta_j) = \sum\limits_{t=1}^{n} \{ \frac{Y_t - \mu_t}{\mu_t} \} X_{tj} \text{ if } g = \text{identity} \\ \nabla L(\beta_j) = 2 \sum\limits_{t=1}^{n} \{ \frac{Y_t - \mu_t}{\sqrt{\mu_t}} \} X_{tj} \text{ if } g = \sqrt{.} \end{cases} \tag{B.6}$$

The second derivative of the log-likelihood function is named the Hessian and denoted by Hess.
If $g = \log$, we have

$$\begin{aligned} \text{Hess}\, L(\beta) &= \frac{\partial \nabla L(\beta)}{\partial \beta} \\ &= \frac{X^T \{Y - \mu\}}{\partial \beta} \\ &= -X^T \begin{bmatrix} \frac{\partial \mu_1}{\partial \eta_1} & 0 & \dots & 0 \\ 0 & \frac{\partial \mu_2}{\partial \eta_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{\partial \mu_n}{\partial \eta_n} \end{bmatrix} \frac{\partial \eta}{\partial \beta} \\ H(\beta) &= -X^T W X \end{aligned}$$

where $W = \frac{\partial \mu}{\partial \eta}$ is the diagonal matrix with $[\frac{\partial \mu_t}{\partial \eta_t}]_{tt} = [\mu_t]_{tt} = [\exp(\eta_t)]_{tt}$ and called the iterative weights.

Then, if g=identity

$$\begin{aligned}
\text{Hess}\, L(\beta_j) &= \frac{\partial \nabla L(\beta_j)}{\partial \beta_j} \\
&= \frac{\partial \left[\sum_{t=1}^{n} \{ \frac{Y_t}{\eta_{tj}} - 1 \} X_{tj} \right]}{\partial \beta_j} \\
&= \frac{\partial \left[\sum_{t=1}^{n} \{ \frac{Y_t}{\eta_{tj}} - 1 \} X_{tj} \right]}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\
&= -\sum_{t=1}^{n} \{ \frac{Y_t}{\eta_{tj}^2} X_{tj} \} X_{tj} \\
&= -\frac{1}{\beta_j^2} \sum_{t=1}^{n} Y_t.
\end{aligned}$$

Finally, if $g = \sqrt{}$

$$\begin{aligned}
\text{Hess}\, L(\beta_j) &= \frac{\partial \nabla L(\beta_j)}{\partial \beta_j} \\
&= 2\frac{\partial \left[\sum_{t=1}^{n} \{ \frac{Y_t}{\eta_{tj}} - \eta_{tj} \} X_{tj} \right]}{\partial \beta_j} \\
&= 2\frac{\partial \left[\sum_{t=1}^{n} \{ \frac{Y_t}{\eta_{tj}} - \eta_{tj} \} X_{tj} \right]}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\
&= -2 \sum_{t=1}^{n} \{ \frac{Y_t}{\mu_t} + 1 \} X_{tj}^2.
\end{aligned}$$

We summarize the Hessian expressions depending on the link function:

$$\begin{cases}
\text{Hess}\, L(\beta_j) = -X^T W X \ \text{ if } \ g = \log \\
\text{Hess}\, L(\beta_j) = -\frac{1}{\beta_j^2} \sum_{t=1}^{n} Y_t \ \text{ if } \ g = \text{identity} \\
\text{Hess}\, L(\beta_j) = -2 \sum_{t=1}^{n} \{ \frac{Y_t}{\mu_t} + 1 \} X_{tj}^2 \ \text{ if } \ g = \sqrt{}.
\end{cases} \tag{B.7}$$

B.2 Maximum likelihood estimation method for NB

As in [ADA20, HDHS21, MvdHP08], the NB-likelihood function is defined by

$$\begin{aligned}\ell(Y_t, \mu_t, \alpha) &= \prod_{t=1}^{n} p(Y_t; \mu_t, \alpha) \\ &= \prod_{t=1}^{n} [\frac{\Gamma(Y_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(Y_t + 1)} (\frac{1}{1+\alpha\mu_t})^{\frac{1}{\alpha}} (\frac{\alpha\mu_t}{1+\alpha\mu_t})^{Y_t}].\end{aligned}$$

