POLITECNICO DI TORINO

Dipartimento di Scienze Matematiche
"Giuseppe Luigi Lagrange"

Corso di Laurea Magistrale in Ingegneria Matematica

Tesi di Laurea Magistrale

Population dynamic and statistical approach for the analysis of growth of *Varroa destructor* in *Apis mellifera* colonies

Relatore
Gianluca Mastrantonio

Correlatore
Ezio Venturino

Candidato
Emiliano Traini

July 10, 2018

Abstract

The ectoparasitic mite Varroa destructor has become one of the major threats for apiculture worldwide. Varroa destructor attacks the honey bee Apis mellifera weakening its host by sucking hemolymph. However, the damage to bee colonies is not strictly related to parasitic action of the mite but it derives, above all, from the increased trasmission of many viral diseases vectored by it. In this thesis an analysis of this phenomenon is carried out by approaching it in two ways: the first is a study from the point of view of population dynamics, while the second is a statistical analysis.

The data used were collected during an experiment aimed at Ciriè (TO) from the Cooperative DSP. For the reproduction period of the bees, eleven beehives were observed, in which initially adult varroa were introduced, taking samples about every 15 days. These hives were treated with oxalic acid, the most common treatment used nowaday to combat varroa, that is not harmless for the life of the bees. In the first part of this Thesis a dynamics model is developed that describes the growth of four populations that coexist within a beehive: adult bees, larvae (closed bees in the cells undergoing growth), varroa in phoretic phase (parasite for adult bees) and the varroa in the reproduction phase in the cells occupied by the larvae. The final model was conceived starting from growth, death, SIS and Lotka-Volterra models and it depends on ten parameters. Seven of these parameters were taken from the literature while the remaining three (varroa growth rate and varroa transition rates from reproductive to phoretic phase and from phoretic to reproductive phase) were estimated using data at the disposition.

In the second part a statistical analysis of the phenomenon is performed. In addition to data sets about bees, climatic data collected by surveyors in the locality of Caselle Torinese (TO) were added to study the effect of climate variables on the growth of the parasite. Varroa are studied in both the reproductive and the foretic phases. Initially, linear models with interactions, structured variances and random effects were used, and zero-inflated mellows to manage the presence of a large number of zeros in the response variable, most probably due to experimental limits.

In the final part of this Thesis, we compare the results obtained from these two approaches, i.e. deterministic and stochastic, giving an interpretation of the results that can be used to improve the beekeeping activity.

Ringraziamenti

Questo spazio dovrebbe essere dedicato a tutte le persone che mi sono state vicino nel percorso universitario, ma...

...Gianluca e professori Gasparini, Venturino, Preziosi, Rolando, Rondoni e Vaccarino, anche se spero nel futuro di poter lavorare con persone così genuine e disponibili, questa pagina non è per voi...

...Marco, Giorgio, Giacomo, Stefano, Edoardo e praticamento anche tu Deggio, anche se la nostra convivenza credo sia stata la miglior droga che io abbia mai assunto, questa pagina non è per voi...

...tutti voi amici miei, anche se siete costantemente il mio punto di partenza dopo ogni sconfitta, questa pagina non è per voi...

...Sara, Monica, anche se il nostro concetto di famiglia è un modello che i nostri modelli avrebbero tanto dovuto seguire, questa pagina non è per voi...

...Francesca, anche se ancora non mi capacito di quanto ci siamo potuti avvicinare in questi ultimi anni, questa pagina non è per te...

...Annie, anche se non sarei in grado di mettere in parole cosa vogliano dire questi ultimi anni passati insieme a te, questa pagina non è per te...

...Lucia, Silvano ...mamma, papà, questa pagina è solo vostra.

Mi è capitato spesso di sentirmi in difetto verso il mondo cercando di ricevere da esso un grazie sincero. Bene, per quanto possa valere poco questo gesto, da oggi non dimenticate che almeno una persona in questo mondo ha tentato di esprimervi il più riconoscente grazie che un gesto possa esprimere.

Emiliano Traini

Contents

Chapter 1

Introduction

The life on the planet will end if the bees disappeare. Plants need to receive the pollen from similar plants to reproduce, a function that is carried out mainly by the bees, the principal pollinator. This relationship is as old as their existence, as for the egg and for the hen we can't ask who was born before: the plants, in their extensive biodiversity or the bees pollinators. The humid secretion that the plants emit is used to attract bees and to be fertilized with the pollen that the bees are carry from on the flowers of the same variety. The bees are responsible for about 70% of the pollination of all living plant species on the planet, guaranteeing about 35% of global food production, so we must protect this small insect, threatened today along with many other pollinating insects. The beekeepers saw, year by year, their breeding and the production of honey beeing reduced.

"If the bee disappeared off the surface of the globe then man would only have four years of life left. No more bees, no more pollination, no more plants, no more animals, no more man.

If the bee disappears from the surface of the earth, man would have no more than four years to live"

Albert Einstein

Recent studies have reported an epidemic presence of some viruses in many species of bees and this increase is due to the presence of a new transmission vector: the *varroa destructor*, also called varroa ([1], [2]). The varroa is a parasite that has as its natural host the oriental *apis cerana* which is not hardly damaged by it. In the 40s, however, the European cousin *apis mellifera* was introduced in Southeast Asia to increase honey production and something unexpected happened: in 1958 the first cases of varroa aggression to this species of bees were reported in China. Subsequently due to the uncontrolled trade of the

mellifera queens, the varroa spread practically all over the world. In Italy the fight began in 1981. The difference between the oriental and European species is that the second one leads to mite populations much larger. Therefore, this fact accelerates the spread of some viruses that often causes the death of the colony.

Given the vastness of the problem, the scientific community that has been conducting research on this parasite for many years is not limited only to biologists. It includes also mathematicians who try to model the evolution of this parasitic population to optimize human intervention to rescue the bees.

This Thesis is based on experimental activity followed by beekeepers and biologists in Ciriè, Torino. This experiment consisted in monitoring for the bee activity season the presence of the varroa in different artificial hives. The interest of beekeepers is to find a way to have the better result of treatment, which for now is done based on oxalic acid. Obviously this remains exclusively an ideal goal.

Using the data collected in the experience, two different approaches were adopted: one deterministic and one statistical. The goal of the deterministic approach is to develope a mathematical model able to describe the dynamic of the evolution of the varroa. This is possible due to the fact that the model is able to estimate different parameters from the data experimentally collected. Instead the statistical approach is finalized to the development of a predictive model that describes the varroa growth over time and furthermore analyses the climatic influence on this phenomenon.

For this Thesis R programming language [23] was used for both the deterministic and the statistical part. R is a programming language and free software environment for statistical computing and graphics that is widely used among statisticians and data miners for developing statistical software and data analysis. The capabilities of R are extended through user-created packages, which allow specialized statistical techniques, graphical devices, import/export capabilities, reporting tools, etc. These packages are developed primarily in R, and sometimes in Java, C, C++, and Fortran. All R packages used for this project are shown in the dedicated appendix part, including those used for solving differential equation systems.

To better understand the subject of this study the following chapter, a biological introduction, is splitted in two parts. The first one is about the so called host, the bee specie; the second part presents the vector of pathogen agents, the acarus.

Chapter 2

Biological background

2.1 Host: the honey bee *Apis mellifera*

Called *Apis mellifica* from *Carl Nilsson Linnaeus* in 1758, the *Apis mellifera* is the most widepread species of bee in the world. Originally widespread in Europe, Africa and part of Asia, this species was introduced on every continent for business. Usually in a beehive live a queen, the only fertile female, from 40 to 100 thousand workers, sterile females destined to the maintenance and to the defense of the colony, and between April and July (in Europe) from 500 to 2000 males (also called drones or pigeons), which are destined exclusively for reproduction. The species is polymorphic because the three castes have different morphological conformations.

Apis mellifera.

Male pupae.

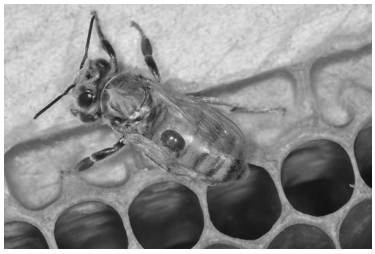
The queen can live up to 4 years and it is fertilized once in her whole life by about 8 drones. In the period in which the harvest of nectar is abundant, a queen arrives to deposit up to 3 thousand eggs a day, which if are fertilized they will generate worker bees or more rarely queens, otherwise they will become drones. Once the egg is attached to the bottom of a cell, it opens up later about 3 days from the deposition and emerges a tiny vermiform larva, apoda and anophthalma (without compound eyes). The pupa suffers one complete metamorphosis, and finally cut the operculum of the cell with its own jaws to flicker like a young bee. Development time for each caste it is standardized, thanks to the thermoregulation in the hive. In the Table 2.1 are shown the times estimated for each phase of life of each caste of this species.

	Prima del taglio dell'opercolo				Sfarfallamento
	Uovo	Larva	Pupa	Totale	Adulto
Regina	3	5.5	7.5	16	3.5 anni
Operaia	3	6	12	21	30-45
Fuco	3	6.5	14.5	24	-

Table 2.1: Time estimatesare are expressed in day if not specified.

2.2 Vector: the mite *Varroa destructor*

The varroa mite can reproduce only in the apis mellifera colonies. It has two life stages, phoretic and reproductive.

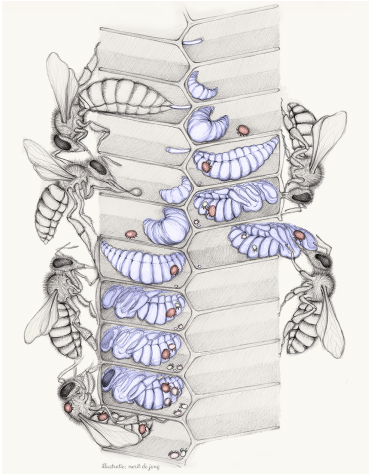


Bee with Varroa Mite.

Credit: Bayer Bee Health.

The phoretic stage is when a mature varroa mite is attached to an adult bee and survives taking hemolymph from it. During this stage the mite may change hosts often transmitting viruses by picking up the virus on one infected individual and injecting it to another during feeding. Phoretic mites may fall off the host for its grooming activity or be bitten by another bee. This mites and those dead due to natural causes rest in a bottom plate under the hive and they are called the *natural mite drop.*

The reproductive life stage of Varroa begins when an adult female mite is ready to lay eggs and moves from an adult bee into the cell of a developing larval bee. After the brood cell is capped and the larva begins pupating, the mite begins to feed. After about three days from capping, the mite lays its eggs, one unfertilized egg (male) and more or less 4 fertilized (female) eggs. After the eggs hatch, the female mites feed on the pupa, mate with the male mite and the surviving mature female mites stay attached to the host bee when it emerges as an adult. The varroe perform up to 7 reproductive cycles so they die because old.



Reproduction activity of varroa in the hive.

The hemolymph suction causes lacerations in which pathogens can penetrate causing clinical effects to the bee as well as its anomalous development so that sometimes the bee is already deformed. The symptoms occur more in the period in which the drones are no longer raised because they attract the varroa more tha the female bees. The symptoms are:

- reduction of the number of bees, bees with flight difficulties, substitution

of the queen, abandonment of the hive;

- irregular brood, larvae out of place in the cell and liquefied, brown-colored larvae.

The infestation spreads from one hive to another with drifts, looting, trade in swarms and queens, gathering swarms, etc. Field operations can also contribute to spread the disease, even if the beekeeping equipment is not a source of contagion because the varroe survive shortly in the absence of the bees. The disease is widespread in 100% of the hives, so the diagnosis of the disease has no sense, but it is needed estimating the degree of the infestation during the year.

2.3 Viral pathogens: DWV e ABPV

Most of honey bee viruses commonly causes covert infections, namely the virus can be detected at low titers within the honey bee population in the absence of obvious symptoms in infected individuals or colonies. However, when injected into the open circulatory system, these diseases are extremely virulent with only few viral particles per bee required to cause death within a few days. The most serious problem caused by Varroa destructor. On one hand, when a mite carrying virus attach to a healthy bee, it can transmit the virus to the bee. Further, viral diseases are also transmitted among bees through food, feces, from queen to egg, and from drone to queen. On the other hand, a virus free phoretic mite can begin carrying a virus when it moves from an uninfected to an infected bee but it can also acquire it horizontally from other infected mites. Therefore, the management of varroa activities in a hive will control the associated viruses because often presence of viral symptoms diseases is indicative of an invasion of varroa in the colony.

Since 1963, year of isolation of the first virus (CPV) to date, they have been identified and characterized no less of 21 viruses. Among all these viruses many are associated with pathogens, specifically DWV and ABPV are the two most associated with varroa [5].

2.3.1 Deformed wing virus (DWV)

The virus was first isolated from a sample of symptomatic honeybees from Japan in the early 1980s and is currently distributed worldwide. It is found also in pollen baskets and commercially reared bumblebees. Deformed wing virus is suspected of causing the wing and abdominal deformities often found on adult honeybees in colonies infested with Varroa mites [6]. These symptoms include damaged appendages, particularly stubby, useless wings, shortened, rounded abdomens, miscoloring and paralysis of the legs and wings. Symptomatic bees have severely reduced life-span (less than 48 hours usually) and are typically expelled from the hive. In the absence of mites the virus is thought to persist in the bee populations as a covert infection, transmitted orally between adults (nurse bees) since the virus can be detected in hypo-pharyngeal secretions (royal jelly) and brood-food and also vertically through the queen's ovaries and through drone sperm. The virus may replicate in the mite but this is not certain.

2.3.2 Acute paralysis virus (ABPV) [3]

For many years on, ABPV was shown to exist in low concentrations as a covert infection in adult bees, never producing outbreaks of paralysis. Shortly after the establishment of varroa in Europe, the virus was then isolated from healthy adult from most regions all over the world: France, Italy, Canada, China, the USA, New-Zealand. ABPV is known today to have a geographical distribution similar to that of *A. mellifera*. Bees affected by this virus tremble uncontrollably. The virus has been suggested to be a primary cause of bee mortality. Infected pupae and adults suffer rapid death.

2.4 Treatment: oxalic acid [7]

Oxalic acid is a colorless crystalline solid that forms a colorless solution in water. It occurs naturally in many foods, but excessive ingestion of oxalic acid or prolonged skin contact can be dangerous. Oxalic acid is used against varroa, because it is the most convenient: it leaves no residue in honey, it is well tolerated by bees in any way it is propelled in the hive, it is considered a "natural" active ingredient, currently it is not yet contemplated from the EC Regulation 2377/90 on MRLs (maximum residual limits). The oxalic acid solution should be prepared before use by stirring the distilled water, dissolving the oxalate inside and then adding the sugar. The suspension acts by contact and currently it is administered to the bees by dripping and spraying. Considering that during the bee season the number of varroe has been estimated for about 2/3 in the brood and for a 1/3 on the adult bees, the application of the above techniques is limited to the autumn and winter period or in absence of brood. Waiting a new brood cycle, a second antivarroa cleaning operation is done.

The autumn treatment is defined as "radical cleaning" and reaches an efficacy even higher than 95%, provided that in the colonies treated at least one month has elapsed since the last solid feeding, since the absorption of candied fruit, which occurs slowly from part of the bees, causes the queen to stimulate the deposition.

It is possible to treat the colonies even when they are fed with a very concentrated syrup, in large quantities and for a short period, because the nutrition concentrated for a short cycle does not stimulate the queen to lay down but has the function of integrating only the stocks.

The fall of the varroe, after treatment, occurs approximately after 24-48 hours. During the treatments always remain on the bees a variable percentage of varroe, the important thing is to know how much. To ascertain the percentage of varroe in the phoretic phase, a control treatment must be carried out with a tested formula: the percentage of fallen varroe is $(AC/(AC+AT)) \times 100$, where AC is the mites fallen into followed by treatment with oxalic acid and AT is the mites fallen after treatment of a tested product.

The monitoring of the fall rate, in the absence of brood, allows to ascertain in time the increase in the number of mites and to adopt the necessary fighting subtleties. When it is not possible with two treatments to arrive at a high percentage of fall, greater than 95%, one must change the product because one is in the presence of the addiction of the varroa.

Chapter 3

Collected data

For this project two different datasets are used and in this chapter they are showed separatly. The first one are measuremets about hives, bees and varroe, instead the second one are data about weather conditions.

3.1 Varroa data

The first part of data used in this project come from an experiment started on May 30, 2016 in Cirié (TO) and for this in twelve hives, emptied and sterilized, bee colonies were introduced for the creation of a beehive. Furthermore, after treating all the hives with oxalic acid so as to consider the presence of the varroa equal to zero, in each beehive a different quantity of adult varroa specimens was introduced (inoculum), to create a characterization of each hive (table 3.1).

Hive code	01	02	03	04	05	06	07	08	09	10	11	12
Inoculum	9	11	15	10	18	9	22	37	12	15	18	10

Table 3.1: Initial number of varroa in phoretic phase per hive.

During the experiment, about every fifteen days the biologists took two type of measurements. The first one consisted of a small part of the hive, extracting from it 100 cells and, after friezed them, to count how many varroa are in each one. The second measurement, instead, was made by taking a certain number of bees, freezing them and subsequently detecting the presence or absence of varroa.

Figure 3.1 shows all measurements dates but the data present some problems:

- in hives 5 and 8 there are two outsider dates that will not be considered;
- in hives 8, 9, 11 and 12 some dates are missing but these hives will be considered equally;
- in hives 10 three dates are missing and, at the start of the experiment, the queen died, so this hive will not be considered.

So after these adjustments seven measurements dates and eleven hives will be available for the analysis.

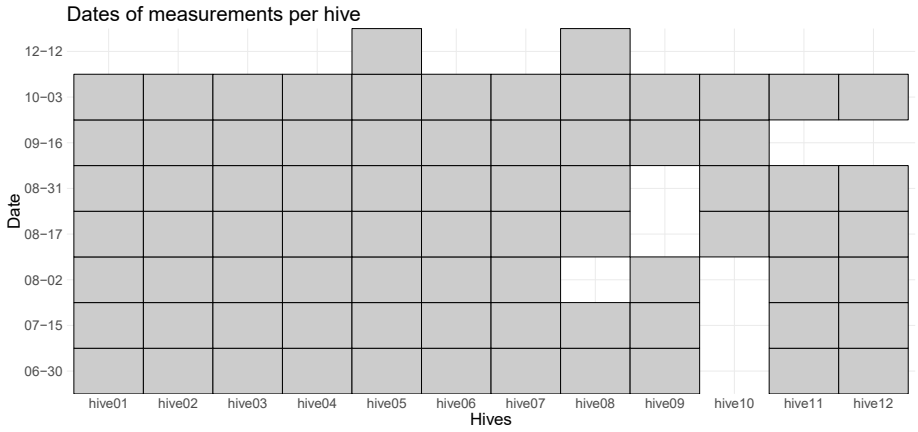


Figure 3.1: The presence of a square per hive and per date indicates that the measurement was taken.

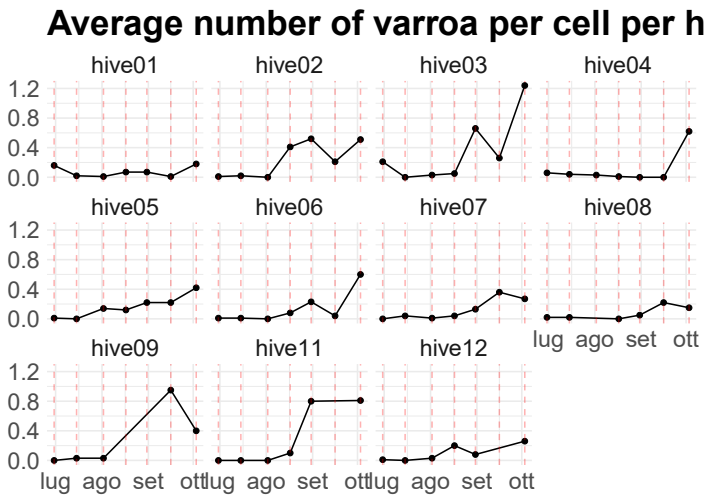


Figure 3.2: For each hive this plot shows the trend of the average number of varroa found in the 100 cells taken on the measurement date. The plotted values are obtained by adding together all the varroa found in the sample and dividing by 100.

Hive code	01	02	03	04	05	06	07	08	09	11	12
Dead varroa	682	1204	974	127	1899	1679	496	1872	112	2276	583
Dead hive	N	N	N	N	Y	N	Y	Y	N	N	N

Table 3.2: This Table shows the initial number of varroa per hive and if an hive is dead or not.

Another type of measurement was made: the treatment of oxalic acid has been done in autumn and a debris collector was placed under the hive to collect the varroa once dead. After having carried out the treatment against the varroa,

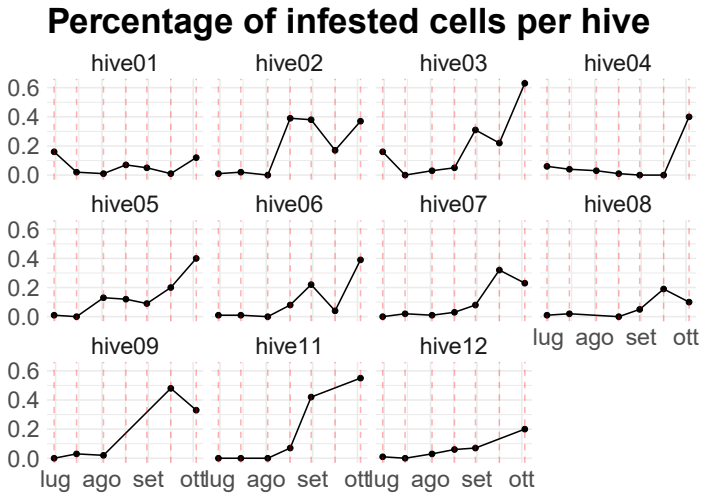


Figure 3.3: For each hive this plot shows the trend of the percentage of cells infested on the 100 cells taken.

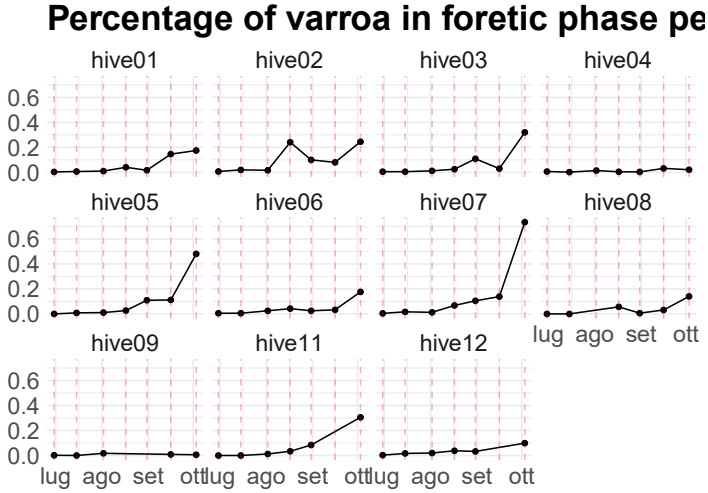


Figure 3.4: For each hive this plot shows the trend of the percentage of varroa in the phoretic phase on the number of bees in circulation in the hive.

at the end of the experiment the mites found in this collector were counted. Table 3.2 shows the data about the final dead varroa and if the hive is dead or not.

Figures 3.2, 3.3, 3.4 e 3.5 shows the other principal outputs from the exper-

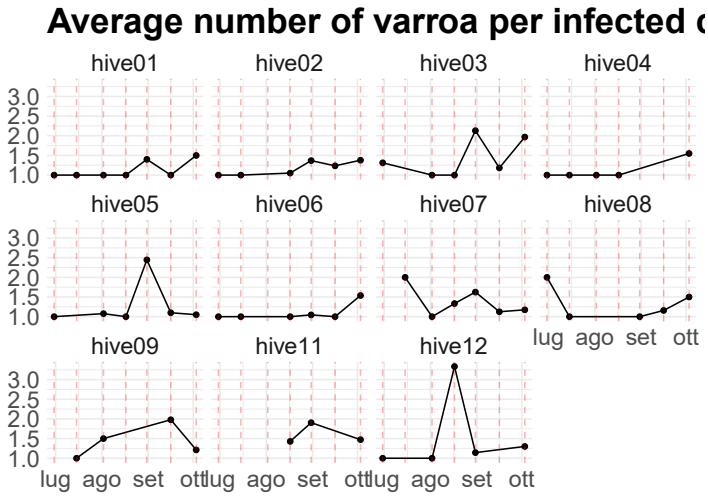


Figure 3.5: For each hive this plot shows the trend of the average number of varroa but it is calculated only on the found infested cells and not on the total of 100.

iments. An infested cell means that at least one varroa has been found in it. The increasing trend of infested cells (Figure 3.3) in almost all hives is a very important fact that shows how over time the mite increases its impact on the brood. From Figure 3.3 we see that the percentage of varroa in phoretic stage was obtained by taking a variable number of bees at each measurement date, which were frozen and then for each sample the total number of mites attached to the body of these bees was counted. In Figure 3.5 no growing trend is observed and this fact shows how the reproduction of a single varroa inside the operculated cell is a phenomenon independent of the dynamics of populations outside the cell.