Then,

$$\log \ell(Y_t, \mu_t, \alpha) = \sum_{t=1}^{n} [Y_t \log(\frac{\alpha\mu_t}{1+\alpha\mu_t}) - \frac{1}{\alpha}\log(1+\alpha\mu_t) + \log\Gamma(Y_t + \frac{1}{\alpha}) - \log\Gamma(Y_t + 1) - \log\Gamma(\frac{1}{\alpha})]. \tag{B.8}$$

Abusing notation, we omit the term $\log(\Gamma(y_i + 1))$ since it does not depend on the parameters to be estimated. This yields

$$\log \ell = \sum_{t=1}^{n} [Y_t \log(\frac{\alpha\mu_t}{1+\alpha\mu_t}) - \frac{1}{\alpha}\log(1+\alpha\mu_t) + \log(\Gamma(Y_t + \frac{1}{\alpha}) - \log(\Gamma(\frac{1}{\alpha}))].$$

According to [CT13, p. 81], and if Y_t is an integer, we have

$$L(Y_t) := \sum_{j=0}^{Y_t} \log(j + \frac{1}{\alpha}) = \log(\Gamma(Y_t + \frac{1}{\alpha}) - \log(\Gamma(\frac{1}{\alpha})).$$

We mention it in order to avoid using the gamma function that gives infinity when Y_t is high. Finally, the formula equation (B.8) yields

$$\log \ell(\alpha, \beta; Y, X) = \sum_{t=1}^{n} [Y_t \log(\frac{\alpha\mu_t}{1+\alpha\mu_t}) - \frac{1}{\alpha}\log(1+\alpha\mu_t) + L(Y_t)], \tag{B.9}$$

where μ_t depends on X and β according to equation (1.16).

Bibliography

[AA13] Ratnadip Adhikari and R. K. Agrawal. An introductory study on time series modeling and forecasting. *CoRR*, abs/1302.6613, 2013. URL: http://arxiv.org/abs/1302.6613, arXiv:1302.6613.

[ADA20] Sehlabana Makwelantle Asnath, Maposa Daniel, and Boateng Alexander. Modelling Malaria Incidence in the Limpopo Province, South Africa: Comparison of Classical and Bayesian Methods of Estimation. *International Journal of Environmental Research and Public Health*, 17(14), 2020. URL: https://www.mdpi.com/1660-4601/17/14/5016, doi:10.3390/ijerph17145016.

[ADB+13] Folashade B. Agusto, Sara Y. Del Valle, Kbenesh W. Blayneh, Calistus N. Ngonghala, Maria J. Goncalves, Nianpeng Li, Ruijun Zhao, and Hongfei Gong. The impact of bed-net use on malaria prevalence. *Journal of Theoretical Biology*, 320:58–65, 2013. URL: https://www.sciencedirect.com/science/article/pii/S0022519312006315, doi:https://doi.org/10.1016/j.jtbi.2012.12.007.

[ADD21] P.-A. Absil, Ousmane Diao, and Mouhamadou Diallo. Assessment of COVID-19 Hospitalization Forecasts from a Simplified SIR Model. *Letters in Biomathematics*, 8(1):215–228, 2021. URL: https://lettersinbiomath.journals.publicknowledgeproject.org/index.php/lib/article/view/403.

[ADH+06] Sonia Altizer, Andy Dobson, Parviez Hosseini, Peter Hudson, Mercedes Pascual, and Pejman Rohani. Seasonality and the Dynamics of Infectious Diseases. *Ecology letters*, 9:467–84, 05 2006. doi:10.1111/j.1461-0248.2005.00879.x.