3.2 Climatic data

To analyse the climatic effect on the varroa growth, we need data about the weather during the experiment. In this section we speak about data concerning this.

The second part of data used in this project comes from a weather website [11]. The data were collected by the Torino Caselle weather station, a locality close to Ciriè (about 7 km); we assume they are a good approximation of the weather of Ciriè. They consist in a dataset with the following attributes:

- $DATE$ **(YYYY-MM-DD)** is the day to which the measurements refer;
- $AVRG\_TEMP$ **(°C)** is the average temperature during the day;
- MIN_TEMP **(°C)** is the minimum temperature during the day;
- MAX_TEMP **(°C)** is the maximum temperature during the day;

- $INT\_TEMP = MAX\_TEMP$ - MIN_TEMP (°**C**);
- DEW_POINT (°**C**) is the thermodynamic state in which a two-phase liquid-vapor mixture becomes saturated with vapor, or rather above the dew point there is only the presence of steam;
- $HUMIDITY$ (**%**) is the average percentage of humidity during the day;
- $VISIBILITY$ (**km**) is the average visibility during the day;
- $AVRG\_WIND$ (**km/h**) is the wind average speed during the day;
- MAX_WIND (**km/h**) is the wind maximum speed during the day;
- $GUST$ (**km/h**) shows the presence of a gust of wind during the day, that means an instant of time in which the wind speed reaches a peak respect the rest of the day;
- $PRESS\_OSL$ (**mb**) is the average daily pressure referred to a zero altitude, i.e. the sea level;
- $AVRG\_PRESS$ (**mb**) is the average daily pressure referred to the altitude of the place.

3.3 Final data frame

A general single data frame was obtained from the previous two, after making the following changes:

- the variables **AVRG_PRESS** and **GUST** have been deleted because they are almost always null;
- only the 7 dates referring to the measurements concerning the hives were considered;
- the varible **MIN_TEMP** referred to a certain date now represents the minimum value that this variable assumes in the previous 15 days;
- the varibles **AVRG_TEMP**, **DEW_POINT**, **HUMIDITY**, **VISIBILITY**, **AVRG_WIND** and **PRESS_OSL** referred to a certain date now represents the mean value that this variables assume in the previous 15 days;
- the varibles **MAX_TEMP** and **MAX_WIND** referred to a certain date now represents the maximum value that this variables assume in the previous 15 days;
- the variable **t** is an alternative way to show the time: in the start data of experiment ("$2016 - 5 - 30$") the value of it is $t = 0$ and $t = n$ indicates the n-th day since the beginning of experiment.

Chapter 4

Deterministic model

4.1 Stable and symptotically stable equilibrium points

A good place to start analyzing the nonlinear system

$$\dot{x} = f(x), \quad \text{with} \quad x \in \Omega,$$

is to determine the equilibrium points of 4.1. Equilibrium points represent the simplest solutions to differential equations.

Definition 4.1.1 *Suppose an* ***autonomous*** *system of ordinary differential equations, that is a system of the form 4.1 in which the right side does not contain the independent variable t and $\Omega \subseteq \mathbb{R}$ is the domain of x. An* ***equilibrium point*** *of the differential equation system 4.3 is a point x_∞ such that $f(x_\infty) = 0$.*

In addition, an equilibrium point x_∞ is *feasible* if and only if $x_\infty \in \Omega$. In this way x_∞ is a solution for all t.

It is often important to know whether an equilibrium point is stable, i.e. whether it persists essentially unchanged on the infinite interval $[0, \infty]$ under small changes in the initial data. This is particularly important in applications, where the initial data are often known imperfectly.

Definition 4.1.2 *An equilibrium x_∞ is said to be* ***stable*** *if for every $\epsilon > 0$ there exists $\delta > 0$ such that*

$$|x(0) - x_\infty| < \delta \quad \textit{implies} \quad |x(t) - x_\infty| < \epsilon, \quad \forall t > 0.$$

It is implicit in definition 4.1 that the existence of the solution $x(t)$ is required for $0 \leq t \leq \infty$. The definition is restricted to *Lyapunov* stability, wherein only perturbations of the initial data are contemplated, and thereby exclude consideration of *structural* stability, in which one considers perturbations of the vector field.

Definition 4.1.3 *An equilibrium x_∞ is said to be* ***asymptotically stable*** *if it is stable and if in addition*

$$|x(0) - x_\infty| < \delta \quad implies \quad \lim_{t\to\infty} x(t) = x_\infty.$$

Thus, stability means roughly that a small change in initial value produces only a small effect on the solution, and this condition is a natural requirement for an equilibrium to be biologically meaningful. In biological equations the asymptotic stability rather than stability is usually required, both because asymptotic stability can be determined from the linearization technique while stability cannot, and because an asymptotically stable equilibrium is not disturbed greatly by a perturbation of the differential equation.

If a system of n ODE

$$\dot{x} = f(x) \quad x \in \Omega \subseteq \mathbb{R}^n, \quad f \in \mathbb{R}^n$$

is *linearized* about the equilibrium point $x_\infty \in \Omega$, with perturbation variable $z = x - x_\infty$, then the linear system of differential equation is

$$\dot{z} = Jz,$$

where J is the Jacobian matrix of the system 4.1 evaluated at the equilibrium x_∞. That is

$$J = (J_{ij}) = \left(\frac{\partial f_i}{\partial x_j}(x_\infty\right). \tag{4.1}$$

Trough the linearization technique, the stability of an equilibrium x_∞ can be determined from the eigenvalues of the Jacobian matrix evaluated at the equilibrium. It follows from the next results.

Theorem 4.1.4 *For the linear first order constant coefficient system of ODE*

$$\dot{z} = Az, \quad z \in \mathbb{R}^n, \quad A \in R^{n\times n},$$

the zero vector $z \equiv 0$ is ***stable*** *or* ***unstable*** *as follows:*

- *if all eigenvalues of A have not positive real parts and all those with zero real parts are simple, the $z \equiv 0$ is stable;*
- *if and only if all eigenvalues of A have negative real parts, then $z \equiv 0$ is asymptotically stable;*
- *if one or more eigenvalues of A have a positive real parts, then $z \equiv 0$ is unstable.*

For the general autonomous ODE system 4.1, the analysis of the stability of an equilibrium point x_∞ reduces to the study of stability of the corresponding linearized system in the neighborhood of the equilibrium point, as stated in the theorem 4.1.5.

Theorem 4.1.5 (Lyapunov theorem) *An equilibrium point x_∞ of the ODE system 4.1 is* ***stable*** *if all the eigenvalues of J (Jacobian matrix evaluated in x_∞) have negative real parts. The equilibrium point is* ***unstable*** *if at least one of the eigenvalues has a positive real part.*

In order to determine the eigenvalues of (4.1), it is necessary to find the roots of the characteristic equation (4.2).

$$det(J - \lambda I) = 0. \tag{4.2}$$

However, the characteristic equation for an n-dimensional system is a polynomial equation of degree n for witch it may be difficult or impossible to find all root explicitly. In this regard, the theorem 4.1.6 is a general criterion for determining whether all roots of a polynomial equation have negative real part.

Theorem 4.1.6 (Routh-Hurwitz criterion) *Given the polynomial*

$$P(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0,$$

where the coefficients a_i are real constants $\forall i = 1, \ldots, n$. The ***n Hurwitz matrices*** *are defined using the coefficients a_i of the characteristic polynomial:*

$$H_1 = (a_1), \quad H_2 = \begin{pmatrix} a_1 & a_3 \\ 1 & a_2 \end{pmatrix}, \quad H_3 = \begin{pmatrix} a_1 & a_3 & a_5 \\ 1 & a_2 & a_4 \\ 0 & a_1 & a_3 \end{pmatrix},$$

$$H_k = \begin{pmatrix} a_1 & a_3 & a_5 & \cdot & \cdot & \cdot \\ 1 & a_2 & a_4 & \cdot & \cdot & \cdot \\ 0 & a_1 & a_3 & \cdot & \cdot & \cdot \\ 0 & 1 & a_2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & a_k \end{pmatrix}, \quad k = 1, \cdots, n.$$

All of the roots of the polynomial $P(\lambda)$ are negative or have a negative real part if and only if the determinants of all Hurwitz matrices are positive:

$$det(H_j) > 0, \quad \forall j = 1, \cdots, n.$$

A remark is that for $n = 2$, the criterion 4.1.6 simplify to

$$det(H_1) = a_1 > 0, \quad det(H_2) = a_1 a_2 > 0,$$

that is $a_1 > 0$ and $a_2 > 0$.

For the stability analysis, we can equivalently require that the trace of the matrix J be negative and the determinant of the same matrix be positive. In fact, in this case the characteristic polynomial can be written as

$$P(\lambda) = \lambda^2 - tr(J)\lambda + \det(J).$$

Similarly, when $n = 3$ we get the following Routh-Hurwitz conditions:

$$a_3 > 0, \quad a_1 > 0, \quad a_1 a_2 > a_3.$$

4.2 The model

Varroa destructor attacks the honey bee Apis mellifera sucking hemolymph from both the adult bees and the brood. However, the honey bee mortality induced by the subtraction of hemolymph and tearing of tissues in the act of sucking

is very insignificant. Therefore, the damage to bee colonies derives from the parasitic action of the mite but, above all, from its action of vector for many viral diseases even seriously harmful.

In this work discusses an SI model that describes how the presence of the mite varroa affects the epidemiology of these viruses on adult bees and larvae. Let B denotes the number of bees, L the number of larvae (bees in growth phase), R the mites in reproductive stage and P the mites in the phoretic stage. The model reads as follows:

$$\begin{cases} \dot{L} = b\frac{B^2}{k^2+B^2} - cL \\ \dot{B} = cL - mB - \mu P \\ \dot{R} = raP - gcRL \\ \dot{P} = gcRL - nP - ePB - aP. \end{cases}$$

In the model each term has a precise meaning for the description of the dynamics of the three populations. Precisely:

- $bB^2(k^2 + B^2)^{-1}$ is the growth term for bees. It is such that for a enough large population compared to the parameter k, a linear growth with coefficient b is obtained. It does not depend from L because the queen deposes eggs and the worker bees bring them to the cells still to be operculated;
- the term cL models the birth of bees, i.e. the larvae that leave the cells $(-cL)$ because now they are adult bees $(+cL)$;
- the term mB represents the natural death of the bees;
- the term μP models the death of the bees due to the vector action of the parasite;
- daily, a portion of mites in the phoretic phase, aP, is introduced into the unoperculated cells to begin the reproduction activity;
- this portions of varroa reproduces with a rate r;
- therefore the varroa leave the cell with rate g, according to a term proportional to both how many varroa are in the reproductive phase and to the birth term of the bees: so the term $gcRL$ is subtracted from the dynamics of the varroa in reproduction and added to the phoretic one;
- the varroa in phoretic stage die naturally with rate n;
- finally the term ePB is related to the grooming behavior that occurs at rate e in bees.

Note that all parameters of the model are positive.

4.3 Equilibrium points of the model

The beginning is the research of the constant solutions of the model 4.2. From definition 4.1.1 we deduce that to find the equilibrium points it is necessary to solve the system 4.3. In order to simplify the analysis of the solutions, we

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
L	0	$*$	0	0	0	$*$	$*$	$*$	0	0	0	$*$	$*$	$*$	0	$*$
B	0	0	$*$	0	0	$*$	0	0	$*$	$*$	0	$*$	$*$	0	$*$	$*$
R	0	0	0	$*$	0	0	$*$	0	$*$	0	$*$	$*$	0	$*$	$*$	$*$
P	0	0	0	0	$*$	0	0	$*$	0	$*$	$*$	0	$*$	$*$	$*$	$*$

Table 4.1: The symbol $*$ denotes a population not necessairly equal to zero.

discuss, one by one, all the possible configurations of the three populations, shown in Table 4.1.

$$\begin{cases} b\frac{B^2}{k^2+B^2} - cL = 0 \\ cL - mB - \mu P = 0 \\ raP - gcRL = 0 \\ gcRL - nP - ePB - aP = 0. \end{cases}$$

1. (L, B, R, P) = **(0, 0, 0, 0)**
   The solution
   $$\boxed{E_1 = (0, 0, 0)}$$
   is an equilibrium point and it is feasible.

2. (L, B, R, P) = **(L, 0, 0, 0)**
   From the first equation of 4.3 we found $L = 0$ and so for $L > 0$ there are not any equilibrium points.

3. (L, B, R, P) = **(0, B, 0, 0)**
   From the first equation of 4.3 we found $B = 0$ and so for $B > 0$ there are not any equilibrium points.

4. (L, B, R, P) = **(0, 0, R, 0)**
   From the third equation of 4.3 we found $R = 0$ and so for $R > 0$ there are not any equilibrium points.

5. (L, B, R, P) = **(0, 0, 0, P)**
   From the first equation of 4.3 we found $P = 0$ and so for $P > 0$ there are not any equilibrium points.

6. (L, B, R, P) = **(L, B, 0, 0)**
   The first equation of 4.3 becomes
   $$cL = b\frac{B^2}{k^2 + B^2},$$
   so the second one becomes
   $$b\frac{B^2}{k^2 + B^2} = mB,$$
   that is
   $$mB^3 - bB^2 + mk^2B = 0.$$

Since $B = 0$ is a case already considered, we consider $B > 0$ and divide both members by B, obtaining

$$mB^2 - bB + mk^2 = 0.$$

So we have three cases corresponding to cases in which $\Delta = b^2 - 4m^2k^2$ is positive, null or negative:

- if $b > 2mk$ then there are two equilibrium points (not considering $(0, 0, 0, 0)$):

$$\left(\frac{b \pm \sqrt{b^2 - 4m^2k^2}}{2m}, \frac{b \pm \sqrt{b^2 - 4m^2k^2}}{2c}, 0, 0 \right)$$

  and both cases are feasible;

- if $b = 2mk$ then there is an only equilibrium point (not considering $(0, 0, 0, 0)$):

$$\left(\frac{b}{2m}, \frac{b}{2c}, 0, 0 \right);$$

- if $b < 2mk$ then there are not any equilibrium points (not considering $(0, 0, 0, 0)$).

So the solutions

$$\boxed{E_2 = \left(\frac{b + \sqrt{b^2 - 4m^2k^2}}{2m}, \frac{b + \sqrt{b^2 - 4m^2k^2}}{2c}, 0, 0 \right)}$$

and

$$\boxed{E_3 = \left(\frac{b - \sqrt{b^2 - 4m^2k^2}}{2m}, \frac{b - \sqrt{b^2 - 4m^2k^2}}{2c}, 0, 0 \right)}$$

are equilibrium points and for the positivity of parameters they are feasible for $b > 2mk$ (the solution E_3 is always positive), while for $b = 2mk$ they degenerate in a single feasible solution: $E_2 = E_3$.

7. (L, B, R, P) = **(L, 0, R, 0)**
   From the first equation of 4.3 we found $L = 0$ and so for $L > 0$ there are not any equilibrium points.

8. (L, B, R, P) = **(L, 0, 0, P)**
   From the first equation of 4.3 we found $L = 0$ and so for $L > 0$ there are not any equilibrium points.

9. (L, B, R, P) = **(0, B, R, 0)**
   From the first equation of 4.3 we found $B = 0$ and so for $B > 0$ there are not any equilibrium points.

10. (L, B, R, P) = **(0, B, 0, P)**
    From the first equation of 4.3 we found $B = 0$ and so for $B > 0$ there are not any equilibrium points.

	Equilibrium	L	B	R	P	Feasibility conditions
1	E_1	0	0	0	0	always
6	E_2	$*$	$*$	0	0	$b \geq 2mk$
6	E_3	$*$	$*$	0	0	$b \geq 2mk$
8	E_*	$*$	$*$	$*$	$*$	

Table 4.2: Summering table of equilibria and existence conditions (the symbol $*$ denotes a population not necessarily equal to zero).

11. (L, B, R, P) = **(0, 0, R, P)**
    From the second equation of 4.3 we found $P = 0$ and so for $P > 0$ there are not any equilibrium points.

12. (L, B, R, P) = **(L, B, R, 0)**
    From the second equation of 4.3 we found $P = 0$ and so for $P > 0$ there are not any equilibrium points.

13. (L, B, R, P) = **(L, B, 0, P)**
    From the fourth equation of 4.3 we found $R = 0$ or $L = 0$, so for $R > 0$ or $L > 0$ there are not any equilibrium points.

14. (L, B, R, P) = **(L, 0, R, P)**
    From the first equation of 4.3 we found $L = 0$ and so for $L > 0$ there are not any equilibrium points.

15. (L, B, R, P) = **(0, B, R, P)**
    From the first equation of 4.3 we found $B = 0$ and so for $B > 0$ there are not any equilibrium points.

16. (L, B, R, P) = **(L, B, R, P)**
    Finally, we consider the case for which all populations do not vanish, namely the system exhibits coexistence. However, we are unable to find this equilibrium analytically by solving 4.3. This equilibrium will be not study in this Thesis.

Table 4.2 lists the result obtained. In particular, we summarize all the possible equilibrium points with their feasibility conditions.

4.4 Stability of equilibrium points

The aim now is to verify the stability of equilibria determined in the previous section (the structure of this part is taken from [3]).

We proceed with the analysis of the stability for each equilibrium point. In the following part *stability* means *asymptotically stability*.

The Jacobian matrix for the system 4.2 at a generic point is the equation

4.4.

$$J = \begin{pmatrix} -c & \frac{2bk^2B}{(k^2+B^2)^2} & 0 & 0 \\ c & -m & 0 & -\mu \\ -gcR & 0 & -gcL & ra \\ gcR & -eP & gcL & -n-a-eB \end{pmatrix}.$$

4.4.1 Stability analysis for E_1

For the equilibrium point $E_1 = (0, 0, 0, 0)$, the Jacobian matrix is

$$J(E_1) = \begin{pmatrix} -c & 0 & 0 & 0 \\ c & -m & 0 & -\mu \\ 0 & 0 & 0 & ra \\ 0 & 0 & 0 & -n-a \end{pmatrix}.$$

Given the property in A.1, the eigenvalues of 4.4.1 are the diagonal elements. These are:

$$\begin{aligned} \lambda_1 &= -c, \\ \lambda_2 &= -m, \\ \lambda_3 &= 0, \\ \lambda_4 &= -n-a. \end{aligned}$$

Because $lambda_3 = 0$, there is an eigenvalue with no real negative part and so, according to the Lyapunov theorem, the equilibrium E_1 is not stable.

4.4.2 Stability analysis for E_2 and E_3

From the second equation of the system 4.3 for the equilibrium points, we have that $cL = mB$, so from the first equation of 4.3 we have

$$\frac{b^2B^4}{(k^2+B^2)^2} = c^2L^2 = c^2B^2 \quad \text{so} \quad \frac{2bk^2B}{(k^2+B^2)^2} = \frac{2k^2m^2}{bB},$$

and the Jacobian becomes

$$J_2 = \begin{pmatrix} -c & \frac{2k^2m^2}{bB} & 0 & 0 \\ c & -m & 0 & -\mu \\ -gcR & 0 & -gcL & ra \\ gcR & -eP & gcL & -n-a-eB \end{pmatrix}.$$

For the equilibrium points $(\omega_\pm/m, \omega_\pm/c, 0, 0)$, such that

$$\omega_\pm = \frac{1}{2}\left(b \pm \sqrt{b^2 - 4m^2k^2}\right),$$

the new Jacobian matrix J_2, without explaining ω, is

$$J_2(\omega_\pm/m, \omega_\pm/c, 0, 0) = \begin{pmatrix} -c & \frac{2k^2m^3}{b\omega_\pm} & 0 & 0 \\ c & -m & 0 & -\mu \\ 0 & 0 & -g\omega_\pm & ra \\ 0 & 0 & g\omega_\pm & -n-a-\frac{e}{m}\omega_\pm \end{pmatrix}.$$

The matrix 4.4.2 is a block triangular matrix and so for A.1 the eigenvalues of the Jacobian are the eigenvalues of the two square matrices A and B.

$$A = \begin{pmatrix} -c & \frac{2k^2m^3}{b\omega_\pm} \\ c & -m \end{pmatrix}$$

$$B = \begin{pmatrix} -g\omega_\pm & ra \\ g\omega_\pm & -n-a-\frac{e}{m}\omega_\pm \end{pmatrix}$$

Considering the matrix A, for the theorem 4.1.6 it has two eigenvalues with negative real part if and only if

$$tr(A) = -m - c < 0 \quad \text{and} \quad det(A) = cm - \frac{2ck^2m^3}{b\omega_\pm} > 0.$$

While the first one is verified for the positivity of parameters, the second expression is equivalent to

$$\omega_\pm > \frac{2k^2m^2}{b}.$$

In the case of E_2 the relation 4.4.2 becomes

$$\frac{1}{2}\left(b + \sqrt{b^2 - 4m^2k^2}\right) > \frac{2k^2m^2}{b}.$$

from that the following expression is obtained

$$b^2 > 4k^2m^2,$$

that is the condition for existence of E_2.

In the case of E_3 the relation 4.4.2 becomes

$$\frac{1}{2}\left(b - \sqrt{b^2 - 4m^2k^2}\right) > \frac{2k^2m^2}{b}.$$

from that the following expression is obtained

$$b^2 < 4k^2m^2,$$

that is in contrast with the condition for existence of E_3 and so it is not a stable equilibrium.

The remaining candidate $E_2 = (\omega_+/m, \omega_+/c, 0, 0)$ is stable if B has two eigenvalues with negative real part. Because we have $tr(B) < 0$, E_2 is stable if and only if

$$det(B) = g\omega_+\left(n + a + \frac{e}{m}\omega_+\right) - gra\omega_+ = g\omega_+\left(n + a + \frac{e}{m}\omega_+ - ra\right) > 0,$$

	Expression	Feasibility	Stability
E_1	$(0,0,0,0)$	always	never
E_2	$\left(\frac{b+\sqrt{b^2-4m^2k^2}}{2m}, \frac{b+\sqrt{b^2-4m^2k^2}}{2c}, 0, 0\right)$	$b \geq 2mk$	$\omega_+ > m(ra-n-a)/e$
E_3	$\left(\frac{b-\sqrt{b^2-4m^2k^2}}{2m}, \frac{b-\sqrt{b^2-4m^2k^2}}{2c}, 0, 0\right)$	$b \geq 2mk$	never
E_*	(L_*, B_*, R_*, P_*)		

Table 4.3: Summaring Table of equilibria: existence and stability conditions. Remember that $\omega_+ = (b \pm \sqrt{b^2 - 4m^2k^2})/2$.

that is

$$\begin{cases} \omega_+ > 0 \\ \omega_+ > \frac{m}{e}(ra-n-a) \end{cases} \quad \text{or} \quad \begin{cases} \omega_+ < 0 \\ \omega_+ < \frac{m}{e}(ra-n-a). \end{cases}$$

Because for E_2 we have $\omega_+ > 0$, the stability condition began

$$\omega_+ > m(ra-n-a)/e.$$

4.4.3 Stability analysis for E_*

As we have seen in the previous section, the coexistence equilibrium E_* is not analytically tractable.

Finally, the Table 4.3 summarizes the feasibility and stability conditions of the equilibrium points.

4.5 Model parameters: assumptions and estimates

Some parameters of the model are taken from literature, like shown in Table 4.4, and the remaining ones are estimated using the experimental data.

The birth rate of larvae, specified as the number of larvae born per day, is proposed being $b = 2500$, while the bee maturation rate, specified as the number of bees that come out of the cells every day, is $c = 0.05$, equivalent to a 20-day growth cycle. The bee natural mortality rate $m = 0.04$ is equivalent to choose 25 days as life expectancy. Another bee mortality term is due to the varroa action and it is $\mu = 10^{-7}$.

Always in [18] the bees growth is modeled with a sigmoidal Hill function (in our case $N = 2$), i.e.

$$g(B) = \frac{B^N}{k^N + B^N},$$

where the parameter k is the size of the bee colony at which the birth rate is half of the maximum possible rate and the integer exponent $N > 1$. If $k = 0$ is chosen, then the brood is always reared at maximum capacity, independent of the actual bee population size, because $g(B) \equiv 1$. In [18] the value of this parameter is $k = 0.000075$ for spring and autumn and $k = 0.00003125$ for summer, so we choose the value linked to the summer period.

	Parameter meaning	Value	Unit	Ref.
b	Bee daily birth rate	2500	day^{-1}	[13]
c	Bee maturation rate	0.05	day^{-1}	[13]
m	Bee natural mortality rate	0.04	day^{-1}	[14]
μ	Bee mortality rate for varroa action	10^{-7}	day^{-1}	[15]
r	Varroa growth rate	to be estimated	day^{-1}	-
n	Varroa natural mortality rate in phoretic phase	0.007	day^{-1}	[17]
e	Grooming rate of bee	5×10^{-6}	day^{-1}	[13]
g	Varroa transition rate from reproductive to phoretic phase	to be estimated	day^{-1}	-
a	Varroa transition rate from phoretic to reproductive phase	to be estimated	day^{-1}	-
k	Bees minimum number index for which bee growth is linear	3.1255×10^{-5}	-	[18]

Table 4.4: Model parameters. The parameters r, a and g will be estimate like an optimization problem.

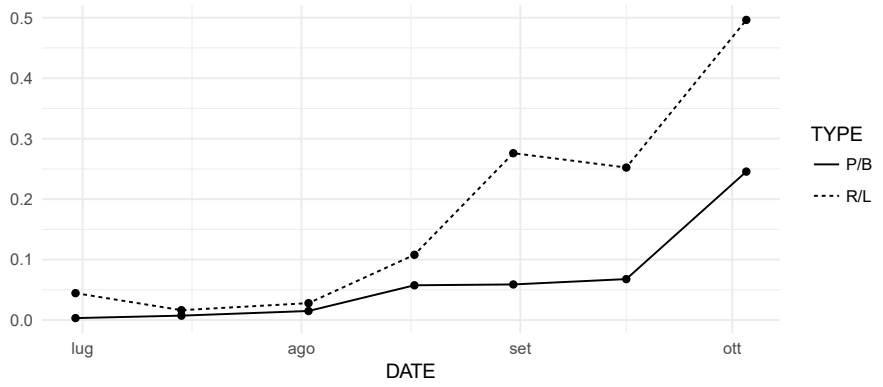


Figure 4.1: The top function represents the percentage of varroa in the cells trend, instead the lower one is the percentage of varroa in phoretic stage trend.

Note that with this parameters values, the equilibrium feasibility condition $b \geq 2mk$ is satisfied.

The literature does not provide a precise value corresponding to the grooming behavior, but from [13] we have a range of reasonable values for this parameter from 10^{-6} to 10^{-5}, so for the simulations we choose $e = 5 \times 10^{-6}$. From [17] varroa natural mortality rate n is taken equal to 0.007.

For parameters estimation we use the data concerning the average situation in the eleven hives, i.e. the average percentages trend shown in Figure 4.1 and the average inoculum 15.54545 for the eleven hives. Obviously to compare experimental data with model data, we can not consider our model populations L, B, R nd P, but their ratio R/L and P/B.

For an optimization problem the initial conditions are necessary and they have been chosen from literature, experiment and considerations on the problem.

	Variable meaning	Value	Initial value meaning
$L(0)$	Number of larvae, i.e of opercolated cells	1	To not have an indeterminate form for R/L
$B(0)$	Number of adult bees	60000	Average value of estimated range for population dimension
$R(0)$	Varroa in reproductive stage	0	Imagine to have 0 opercolated cells
$P(0)$	Varroa in phoretic stage	15.54545	Mean value of the inoculants in hives

Table 4.5: Initial values of the populations in the optimization problem.

The initial number of larvae is supposed equal to 0, but this could cause an undetermined form for the quantity R/L, so we choose conventionally $L(0) = 1$. The choice $B(0) = 60000$ is the average value in the estimated range $[40000, 80000]$, in which the number of apis mellifera in a colony is estimated to move. The initial value of the number of varroa in reproductive stage is 0, for the idea used in the choice of $L(0)$, and for the number of varroa in phoretic stage we use the average inoculum, so $P(0) = 15.54545$. This values are shown in Table 4.5.

The parameters r, a and g will be estimate, but we can calculate relative intervals in which we expect to find their values. The varroa growth rate in the bee cells, r, is a net growth value, i.e. with it the number of varroa that will exit infested cells is modeled. From [18] we know that for the competition within the cell on average each mother varroa produces $1.3 - 1.45$ descendants in the female brood and $2.2 - 2.6$ from the drone brood, and as mentioned in 2.1, usually in a beehive there are from 40 to 100 thousand females and between April and July from 500 to 2000 males. For r range we make a weighted average based on the sex of the larva, first considering the minimum reproduction values and then the maximum ones (shown in 4.5).

$$r_{low} = \frac{40000 \cdot 1.3 + 500 \cdot 2.2}{40500} \simeq 1.311 \quad \text{and} \quad r_{sup} \simeq 1.4725.$$

This means that $r \in [2.311, 2.4725]$ because considering the mother that enter the cell and then exit, we must add 1 to the values found. Regarding the number of varroa that from phoretic stage enter the cells, we know from [18] that the phoretic phase lasts $4 - 10$ days in the presence of brood. Obviously, in the absence of brood conditions, the mites are forced to remain phoretic. For this reason we estimate that this parameter is from $1 \div 10 = 0.1$ to $1 \div 4 = 0.25$, i.e. $a \in [0.1, 0.25]$.

Regarding the optimization problem, the data on the percentages of varroa in cells and in the phoretic phase are used to calculate the quadratic errors respect our model, and to minimize them to obtain the estimate of the parameters. To minimize them, we us an implementation of the method of *Nelder* and *Mead* (1965, defined in A.2), that uses only function values and is robust but relatively slow. It will work reasonably well for non-differentiable functions.

Table 4.6 shows the results of the optimization problem. Note that the a parameter is in the expected range, while r is slightly higher than the expected maximum value.

The Figure 4.2 shows the populations trend with new estimated parameters

	Parameter meaning	Value	Unit
r	Varroa growth rate	2.77	day^{-1}
g	Varroa transition rate from reproductive to phoretic phase	1.723×10^{-4}	day^{-1}
a	Varroa transition rate from phoretic to reproductive phase	0.2463	day^{-1}

Table 4.6: Estimated model parameters values.

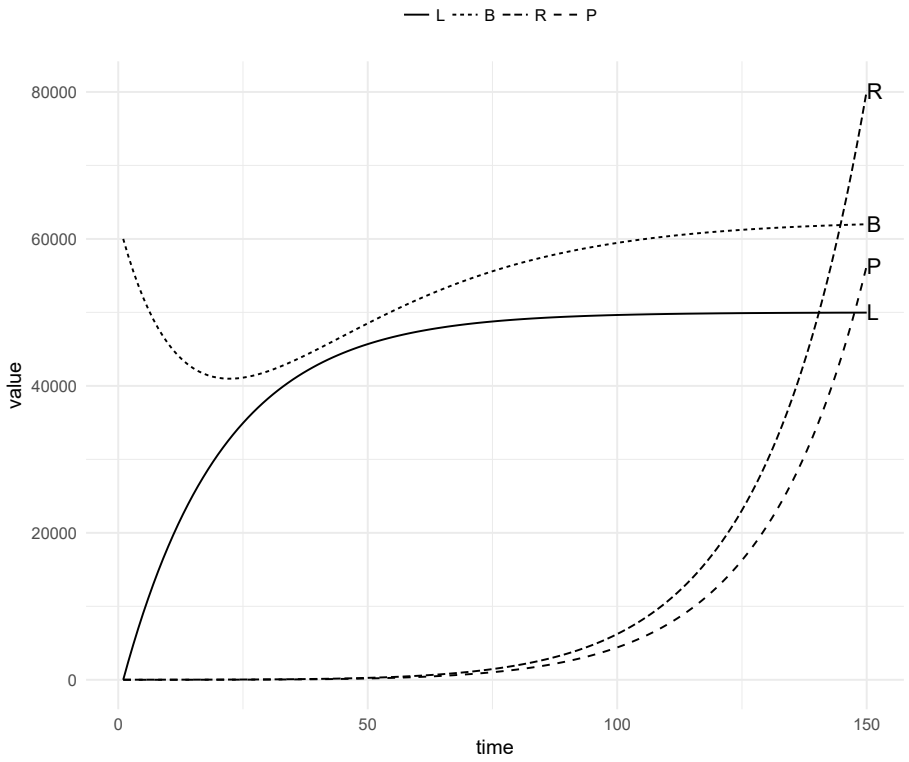


Figure 4.2: The top function represents the percentage of varroa in the cells trend, insted the lower one is the percentage of varroa in phoretic stage trend.

in Table 4.6. The initial negative trend of the number of bees is due to the low number of larvae, that once reached values around 30000 ago, the numbers of bees starts an increasing trend.

Figure 4.3 shows the different between experiment and model values for the average number of varroa per opercolated cell. Almost everywhere the model underestimates the experimental data, even if the last measurement is overestimated.

Figure 4.4 shows the different between experiment and model values for the average number of varroa in phoretic stage per bee. This quantity is much better fitted from the theoretical model, in fact there is a good alternation of positive and negative errors.

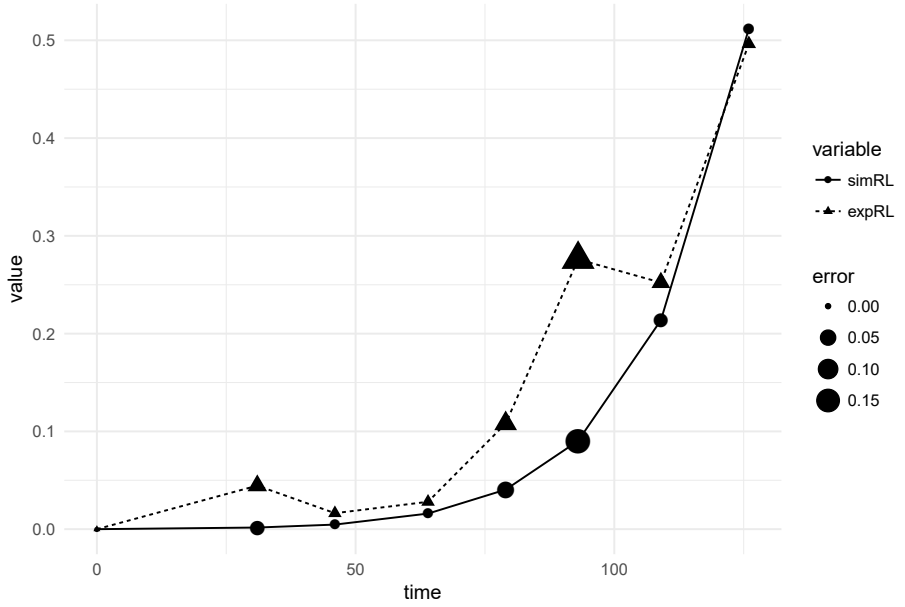


Figure 4.3: The line - - - shows the experimental values for the average number of varroa per opercolated cell, insted the line — is for the theoretical model values.

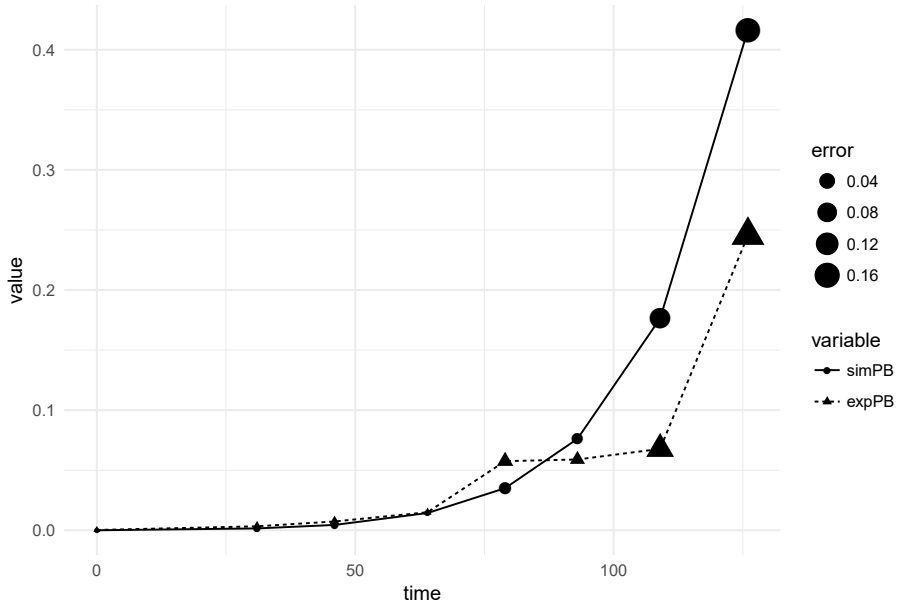


Figure 4.4: The line - - - shows the experimental values for the average number of varroa in phoretic stage per bee, insted the line — is for the theoretical model values.

It is clear that our model works very well for the determination of values in the first 2 months of experiment, and worsens for evaluations from 3 months onwards, even if in the case of the size R/L from 4 months onwards the two curves they rejoin perfectly. Note that the robustness of the model was tested using different initial conditions and the values of the parameters emerging from the optimization method were always the same.

Chapter 5

Statistical background and considerations

Before the introduction of the model, we want to recall that:

- $AVRG_{hi}$ is the percentage of infested cells on general cells, of hth hive at ith time point;
- $PERC\_PHO\_VARR_{hi}$ is the number of varroa in phoretic stage on 100 bees, of hth hive at ith time point;
- $DEAD\_HIVE_h$ is a boolean variable that says if the infestation caused the dead of the hth hive;
- $INOCULUM_h$ is the initial inoculum of hth hive;
- t_i is the time of ith measurement,

where $i = 1, \cdots 7$, $h = 1, \cdots, 12$ ($h \neq 10$) and $k = 1, \cdots, 100$.

The purpose of the statistical analysis is to choose which statistical models to use to model $AVRG$ and $PERC\_PHO\_VARR$, that are the same quantiy modeled in the Chapter 4. In this chaper there are some theoretical backgrounds used to model our response variables.

Before continuing, we wish to underline the fact that most of the statistical approach are based on [4].

5.1 Linear model and GLM

In this section we present the simplest statistical model, i.e. the linear model, that is a particular case of GLM (Generalized Linear Model). But before introducing this family of model, we define the family of distributions used by GLMs: the *exponential family*.

5.1.1 Exponential family

Definition 5.1.1 *Given a measure η, a distribution falls into the exponential family if its distribution function can be written as*

$$f(Y \mid \theta) = h(Y)exp\{\theta^T T(Y) - A(\theta)\},$$

for a parameter vector θ, often referred to as the canonical parameter, and for given functions T and h. The statistic $T(Y)$ is referred to as a sufficient statistic. The function $A(\theta)$ is known as the cumulant function.

Integrating equation 5.1.1 with respect to the measure ν, we have:

$$A(\theta) = log \int h(Y)exp\{\theta^T T(Y)\}\nu(dY)$$

where we see that the cumulant function cab be viewed as the logarithm of a normalization factor. This shows that $A(\theta)$ is not a degree of freedom but it is determined once ν, $T(Y)$ and $h(Y)$ are determined.

Let us now consider computing the first derivative of $A(\theta)$ for a general exponential family distribution. The computation begins as follows:

$$\frac{\partial A}{\partial \theta^T} = \frac{\partial}{\partial \theta^T}\left\{log \int h(Y)exp\{\theta^T T(Y)\}\nu(dY)\right\}$$

To proceed we need to move the gradient past the integral sign. In general derivates can not be moved past integral signs. However, in this case the move is justified but we don't prove it here. Thus we continue our computation:

$$\begin{aligned}\frac{\partial A}{\partial \theta^T} &= \frac{\int T(Y)h(Y)exp\{\theta^T T(Y)\}\nu(dY)}{\int h(Y)exp\{\theta^T T(Y)\}\nu(dY)} = \\ &= \int T(Y)exp\{\theta^T T(Y) - A(\theta)\}h(Y)\nu(dY) = \\ &= \mathbb{E}[T(Y)].\end{aligned}$$

We see that the first derivative of $A(\theta)$ is equal to the mean of the sufficient statistic. Let us now take the second derivative:

$$\begin{aligned}\frac{\partial^2 A}{\partial\theta\partial\theta^T} &= \int T(Y)\left(T(Y) - \frac{\partial}{\partial\theta^T}A(\theta)\right)^T exp\{\theta^T T(Y) - A(\theta)\}h(Y)\nu(dY) = \\ &= \int T(Y)\left(T(Y) - \mathbb{E}[T(Y)]\right)^T exp\{\theta^T T(Y) - A(\theta)\}h(Y)\nu(dY) = \\ &= \mathbb{E}[T(Y)T(Y)^T] - \mathbb{E}[T(Y)]\mathbb{E}[T(Y)]^T = \\ &= Var[T(Y)].\end{aligned}$$

and thus we see that the second derivative of $A(\theta)$ is equal to the variance (i.e. the covariance matrix) of the sufficient statistic.

In the following part we present three distributions that we use in our models: normal, Poisson and negative binomial.

Normal distribution Every normal distribution is a particular distribution of the exponential family in which:

- $\theta = [\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}]^T$;
- $T(Y) = [Y, Y^2]^T$;
- $A(\theta) = \frac{\mu^2}{2\sigma^2} + log(\sigma)$;
- $h(Y) = \frac{1}{\sqrt{2\pi}}$.

Infact a normal variable, $Y \sim N(\mu, \sigma^2)$, is described by this probability density function:

$$f(Y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(Y-\mu)^2}{2\sigma^2}\right\} = \\ = \frac{1}{\sqrt{2\pi}} exp\left\{\frac{\mu Y}{\sigma^2} - \frac{Y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - log(\sigma)\right\}.$$

5.1.2 Linear model and GLM

The *Generalized Linear Model* was originally formulated by John Nelder and Robert Wedderburn as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression [22].

Definition 5.1.2 *Given a univariate response variable Y and some predictor X_i with $i \in \{1, \cdots, p\}$, a **GLM** chooses an exponential family distribution for Y and a **link function** $g(\cdot)$ relating the expected value of Y to the predictor variables via a structure such as*

$$g(\mathbb{E}(Y)) = \eta(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

A GLM consists of two steps:

- an **assumption on the distribution** of the response variable Y;, that defines its mean and variance;
- specification of the **link function**, that is the specification of the relationship between the mean value of Y and the systematic part.

Definition 5.1.3 *A **linear model** is a particula GLM in which:*

- *Y is assumed to be normally distributed;*
- *$g(\cdot)$ is the identity function;*
- $\mathbb{E}(Y) = \eta(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.

Without doubt the linear regression model is the mother of all models. [4] The model is based on a series of assumptions: normality, homogeneity, fixed X, independence and correct model specification. In ecology, the data are seldom modelled adequately by linear regression models. To apply a linear regression

model on data, they must be all verificated and this verification process is called the *model validation process*.

We now introduce the hypothesis of the linear model: normality, Heteroscedasticy, fixed predictors and independence.

Several authors argue that violation of **normality** is not a serious problem [8] as a consequence of the central limit theory. Normality at each X value should be checked by making a histogram of all observations at that particular X value. Very often, we don't have multiple observations (sub-samples) at each X value. In that case, the best we can do is to pool all residuals and make a histogram of the pooled residuals; normality of the pooled residuals is reassuring. The residuals represent the information that is left over after removing the effect of the explanatory variables. However, the raw data Y contains the effects of the explanatory variables. To assess normality of the Y data, it is therefore misleading to base your judgement purely on a histogram of all the Y data.

Heteroscedasticy can be checked by the comparison of the spread of the residuals with respect to the different X and fitted values. The only thing to analyze is to pool all the residuals and plot them against fitted values. The spread should be roughly the same across the range of fitted values and predictors. The easiest option to deal with Heterogeneity is a data transformation. The assessment of the homogeneity purely based on a graphical inspection of the residuals is generally preferred.

Fixed X is an assumption implying that the explanatory variables are deterministic. The values at each sample are know in advance.

Violation of **independence** is the most serious problem as it invalidates important tests such as the F-test and the t-test. A key question is then how do we identify a lack of independence and how do deal with it. You have violation of independence if the Y_i value is influenced by an other one Y_j [9]. In fact, there are two ways that this can happen: either an improper model or dependence structure due to the nature of the data itself. Other causes for violation of independence are due to the nature of the data itself. If it rains at $100m$ in the air, it will also rain at $200m$ in the air. This type of violation of independence can be taken care of by incorporating a temporal or spatial dependence structure between the observations (or residuals) in the model.

Standard model validation graphs are *versus fitted values*, i.e. plotting residuals respect the fitted values, to verify homogeneity, a *Q-Q plot* or histogram of the residuals for normality, and *residuals versus each explanatory variable* to check independence.

5.1.3 Interaction factors [20]

The typical treatment of interactions in linear models is to consider the interaction as a product term of the main effect variables. This takes the form of the equation (5.1).

$$\mathbb{E}[Y_i|X] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}. \tag{5.1}$$

The complete product term is called a *first-order interaction*, where for obviuous reason the order is one less the number of factors. Subject to mild assumpions [19], the sampling distribution of β_3 over its standard error is student's-t with $N - k - 1$ degrees of freedom.

The meaning of interaction in the linear model is actually easier to interpret if equation (5.1) is rearranged as follows:

$$\mathbb{E}[Y_i|X] = \beta_0 + \beta_1 X_{i1} + (\beta_2 + \beta_3 X_{i1}) X_{i2}. \tag{5.2}$$

If one is interested in the consequence from changes in the explanatory variable X_{i2} on the outcome variable, it is necessary to take the first derivate of equation (5.2) with respect to this variable in order to obtain the *marginal* effect as a composite coefficient estimate.

$$\frac{\partial}{\partial X_{i2}} \mathbb{E}[Y_i|X] = \beta_2 + \beta_3 X_{i1}$$

This is useful because it demonstrates that the effect of levels of X_{i2} on the outcome variable is intrinsically tied to specific levels of X_{i1}: the marginal contribution of X_{i2} is conditional on X_{i1}.

Interaction effects are more complicated in generalized linear models due to the link function between the systematic component and the outcome variable. From definition 5.1.2, we know that in GLM's the systematic component is related to the mean of the outcome variable by a smooth, invertible function, $g(\cdot)$, according to (5.3) (writed in matrix form).

$$g(\mu) = X\beta \quad \text{where} \quad \mu = \mathbb{E}[Y|X] = g^{-1}(X\beta) \tag{5.3}$$

Using the link function, it is possible to change equation (5.3) to the more general form expressed by (5.4):

$$\mathbb{E}[Y_i|X] = g^{-1}(\beta_0 + \beta_1 X_{i1} + (\beta_2 + \beta_3 X_{i1}) X_{i2}). \tag{5.4}$$

A less well-understood ramification of interactions in generalized linear models is that by including a link function, the model automatically specifies interactions on the natural scale of the linear predictor (though not necessarily on the transformed scale of the linear predictor). To see that this is true, revisit the calculation of the marginal effect of a single coefficient by taking the derivative of equation (5.4) but without an explicitly specified multiplicative term for the interaction. If the form of the model implied no interactions, then this calculation would produce a marginal effect free of other variables, but this is clearly not so:

$$\frac{\partial}{\partial X_{i2}} \mathbb{E}[Y_i|X] = \frac{\partial}{\partial X_{i2}} g^{-1}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}) = (g^{-1})'(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})\beta_2.$$

From this discussion it is clear that interactions are naturally produced in GLM's, regardless of whether they are recognized or desired. Yet this observation does not really help in testing for the existence and statistical reliability of hypothesized interactions, or in determining overall model quality in the acknowledged presence of such terms.

5.1.4 Structural variance models

In this part some models are presented in which the variance of the residuals is not supposed constant, but dependent on some variables used in the model. An explanatory variable used to model the variance of residuals is called **variance**

covariate. The trick is to find the appropriate structure for the variance of residuals. The easiest approach to choosing the best variance structure is to apply the well knows structures and compare them using the AIC or using biological knowledge combined with some informative plots. The AIC is defined in A.4: lower values are better. Some of the variance functions are nested, and a likelihood ratio test can be applied to judge which one performs better for your data.

Fixed variance structure The first option is called *fixed variance* and it assumes that

$$Var(\epsilon_i) = \sigma^2 X_k, \quad i = 1, \cdots, n, \quad k = 1, \cdots, p,$$

for k taken in $\{1, \cdots, n\}$ and so X_k is a particular predictor chosen among those chosen in the model.

VarIdent variance structure The second one option is called *varIdent variance*. This variance structure is used in specific cases, such as longitudinal data (defined in 5.2), where the modeling variable is Y_{it}, where t indicates the measurement date made on the i-th subject. In this case it is assumed that the error variance is different per subject, i.e.:

$$\epsilon_{it} \sim N(0, \sigma_i^2), \quad i = 1, \cdots, n.$$

VarPower variance structure Then we look at the *power of the covariate* variance structure, that is

$$\epsilon_i \sim N(0, \sigma^2 |X_i|^{2\delta}), \quad i = 1, \cdots, n.$$

The variance of the residuals is modelled as σ^2 multiplied with the power of the absolute value of the variance covariate X. The parameter δ is unknown and needs to be estimated. If $\delta = 0$, we obtain the simple linear regression model, so this model is nested with the simple linear one, and therefore the likelihood ratio test can be applied to judge which one is better. If the variance covariate has values equal to 0, the variance of the residuals is 0 as well. This causes problems in the numerical estimation process, and if the variance covariate has values equal to zero, the varPower should not be used.