[ALPP16] Mohammad Y. Anwar, Joseph A. Lewnard, Sunil Parikh, and Virginie E. Pitzer. Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence. *Malaria Journal*, 2016. doi:10.1186/s12936-016-1602-1.

[AMA+19] Gbenga J. Abiodun, Olusola S. Makinde, Abiodun M. Adeola, Kevin Y. Njabo, Peter J. Witbooi, Ramses Djidjou-Demasse, and Joel O. Botai. A Dynamical and Zero-Inflated Negative Binomial Regression Modelling of Malaria Incidence in Limpopo Province, South Africa. *International Journal of Environmental Research and Public Health*, 16(11), 2019. URL: https://www.mdpi.com/1660-4601/16/11/2000, doi:10.3390/ijerph16112000.

[Atk20] Andrew Atkeson. What will be the Economic Impact of COVID-19 in the US? Rough Estimates of Disease Scenarios. Working Paper 26867, National Bureau of Economic Research, 3 2020. doi:10.3386/w26867.

[AWS+21] Steven Abrams, James Wambua, Eva Santermans, Lander Willem, Elise Kuylen, Pietro Coletti, Pieter Libin, Christel Faes, Oana Petrof, Sereina A. Herzog, Philippe Beutels, and Niel Hens. Modelling the early phase of the Belgian COVID-19 epidemic using a stochastic compartmental model and studying its implied future trajectories. *Epidemics*, 35:100449, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1755436521000116, doi:https://doi.org/10.1016/j.epidem.2021.100449.

[BD20] Gyan Bhanot and Charles DeLisi. Predictions for Europe for the COVID-19 pandemic from a SIR model. *medRxiv*, 2020. doi:10.1101/2020.05.26.20114058.

[BFG+20] Jackie Baek, Vivek F. Farias, Andreea Georgescu, Retsef Levi, Tianyi Peng, Deeksha Sinha, Joshua Wilde, and An-

drew Zheng. The Limits to Learning an SIR Process: Granular Forecasting for COVID-19, 2020. arXiv:2006.06373.

[BN16] E. A. Bakare and C. R. Nwozo. On the Mathematical Analysis of the Influence of Chemoprophylaxis on the Malaria Epidemic Model. *International Journal of Contemporary Mathematical Sciences*, 11(2):45–63, 2016. URL: http://dx.doi.org/10.12988/ijcms.2016.4544.

[Bro04] R.G. Brown. *Smoothing, Forecasting and Prediction of Discrete Time Series*. Dover Phoenix Editions. Dover Publications, 2004. URL: https://books.google.be/books?id=XXFNW_QaJYgC.

[BVG+08] Olivier JT Briët, Penelope Vounatsou, Dissanayake M Gunawardena, Gawrie NL Galappaththy, and Priyanie H Amerasinghe. Models for short term malaria prediction in Sri Lanka. *Malaria Journal*, 2008. doi:10.1186/1475-2875-7-76.

[BW20] Nathaniel S. Barlow and Steven J. Weinstein. Accurate closed-form solution of the SIR epidemic model. *Physica D: Nonlinear Phenomena*, 408:132540, 2020. doi:10.1016/j.physd.2020.132540.

[CFP20] Timoteo Carletti, Duccio Fanelli, and Francesco Piazza. COVID-19: The unreasonable effectiveness of simple models. *Chaos, Solitons & Fractals: X*, 5:100034, 2020. doi:10.1016/j.csfx.2020.100034.

[CGG20] Alessandro Comunian, Romina Gaburro, and Mauro Giudici. Inversion of a SIR-based model: A critical analysis about the application to COVID-19 epidemic. *Physica D: Nonlinear Phenomena*, page 132674, 2020. doi:10.1016/j.physd.2020.132674.

[Cha16] C. Chatfield. *The Analysis of Time Series: An Introduction, Sixth Edition*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press, 2016. URL: https://books.google.be/books?id=qKzyAbdaDFAC.

[Cle19] Lieven Clemen. Poisson IRWLS, 2019. URL: https://statomics.github.io/SGA2019/assets/poissonIRWLS-implemented.html.