It is also possible to allow multiple variables in the form argument. This extension makes it possible to model a case like longitudinal data (defined in 5.2), infact this structure model an increase in spread for larger t values, but only in certain subjects. The structure for the residuals is now the following one:

$$\epsilon_{it} \sim N(0, \sigma^2 |X_{it}|^{2\delta}), \quad i = 1, \cdots, n, \quad t = 1, \cdots, T.$$

VarExp variance structure If the variance covariate can take the value of zero, the *exponential variance structure* is a better option. In this case the

residual variance takes the following form:

$$\epsilon_{it} \sim N\left(0, \sigma^2 e^{2\delta X_{it}}\right), \quad i = 1, \cdots, n, \quad t = 1, \cdots, T.$$

VarConstPower variance structure Getting to the point, the model is

$$\epsilon_{it} \sim N\left(0, \sigma^2(\delta_1 + |X_{it}|^{\delta_2})^2\right), \quad i = 1, \cdots, n, \quad t = 1, \cdots, T.$$

VarComb variance structure With this last variance structure, we can allow for both an increase in residual spread for larger t values as well as a different spread per subject. This variance structure is of the form:

$$\epsilon_{it} \sim N\left(0, \sigma_i^2 e^{2\delta X_{it}}\right), \quad i = 1, \cdots, n, \quad t = 1, \cdots, T.$$

Note that σ has an index i like subject. This is a combination of varIdent and varExp.

5.2 Mixed models and longitudinal data

In this section the longitudinal data are introduced and after a part that speaks about the mixed model, i.e. the model used in cases of longitudinal data.

Longitudinal data Longitudinal data, sometimes referred to as panel data, track the same sample at different points in time. The sample can consist of individuals, households, establishments, and so on. They are often used in social-personality and clinical psychology, in developmental psychology (to study developmental trends across the life span), in sociology, in medicine (to uncover predictors of certain diseases), in advertising (to identify the changes that advertising has produced in the attitudes and behaviors of people). The reason for this is that unlike cross-sectional studies, in which different individuals with the same characteristics are compared, longitudinal studies track the same individuals and so the differences observed in them are less likely to be the result of particular characteristics between two indivisuals. Longitudinal studies thus make observing changes more accurate. When longitudinal studies are observational, in the sense that they observe the state of the world without manipulating it, it has been argued that they may have less power to detect causal relationships than experiments. However, because of the repeated observation at the individual level, they have more power than cross-sectional observational studies, by virtue of being able to exclude time-invariant unobserved individual differences and also of observing the temporal order of events. Types of longitudinal studies include cohort studies, which sample a cohort (a group of people who share a defining characteristic, typically who experienced a common event in a selected period) and perform cross-section observations at intervals through time.

Repeated measures data consist of measurements of a response on several experimental (or observational) units. Considering the case of experiment about varroa, the response variables are the varroa in the cells and the percentage

of varroa in phoretic stage, and the observational units are the measurements in the same beehives around the time. This has already been shown by the Figures 3.2, 3.3, 3.4 and 3.5 in the chapter where data are presented. Even if some measurement is missing, these data are *balanced* in that each subject is measured the same number of times and on the same occasions.

Mixed-effects models Longitudinal data generally result in the correlated errors that are explicitly forbidden by regression models. Mixed model analysis provides a general, flexible approach for correlated data, because it allows a wide variety of correlation patterns (or variance- covariance structures) to be explicitly modeled. The term mixed model refers to the use of both fixed and *random effects* in the same analysis. These repeated measures approaches discard all results on any subject with even a single missing measurement, while mixed models allow other data on such subjects to be used as long as the missing data meets the so-called missing-at-random definition. Another advantage of mixed models is that they naturally handle uneven spacing of repeated measurements, whether intentional or unintentional. Also important is the fact that mixed model analysis is often more interpretable than classical repeated measures.

In a random effects model, the unobserved variables are assumed to be uncorrelated with (or, more strongly, statistically independent of) all the observed variables. An effect is classified as a random effect when you want to make inferences on an entire population, and the levels in your experiment represent only a sample from that population and for this for some coefficients you want different values for each levels values. .

Definition 5.2.1 *A **random intercept model** has the form*

$$Y_{hi} = \beta_0 + \beta_1 X_{hi} + v_h + \epsilon_{hi},$$

where

- $h = 1, \cdots, H$ *with* H *that indicates the number of subjects (beehives in our case);*
- $i = 1, \cdots, N$ *indicates the* i*th measurement;*
- $Y_{hi} \in \mathbb{R}$ *is the response for* i*th measurement of* h*th subject;*
- $\beta_0 \in \mathbb{R}$ *is the fixed intercept for the regression model;*
- $\beta_1 \in \mathbb{R}$ *is the fixed slope for the regression model;*
- $X_{hi} \in \mathbb{R}$ *is the predictor for* i*th measurement of* h*th subject;*
- $v_h \overset{iid}{\sim} N(0, \sigma_v^2)$ *is the random intercept for the* h*th subject;*
- $\epsilon_{hi} \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$ *is the normal error term.*

Note that v_h allows each subject to have unique regression intercept and $(Y_{hi}|X_{hi}) \sim N(\beta_0 + \beta_1 X_{hi}, \sigma_v^2 + \sigma_\epsilon^2)$ is an assumption.

The covariance between any two observations is another thing to note. Taken two different observations of the same subject in different times, the covariance is

$$\begin{aligned} Cov(Y_{hi}, Y_{hj}) &= \mathbb{E}\left[(Y_{hi} - \mathbb{E}[Y_{hi}])(Y_{hj} - \mathbb{E}[Y_{hj}])\right] = \\ &= \mathbb{E}\left[(v_h + \epsilon_{hi})(v_h + \epsilon_{hj})\right] = \\ &= \mathbb{E}[v_h^2] + \mathbb{E}[v_h]\mathbb{E}[\epsilon_{hj}] + \mathbb{E}[v_h]\mathbb{E}[\epsilon_{hi}] + \mathbb{E}[\epsilon_{hi}]\mathbb{E}[\epsilon_{hj}] = \\ &= \mathbb{E}[v_h^2] = \sigma_v^2, \end{aligned}$$

while, for two different observations of different subjects in different times,

$$\begin{aligned} Cov(Y_{h_1 i}, Y_{h_2 j}) &= \mathbb{E}\left[(Y_{h_1 i} - \mathbb{E}[Y_{h_1 i}])(Y_{h_2 j} - \mathbb{E}[Y_{h_2 j}])\right] = \\ &= \mathbb{E}\left[(v_{h_1} + \epsilon_{h_1 i})(v_{h_2} + \epsilon_{h_2 j})\right] = \\ &= \mathbb{E}[v_{h_1} v_{h_2}] + \mathbb{E}[v_{h_1}]\mathbb{E}[\epsilon_{h_2 j}] + \mathbb{E}[v_{h_2}]\mathbb{E}[\epsilon_{h_1 i}] + \mathbb{E}[\epsilon_{h_1 i}]\mathbb{E}[\epsilon_{h_2 j}] = 0. \end{aligned}$$

For equations 5.2 and 5.2 note that $\mathbb{E}[v_h] = \mathbb{E}[\epsilon_{hi}] = 0$, $\forall h, i$ for definiton, and $v_h \sim N(0, \sigma_v^2)$, so $v_h^2 \sim \sigma_v^2 \chi^2(1) = Gamma(1/2, 2\sigma_v^2)$, that has expected value equal to σ_v^2.

Finally the covariance between any two observations is

$$Cov(Y_{h_1 i}, Y_{h_2 j}) = \begin{cases} 1 & \text{if} \quad h_1 = h_2, \quad i = j \\ \sigma_v^2 & \text{if} \quad h_1 = h_2, \quad i \neq j \\ 0 & \text{if} \quad h_1 \neq h_2. \end{cases}$$

Definition 5.2.2 *A **random intercept and slope model** has the form*

$$Y_h i = \beta_0 + \beta_1 X_{hi} + v_{h0} + v_{h1} X_{hi} + \epsilon_{hi},$$

where

- $v_{h0} \overset{iid}{\sim} N(0, \sigma_0^2)$ *is the random intercept for the hth subject;*
- $v_{h1} \overset{iid}{\sim} N(0, \sigma_1^2)$ *is the random slope for the hth subject.*

The fundamental assumptions of the random intercept and slope model are:

- $(v_{h0}, v_{h1}) \overset{iid}{\sim} N(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{pmatrix},$$

  where v_{h0} and v_{h1} can be independents ($\sigma_{01} = 0$);
- $(Y_{hi} | X_{hi}) \sim N(\beta_0 + \beta_1 X_{hi}, \sigma_0^2 + 2\sigma_{01} X_{hi} + \sigma_1^2 X_{hi}^2 \sigma_\epsilon^2)$.

Now, without showing calculus, the covariance between any two observations is

$$Cov(Y_{h_1 i}, Y_{h_2 j}) = \begin{cases} 1 & \text{if} \quad h_1 = h_2, \quad i = j \\ \sigma_0^2 + \sigma_{01}(X_{h_1 i} + X_{h_2 j}) + \sigma_1^2 X_{h_1 i} X_{h_2 j} & \text{if} \quad h_1 = h_2, \quad i \neq j \\ 0 & \text{if} \quad h_1 \neq h_2. \end{cases}$$

Now we give the general definition of a linear mixed effects model.

Definition 5.2.3 *A **linear mixed effects model** has the form*

$$Y_h i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{hik} + v_{h0} + \sum_{k=1}^{q} v_{hk} Z_{hik} + \epsilon_{hi},$$

where

- $\beta_k \in \mathbb{R}$ *is the fixed slope for the kth predictor;*
- $X_{hik} \in \mathbb{R}$ *is the ith measurement of kth fixed predictor for hth subject;*
- $v_{hk} \overset{iid}{\sim} N(0, \sigma_k^2)$ *is the random slope for kth predictor of hth subject;*
- $Z_{hik} \in \mathbb{R}$ *is the ith measurement of kth random predictor for hth subject.*

Note that a predictor can be used both fixed and random predictor and using matrix notation, we can write the mixed effects model as

$$\mathbf{Y}_h = \mathbf{X}_h \beta + \mathbf{Z}_h v_h + \epsilon_h.$$

The covariance between any two observations is

$$Cov(Y_{h_1 i}, Y_{h_2 j}) = \begin{cases} 1 & \text{if} \quad h_1 = h_2, \quad i = j \\ \sigma_0^2 + \sigma_{01}(X_{h_1 i} + X_{h_2 j}) + \sigma_1^2 X_{h_1 i} X_{h_2 j} & \text{if} \quad h_1 = h_2, \quad i \neq j \\ 0 & \text{if} \quad h_1 \neq h_2. \end{cases}$$

5.3 Hurdle-at-zero models

In ecological research, most count data are zero inflated. This means that the response variable contains more zeros than expected, based on a particular distribution. For example, if we suppose we want to model N_{hi}, that is the number of varroa in a cell taken at time t_i from the h-th hive, we will find ourselves in front of an almost always zero variable, as shown by the Figure 5.1. Ignoring zero inflation can have two consequences: firstly, the estimated parameters and standard errors may be biased, and secondly, the excessive number of zeros can cause overdispersion. Before discussing the technique that can cope with all these zeros, we need to ask the question: why do we have all these zeros? Basically the data are divided in two imaginary group:

- **zero mass observations** that contain only zeros;
- **positive obsevations** that contain values larger than zero.

Let H a stochastic variable that we model with an hurdle-at-zero model. The probability to have a zero count is

$$P(H_i = 0) = \pi_i,$$

where we assume that the probability that H assumes a zero value is Bernoulli distributed with probability π_i, and automatically $1 - \pi_i$ is the probability to have a true zero, i.e. a zero that comes out of the principal phenomenon and not of the phenomenon that generates only zeros.

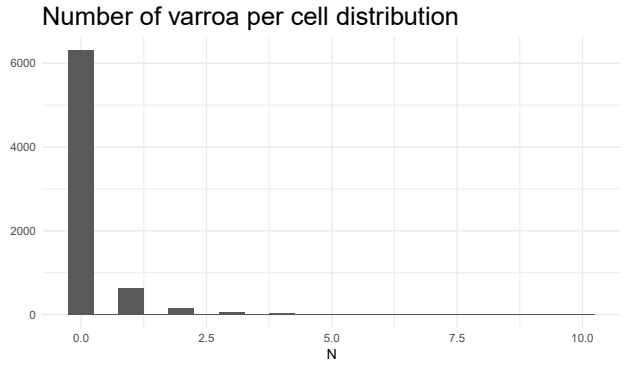


Figure 5.1: Most cells taken in the experiment are healthy, i.e. they do not count varroa ($N_{hi} = 0$).

Instead probabilities that H assumes positive values are give by

$$P(H_i = n \mid n > 0) = (1 - \pi_i)P(\text{count process gives no-zero}).$$

If we suppose that the main process for positive values of H follows a normal distribution, we have a *normal hurdle-at-zero* model (**NHZ**).

Let us assume that the positive values of H follows a normal distribution with density function $f_N(\cdot)$, so we have

$$P[H_i = n] = \begin{cases} (1 - \pi_{hi}) f_N(n) & \text{if } n > 0 \\ \pi_{hi} & \text{if } n = 0 \end{cases}.$$

The last step we need is to introduce covariates, that are used in GLMs. To model the probability of having a false zero, π_i, the easiest approach is to use a logistic regression with covariates:

$$\pi_i = \frac{e^{\alpha_0 + \alpha_1 Q_{i1} + \cdots + \alpha_1 Q_{iq}}}{1 + e^{\alpha_0 + \alpha_1 Q_{i1} + \cdots + \alpha_1 Q_{iq}}},$$

where the symbol Q for the covariates as these may be different to the covariates that influence the positive counts, and α are regression coefficients.

Chapter 6

Model REP: relationship between varroa in reproductive stage and larvae

In this section, relationship between varroa in reproductive stage and larvae is modeled. For each hive in each time we know the number of varroa founded in one hundred operculated cells. One hundred operculated cells means one hundred larvae, for this if we divide this number by hundred, we have exactly the ratio between varroa and larvae. This ratio is expressed by the variable $AVRG$ of our data frame and this variable is the one that we want to model.

6.1 Multicollinearity analysis

The aim of this section is to reduce the number of variables take in account in the models. This work is useful for understanding what kind of data we have, but it is a fundamental step before using a statistical regressive model. This work can be done before the choice of the model because it is independent of it. To reach this aim, the first step is to do a check for the multicollinearity, that is when two or more predictor variables in a multiple regression model are highly correlated, this means that one can be linearly predicted from the others with a substantial degree of accuracy. The concept of linear regression will then be presented in depth in the following chapters.

Definition 6.1.1 *Two variables are* ***perfectly collinear*** *if there is an exact linear relationship between them:*

$$X_{2i} = \beta_0 + \beta_1 X_{1i}.$$

A set of variables is ***perfectly multi-collinear*** *if there are one or more exact linear relationships among some of the variables*

$$\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} = 0.$$

There are two types of multicollinearity (not necessarily perfect):

- *structural multicollinearity* that is a mathematical artifact caused by creating new predictors from other predictors;
- *data-based multicollinearity* that is a result of a poorly designed experiment, reliance on purely observational data, or the inability to manipulate the system on which the data are collected.

In the case of structural multicollinearity, the multicollinearity is induced by what you have done. Data-based multicollinearity is the more troublesome of the two types of multicollinearity. Unfortunately it is the type that researchers encounter most often.

The effects of multicollinearity on the regression analyses are:

- the estimated regression coefficient of any one variable depends on which other predictor variables are included in the model;
- the precision of the estimated regression coefficients decreases as more predictor variables are added to the model;
- the marginal contribution of any one predictor variable in reducing the error sum of squares varies depending on which other variables are already in the model;
- hypothesis tests for $\beta_k = 0$ may yield different conclusions depending on which predictor variables are in the model (this effect is a direct consequence of the three previous effects).

Firstly, to study the collinearity, the correlation matrix of predictor variables is considered and shown in the Figure 6.1. The high correlations among some of the predictors suggest that multicollinearity exists. This matrix is usefull to delete from the model the variables that are strongly correlated. In this way the high correlation couples are:

- Corr(DEW_POINT, AVRG_TEMP) = 0.97;
- Corr(MAX_TEMP, AVRG_TEMP) = 0.94;
- Corr(MAX_TEMP, DEW_POINT) = 0.97;
- Corr(MAX_TEMP, AVRG_WIND) = 0.95.

Depend on this consideration, the variables AVRG_TEMP, DEW_POINT and AVRG_WIND are candidates to be not used in the model. They will be analyzed in descending order of correlation by observing the *Variation Inflation Factor* (*VIF*), defined by 6.1.2.

Let $(X_1, \cdots, X_p)$ the predictors sample for the response variable Y. For the model with only the predictor X_k

$$\hat{Y} = \beta_0 + \beta_k X_k, \quad k = 1, \cdots, p,$$

the variance of the estimated coefficient β_k is

$$Var(\beta_k)_{min} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2},$$

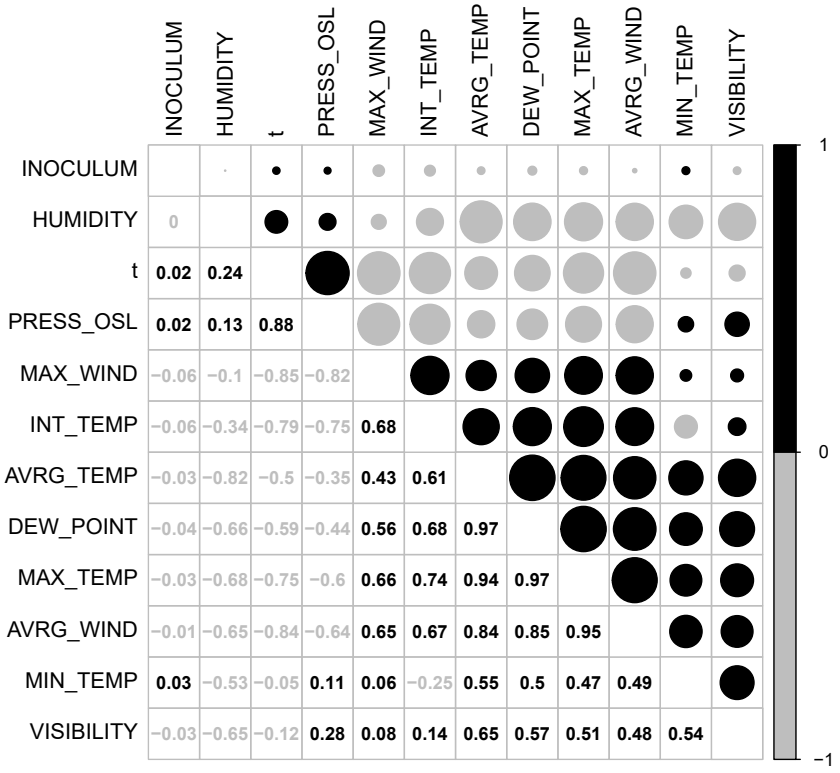


Figure 6.1: Correlation matrix of the predictor variables.

where the subscript *min* denotes that it is the smallest the variance can be. Now let's consider that the predictors are correlated. In the model

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \quad i = 1, \cdots, n.$$

if some of the predictors are correlated with X_k, then the variance of β_k is inflated. This can be shown by the fact that the variance of β_k is

$$Var(\beta_k) = Var(\beta_k)_{min} \frac{1}{1 - R_k^2},$$

where R_k^2 is the R^2 value (it is defined in A.3.1) obtained by regressing the kth predictor on the remaining predictors. Of course, the greater the linear dependence among the predictor X_k and the other predictors, the larger the R_k^2 value. And, as the above formula suggests, the larger the R_k^2 value, the larger the variance of β_k.

Definition 6.1.2 *Given*

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \quad i = 1, \cdots, n,$$

the ***variation inflation factor*** *is what is deemed the variance inflation factor for the kth predictor, that is*

$$VIF_k = \frac{Var(\beta_k)}{Var(\beta_k)_{min}} = \frac{Var(\beta_k)_{min}\frac{1}{1-R_k^2}}{Var(\beta_k)_{min}} = \frac{1}{1-R_k^2},$$

where R_k^2 is the R^2 value obtained by regressing the kth predictor on the remaining predictors.

Note that a variance inflation factor exists for each of the k predictors in a multiple regression model.

The VIF_k values were calculated by considering the linear model

$$X_k = \alpha_0 + \sum_{j \neq k} \alpha_j X_j, \quad i = 1, \cdots, n,$$

where X_k is the variable most correlated in module (the values nearest $\pm$1with with another variable X_j. If the value of VIF_k is infinity, this variable is no longer considered in the regression model and the calculation of the VIF_j relative to the second most correlated variable is carried out, without considering the variable considered previously. The values of VIF_j are:

- $VIF_{AVRG\_TEMP} = \infty$;
- $VIF_{DEW\_POINT} = \infty$;
- $VIF_{AVRG\_WIND} = \infty$;
- $VIF_{PRESS\_OSL} = \infty$;
- $VIF_{MIN\_TEMP} = \infty$;
- $VIF_{MAX\_WIND} = 4.32074$.

Therefore the variables $AVRG\_TEMP$, DEW_POINT, $AVRG\_WIND$, $PRESS\_OSL$ and MIN_TEMP will no longer be considered in any future model.

Finally, $INOCULUM$ is another variable to be not considered, because it obviously explains a percentage of variance of the response variable, already explained by the $HIVE$ variable.

6.2 Linear model: normal assumption

Our variable, $AVRG$, is a ratio that takes value in $\mathbb{R}^2$. More precisely we know that in our data frame $AVRG \in [0, 1.24]$ and Figure 6.2 shows the experiment distribution.

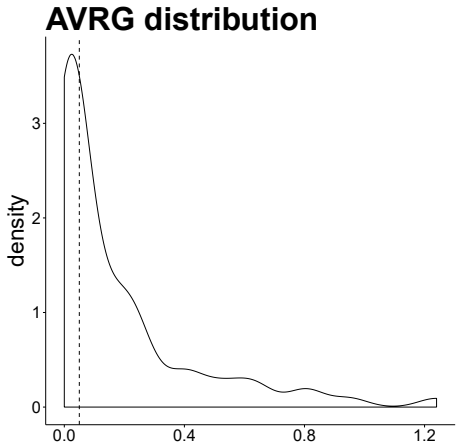


Figure 6.2: This Figure shows the distribution of $AVRG$, the ration between varroa in reproductive stage and larvae. The vertical line represents the median value.

After analyzing the Figure 6.2, to model our response variable: we assume that the transformation of our response variable $log(AVRG)$ is normally distributed and consequently $AVRG$ is log-normal distribuited, with logarithmic link function.

One last consideration is missing. For $AVRG = 0$ the value of the logarithmic link function is not defined. Since at the beginning of the experiment in each hive was inserted a population of varroa definitely active from the reproductive point of view, the null values of $AVRG$ could be linked to the fact that, especially in the first measurements, it was very likely to take a healthy sample, taking only one hundred cells from the whole hive. We then define $AVRG_m := \min_{hi}\{AVRG_{hi} > 0\}$, where $h = 1, \cdots, 9, 11, 12$ indicates the h-th hive and $i = 1, \cdots, 7$ indicates the value $DATE_i$. The new response variable is

$$Z_{hi} = \begin{cases} AVRG_m & \text{if } AVRG_{hi} = 0 \\ \\ AVRG_{hi} & \text{otherwise.} \end{cases}$$

We try to model Z using the variables $HUMIDITY$, INT_TEMP, $DATE$, MAX_TEMP, $HIVE$, $VISIBILITY$ and MAX_WIND. The variables selection is made, as a first step, through the *stepwise procedure* by AIC, defined in A.4. It is know that not always the stepwise produce the real better model then after it, we try to adjust it incorporating what we think are important variables.

The model found after the stepwise selection is

$$log(Z_{hi}) = \beta_0 + \beta_1 DATE_i + \beta_2 INT\_TEMP_i + \epsilon_{hi}, \tag{6.1}$$

and the estimates are shown in Table 6.1. A parameter to evaluate the goodness of the model is the adjusted R-squared defined in A.3.2.

The model (6.1) predicts only seven values, as the Figure 6.3 shows, that correspond to the seven measurements date. But we want to see if there are any

Variable	Estimate	p-value
β_0	-344.8	0.0033
β_1	0.0204	0.0028
β_2	-0.2537253.	0.0143

Table 6.1: Results of (6.1). The residual standard error is 1.091 on 69 degrees of freedom. The adjusted R-squared is 0.5091 and the p-value of the model is $8.144e-12$. AIC is 221.8223.

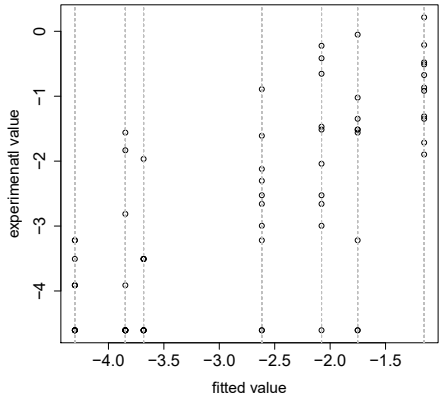


Figure 6.3: This plot shows fitted values for the model (6.1). Each vertical line refers to a single value predicted by the model.

Variable	Estimate	p-value	Variable	Estimate	p-value
β_0	201.9	0.3770	$\beta_{2,1}$	-0.0117	0.3799
$\beta_{1,2}$	-750.1	0.0146	$\beta_{2,2}$	0.0323	0.0184
$\beta_{1,3}$	-537.3	0.0760	$\beta_{2,3}$	0.0199	0.1407
$\beta_{1,4}$	-920.5	0.7574	$\beta_{2,4}$	-0.0064	0.6335
$\beta_{1,5}$	-659.8	0.0306	$\beta_{2,5}$	0.0270	0.0469
$\beta_{1,6}$	-667.2	0.0289	$\beta_{2,6}$	0.0274	0.0439
$\beta_{1,7}$	-607.2	0.0458	$\beta_{2,7}$	0.0239	0.0774
$\beta_{1,8}$	-421.3	0.1652	$\beta_{2,8}$	0.0130	0.3347
$\beta_{1,9}$	-734.6	0.0177	$\beta_{2,9}$	0.0314	0.0231
$\beta_{1,11}$	-936.1	0.0040	$\beta_{2,11}$	0.0433	0.0040
$\beta_{1,12}$	-609.5	0.0550	$\beta_{2,12}$	0.0241	0.0991
			β_3	-0.2659	0.0073

Table 6.2: Results of (6.2). The residual standard error is 1.024 on 49 degrees of freedom. The p-value of the model is $7.154e-07$. AIC is 227.9793

differences between hives. For this reason, we consider all predictors, as done before, and the interaction factor $DATE \times HIVE$ and we perform a variables selection. $DATE \times HIVE$, $DATE$, $HIVE$, INT_TEMP, $HUMIDITY$ and

MAX_WIND are chosen by the algorithm of variables selection, but since observing the p-values we realize that not all these variables are significant. The model we consider is the following one:

$$log(Z_{hi}) = \beta_0 + \beta_{1h} HIVE_h + \beta_{2h} DATE_i \times HIVE_h + \\ + \beta_3 INT\_TEMP_i + \epsilon_{hi}. \quad (6.2)$$

Note that every $HIVE$ or $DATE \times HIVE$ predictors has a different β_1 and β_2 coefficients. This is because $HIVE$ is a factor and not numeric variable, so the model finds a single parameter for each value of $HIVE$. The model is set such that the intercept is $\beta_0 + \beta_{1h}$, with $\beta_{11} = 0$. Table 6.2 shows the output of (6.2) and the Figure 6.4 shows plots to test normality of residuals and heterogeneity. The Figure 6.4(a) shows that the residual distribution is similarly normal and it is confirmed by the Figure 6.4(a) that is a normal Q-Q plot that says that a quantity is normally distributed if the points follow the bisector. We consider this last one model, also if its AIC value is greater than the previous simple linear model. However the model (6.2) has an high adjusted R-squared (defined in A.3.2) value, that is 0.568.

However, heterogeneity remains a problem and in the first predictions some residuals have a predictable behavior, see the points highlighted by a dotted line in Figure 6.4(c). This is caused by the model that is not able to handle $AVRG$ values close to zero, which are very negative Z values. We will try to resolve this issue with a hurdle-at-zero model in Section 6.5 and modeling the variance in the next section.

6.3 Model with structural variance

In this section we model the variance of GLM. We do this to deal with heterogeneity shown in Figure 6.4. We try and compare all structural variance model described in Section 5.1.4. Firstly from Figure 6.4(c) we see that the residual spread increases for higher fitted values, so for this we suppose that time plays a fundamental role for residual variance. Figure 7.3 shows that in addition to time, even belonging to a particular hive can explain a different variance. To compare structural variance models between them we use the AIC.

After comparing all the models, including that of the equation (6.2), the following fixed variance structure model is the one with lower AIC values:

$$log(Z_{hi}) = \beta_0 + \beta_{1h} HIVE_h + \beta_{2h} DATE_i \times HIVE_h + \beta_3 INT\_TEMP_i + \epsilon_{hi} \\ \text{with} \quad Var(\epsilon_{hi}) = \sigma^2 DATE_i. \quad (6.3)$$

Table 6.3 shows the results for the model (6.3). Figure 6.6 shows that problems about heterogeneity are not solved.

6.4 Mixed effect model

Another way to introduce the $HIVE$ variable is like a random effect in a mixed model for intercept and time. We do this also because we hypothesize that observations in the same hive are correlated. We hypothesize that this correlation

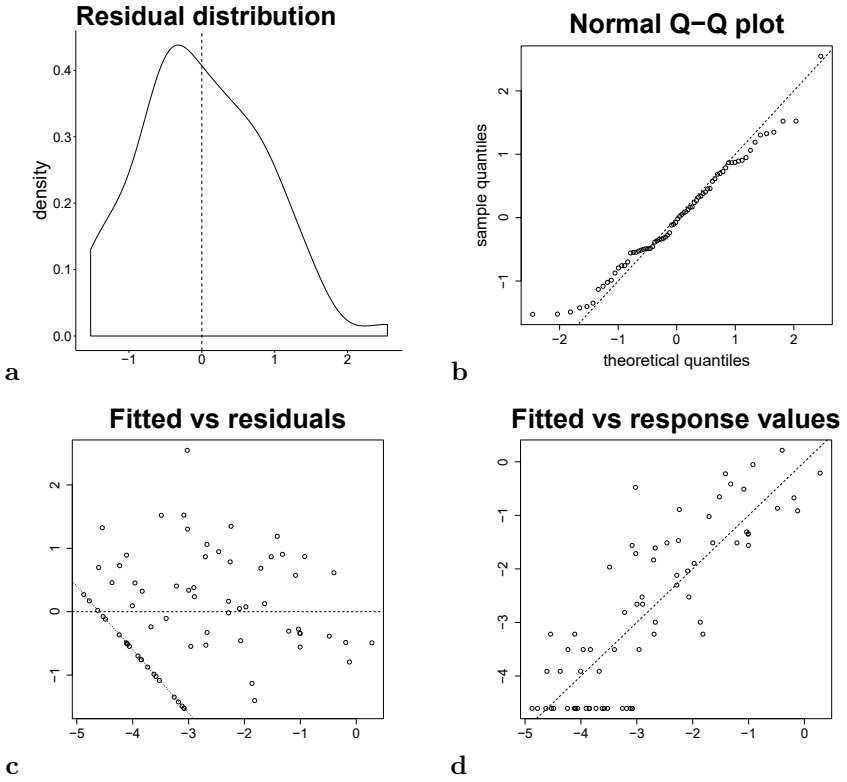


Figure 6.4: Plot analysis for the model (6.2). In this plots the ***X vs Y*** title indicates which quantities are on the axis. 6.4(a) and 6.4(b) show the normal distribution of residuals. Furthermore in 6.4(a) the zero mean is plotted. 6.4(c) shows the trend of residuals according to the predicted values. 6.4(d) shows fitted values compared to experimental values.

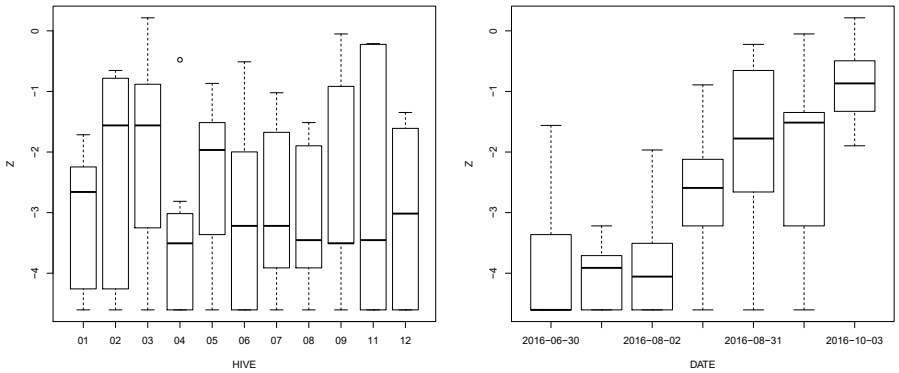


Figure 6.5: Box-plots for the response variable Z, with respect to $HIVE$ and $DATE$.

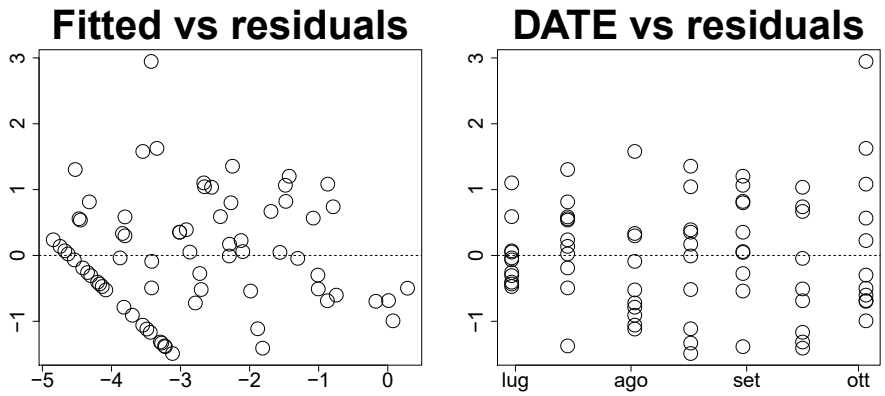


Figure 6.6: Plot results for model (6.3).

Variable	Estimate	p-value	Variable	Estimate	p-value
β_0	2.7731	0.1312	$\beta_{2,1}$	-0.0174	0.1539
$\beta_{1,2}$	-3.5834	0.0033	$\beta_{2,2}$	0.0376	0.0029
$\beta_{1,3}$	-1.1261	0.3364	$\beta_{2,3}$	0.0112	0.3581
$\beta_{1,4}$	-0.5591	0.6319	$\beta_{2,4}$	-0.0136	0.2623
$\beta_{1,5}$	-3.3922	0.0052	$\beta_{2,5}$	0.0347	0.0058
$\beta_{1,6}$	-3.5445	0.0036	$\beta_{2,6}$	0.0287	0.0210
$\beta_{1,7}$	-3.0806	0.0106	$\beta_{2,7}$	0.0255	0.0388
$\beta_{1,8}$	-2.3237	0.0524	$\beta_{2,8}$	0.0107	0.3781
$\beta_{1,9}$	-3.3849	0.0056	$\beta_{2,9}$	0.0366	0.0065
$\beta_{1,11}$	-4.1721	0.0010	$\beta_{2,11}$	0.0445	0.0016
$\beta_{1,12}$	-3.3983	0.0061	$\beta_{2,12}$	0.0302	0.0281
			β_3	-0.2451	0.0017

Table 6.3: Results of (6.3). The residual standard error is 0.118086 on 49 degrees of freedom. AIC is 225.9793

is due in part to the different quantities of varroa introduced in the hives at the beginning of the experiment: a different initial condition generates a different growth trend. We consider all predictors except $HIVE$ because it is introduced like a random effect and after a variables selection we find the following model:

$$log(Z_{hi}) = \beta_0 + v_{0h} + (\beta_1 + v_{1h})DATE_i + \beta_2 INT\_TEMP_i + \epsilon_{hi}. \qquad (6.4)$$

Since $HIVE$ is used as a random effect, for every h-th hive there is a different intercept $\beta_0 + v_{0h}$, where also v_h is assumed normally distributed, and a different coefficient $\beta_1 + v_{1h}$ for $DATE$, where v_{1h} is the random effect.

6.7(a) and 6.7(b) show the normal distribution of residuals (in a normal Q-Q plot a quantity is normally distributed if the points follow the bisector). Furthermore in 6.7(a) the zero mean is plotted. 6.7(c) shows the expected heterogeneity behavior of the residuals, highlighted by the oblique line. 6.7(d) shows good fitted values compared to experimental values. Figures 6.7(e) and 6.7(f) show the residuals distribution on time and per hive. Table 6.4 shows the

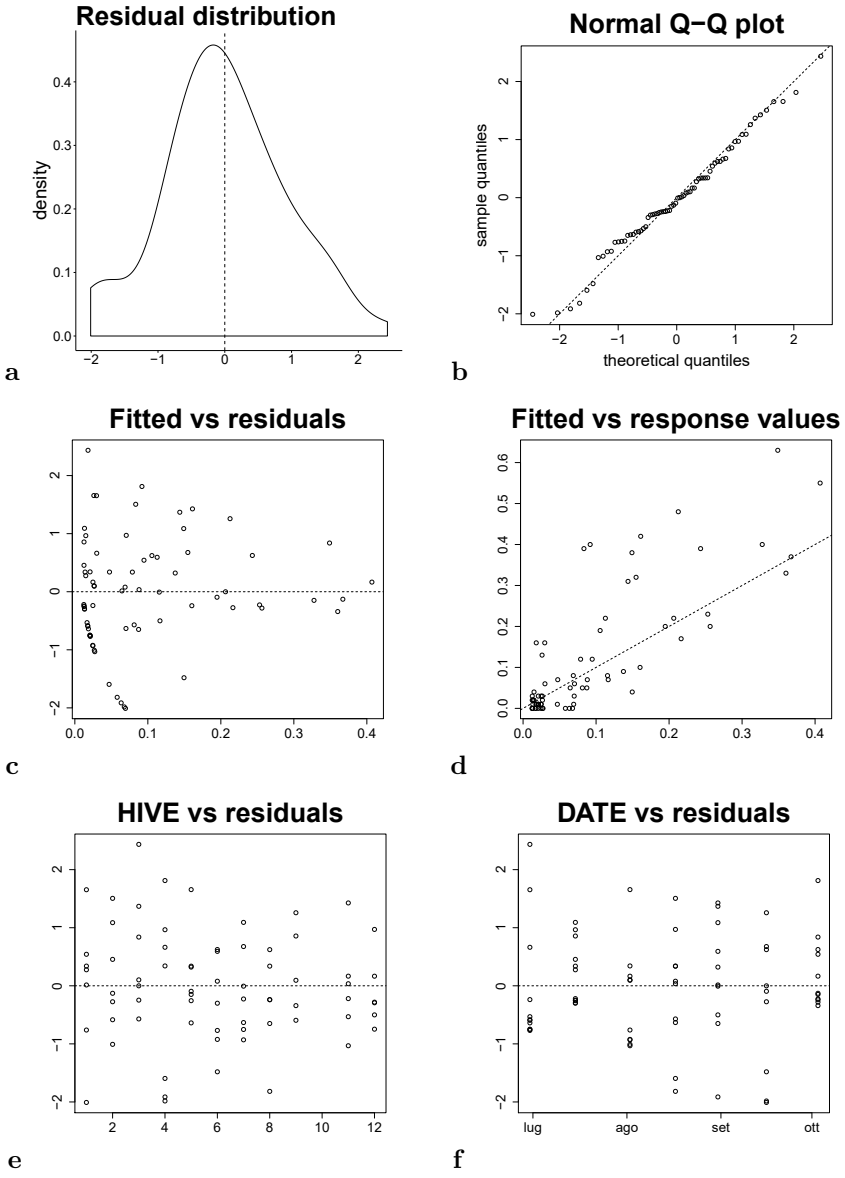


Figure 6.7: Plot analysis for the model (6.4). In this plots the ***X vs Y*** title indicates which quantities are on the axis.

results for (6.4) and we see that it has the clearly lower value of AIC found so far, but Figure 6.7 shows that the problem of heterogeneity is not solved and for this we try with a zero-inflated model.

6.5 NHZ model

In this section we hypothesize that there are two distinct processes that we have to consider to explain the phenomenon related to varroa: one that generates

Fixed effects			Random effects	
Variable	Estimate	p-value	Variable	Std.Dev.
β_0	0.4191	0.8384	v_{0h}	0.6162
β_1	0.0205	0.0046	v_{1h}	0.0115
β_2	-0.2577	0.0053		

Table 6.4: Results of (6.4). AIC is 222.9723.

Fixed effects			Random effects	
Variable	Estimate	p-value	Variable	Std.Dev.
α_0	10.9073	0.0116	α_{0h}	2.1012
α_1	-0.4752	0.0254	α_{1h}	0.0315

Table 6.5: Results of (6.5). AIC is 69.89667.

zeros ($AVRG = 0$, that is the situation without the slightest infestation) and ones ($AVRG > 0$), while another one that explains the proportion of infected cells. To do this, we use an hurdle-at-zero log-normal model. In this way, model Z is equivalent to model B and $(Z|B = 1)$ (Z in the space where $AVRG > 0$), where

$$B_{hi} = \begin{cases} 0 & \text{if } AVRG_{hi} = 0 \\ 1 & \text{otherwise} \end{cases},$$

with $B_{hi} \sim Bin(\pi_{hi})$ is supposed Bernoulli distributed with $\pi = P[AVRG_{hi} = 0]$.

We start from the regression of B, where the Bernoulli probability π_{hi} is modeled with a GLM with the binomial distribution and the logit link function, that is

$$logit(\pi_{hi}) = log\left(\frac{\pi_{hi}}{1 - \pi_{hi}}\right)$$

After testing various models like simple linear model, linear model with interaction factors, models with structural variance and the mixed effects ones, and after variables selections on them, we have two significant candidates:

$$logit(\pi_{hi}) = \alpha_0 + \alpha_{1h} DATE_i \times HIV E_h + \epsilon_{hi}$$

and the mixed effect one

$$logit(\pi_{hi}) = \alpha_0 + u_{0h} + \alpha_1 INT\_TEMP_i + u_{2h} DATE_i + \epsilon_{hi}, \tag{6.5}$$

where there are a random (u_{0h}) and a fixed (α_0) intercept and only a random coefficient (u_{2h}) for $DATE$ and the random effects are linked to $HIV E$ values.

From this two model the best one, i.e. the one with minor AIC, is the model (6.5) and Table 6.5 shows the results.

For the variable $(Z|B = 1)$, that is Z restricted to the subsets of positive values of $AVRG$, i.e. $B = 1$, we have to choose from three groups of model:

- linear models (also with interactions) of which the best one, highest AIC, is

$$log(Z1_{ih}|B_{ih} = 1) = \beta_0 + \beta_1 DATE_i + \beta_2 INT\_TEMP_i + \epsilon_{ih}; \tag{6.6}$$

Model	AIC
(6.6)	172.0292
(6.7)	173.8240
(6.8)	186.5028

Table 6.6: AIC values for model (6.6), (6.7) and (6.8)

- mixed models of which the best one is

$$log(Z1_{ih}|B_{ih}=1)=\beta_0+v_{0h}+\beta_1 DATE_i+\beta_2 INT\_TEMP_i+\epsilon_{ih}, \quad (6.7)$$

  where there is a random intercept v_{0h} dependent from $HIVE$ values;

- models with structural variance of which the best one is

$$log(Z1_{ih}|B_{ih}=1)=\beta_0+\beta_1 DATE_i+\epsilon_{ih} \quad \text{with} \quad Var(\epsilon_{ih})=\sigma_i^2. \quad (6.8)$$

Table 6.6 shows the AIC values for this three best models.

Even if in the model (6.6) $HIVE$ is not used, to predict differently for each hive we can choose it because $HIVE$ is significant for the estimation of the probability π_{hi}. The final zero-inflated model is

$$\begin{aligned} &Z_{hi}=B_{hi}\times exp\{\beta_0+\beta_1 DATE_i+\beta_2 INT\_TEMP_i+\epsilon_{ih}\} \quad \text{where} \\ &B_{hi}\sim Bin(\pi_{hi}) \quad and \\ &\pi_{hi}=\frac{exp(\alpha_0+u_{0h}+\alpha_1 MAX\_TEMP_i+u_{2h}DATE_i+\epsilon_{hi})}{1+exp(\alpha_0+u_{0h}+\alpha_1 MAX\_TEMP_i+u_{2h}DATE_i+\epsilon_{hi})} \end{aligned} \quad (6.9)$$

Table 6.7 shows the results for (6.9). Note that the global AIC value is obtained adding the single AIC values of the normal and Bernoulli regression. In fact from A.4 we know that the definition of AIC is

$$AIC=2p-2log(L).$$

where p is the number of parameters of the model plus one and L is the maximum estimated value for the likelihood function $L(\mu,\sigma^2,\pi;\cdot)$, with μ, σ^2 and π that are the parameters of our distributions. The likelihood function is a function of the parameters μ, σ^2 and π, equal to the density of the observed data. Let $f_{\mu,\sigma^2,\pi}(\cdot)$ be the hurdle distribution of our sample, then the likelihood function is

$$L(\mu,\sigma^2,\pi;z_1,z_2,\cdots)=\prod_i f_{\mu,\sigma^2,\pi}(z_i).$$

Because we want pay attention to the zero values, we have to consider b_i values

$$\prod_i f_{\mu,\sigma^2,\pi}(z_i)=\prod_i f_{\mu,\sigma^2,\pi}(b_i)\times\prod_{b_i=1} f_{\mu,\sigma^2,\pi}(z_i|b_i=1)$$

Normal regression			Bernoulli regression		
AIC		172.0292	AIC		67.1
Fixed effects			Fixed effects		
Variable	Estimate	p-value	Variable	Estimate	p-value
β_0	-319	0.0098	α_0	10.9073	0.0116
β_1	0.0189	0.0085	α_1	-0.4752	0.0254
β_2	-0.2486	0.0276			
			Random effects		
			Variable		Std.Dev.
			u_{0h}		2.1012
			u_{2h}		0.0316

Table 6.7: Results of (6.9). AIC is $239,1292$.

and so the total AIC value of our model is

$$
\begin{aligned}
AIC &= 2p - 2log\left[\prod_i f_{\mu,\sigma^2,\pi}(z_i)\right] = \\
&= 2p - 2log\left[\prod_i f_{\mu,\sigma^2,\pi}(b_i) \times \prod_{b_i=1} f_{\mu,\sigma^2,\pi}(z_i|b_i=1)\right] = \\
&= 2(p_b + p_{z|b=1}) - 2log\left[\prod_i f_{\mu,\sigma^2,\pi}(b_i)\right] - 2log\left[\prod_{b_i=1} f_{\mu,\sigma^2,\pi}(z_i|b_i=1)\right] = \\
&= \left\{2p_b - 2log\left[\prod_i f_{\mu,\sigma^2,\pi}(b_i)\right]\right\} + \left\{2p_{z|b=1} - 2log\left[\prod_{b_i=1} f_{\mu,\sigma^2,\pi}(z_i|b_i=1)\right]\right\} = \\
&= AIC_b + AIC_{z|b=1},
\end{aligned}
$$

where p_b is the number of parameters in the model B and $p_{z|b=1}$ is the number of parameters used to model $(Z|B=1)$.

However the mixed effects model is the best one, i.e. the one with lower AIC.

6.6 Discussion of the chosen model

Finally, the best model found in the previous section is the (6.4) We expected the mixed model as the better model, in fact our experimental data respect the definition of longitudinal data for which mixed models are ideal. The hurdle-at-zero model is not so performing and it may seem strange considering the presence of a large amount of zeros and of over-dispersion (variation greater than predicted by model). A reason to explain this is simply that not always a high presence of zeros means that in the phenomenon under study there is some sub-phenomenon that generates zeros (such as those generated by the limited accuracy of the experiment). Most likely, above all, in experimental data describing a growth over time, we often find many zeros, mainly for small

h ($HIVE$)	$\beta_0 + v_{0h}$	$\beta_1 + v_{1h}$	β_2
1	1.2602	0.0048	-0.2577
2	-0.0070	0.0285	$''$
3	0.0517	0.0274	$''$
4	1.2845	0.0043	$''$
5	0.1220	0.0260	$''$
6	0.4301	0.0203	$''$
7	0.3909	0.0210	$''$
8	0.8155	0.0131	$''$
9	0.01542	0.0280	$''$
11	-0.1312	0.0308	$''$
12	0.3784	0.0213	$''$

Table 6.8: Results of (6.4). AIC is 222.9723. For β_2 the symbol $''$ is used to indicate that this coefficient is constant.

times, but this could simply be caused by a slow or even null growth of the population under investigation.

Using $HIVE$ as a random effect, the interpretation of all the coefficients drastically changes. In fact the model (6.4) gives an intercept for each hive $\beta_0 + v_{0h}$. The same thing is for $DATE$. Table 6.8 show all coefficients for the model (6.4). The intercept and the coefficient of $DATE$ are different for each hive and they comes from the formulas $\beta_{intercept,h} = \beta_0 + v_{0h}$ and $\beta_{DATE,h} = \beta_1 + v_{1h}$.