[CLP+21] Pietro Coletti, Pieter Libin, Oana Petrof, Lander Willem, Steven Abrams, Sereina A. Herzog, Christel Faes, Elise Kuylen, James Wambua, Philippe Beutels, and Niel Hens. A data-driven metapopulation model for the Belgian COVID-19 epidemic: assessing the impact of lockdown and exit strategies. *BMC Infectious Diseases*, 21(1), 2021. doi:10.1186/s12879-021-06092-w.

[Com22] European Commission. Des vaccins sûrs contre la COVID-19 pour les Européens, 02 2022. URL: https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans_fr.

[con22] Wikipedia contributors. Least squares, 2022. Online; accessed 2-December-2022. URL: https://en.wikipedia.org/w/index.php?

[CT13] A. Colin Cameron and Pravin K. Trivedi. *Regression Analysis of Count Data-Second Edition*. Econometric Society Monographs. Cambridge University Press, 2 edition, 2013. doi:10.1017/CBO9781139013567.

[dAD20] Gustavo M. de Athayde and Ruben M. Damião. A SIR Model with Time-Varying Parameters. 2020. URL: https://www.insper.edu.br/wp-content/uploads/2020/10/A-SIR-Model-With-Time-Varying-Parameters.pdf.

[DAD23] Ousmane Diao, P.-A. Absil, and Mouhamadou Diallo. Generalized linear models to forecast malaria incidence in three endemic regions of senegal. *International Journal of Environmental Research and Public Health*, 20(13), 2023. URL: https://www.mdpi.com/1660-4601/20/13/6303, doi:10.3390/ijerph20136303.

[Dat20] Sachin Date. Negative Binomial Regression: A Step by Step Guide, 2020. Online; accessed 28-April-2023. URL: https://towardsdatascience.com/negative-binomial-regression-f99031bb25b4.

[DDAAB20a] Ramsés Djidjou-Demasse, Gbenga J. Abiodun, Abiodun M. Adeola, and Joel O. Botai. Development and analysis of a malaria transmission mathematical model with seasonal mosquito life-history traits. *Studies in Applied Mathematics*, 144(4):389–411, 2020. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sapm.12296, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/sapm.12296, doi:https://doi.org/10.1111/sapm.12296.

[DDAAB20b] Ramsés Djidjou-Demasse, Gbenga J. Abiodun, Abiodun M. Adeola, and Joel O. Botai. Development and analysis of a malaria transmission mathematical model with seasonal mosquito life-history traits. *Studies in Applied Mathematics*, 144(4):389–411, 2020. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/sapm.12296, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/sapm.12296, doi:https://doi.org/10.1111/sapm.12296.

[dJDP+07] G. de Jong, A. Daly, M. Pieters, S. Miller, R. Plasmeijer, and F. Hofman. Uncertainty in traffic forecasts: literature review and new results for the Netherlands. *Transportation*, (34):375–395, 7 2007. URL: https://doi.org/10.1007/s11116-006-9110-8.

[dLclP16] Programme National de Lutte contre le Paludisme. BULLETIN EPIDEMIOLOGIQUE ANNUEL 2016 DU PALUDISME AU SENEGAL, 2016.

[doc] Internet document. From AR to SARIMAX: Mathematical Definitions of Time Series Models. Online; accessed 28-April-2023. URL: https://phosgene89.github.io/sarima.html.

[FCG+18] Sophie Faye, Altea Cico, Alioune Badara Gueye, Elaine Baruwa, Benjamin Johns, Médoune Ndiop, and Martin Alilio. Scaling up malaria intervention "packages" in Senegal: using cost effectiveness data for improving allocative efficiency and programmatic decision-making. *Malaria Journal*, 2018. URL: https://doi.org/10.1186/s12936-018-2305-6.

[FP20] Duccio Fanelli and Francesco Piazza. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons & Fractals*, 134:109761, 2020. doi:10.1016/j.chaos.2020.109761.