From Table 6.8 we can get some information. Obviously all $\beta_{DATE,h} > 0$, in fact over time the number of varroa in the cells grows, proof of the fact that the varroa finds in the *apis mellifera* the perfect host to reproduce. To analyze the effect of the time we consider the quantity

$$\bar{\beta}_{DATE} = \frac{1}{11}\sum_h \beta_{DATE,h} \simeq 0.02050,$$

that is the mean of coefficients for $DATE$ and expresses the effect of time on average between all the hives. This means that at with increase of $DATE$ of one unite, that is "after a day", corresponds an increase in the percentage of phoretic varroa compared to adult bees of 2.07%. To calculate this percentage we suppose to have a fixed values for predictors for $h = h_0$ and $i = i_0$ and consequently we obtain a value for Z:

$$Z_0 = Z_{h_0,i_0} = exp\{\beta_0 + v_{0,h_0} + (\beta_1 + v_{1,h_0})DATE_{i_0} + \beta_2 INT\_TEMP_{i_0}\},$$

and we call the linear part inside the exponential

$$\eta_0 = \beta_0 + v_{0,h_0} + (\beta_1 + v_{1,h_0})DATE_{i_0} + \beta_2 INT\_TEMP_{i_0},$$

and so $Z_0 = e^{\eta_0}$. Now we suppose an increase of one day for date and the new value of Z is (remember that we used the mean value of the time coefficients)

$$Z_1 = e^{\eta_0 + \bar{\beta}_{DATE}\times 1} = e^{\eta_0 + 0.0205} = 1.0207 e^{\eta_0} = (1 + 2.07\%)Z_0.$$

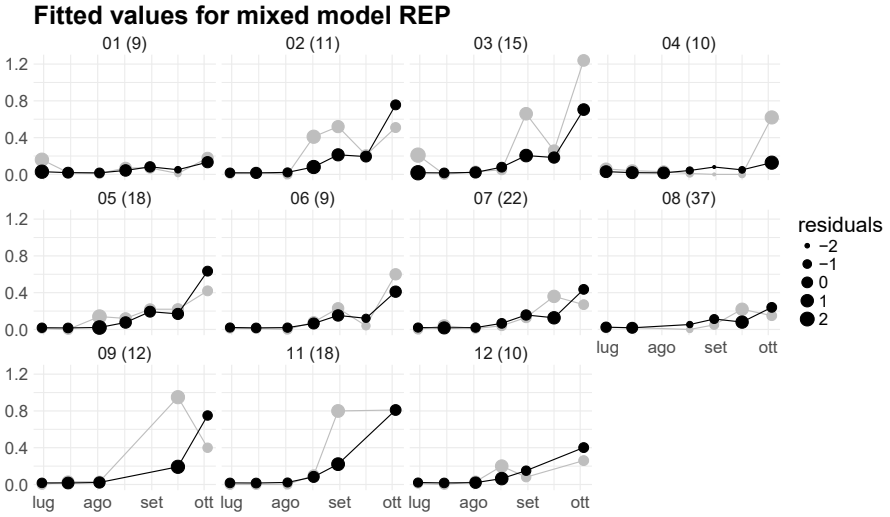


Figure 6.8: Fitted values compared to experimental values for the model (6.4). Each plot is referred to the h-th hive and in brackets the initial inoculum quantity is indicated. The points dimension in the measurement dates are proportionally to the value of residual in that date.

$\beta_2 < 0$ means that in the reproductive season (spring and summer) a high daily temperature range disinhibit the growth of varroa. Quantitatively this means that if the range between maximum and minimum temperature increases by $1°C$, than the percentage of varroa decreases of 22.7%.

Figure 6.8 shows fitted values trend respect experimental values for hive.

Chapter 7

Model PHO: relationship between varroa in phoretic stage and adult bees

In this section, relationship between varroa in phoretic stage and adult bees is modeled. For each hive in each time we have a number of taken bees from the hive and the number of varroa attached to them (parasitic activity). The ratio between phoretic varroa and adult bees is expressed by the variable $PERC\_PHO\_VARR$ of our data frame and this variable is the one that we want to model in this section. In this chapter we follow steps similare to those used in Chapter 6 for the variable $AVRG$. For this we omit the explanations of various passages already explained in Chapter 6.

7.1 Linear model: normal assumption

Our variable, $PERC\_PHO\_VARR$, is a ratio that takes value in $\mathbb{R}^2$. More precisely we know that in our data frame $PERC\_PHO\_VARR \in [0, 0.7343]$ and Figure 7.1 shows the experiment distribution.

After analyzing the Figure 7.1, to model our response variable we use the same hypothesis for the response variable $AVRG$. Also $PERC\_PHO\_VARR$ takes null values where the logarithmic function is not defined. We then define $PERC\_PHO\_VARR_m := \min_{hi}\{PERC\_PHO\_VARR_{hi} > 0\}$, where $h = 1, \cdots, 9, 11, 12$ indicates the h-th hive and $i = 1, \cdots, 7$ indicates the value $DATE_i$. The new response variable is

$$W_{hi} = \begin{cases} PERC\_PHO\_VARR_m & \text{if } PERC\_PHO\_VARR_{hi} = 0 \\ \\ PERC\_PHO\_VARR_{hi} & \text{otherwise.} \end{cases}$$

To model W we use directly a linear model with $VISIBILITY$, $DATE$, INT_TEMP, MAX_TEMP, $HIVE$, $HUMIDITY$ and MAX_WIND, plus an interaction factor between $DATE$ and $HIVE$. The variables selection is made, as a first step, through the *stepwise procedure* by AIC, defined in A.4. $DATE \times HIVE$, $DATE$, $HIVE$, MAX_TEMP and $VISIBILITY$ are chosen

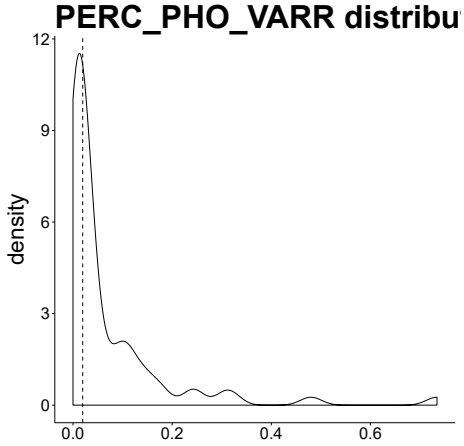


Figure 7.1: This Figure shows the distribution of $PERC\_PHO\_VARR$, the ratio between the number of varroa in phoretic stage and the number of adult bees. The vertical line represents the median value.

by the algorithm of variables selection and all have significant p-values. The model we consider is the following one:

$$W_{hi} = exp\{\beta_0 + \beta_{1h} HIVE_h + \beta_{2h} DATE_i \times HIVE_h + \beta_3 DATE_i + \\ + \beta_4 MAX\_TEMP_i + \beta_5 VISIBILITY_i + \epsilon_{hi}\}. \quad (7.1)$$

Note that every $HIVE$ or $DATE \times HIVE$ predictor has a different β_1 and β_2 coefficients. This is because $HIVE$ is a factor and not numeric variable, so the model finds a single parameter for each value of $HIVE$. The model is set such that the intercept is $\beta_0 + \beta_{1h}$, with $\beta_{11} = 0$. Table 7.1 shows the output of (7.1). The Figure 7.2 shows graphical tests for the residuals. Figure 7.2(a) shows that the density distribution of the residual follows the bell trend of the normal distribution even if the Figure 7.2(b) shows that residuals are not concentrated on the mean value (in a normal Q-Q plot a quantity is perfectly normally distributed if the points follow the bisector).

Having inserted an interaction factor, we have solved the problem of predicting only a different number of values equal to the number of experimental measurements. Unfortunately heterogeneity is the problem of the model (7.1) (Figure 7.2(c)). 7.2(a) and 7.2(b) show the normal distribution of residuals. Furthermore in 7.2(a) the zero mean is plotted. 7.2(c) puts the fitted values on the x axis and the residuals on the y axis. 7.2(d) shows good fitted values compared to experimental values. Figures 7.2(e) and 7.2(f) show the residuals distribution on time and per hive. As done in the previous chapter, we try to solve this problem by trying new models, aware of the fact that the results obtained are already acceptable.

To compare the results obtained for Z with those for now obtained for W, we can not use the AIC value, as obviously the response variable changes. Surely it is worth noting the very high value of the adjusted R^2 of the model (7.1): 0.8146. This value tells us that already with the first attempt we managed to explain more than 80% of the deviance of our experimental data. Another little problem

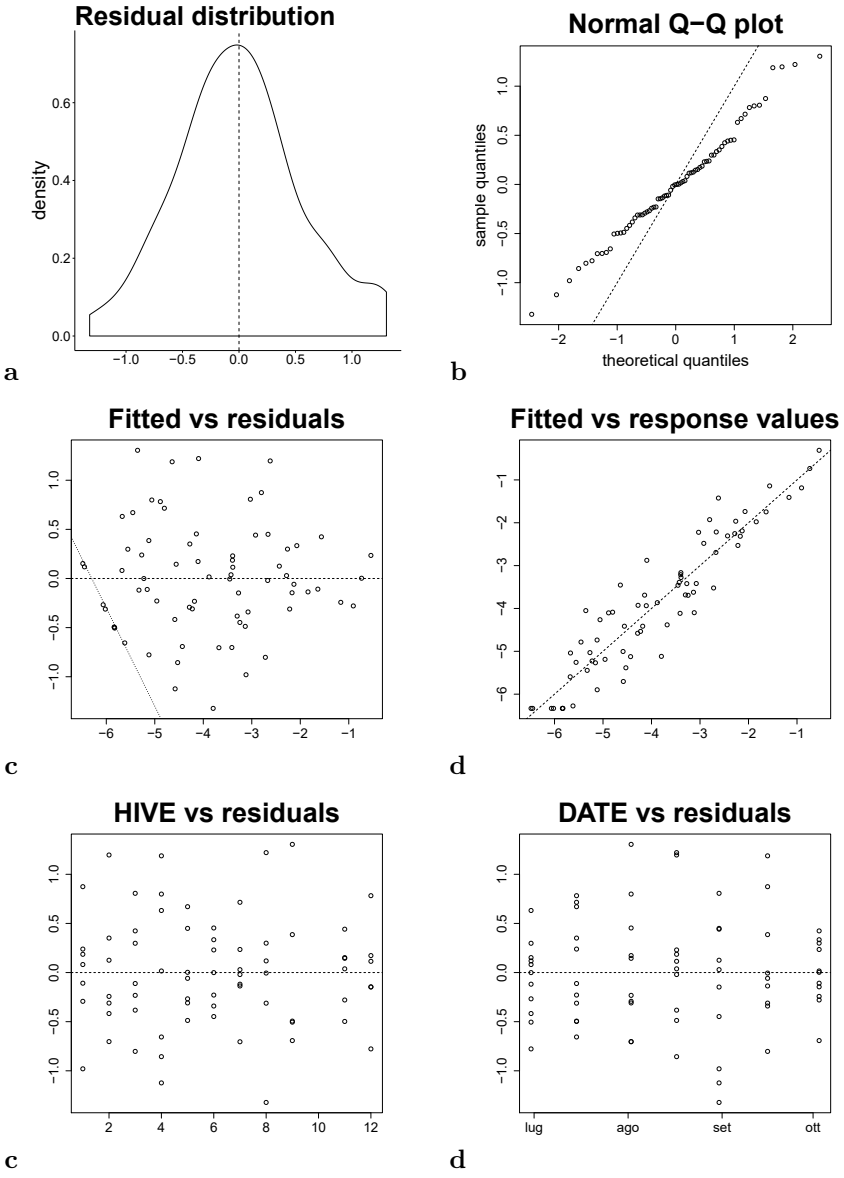


Figure 7.2: Plot analysis for the model (7.1). In this plots the ***X vs Y*** title indicates which quantities are on the axis.

of this model, shown by the Figure 7.2(b), is that the normal distribution of the residuals is a fat-tailed distribution: in this case, this property, already seen in a minor way for the models referred to Z, is shown by the kurtosis phenomenon and is due to the high variance of the experimental data. However from the Figure 7.2(d) we notice that our model estimates very well the experimental values.

Figure 7.3 shows that maybe there is a strong variability of the variance linked to the different hives. For this a structural variance model based on

Variable	Estimate	p-value	Variable	Estimate	p-value
β_0	-592.5	0.0003	$\beta_{2,2}$	-0.0065	0.5751
β_3	0.0474	0.0002	$\beta_{2,3}$	-0.0005	0.9655
$\beta_{1,2}$	111.4	0.5725	$\beta_{2,4}$	-0.0237	0.0452
$\beta_{1,3}$	8.609	0.9651	$\beta_{2,5}$	0.0136	0.2441
$\beta_{1,4}$	401.9	0.0458	$\beta_{2,6}$	-0.0093	0.4222
$\beta_{1,5}$	-230.8	0.2447	$\beta_{2,7}$	0.0078	0.4996
$\beta_{1,6}$	158.7	0.4221	$\beta_{2,8}$	0.0016	0.8940
$\beta_{1,7}$	-132.6	0.5019	$\beta_{2,9}$	-0.0278	0.0207
$\beta_{1,8}$	-27.20	0.8912	$\beta_{2,11}$	0.0162	0.1861
$\beta_{1,9}$	472.4	0.0210	$\beta_{2,12}$	-0.0115	0.3474
$\beta_{1,11}$	-275.8	0.1860	β_4	-0.1463	0.0414
$\beta_{1,12}$	195.1	0.3473	β_5	0.0998	0.0391

Table 7.1: Results of (7.1). The residual standard error is 0.6769 on 48 degrees of freedom. The adjusted R-squared is 0.8146 (defined in A.3.2) and the p-value of the model is $1.395e-14$. AIC is 168.9445

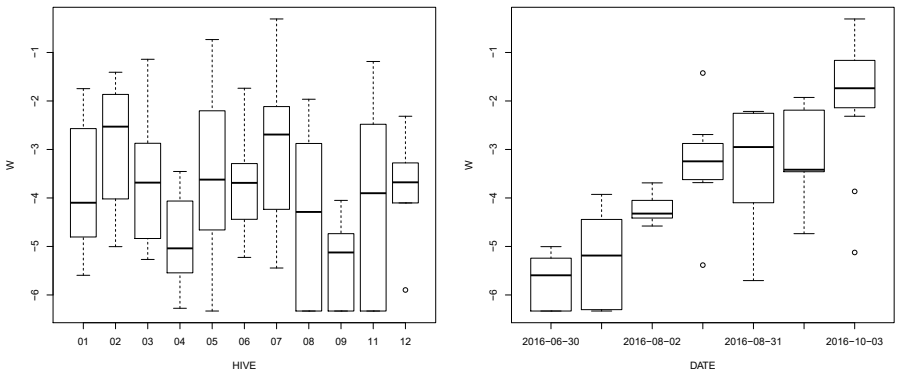


Figure 7.3: Box-plots for the response variable W, respect $HIVE$ and $DATE$.

$HIVE$ might seem like the solution, but by trying out various models and comparing them we find that in this case, as for Z, a structured variance model is useless. Now we pass directly to a mixed model.

7.2 Mixed effect model

Another way to introduce the $HIVE$ variable is like a random effect in a mixed model for intercept and time. We consider all predictors except $HIVE$ because it is introduced like a random effect. The choice of random on intercept is linked to the different quantities of varroa introduced in the hives at the beginning of the experiment. We perform a variables selection and we find the following

Fixed effects			Random effects	
Variable	Estimate	p-value	Variable	Std.Dev.
β_0	-3.3573	0.1015	u_{0h}	0.3831
β_1	0.0308	<3e-06	u_{1h}	0.0104
β_2	-0.1509	0.0319		
β_3	0.1080	0.0229		

Table 7.2: Results of (7.2). AIC is 186.4597.

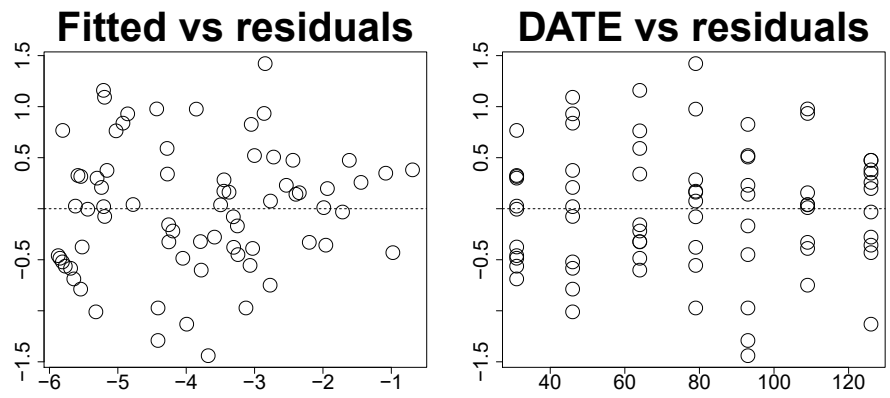


Figure 7.4: The Figure shows result for model (7.2).

model:

$$W_{hi} = exp\{\beta_0 + u_{0h} + (\beta_1 + u_{1h})DATE_i + \\ + \beta_2 MAX\_TEMP_i + \beta_3 VISIBILITY + \epsilon_{hi}\}. \quad (7.2)$$

Since $HIVE$ is treated as random effect, for every h-th hive there is a different intercept $\beta_0 + u_{0h}$, where also u_h is assumed normally distributed, and a different coefficient $\beta_1 + u_{1h}$ for $DATE$.

Table 7.2 shows the results for (7.2) and we see that it has the clearly lower value of AIC found so far, but Figure 7.4 shows that the problem of heterogeneity is not solved and for this we try with a zero-inflated model.

This time, however, we obtain a higher AIC value than the linear model.

7.3 NHZ model

We consider in this section that considering two distinct processes can help us to explain the phenomenon related to varroa: one that generates zeros ($PERC\_PHO\_VARR = 0$) and ones ($PERC\_PHO\_VARR > 0$), while another one that explains the proportion of phoretic varroa and bees. To do this, we use an hurdle-at-zero log-normal model and we consider the following new variable:

$$C_{hi} = \begin{cases} 0 & \text{if} PERC\_PHO\_VARR_{hi} = 0 \\ 1 & \text{otherwise} \end{cases},$$

Fixed effects			Random effects	
Variable	Estimate	p-value	Variable	Std.Dev.
α_0	164.65	<2e-16	u_{0h}	908.76
α_1	-7.0055	<2e-16	u_{1h}	19.58

Table 7.3: Results of (7.3). AIC is 26.27603.

where $C \sim Bin(\bar{\pi})$ is supposed Bernoulli distributed with $\bar{\pi}$ that is the probability of finding zero varroa.

We start from the regression of C, or better, we use a GLM with a binomial distribution and the logit link function. After considering various models and the variables selections on them, we have two significant candidates:

$$\bar{\pi}_{hi} = \frac{exp\{\alpha_0 + \alpha_{1h} DATE_i \times HIVE_h\}}{1 + exp\{\alpha_0 + \alpha_{1h} DATE_i \times HIVE_h\}}$$

and the mixed effect one

$$\bar{\pi}_{hi} = \frac{exp\{\alpha_0 + u_{0h} + \alpha_1 INT\_TEMP_i + u_{2h} DATE_i + \epsilon_{hi}\}}{1 + exp\{\alpha_0 + u_{0h} + \alpha_1 INT\_TEMP_i + u_{2h} DATE_i + \epsilon_{hi}\}}, \tag{7.3}$$

where there are a random (u_{0h}) and a fixed (α_0) intercept and only a random coefficient (u_{2h}) for $DATE$ and the random effects are linked to $HIVE$ values.

From this two model the best one, i.e. the one with minor AIC, is the model (7.3) and in Table 7.3 we show the results.

For the variable $(W|C = 1)$, that is W restricted to the subsets of positive values of $PERC\_PHO\_VARR$, i.e. $C = 1$, we have to choose from three groups of model:

- linear models (also with interactions) of which the best one, highest AIC, is
$$(W|C = 1)_{ih} = exp\{\beta_0 + \beta_{1h} DATE_i \times HIVE_h + \epsilon_{ih}\}; \tag{7.4}$$
- mixed models of which the best one is
$$(W|C = 1)_{ih} = exp\{\beta_0 + v_{0h} + (\beta_1 + v_{1h}) DATE_i + \epsilon_{ih}\}, \tag{7.5}$$
where there is a random intercept v_{0h} and a random coefficient v_{1h} for $DATE$ dependent from $HIVE$ values;
- models with structural variance of which the best one is
$$(W|C = 1)_{ih} = exp\{\beta_0 + \beta_{1h} DATE_i \times HIVE_h + \epsilon_{ih}\} \quad \text{with} \quad Var(\epsilon_{ih}) = \sigma^2 DATE_i. \tag{7.6}$$

Table 7.4 shows the AIC values for this three best models.
We choose the model (7.4) and the final zero-inflated model is

$$\begin{aligned} W_{hi} &= C_{hi} \times exp\{\beta_0 + \beta_{1h} DATE_i \times HIVE_h + \epsilon_{ih}\} \quad \text{where} \\ C_{hi} &= \frac{exp\{\alpha_0 + u_{0h} + \alpha_1 INT\_TEMP_i + u_{2h} DATE_i + \epsilon_{hi}\}}{1 - exp\{\alpha_0 + u_{0h} + \alpha_1 INT\_TEMP_i + u_{2h} DATE_i + \epsilon_{hi}\}}. \end{aligned} \tag{7.7}$$

Model	AIC
(7.4)	165.7393
(7.5)	169.2767
(7.6)	255.0070

Table 7.4: AIC values for model (7.4), (7.5) and (7.6)

Normal regression (AIC: 165.7393)					
Variable	Estimate	p-value	Variable	Estimate	p-value
β_0	-597	<2e-15	$\beta_{1,6}$	0.0348	<2e-15
$\beta_{1,1}$	0.0348	<2e-15	$\beta_{1,7}$	0.0349	<2e-15
$\beta_{1,2}$	0.0349	<2e-15	$\beta_{1,8}$	0.0348	<2e-15
$\beta_{1,3}$	0.0348	<2e-15	$\beta_{1,9}$	0.0347	<2e-15
$\beta_{1,4}$	0.0348	<2e-15	$\beta_{1,11}$	0.0349	<2e-15
$\beta_{1,5}$	0.0349	<2e-15	$\beta_{1,12}$	0.0349	<2e-15
Bernoulli regression (AIC: 26.27603)					
Fixed effects			Random effects		
Variable	Estimate	p-value	Variable		Std.Dev.
α_0	164.6524	<2e-16	u_{0h}		908.76
α_1	-7.0055	<2e-16	u_{2h}		19.58

Table 7.5: Results of (7.7). AIC is 192.0153.

Table 7.5 shows the results for (7.7). Note that the global AIC value is obtained adding the single AIC values of the normal and Bernoulli regression.

However the linear model is the best one, i.e. the one with lower AIC.

7.4 Discussion of the chosen model

In this case the best model is the simple linear model with interactions, i.e. the model (7.1). Adding an interaction term to a model drastically changes the interpretation of all the coefficients. For example, if there were no interaction term, β_3 would be interpreted as the unique effect of $DATE$ on W. But the interaction means that the effect of $DATE$ on W is different for different values of $HIVE$. So the unique effect of $DATE$ on W is not limited to β_3 but also depends on the values of β_{2h} and $HIVE$. The unique effect of $DATE$ is represented by everything that is multiplied by $DATE$ in the model: $\beta_3+\beta_{2h}HIVE_h$. Instead for $HIVE$ we know that in the case in which we have a particular hive, for example the H-hive ($HIVE_h = H$) the model (7.1) becomes

$$Z_{hi} = exp\{(\beta_0 + \beta_{1H}) + (\beta_3 + \beta_{2H})DATE_i + \\ + \beta_4 MAX\_TEMP_i + \beta_5 VISIBILITY_i + \epsilon_{hi}\},$$

where $\beta_{11} = \beta_{21} = 0$, that is that in the case of the first hive the hive and interaction coefficients are null. This is shown in the Table 7.1 where there are not $\beta 11$ and β_{21}.

Now we consider Table 7.1 and we calculate the correlation between inital inoculum per hive, β_{1h} and β_{2h}:

- $Corr(INOCULUM_h, \beta_{1h}) \simeq -0.17$ demonstrates a low correlation between initial inoculum and varroa growth which in any case is negative and therefore means that at high initial inoculum values correspond to lower coefficients and therefore lower intercept of the model. But a priori we hypothesized that this correlations was very high, testifying that the initial inoculum influences the varroa growth through the factor variable $HIVE$. This fact, in contrast with our hypothesis, that a greater intercept correspond to a minor inoculum is most likely due to the fact that initial quantities of varroa in the phoretic phase per hive were too smalls compared to the final mite population and therefore not significants;

- $Corr(INOCULUM_h, \beta_{2h}) \simeq 0.17$ is ever small compared to expectations but in this case the correlation is positive and it means that the effect of time on the varroa growth it is larger in hives with larger initial inoculations.

From Table 7.1 we can get more information. Obviously $\beta_3 > 0$, meaning that the presence of varroa grows, proving the fact that the varroa finds in the *apis mellifera* the perfect host to reproduce.

To analyse the effect of the time we consider the quantity

$$\bar{\beta}_2 = \frac{1}{11} \sum_h \beta_{2h} \simeq -0.00365,$$

that is the mean of interaction coefficients between $HIVE$ and $DATE$, considering that $\beta_{21} = 0$. In this way, the coefficient that expresses the effect of time is on average between all the hives $\bar{\beta}_{DATE} = \bar{\beta}_2 + \beta_3 \simeq 0.0311$. This means that at an increase of $DATE$ of one unite, that is "after a day", corresponds a increase in the percentage of phoretic varroa compared to adult bees of 3.16%. This is very interesting and gives the possibility of a numerical comparison with the population dynamics approach done in the Chapter 4. From literature, specially from [16], we know that during the bee season, i.e. spring and summer, the mite population doubles every month. This means that if we hypothesize a simple model of exponential growth for varroa like $\dot{V} = rV$, where V is the generical number of varroa in the hive, we find that $r = 30^{-1} log 2 = 0.023$, that corresponds to a daily percentage rate of $2,34\%$. Even if this daily growth rate is not comparable with our rate found with a statistical model, the fact that they are so similar is a proof of the coherence of our model with respect to the rest of literature.

$\beta_4 < 0$ means that in the reproductive season (spring and summer) high temperature peaks disinhibit the growth of varroa. Assuming to keep other predictors fixed (of course we do not consider the correlation between the climatic predictors that minimally remains even after the multicollinearity analysis), at an increase of $1°C$ of the maximum temperature corresponds a decrease in the percentage of phoretic varroa compared to adult bees of 13.61%.

Instead $\beta_5 \simeq 0.1$ means that at an increase of $1m$ of the visibility corresponds an increase in the percentage of varroa of 10.5%. However, this increase is to be considered overestimated, as for the decrease generated by the maximum

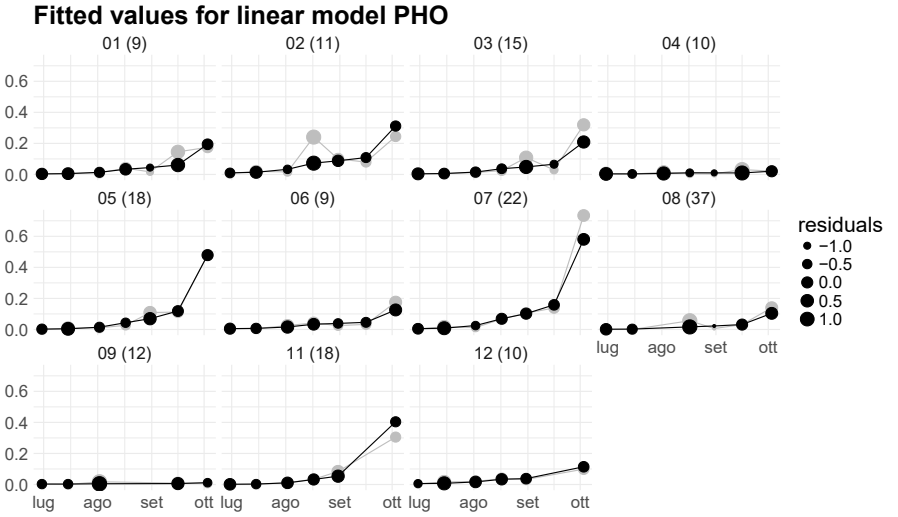


Figure 7.5: Fitted values compared to experimental values for the model (7.1). Each plot is referred to the h-th hive and in brackets the initial inoculum quantity is indicated. The points dimension in the measurement dates are proportionally to the value of residual in that date.

temperature. Infact this two predictors have a positive correlation of about 50% and this means that the effect of the increase in the percentage of varroa due to the increase in visibility is attenuated by the correlated increase in temperature which negatively affects the growth of the varroa.

Figure 7.5 shows fitted values trend respec experimental values for hive. The results are satisfactory and the model estimates very well the different growth in each hive. The only two cases in which experimental data deviate from those modeled are in the second hive, where, however, there is probably an anomaly in experimental measurements, and in the seventh one, where the model can not estimate the final peak. The anomalies are referred to human errors or to outliers.

Chapter 8

Models compared and conclusions

In this Thesis two quantities about the same experiment are been modeled: the percentage of varroa in growth phase that infest the operculated cells of *apis mellifera* (the number of operculated cells is been assumed equal to the number of larvae in the hive) and the percentage of adult varroa in parasitic activity respect the number of workers bees (and drones). This was done with two approaches: population dynamics and statistical analysis.

In the first case four population have been modeled. From two-to-two ratio between them we obtain the two quantities under exam. The differential equation system founded is the following one:

$$\begin{cases} \dot{L} = b\frac{B^2}{k^2+B^2} - cL \\ \dot{B} = cL - mB - \mu P \\ \dot{R} = raP - gcRL \\ \dot{P} = gcRL - nP - ePB - aP, \end{cases} \tag{8.1}$$

where L is the number of larvae, B is the number of adult bees, R is the number of varroa in reproductive phase and P is the number of varroa in phoretic phase. All the parameters in this model have been taken from the literature, except r, g and a, that are respectively the varroa growth rate, the varroa transition rate from reproductive to phoretic phase and the varroa transition rate from phoretic to reproductive phase. Therefore the first result of this thesis is the estimation of this three parameters (Table 4.6). The model (8.1) describes the varroa growth in hives under natural condition. In fact in the experiment held in Ciriè, the oxalic acid treatment was done only after the experimental measurements. This parameters estimate was done on a hypothetical hive, called "mean hive", that is the mean of all data values available for each hive.

In the case of statistical analysis two new considerations have been introduced. The first one is that the climatic conditions can be significant to understand the varroa growth. To analyze this, the climatic data about a locality closed to Ciriè were taken. This location is so close to the experiment location to assume that climatic data are valid for the our model. Subsequently this data have been introduced in statistical modeling to test their meaningfulness.

The second consideration is the hypothesis that also the hive play an important role in the parasite growth. This hypothesis is born mainly from the choice to introduce different quantities of adult varroa in each hive (we have different values for inoculum). Therefore we are interested in what quantity of phoretic varroa at the beginning of the reproductive season of the bees can lead to a level of infestation to be considered dangerous for the hive. In addition we have not introduced the inoculum values in the model, but we have considered the factor hive variable. In this way if this predictor result significant, comparing the results with inoculum values we can understand if there are other unknown properties of the hives that characterize the varroa growth.

After a primary analysis and a management of data, we have found two models that describe our quantities. This models were considered the best using the AIC value and biological consideration about the phenomenon as comparison criterion. The models are as follows:

$$Z_{hi} = exp\{\beta_0 + v_{0h} + (\beta_1 + v_{1h})DATE_i + \beta_2 INT\_TEMP_i + \epsilon_{hi}\} \quad \text{and}$$

$$\begin{aligned} W_{hi} = exp\{&\beta_0 + \beta_{1h} HIVE_h + \beta_{2h} DATE_i \times HIVE_h + \beta_3 DATE_i + \\ &+ \beta_4 MAX\_TEMP_i + \beta_5 VISIBILITY_i + \epsilon_{hi}\}, \end{aligned} \tag{8.2}$$

where

$$Z_{hi} = \begin{cases} AVRG_m & \text{if } AVRG_{hi} = 0 \\ \\ AVRG_{hi} & \text{otherwise} \end{cases}$$

is the percentage of varroa in reproductive phase after the assumption that it can not assume the zero value and

$$W_{hi} = \begin{cases} PERC\_PHO\_VARR_m & \text{if } PERC\_PHO\_VARR_{hi} = 0 \\ \\ PERC\_PHO\_VARR_{hi} & \text{otherwise.} \end{cases}$$

is the percentage of varroa in phoretic phase after the same assumption of Z. The coefficients of $HIVE$ have h has subscript because $HIVE$ is a factor predictor and for this each $HIVE_h$ has a different coefficient. Instead other coefficients with h has subscript are random effects that take values based on the hives.

In both models described in (8.2), $HIVE$ and at least one represents climatic factors are significant. For this reason we expect, even before comparing them, that the statistical model is more accurate than the populations dynamic analysis, simply because it considers more factors. Obviously, introducing the effect of the climate into a system of differential equations like the one described in (8.1) would complicate the model too much, i.e. the model contains too many parameters that could also generate overfitting.

8.1 Results

Figure 8.1 and 8.3 show the results of all this thesis. Three curves (experimental data and statistical and differential fitted values) are shown for each of the eleven hives and also for the average hive. The "virtual" mean hive is the hive whose

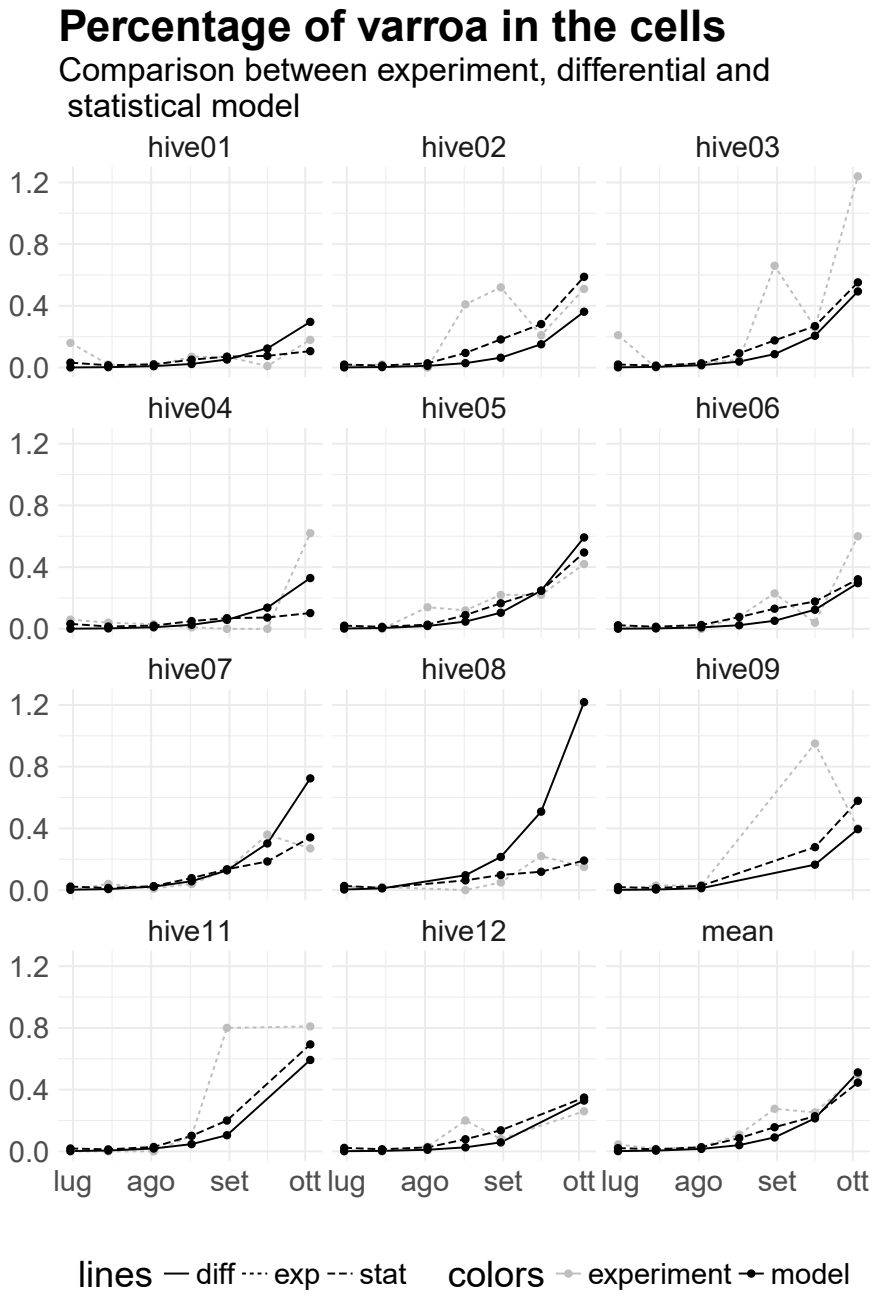


Figure 8.1: This figure shows the results of modeling the percentage of varroa in the cells. The gray dashed line represents the experimental data, the black dashed lines the fitted values by the statistical model (8.2) and the black simple line the fitted values by the differential system (8.1). These three curves are shown for each of the eleven hives and more the average hive is shown.

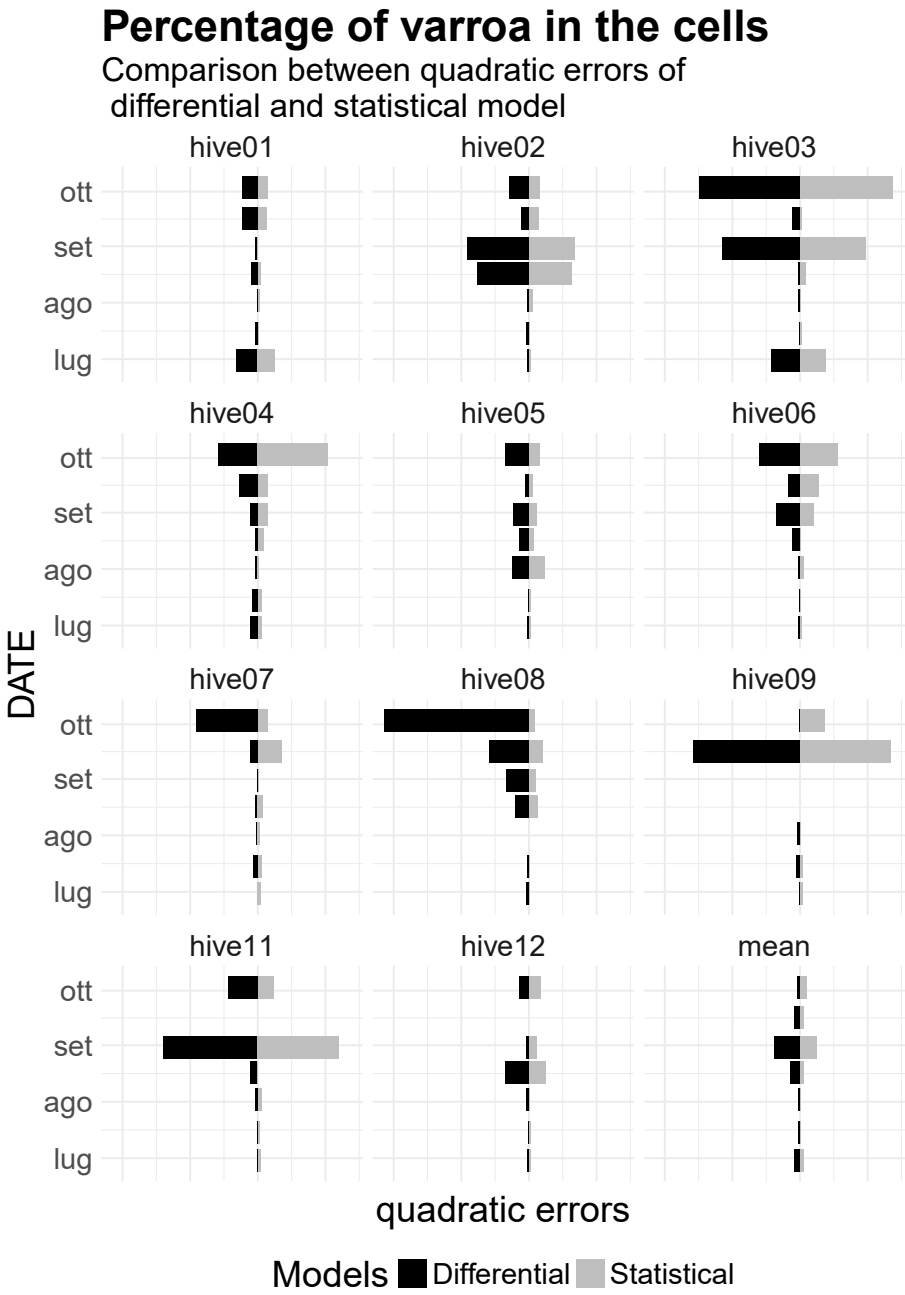


Figure 8.2: This figure shows the quadratic errors in each hive and in each time about modeling the percentage of varroa in the cells for the differential and statistical models. On the left the black bars represents errors for the differential system (8.1) while, on the right, the grey bars represents errors for the statistical model (8.2).

characteristic properties (inoculum and response variable) are the averages of all characteristics of the other hives. This is the hive that we have used for the estimation of parameters in the model (8.1). But in this way the statistical and differential model are not comparable. For this we use statistical models (trained with the real hives) to predict the values related to the mean hive. And in the same way we use the differential model to generate values concerning the real hives: in other words we use the same model but with different initial conditions (the inoculum quantity). From a first observation that the best is the statistical model is already clear.

Figure 8.1 shows the trend of the fitted values with statistical and differential approaches respect the real experimental data about the percentage of varroa in the operculated cells. We remember that curves relating the differential system is obtained considering the ratio in each time between R and L from the system (8.1), i.e. number of varroa in the cells on the number of larvae, that corresponds to the number of operculated cells. Figure 8.2 shows per each hive the trend of the quadratic errors. Regarding the second, the third and the ninth hives, we can see the very irregular trend of the experimental data far away from expectations that varroa trend is a non-decreasing curves in normal conditions. This violation of expectations is most likely caused by the small sample size used for measurements or other factors unknown to us. We also use these hives for the validation of our models, while taking this consideration into account, and we refer to them by calling them "HFE hives"(Hives Far from Expectations).

Considering all hives (also the mean hive) we see that the models fit data very well except for the eleventh hive, where they both fail, for the eighth hive, where the differential model really overestimates the data, for the fourth hive, where the statistical model underestimates the data and for the HFE hives. Particularly we can see how the statistical model well done for a hive not used for training. Although this fact is not reliable because we are talking about a "virtual" beehive dependent on buildings from hives used for training. Obviously having more beehive available, this validation should be done on a small group of real hives. The same goes for the adaptation of the differential model to the other eleven hives. But from Figure 8.2 we can conclude by saying that the statistical model works better than the differential one. To confirm this, Figure 8.5 shows for each time the average of the quadratic errors for each hive and in this way that the statistical approach is the best is clear. A fundamental cause of this is the fact that model (8.1) is constrained to estimate both experimental quantities at the same time. This because we have create a general model that describe completely the phenomenon, considering interactions between larvae, bees, phoretic varroa and varroa in the reproductive phase. This is not a problem for the statistical analysis.

Regarding the modeling of the percentage of varroa in the phoretic phase, from Figure 8.3 we see that the second and third hives can considerated HFE hives. We remember that in Figure 8.3 curves relating the differential system is obtained considering the ratio in each time between P and B from the system (8.1), i.e. number of phoretic varroa on the number of bees. As before, Figure 8.4 shows per each hive the trend of the quadratic errors. Figure 8.4 we see that for deterministic model the worst hives are the second, the eleventh and the eighth hive. By 8.3 we add also the fourth, fifth and seventh hives between the ones that not are fitted very well. The statistical model doesn't work well for the first, the second, the third and the eleventh hives. In this way the

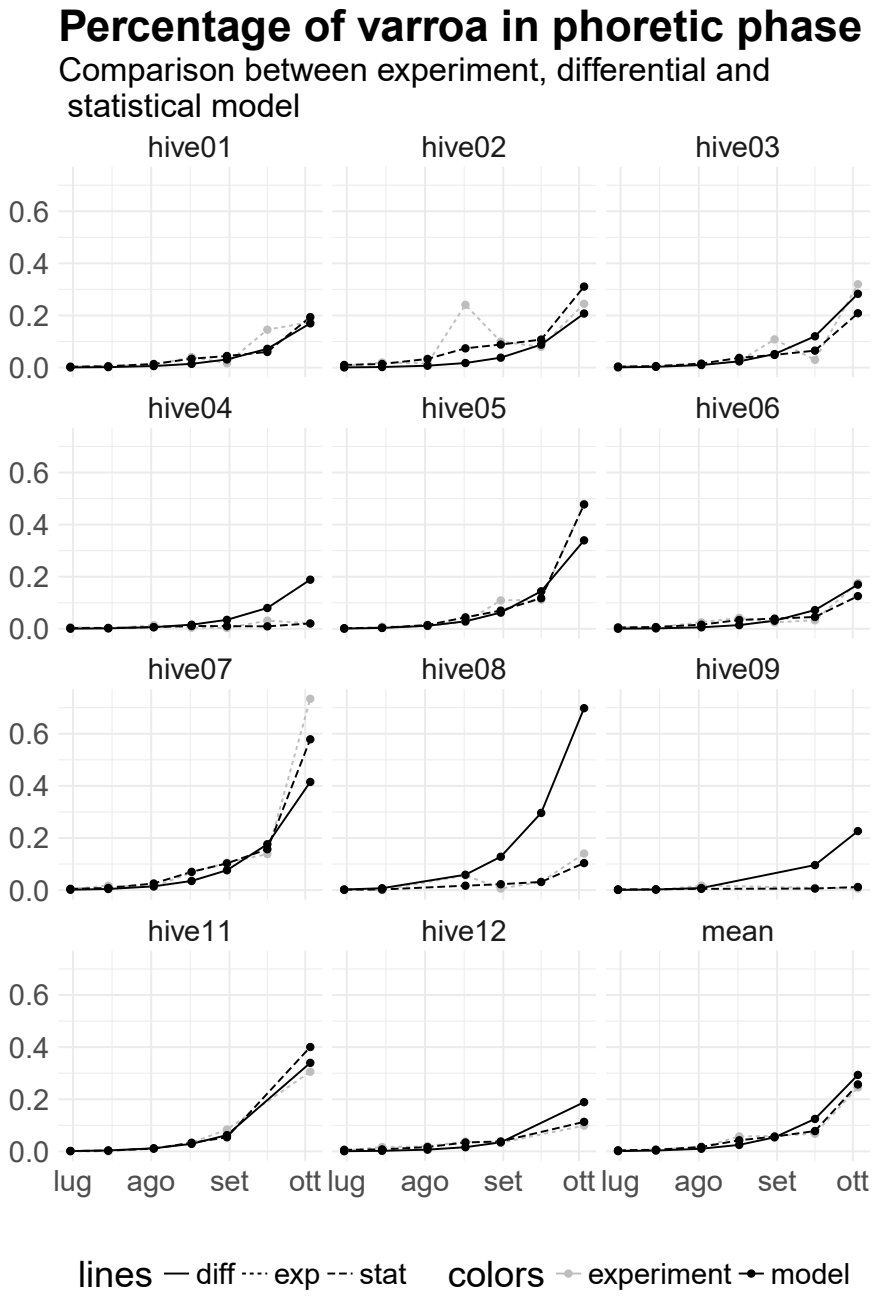


Figure 8.3: This figure shows the results of modeling the percentage of varroa in phoretic action respect the number of bees. The gray dashed line represents the experimental data, the black dashed lines the fitted values by the statistical model (8.2) and the black simple line the fitted values by the differential system (8.1). These three curves are shown for each of the eleven hives and more over the average hive is shown.