[Gab20] Vattay Gabor. Forecasting the outcome and estimating the epidemic model parameters from the fatality time series in COVID-19 outbreaks. *Phys Biol*, 2020. doi:doi:10.1088/1478-3975/abac69.

[GBAS20] Ryad Ghanam, Edward L. Boone, and Abdel-Salam G. Abdel-Salam. COVID-19: SEIRD Model for Qatar COVID-19 Outbreak. *Letters in Biomathematics*, 2020. URL: https://lettersinbiomath.journals.publicknowledgeproject.org/index.php/lib/article/view/323.

[GBMM16] Ethel Gwasira, Claver Bhunu, Mhosisi Masocha, and Emmanuel Mashonjowa. Assessing the Role of Climate Change in Malaria Transmission in Africa. *Malaria Research and Treatment*, 2016:1–7, 01 2016. doi:10.1155/2016/7104291.

[GC20] Indrajit Ghosh and Tanujit Chakraborty. An integrated deterministic-stochastic approach for predicting the long-term trajectories of COVID-19. *medRxiv*, 2020. doi:10.1101/2020.05.13.20101303.

[GGK+12] Federica Giardina, Laura Gosoniu, Lassana Konate, Mame Birame Diouf, Robert Perry, Oumar Gaye, Ousmane Faye, and Penelope Vounatsou. Estimating the Burden of Malaria in Senegal: Bayesian Zero-Inflated Binomial Geostatistical Modeling of the MIS 2008 Data. *PLOS ONE*, 7(3):1–10, 03 2012. doi:10.1371/journal.pone.0032625.

[Gou21] Yannig Goude. Les processus ARIMA. Technical report, Université Paris-Saclay, MAP-STA2: Séries chronologiques, 2021. URL: https://www.imo.universite-paris-saclay.fr/~goude/Materials/time_series/cours6_ARIMA.pdf.

[GRD19] S.R Naffees Gowsar, M Radha, and M Nirmala Devi. A Comparison of Generalized Linear Models for Insect Count Data. *International Journal of Statistics and Analysis*, 9(1):1–9,

2019. URL: http://www.ripublication.com/ijsa19/ijsav9n1_01.pdf.

[GS93] P.J Green and B.W. Silverman. *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach (1st ed.)*. CHAPMAN and HALL/CRC, United States of America, 1993. doi:https://doi.org/10.1201/b15710.

[GWT+20] Ken Junyang Goh, Jolin Wong, Jong-Chie Claudia Tien, Shin Yi Ng, Sewa Duu Wen, Ghee Chee Phua, and Carrie Kah-Lai Leong. Preparing your intensive care unit for the COVID-19 pandemic: practical considerations and strategies. *Critical Care*, 24(215), 2020. URL: https://doi.org/10.1186/s13054-020-02916-4.

[HDHS21] L. H. Hashim, N. K. Dreeb, K. H. Hashim, and Mushtak A. K. Shiker. An Application Comparison of two Negative Binomial Models on Rainfall Count Data. *IOP Publishing*, 1818(1):012100, 3 2021. doi:10.1088/1742-6596/1818/1/012100.

[Het76] Herbert W. Hethcote. Qualitative analyses of communicable disease models. *Mathematical Biosciences*, 28(3):335–356, 1976. URL: https://www.sciencedirect.com/science/article/pii/0025556476901322, doi:https://doi.org/10.1016/0025-5564(76)90132-2.

[Het89] Herbert W Hethcote. Three basic epidemiological models. In *Applied mathematical ecology*, pages 119–144. Springer, 1989.

[Het00] Herbert W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, 2000. doi:10.1137/S0036144500371907.

[HK06] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688, 2006. doi:10.1016/j.ijforecast.2006.03.001.

[HL05] Jason S. Haukoos and Roger J. Lewis. Advanced Statistics: Bootstrapping Confidence Intervals for Statistics with

"Difficult" Distributions. *Academic Emergency Medicine*, 12(4):360–365, 2005. URL: https://onlinelibrary.wiley.com/doi/abs/10.1197/j.aem.2004.11.018, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1197/j.aem.2004.11.018, doi:https://doi.org/10.1197/j.aem.2004.11.018.

[JHX20] Shuang-Lin Jing, Hai-Feng Huo, and Hong Xiang. Modeling the Effects of Meteorological Factors and Unreported Cases on Seasonal Influenza Outbreaks in Gansu Province, China. *Bulletin of Mathematical Biology*, 82(6):73, 2020. doi:10.1007/s11538-020-00747-6.

[JKKPS21] Hildeberto Jardón-Kojakhmetov, Christian Kuehn, Andrea Pugliese, and Mattia Sensi. A geometric analysis of the sir, sirs and sirws epidemiological models. *Nonlinear Analysis: Real World Applications*, 58:103220, 2021. URL: https://www.sciencedirect.com/science/article/pii/S1468121820301383, doi:https://doi.org/10.1016/j.nonrwa.2020.103220.

[KMW27] William Ogilvy Kermack, A. G. McKendrick, and Gilbert Thomas Walker. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721, 1927. doi:10.1098/rspa.1927.0118.

[Koz20] Gregory Kozyreff. Hospitalization dynamics during the first COVID-19 pandemic wave: SIR modelling compared to Belgium, France, Italy, Switzerland and New York City data, 2020. arXiv:2007.01411.

[KPT+10] Wangdi K, Singhasivanon P, Silawan T, Lawpoolsri S, White N.J, and Kaewkungwal J. Development of temporal modelling for forecasting and prediction of malaria infections using time-series and ARIMAX analyses: a case study in endemic districts of Bhutan. *Malaria Journal*, 2010. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2941685/, doi:10.1186/1475-2875-9-251.

[LAA17] Darkoh Esther Love, Larbi John Aseidu, and Lawer Eric Adjei. A Weather-Based Prediction Model of Malaria Prevalence in Amenfi West District, Ghana. *Hindawi Publishing Corporation Malaria Research and Treatment*, 2017. URL: https://doi.org/10.1155/2017/7820454.

[Lee20] Simon CK. Lee. Delta Boosting Implementation of Negative Binomial Regression in Actuarial Pricing. *Risks*, 8(1), 2020. URL: https://www.mdpi.com/2227-9091/8/1/19, doi:10.3390/risks8010019.

[LGWR20] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, 27(2), 02 2020. doi:10.1093/jtm/taaa021.

[Lin97] James K. Lindsey. *Generalized Linear Modelling*, pages 1–26. Springer New York, New York, NY, 1997. doi:10.1007/978-0-387-22730-6_1.

[Mal15] Roll Back Malaria. Mathematical modelling by control and elimination of malaria. *Collection and Impact*, 2015.

[Mas19] Ben Maslen. How to deal with count data? Technical report, Stats Central: Mark Wainwright Analytical Centre, 2019. URL: https://www.analytical.unsw.edu.au/sites/default/files/document_related_files/2019April_Seminar_How.

[MR20] A. Massih-Reza. *Machine learning - 2e édition: Programmes libres (gplv3) essentiels au développement de solutions big data*. Algorithmes. Eyrolles, 2020. URL: https://books.google.be/books?id=f6fYDwAAQBAJ.

[MSH+12] Alemayehu Midekisa, Gabriel Senay, Geoffrey M. Henebry, Paulos Semuniguse, and Michael C. Wimberly. Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. *Malaria Journal*, 165(1):11, 2012. doi:10.1186/1475-2875-11-165.

[MvdHP08] Cruyff MJ and van der Heijden PG. Point and interval estimation of the population size using a zero-truncated

negative binomial regression model. *Biom J*, 2008. doi:10.1002/bimj.200810455.

[Nak97] Eiji Nakashima. Some Methods for Estimation in a Negative Binomial Model. *Annals of the Institute of Statistical Mathematics*, 49:101–115, 1997. doi:10.1023/A:1003114706239.