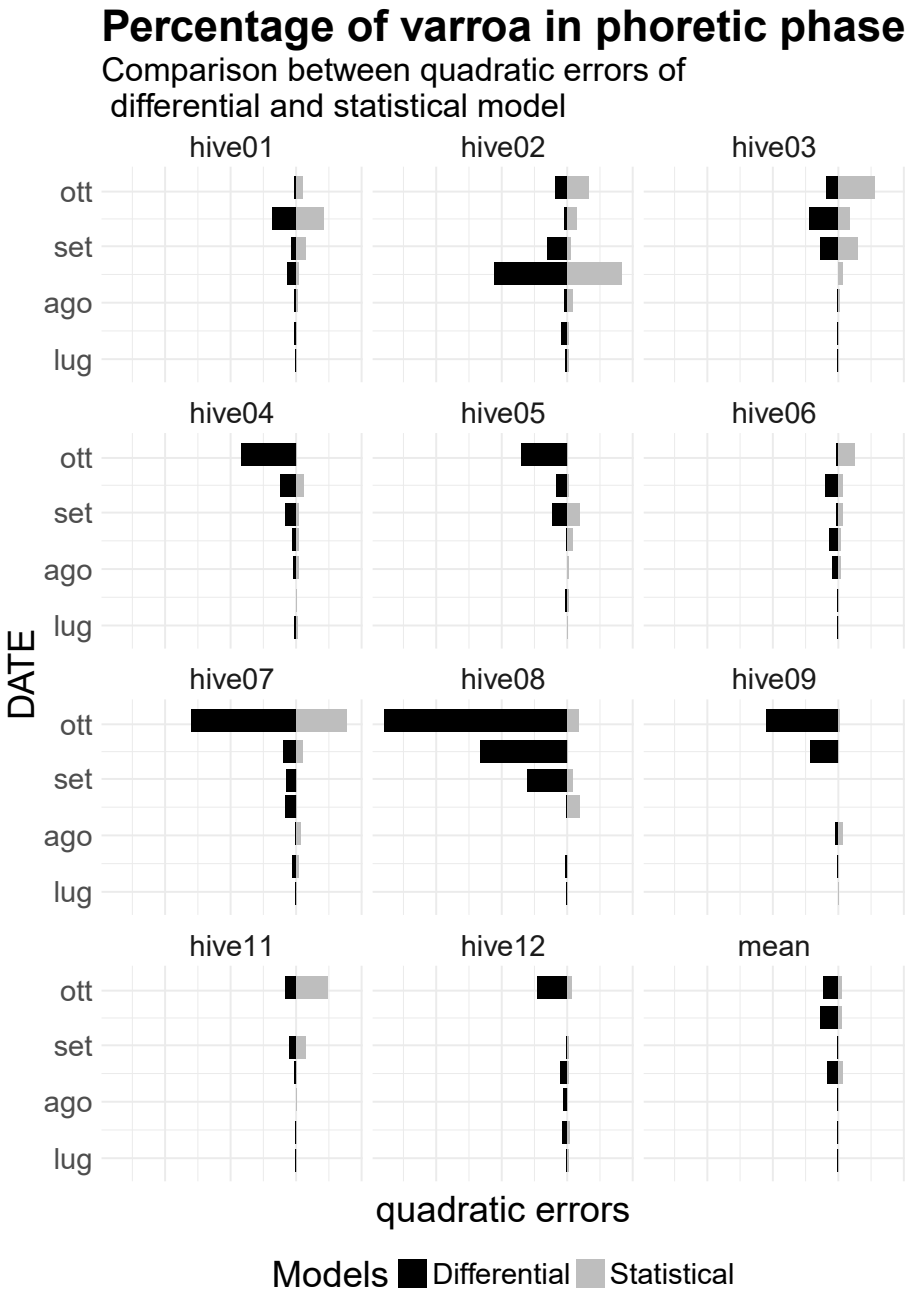


Figure 8.4: This figure shows the quadratic errors in each hive and in each time about modeling the percentage of varroa in phoretic stage for the differential and statistical models. On the left the black bars represents errors for the differential system (8.1) while, on the right, the grey bars represents errors for the statistical model (8.2).

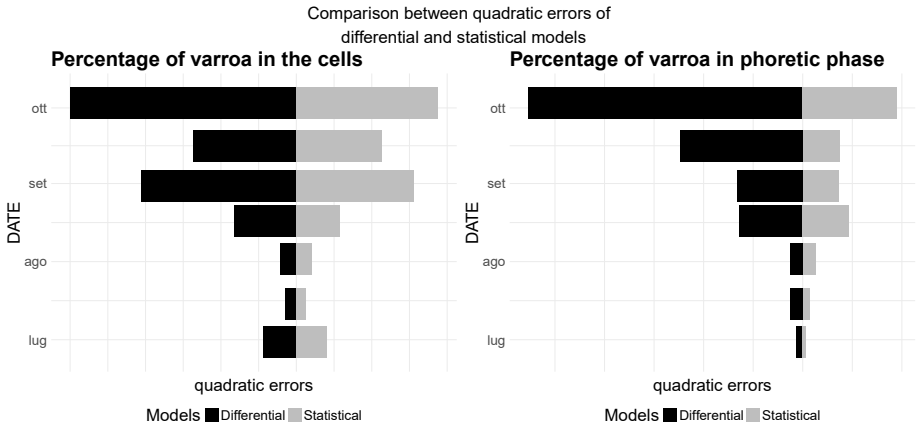


Figure 8.5: This figure shows the average quadratic errors in each time about modeling the percentage of varroa in the cells and in phoretic stage. The values in each time are the average of quadratic errors per hive in each time. On the left the black bars represents errors for the differential system (8.1) while, on the right, the grey bars represents errors for the statistical model (8.2).

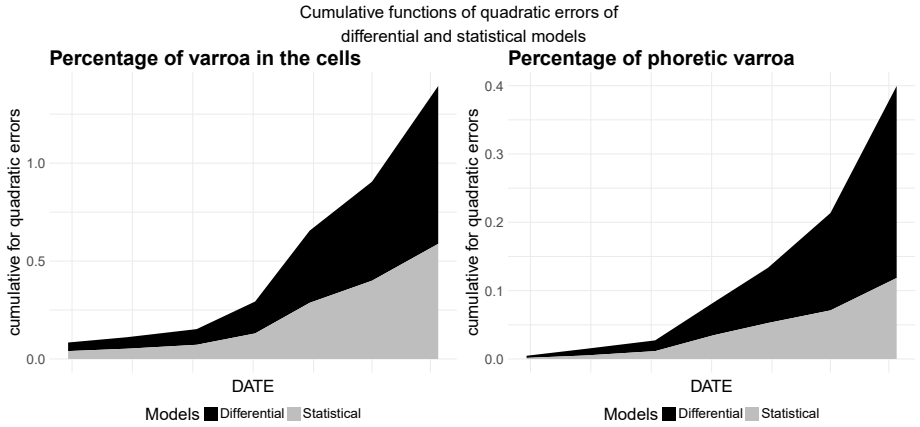


Figure 8.6: This figure shows cumulative functions for average quadratic errors about modeling the percentage of varroa in the cells and in phoretic stage. The black area is referred to the differential system (8.1) while the grey area is referred to the statistical model (8.2).

statistical model works better than the differential model, with an acceptable fitting percentage of 67% against an acceptable fitting percentage of 42% of the differential model. This is clear from Figure 8.4, where we can see immediately how the bars relative to the differential model are much wider than those of the statistical one. We see that also in the eighth hive the performance of the differential model is very low.

Figure 8.5 generally shows the best predictive ability of the statistical model. This figure shows for each measurement date the average of the square errors committed in each hive.

The last instrument to compare our models is Figure 8.6. It shows the cumulative functions for the average square errors over time. We can see that with time the cumulative error of the differential model increases quickly than the cumulative error of the statistical one and more the final values of the differential error are very higher than the statistical one, especially in the case of the percentage of varroa in phoretic stage. In the end we can conclude by saying that the statistical model works better than the differential one also for the percentage of phoretic varroa.

8.2 Biological consideration

We have found that belonging to a particular beehive plays an important role to model the varroa growth but we can not say if this is due to the different quantity of inoculum. The problem is that this different quantities are very small and similar and they don't represent a discriminant data to have a very infested hive or not. The estimation of the parameters for the model (8.1) could be an important biological result but now we can not discuss this because in literature there are estimations for simpler growth models that don't consider most phenomena that occur during the life cycle of varroa. Our estimations are additional comparison values for future models.

From [3] we know that during the bee season, i.e. spring and summer, the mite population doubles every month, so a fast estimate for the growth rate r_{lit} is $r_{lit} \approx 30^{-1} ln2 = 0.02$. This quantity is considered as the rate with which adult varroa are born, i.e. the rate with which adult varroa go out from the cells. In our differential system (8.1) we don't have a comparable parameter because we consider more factors that influence the varroa growth. However we can compare this parameter form literature with the result obtain from the statistical model that consider the increase of phoretic varroa and varroa in reproductive stage separately. As we have seen in discussion parts of Chapters 6 and 7, we can consider two mean parameters: $\bar{\beta}_{REP} \simeq 0.0205$, that describes the growth of varroa in the cells over time, and $\bar{\beta}_{PHO} \simeq 0.0311$, that describes the phoretic varroa growth over time. These two parameters say that in reproductive phase varroa, during the bee season, duplicate their number in almost 34 days, instead the phoretic varroa, that more precisely is the population to which r_{lit} refers, duplicate their number in less than 23 days. For this, our work says that the phenomenon of varroa growth in phoretic activity is underestimated.

Another important consideration, already presented in Chapters 6 and 7, is the climatic influence for the varroa growth. As we have seen an high daily temperature range disinhibit the growth of varroa in the cells and this could be a first indication for beekeepers: if an year shows particularly low temperature ranges, anticipating the date of administration of the treatment based on oxalic acid could be necessary. The problem is that this consideration is valid for the varroa growth in the cells, so if the treatment is inefficient for operculated cells this consideration is wrong. But from Chapter 7 we know that high temperature peaks disinhibit the growth of phoretic varroa. For this pay attention to the daily temperature range remains a good advice. A last climatic consideration is about the visibility. We know that to an increase of visibility corresponds an increase in the percentage of varroa. But probably this is not an useful fact.

8.3 Problems and possible future works

The largest limitation of this thesis is the small number of measurements taken in the hives (unfortunately we know that biological data is the most difficult data to collect), so the analysis could be redone taking measurements at least every two weeks, in way to duplicate the number of time data. An other consideration to understand better the climatic effect on varroa is to position hives in different localities. In fact in our thesis the climatic data, in addition to not being referred precisely to the location of the experiment, are the same for each hive. In this way we don't have a diversification of the data that remain related to the time: obviously in the same locality the temperature has a predictive trend over time. This diversification of hives helps to solve the third problem: low characterization of hives. The only different data from an hive and an other is the inoculum quantities. Also if in this experiment these quantities are not so decisive in the models, the predictor $HIVE$ plays a fundamental role. This means that there are other characteristics of the hives that influence the varroa growth. A way to capture this information is to characterize better the hives: position of the hive, if this position influences the exposure to climate factors, if the hives are closed or not (obviously, the more two hives are close together, the higher the number of bees they exchange). These ideas are simple examples of a more careful characterization of hives. A obvious consideration is that a larger number of hives helps the accuracy of the model and above all it allows to be able to carry out inocula with a greater variance, ranging from not introducing phoretic varroa to insert a really high quantity of them.

Appendix A

Definitions and methods

A.1 Properties of eigenvalues

The relationships that bind the coefficients and roots of an algebraic equation and the definition of characteristic polynomial imply that

$$\sum_{i=1}^{n} \lambda_i = tr(A) \quad \text{and} \quad \prod_{i=1}^{n} \lambda_i = det(A),$$

where λ_i are the eigenvalues of A.

The eigenvalues of a matrix A diagonal or triangular (upper or lower) are equal to the main elements, i.e. the elements on the diagonal. In fact, the matrix $A - \lambda I$ still diagonal or triangular and then

$$det(A - \lambda I) = \prod_{i=1}^{n} (a_{ii} - \lambda), \quad \text{with} \quad a_{ii} = (A)_{ii}.$$

A block matrix or a partitioned matrix is a matrix that is interpreted as having been broken into sections called blocks or submatrices. A **block triangular matrix** (upper case) is a matrix in the form

$$A = \begin{bmatrix} A_1 & * & \cdots & * \\ 0 & A_2 & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_r \end{bmatrix},$$

where A_i are square matrices and the symbol $*$ is for matrices not necessarily equal to zero. For the determinant and trace, the following properties hold:

$$tr(A) = \sum_{i=1}^{r} tr(A_i) \quad \text{and} \quad det(A) = \prod_{i=1}^{r} det(A_i).$$

For even though A is a block triangular matrix, $A - \lambda I$ is so, from what has been said, it is clear that

$$det(A - \lambda I) = \prod_{i=1}^{r} det(A_i - \lambda_i I),$$

and so that the eigenvalues of a block triangular matrix are the eigenvalues of all its blocks.

A.2 Downhill simplex method, i.e. Nelder–Mead method [21]

The *Nelder–Mead method* or *downhill simplex method* is a commonly applied numerical method used to find the minimum or maximum of an objective function in a multidimensional space. It is applied to nonlinear optimization problems for which derivatives may not be known. However, the Nelder–Mead technique is a heuristic search method that can converge to non-stationary points on problems that can be solved by alternative methods.

To define this method we consider the case in which we have only two variables, where a simplex is a triangle and the method is a pattern search that compares function values at the three vertices of a triangle. The worst vertex, where the function is largest, is rejected and replaced with a new vertex. A new triangle is formed and the search is continued. The process generates a sequence of triangles (which might have different shapes), for which the function values at the vertices get smaller and smaller. The size of the triangles is reduced and the coordinates of the minimum point are found.

Let $f(x, y)$ be the function that is to be minimized. To start, we are given three vertices of a triangle: $V_k = (x_k, y_k)$ with $k = 1, 2, 3$. The function $f(x, y)$ is then evaluated at each of the three points: $z_k = f(x_k, y_k)$ for $k = 1, 2, 3$. The subscripts are then reordered so that $z_1 \leq z_2 \leq z_3$. We use the notation

$$B = (x_1, y_1), \quad N = (x_2, y_2), \quad W = (x_3, y_3),$$

to help remember that B is the best vertex, N is good (next to best) and W is the worst vertex.

The construction process uses the midpoint of the line segment joining B and N. It is found by averaging the coordinates:

$$M = \frac{B + N}{2} = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right).$$

The function decreases as we move along the side of the triangle from W to B, and it decreases as we move along the side from W to G. Hence it is feasible that $f(x, y)$ takes on smaller values at points that lie away from W on the opposite side of the line between B and G. We choose a test point R that is obtained by "reflecting" the triangle through the side $\overline{BG}$. To determine R, we first find the midpoint M of the side $\overline{BG}$. Then draw the line segment from W to M and call its length d. This last segment is extended a distance d through M to locate the point R. The vector formula for R is

$$R = M + (M - W) = 2M - W.$$

Now there are two cases:

1. If the function value at R is smaller than the function value at W, then we have moved in the correct direction toward the minimum. Perhaps

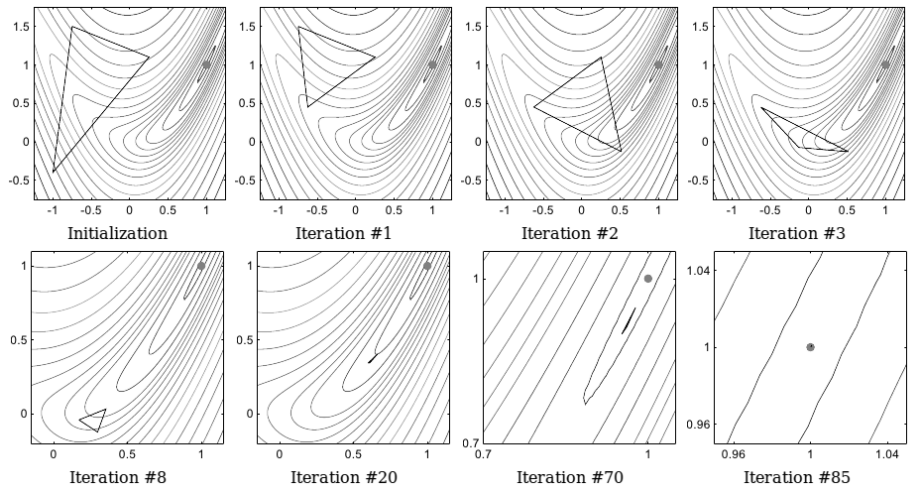


Figure A.1: Illustration of the downhill simplex method of with the Rosenbrock function $f(x, y) = (1-x)^2+10(y-x^2)^2$. The minimum of this function is at the point of coordinate $(1, 1)$, represented by the grey point on the Figures. Here, the convergence is reached in 85 iterations.

the minimum is just a bit farther than the point R. So we extend the line segment through M and R to the point E. This forms an expanded triangle BGE. The point E is found by moving an additional distance d along the line joining M and R. If the function value at E is less than the function value at R, then we have found a better vertex than R. The vector formula for E is

$$E = R + (R - M) = 2R - M.$$

2. If the function values at R and W are the same, another point must be tested. Perhaps the function is smaller at M, but we cannot replace W with M because we must have a triangle. Consider the two midpoints C_1 and C_2 of the line segments $\overline{WM}$ and $\overline{MR}$, respectively. The point with the smaller function value is called C, and the new triangle is BGC. Note that the choice between C_1 and C_2 might seem inappropriate for the two-dimensional case, but it is important in higher dimensions. If the function value at C is not less than the value at W, the points G and W must be shrunk toward B. The point G is replaced with M, and W is replaced with S, which is the midpoint of the line segment joining B with W.

A computationally efficient algorithm should perform function evaluations only if needed. In each step, a new vertex is found, which replaces W. As soon as it is found, further investigation is not needed, and the iteration step is completed.

A.3 R^2 and R^2_{adj}

The coefficient of determination, denoted R^2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

Definition A.3.1 *A data set has n values marked $y_1, \cdots, y_n$, each associated with a predicted (or modeled) value $\hat{y}_1, \cdots, \hat{y}_n$. Define the residuals as $e_i = y_i - \hat{y}_i$. If $\bar{y}$ is the mean of the observed data:*

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

then the variability of the data set can be measured using three sums of squares formulas:

- *the **total sum of squares** (proportional to the variance of the data):*

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2;$$

- *the regression sum of squares, also called the **explained sum of squares**:*

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2;$$

- *the sum of squares of residuals, also called the **residual sum of squares**:*

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2.$$

The most general definition of the coefficient of determination is

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

R^2 shows how well terms fit a curve or line. Adjusted R^2, one common notation is R^2_{adj}, also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, R^2_{adj} will decrease. If you add more useful variables, R^2_{adj} will increase.

Definition A.3.2 *The adjusted R^2 is defined as*

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

where p is the total number of explanatory variables in the model (not including the constant term), and n is the sample size. Adjusted R^2 can also be written as

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}.$$

R^2_{adj} will always be less than or equal to R^2.

A.4 Variables selection: stepwise model by AIC

Stepwise regression is a method of choicing a subset of predictive variables by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Usually, this takes the form of a sequence of F-tests or t-tests, but other techniques are possible, such as adjusted R^2, Akaike information criterion (AIC), Bayesian information criterion (BIC), Mallows's Cp, PRESS, or false discovery rate.

Definition A.4.1 *The **backward elimination** method follows the following procedure:*

1. *start with all the predictors in the model;*
2. *remove the predictor with highest p-value greather than α_{critc};*
3. *refit the model and goto point 2;*
4. *stop when all p-values are less than α_{critc}.*

Definition A.4.2 *The **forward selection** method follows the following procedure (it just reverses the backward method):*

1. *start with no predictors in the model;*
2. *for all predictors not in the model, check their p-value if they are added to the model anf choose the one with lowest p-value less than α_{critc};*
3. *continue until no new predictors can be added.*

The **Akaike information criterion** (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection. AIC offers an estimate of the relative information lost when a given model is used to represent the process that generated the data. It does not provide a test of a model in the sense of testing a null hypothesis. It tells nothing about the absolute quality of a model, only the quality relative to other models. Thus, if all the candidate models fit poorly, AIC will not give any warning of that.

Definition A.4.3 *Suppose that we have a statistical model of some data. Let p be the number of estimated parameters in the model. Let $\hat{L}$ be the maximum value of the likelihood function for the model. Then the **AIC** value of the model is*

$$AIC = 2p - 2log(\hat{L}).$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages overfitting, because increasing the number of parameters in the model almost always improves the goodness of the fit.

The **stepwise regression** is a combination of backward elimination and forward selection. This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later. At each stage a variable may be added or removed and there are several variations on exactly how this is done. The stepwise regression using in this thesis does not use the p-value like criterium to chosen the variables, but thee AIC.

Appendix B

R packages

In this part all packages used in thesis are fastly presented.

- **dplyr:** a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges;
- **stringr:** a tool to easily manage string objects;
- **openxlsx:** it simplifies the creation of Excel .xlsx files by providing a high level interface to writing, styling and editing worksheets;
- **ggplot2:** a system for "declaratively" creating graphics, based on "The Grammar of Graphics": you provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details;
- **corrplot:** it is a graphical display of a correlation matrix or general matrix and it also contains some algorithms to do matrix reordering;
- **lme4:** fit linear and generalized linear mixed-effects models;
- **fmsb:** several utility functions for the book entitled "Practices of Medical and Health Data Analysis using R" (Pearson Education Japan, 2007) with Japanese demographic data and some demographic analysis related functions;
- **mgcv:** it provides functions for GAM and generalized additive mixed modelling (GAMM);
- **nlme:** fit and compare Gaussian linear and nonlinear mixed-effects models;
- **ggpubr:** it provides some easy-to-use functions for creating and customizing "ggplot2", based publication ready plots;
- **MASS:** functions and datasets to support *Venables* and *Ripley*, "Modern Applied Statistics with S" (4th edition, 2002);
- **car:** functions to accompany *J. Fox* and *S. Weisberg*, "An R Companion to Applied Regression";

- **lmerTest:** different kinds of tests for linear mixed effects models as implemented in "lme4" package are provided: the package provides the calculation of population means for fixed factors with confidence intervals and corresponding plots and the backward elimination of non-significant effects is implemented;

- **pscl:** bayesian analysis of item-response theory (IRT) models, roll call analysis; computing highest density regions; maximum likelihood estimation of zero-inflated and hurdle models for count data; goodness-of-fit measures for GLMs; data sets used in writing and teaching at the Political Science Computational Laboratory; seats-votes curves;

- **glmm:** approximates the likelihood of a generalized linear mixed model using Monte Carlo likelihood approximation, then maximizes the likelihood approximation to return maximum likelihood estimates, observed Fisher information, and other model information;

- **lubridate:** functions to work with date-times and time-spans: fast and user friendly parsing of date-time data, extraction and updating of components of a date-time (years, months, days, hours, minutes, and seconds), algebraic manipulation on date-time and time-span objects;

- **deSolve:** functions that solve initial value problems of a system of first-order ordinary differential equations (ODE), of partial differential equations (PDE), of differential algebraic equations (DAE), and of delay differential equations;

- **minpack.lm:** the nls.lm function provides an R interface to lmder and lmdif from the MINPACK library, for solving nonlinear least-squares problems by a modification of the Levenberg-Marquardt algorithm, with support for lower and upper parameter bounds;

- **directlabels:** an extensible framework for automatically placing direct labels onto multicolor lattice or ggplot2 plots: label positions are described using Positioning Methods which can be re-used across several different plots and there are heuristics for examining trellis and ggplot objects and inferring an appropriate Positioning Method;

- **reshape2:** flexibly restructure and aggregate data.

Bibliography

[1] Elke Genersch (2010) *Honey bee pathology: current threats to honey bees and beekeeping*

[2] Ernesto Guzmán-Novoa, Leslie Eccles, Yireli Calvete, Janine Mcgowan, Paul G. Kelly and Adriana Correa-Benítez (2010) *Varroa destructor is the main culprit for the death and reduced populations of overwintered honey bee (Apis mellifera) colonies in Ontario, Canada*

[3] Sara Bernardi, Ezio Venturino (2014) *Epidemiology of viruses in adult Apis Mellifera by Varroa destructor mite.*

[4] Alain F. Zuur, Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, Graham M. Smith *Mixed Effects Models and Extensions in Ecology with R*

[5] Antonio Lavazza (2017) *I virus delle api e interazioni con altri patogeni.*

[6] Gunn, Alan, Bowen Walker PL, Martin SJ (1999). *The transmission of deformed wing virus between Honeybees (Apis melliferaL.) by the ectoparasitic mite varroa jacobsoni oud.*

[7] *http://www.apicoltura2000.it/esperienze/ossalicoangri.htm*

[8] Sokal, R. R., Rohlf, F. J. (1995). *Biometry: the principles and practice of statistics in biological sciences.*

[9] Gerry P. Quinn, Michael J. Keough. *Experimental Design and Data Analysis for Biologists.*

[10] David Posada, Thomas R. Buckley *Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests.*

[11] *https://www.ilmeteo.it*

[12] Hastie, Tibshirani (1990). *Generalized Additive Models.*

[13] Pereira, Lopes, Camargo, Vilela (2002). *Social e desenvolvimento das abelhas apis mellifera.*

[14] Khoury, Myerscough, Barron (2011). *A quantitative model of honey bee colony population dynamics.*

[15] Ratti, Kevan, Eberl (2012). *A mathematical model for population dynamics in honeybee colonies infested with varroa destructor and the acute bee paralysis virus.*

[16] C. Costa, E. Carpana, M. Lodesani. *Patologie e avversità delle api: tecniche diagnostiche e di campionamento.*

[17] M. A. Benavente, R. R. Deza, M. Eguaras (2009). *Assessment of Strategies for the Control of the Varroa descructor mite in Apis mellifera colonies, trough a simple model.*

[18] Fries, Ingemar, Scon Camazine, and James Sneyd (1994). *Population Dynamics Of Varroa Jacobsoni: A Model And A Review.*

[19] William H. Greene (2003). *Econometric Analysis.*

[20] Tsung-han Tsai, Jeff Gill (2013). *Interactions in Generalized Linear Models: Theoretical Issues and an Apllication to Personal Vote-Earning Attributes.*

[21] Nelder John A., R. Mead (1965). *A simplex method for function minimization.*

[22] Nelder, John; Wedderburn, Robert (1972). *Generalized Linear Models.*

[23] R Development Core Team, R Foundation for Statistical Computing (2008) *R: A Language and Environment for Statistical Computing*