[Nes20a] Yurii Nesterov. Online analysis of epidemics with variable infection rate, 2020. arXiv:2007.11429.

[Nes20b] Yurii Nesterov. Online prediction of COVID19 dynamics. Belgian case study. CORE Discussion Paper 2020/22, UCLouvain, 2020. URL: https://uclouvain.be/en/research-institutes/lidam/core/core-discussion-papers.html.

[NFT+14] El Hadji Malick Ngom, Ndèye Diango Faye, Cheikh Talla, El Hadji Ndiaye, Jacques-André Ndione, Ousmane Faye, Yamar Ba, Mawlouth Diallo, and Ibrahima Dia. Anopheles arabiensisseasonal densities and infection rates in relation to landscape classes and climatic parameters in a Sahelian area of Senegal. *BMC Infectious Diseases*, 14, 2014. doi:10.1186/s12879-014-0711-0.

[NHE+01] O. Ndiaye, J-Y. Le Hesran, J-F. Etard, A. Diallo, F. Simondon, M. N. Ward, and V. Robert. Variations climatiques et mortalité attribuée au paludisme dans la zone de Niakhar, Sénégal, de 1984 à 1996. *Cahiers Santé*, 2001.

[NNB+13] El Hadji Malick Ngom, Jacques-André Ndione, Yamar Ba, Lassana Konaté, Ousmane Faye, Mawlouth Diallo, and Ibrahima Dia. Spatio-temporal analysis of host preferences and feeding patterns of malaria vectors in the sylvo-pastoral area of Senegal: impact of landscape classes. *Parasites & Vectors*, 6, 2013. doi:10.1186/1756-3305-6-332.

[NPC21] Odu Nkiruka, Rajesh Prasad, and Onime Clement. Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, 22:100508, 2021. URL: https://www.sciencedirect.com/science/article/pii/S2352914820306596, doi:https://doi.org/10.1016/j.imu.2020.100508.

[OG17] Kamaldeen Okuneye and Abba B. Gumel. Analysis of a temperature- and rainfall-dependent model for malaria transmission dynamics. *Mathematical Biosciences*, 287:72–92, 2017. 50th Anniversary Issue. URL: https://www.sciencedirect.com/science/article/pii/S0025556416300177, doi:https://doi.org/10.1016/j.mbs.2016.03.013.

[OSJF13] Okumu Fredros O., Kiware Samson S., Moore Sarah J., and Killeen Gerry F. Mathematical evaluation of community level impact of combining bed nets and indoor residual spraying upon malaria transmission in areas where the main vectors are anopheles arabiensis mosquitoes. *Parasites & Vectors*, 2013. doi:10.1186/1756-3305-6-17.

[PJ83] McCullagh P. and Nelder J.A. *Generalized Linear Models*. Routledge, 1983. URL: https://doi.org/10.1201/9780203753736.

[PJB14] Randita Gustian Putri, Jaharuddin, and Toni Bakhtiar. SIRS-SI Model of Malaria Disease with Application of Vaccines, Anti-Malarial Drugs, and Spraying. *IOSR Journal of Mathematics*, 10:66–72, 01 2014. doi:10.9790/5728-10526672.

[RC04] Christian P. Robert and George Casella. *Deterministic numerical methods*, page 649. Number 2. Springer New York, New York, NY, 2004. doi:https://doi.org/10.1007/978-1-4757-4145-2.

[Res17] Resmawan. Seirs-sei model of malaria disease with application of vaccines and anti-malarial drugs. *IOSR Journal of Mathematics (IOSR-JM)*, 13, 2017. URL: http://www.iosrjournals.org/iosr-jm/papers/Vol13-issue4/Version-2, doi:10.9790/5728-1304028591.

[Rig20] Michel Rigo. Modèles mathématiques et confinement. *Losanges*, 49, 2020. URL: https://hdl.handle.net/2268/249086.

[RVHL20] Weston C. Roda, Marie B. Varughese, Donglin Han, and Michael Y. Li. Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious Disease Modelling*, 5:271 – 281, 2020. doi:10.1016/j.idm.2020.03.001.

[SM20] Sudhansu Sekhar Singh and Dinakrushna Mohapatra. Predictive Analysis for COVID-19 Spread in India by Adaptive Compartmental Model. *medRxiv*, 2020. doi:10.1101/2020.07.08.20148619.

[SMNM20] Farid Saberi-Movahed, Mohammad Najafzadeh, and Adel Mehrpooya. Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: Training group method of data handling using extreme learning machine conceptions. *Water Resources Management*, 34, 2020. doi:10.1007/s11269-019-02463-w.

[TA17] Agostino Tarsitano and Ilaria L. Amerise. Short-term load forecasting using a two-stage sarimax model. *Energy*, 133:108–114, 2017. URL: https://www.sciencedirect.com/science/article/pii/S0360544217308848, doi:https://doi.org/10.1016/j.energy.2017.05.126.

[VS22] Sylvie Van Schendel. *Forecasting an infectious disease: COVID-19*. PhD thesis, UCL - Ecole polytechnique de Louvain, 2022. URL: http://hdl.handle.net/2078.1/thesis:37840.

[VW10] E. Vynnycky and R. White. *An Introduction to Infectious Disease Modelling*. OUP Oxford, 2010. URL: https://books.google.be/books?id=oXoDAAAAQBAJ.

[VYKK+22] H. Vavilala, N. Yaladanda, P. Krishna Kondeti, R. Unissa, R. Mopuri, K. Chandra Gouda, K. Rao Bhimala, M. Rao Kadiri, S. Murty Upadhyayula, and S. Rao Mutheneni. Weather integrated malaria prediction system using Bayesian structural time series model for northeast states of India. *Environ Sci Pollut Res*, 29:68232–68246, 2022. doi:10.1007/s11356-022-20642-y.

[WF20] Meimei Wang and Steffen Flessa. Modelling COVID-19 under uncertainty: what can we expect? *The European Journal of Health Economics*, 2020. doi:10.1007/s10198-020-01202-y.

[Who21] Who. Coronavirus disease (covid-19): How is it transmitted?, 2021. [Online; accessed 19-July-2023]. URL: https:

//www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted.

[WHOHM03] Tuberculosis World Health Organization HIV/AIDS and Malaria Roll Back Malaria. Malaria entomology and vector control (accessed on 2023-07-07), 7 2003. URL: https://apps.who.int/iris/bitstream/handle/10665/67450/WHO_CDS_CPE_SMT_2002.18_Rev.1_PartI.pdf;sequence=1.

[Wik23] Wikipedia contributors. Normal distribution — Wikipedia, the free encyclopedia, 2023. [Online; accessed 18-July-2023]. URL: https://en.wikipedia.org/w/index.php?title=Normal_distribution&oldid=1165579699.

[WKJB20] John Willan, Andrew John King, Katie Jeffery, and Nicola Bienz. Challenges for NHS hospitals during COVID-19 epidemic. *BMJ*, 368, 2020. URL: https://www.bmj.com/content/368/bmj.m1117, arXiv:https://www.bmj.com/content/368/bmj.m1117.full.pdf, doi:10.1136/bmj.m1117.

[YMMA20] Ashenafi A. Yirga, Sileshi F. Melesse, Henry G. Mwambi, and Dawit G. Ayele. Negative binomial mixed models for analyzing longitudinal CD4 count data. *Scientific Reports*, 10, 2020. doi:10.1038/s41598-020-73883-7.

[ZREG21] Jon Zelner, Julien Riou, Ruth Etzioni, and Andrew Gelman. Explained: The Various Sources of Uncertainty in COVID-19 Studies, 10 2021. URL: https://science.thewire.in/the-sciences/how-do-you-account-for-uncertainty-during-a-pandemic/.