NST 1B – MCB 2024/25
Mini Project 3

Dynamic models for dog rabies outbreaks in Tanzania

Deadline for submission of your report: 8th May 2025, 5pm.

You must submit a single Jupyter notebook containing your code, outputs and answers. Please label each part and task using Markdown cells.

If you require any clarification about the tasks in this mini-project, please contact Olivier Restif or226@cam.ac.uk.

Overview

This project is based on the paper by Hampson et al (2009) https://doi.org/10.1371/journal.pbio.1000053. You do not need to read the paper to do this mini-project, but you may find it helpful to understand the context. Please note that there are a few differences in the data and methods, so you may not get the exact same answers as in the original paper.

Objective: design and parametrise a stochastic model for rabies outbreaks in dog populations, and run simulations under different levels of vaccine coverage.

This mini-project is designed to apply some of the concepts and programming skills taught in Block A (random variables, maximum likelihood, confidence intervals) and Block C (deterministic and stochastic epidemic models, Gillespie simulations).

You may find the following Python libraries useful:

- Numpy: arrays, matrices, random numbers.
- Pandas: read and write data, manipulate dataframes (Block A Practical 3).
- Matplotlib.pyplot to plot data (Block A Practical 3).
- Scipy.optimize for maximum likelihood estimation (Block A Practical 4).
- Scipy.stats for probability distributions (Block A Practical 4).
- Statsmodels for generalised linear models (Block A Practical 5).
- Scipy.integrate for differential equation solving (Block C Practicals 1 and 2).

Introduction

Rabies is an infectious disease caused by the rabies virus, which can infect most mammalian species. The main route of transmission is through bite. Upon exposure, the virus slowly travels from the site of the bite to the central nervous system of its new host. This “latent period” can last from a few days to a few months and does not cause any symptoms. Once the virus reaches nerve cells, it starts replicating, causing changes in behaviour. Eventually, the virus travels from the brain to the salivary glands to be excreted into saliva. This active phase of infection only lasts a few days and always ends with the death of the infected host. Rabid dogs often display aggressive behaviour and try to bite other animals, with the potential to transmit the virus. Human victims can be treated with injections of rabies vaccine and antibodies following a dog bite. Prophylactic vaccination of people, dogs and other animals works effectively, but vaccine campaigns remain insufficient in many parts of Africa and Asia to eradicate the disease.

This mini-project is based on data from a long-term study of dog rabies in Tanzania, where rabies outbreaks have occurred regularly, spreading from village to village in rural regions. During one of the largest outbreaks (2002-2007), an intensive epidemiological study was carried out, collecting detailed data on the dynamics of rabies in dogs. The study was done in parallel in two districts, Ngorongoro and Serengeti. Ngorongoro covers a greater area but has a smaller population than Serengeti. The study is presented in full in Hampson et al (2009), and you will be tasked with replicating some of the analyses, leading to the development of a compartmental model to simulate dog rabies outbreaks. **Note that the data has been modified, so your answers will differ (slightly) from the numerical values presented in the original paper.**

You are provided with the following datasets:

- “contact_tracing.csv”: latency period, infectious period, and number of dogs bitten recorded for 200 dogs as part of a contact tracing survey during the outbreak.
- “case_reports.csv”: monthly reports of dog rabies cases in each district for the first two years of the outbreak. The tally of cases recorded in a given month are recorded on the 1st day of that month in the file.

Part 1. Outbreak data analysis

Your first objective is to estimate the values of key epidemiological parameters from the available data. This will enable you to implement a compartmental model in Part 2. Through contact tracing during the outbreak, it was possible to record the dates at which dogs had been bitten, started showing symptoms, and died of rabies. This allows a direct estimation of the duration of the latent period and the infectious period. Although transmission rates are generally harder to measure directly, a few methods are available to estimate them through the basic reproduction number R0.

The first method was introduced in Henrik Salje's lecture 1 in Block C, linking R0 to the exponential growth rate of the epidemic curve in the early stage of an outbreak. It can be shown that:

$$R_0 = e^{r\,T_g}$$

where *r* is the exponential growth rate of the epidemic curve and T_g is the mean generation time (equal to the sum of the average incubation period and the average infectious period).

The following tasks will guide you through the estimation of *r* and T_g from the data.

Task 1.1. Estimate the generation time.

a. Import the data from "contact_tracing.csv", which records four variables for a total of 200 dogs involved in the rabies outbreak: duration of the incubation period (from being bitten to onset of symptoms) in days, duration of the infectious period (from onset of symptoms to death), number of dogs they have bitten during the infectious period, and district (N for Ngorongoro or S for Serengeti).

b. Summarise and plot the data in an appropriate way.

c. Plot the data by district, and test whether any of the variables differ between the two districts.
Hint: use **`pandas.DataFrame.groupby()`** to analyse or plot the data by district. Online tutorial: https://pandas.pydata.org/pandas-docs/stable/user_guide/groupby.html#

d. Visualise pairwise associations between variables using scatter plots. Calculate the Pearson correlation coefficients between each pair of variables, and test for association using the Pearson correlation test (**scipy.stats.pearsonr**). Comment on the validity of the Pearson correlation coefficient for this dataset. Can you propose (no coding required) an alternative statistical method to test for association between these variables?

e. Estimate the mean, standard error and 95% confidence interval of the generation time.

Task 1.2. Estimate the growth rate and R_0.

To estimate the growth rate, we are going to use Poisson regression, a generalised linear model with a log link function. Specifically, we assume that case reports at time *t* follow a Poisson distribution with mean $\mu_t = Y_0\, e^{r\,t}$. Given a vector of case reports $[x_0, x_1, \dots, x_{n-1}]$, the likelihood of the model is:

$$L(Y_0, r) = \prod_{t \geq 0} \frac{\mu_t^{x_t}\, e^{-\mu_t}}{x_t!}$$

a. Import and plot the monthly case reports for the Ngorogoro and Serengeti districts. We will treat them as two separate outbreaks which started around the same time.

b. Create a new column containing the number of days since 1/1/2002, i.e. converting dates into numbers, which can be achieved in Python in two steps using the pandas library: first, a simple subtraction between two dates creates a "Timedelta" object; then, the Timedelta object has a ".days" attribute equal to the number of days. To convert a whole Series of Timedeltas into floats, use the ".dt.days" attribute (see the pandas user guide on Time deltas).

c. Write a function that calculates the **log**-likelihood of the model (with two unknown parameters Y_0 and *r*) for a given dataset consisting of a vector of time points and a vector of rabies case numbers. Remember that a Python implementation of the Poisson distribution is available in the scipy.stats library, with the log of the distribution function given by the function poisson.logpmf(). See Practical 4 from Block A.

d. Estimate Y_0 and *r* for each district by maximum likelihood, using the function minimize() from scipy.optimize with the option method='Nelder-Mead' (as the default method may not always converge).

e. Check your results using the glm() function from the statsmodels library (which you previously learned in Mini-Project 1). Assuming we have the data in an array called "reports" with numerical colummns "cases" and "day", the syntax is:

```
from statsmodels.formula.api import glm
from statsmodels.api import families
my_glm = glm("cases ~ day", data=reports, family=families.Poisson()).fit()
my_glm.summary()
```

   Note that the Intercept parameter is equal to log(Y_0).

f. Plot the predicted (exponential) dynamics with the observed data for each district.

g. Calculate R_0 for rabies in each district.

h. Using the confidence intervals from Task 1.1.e and Task 1.2.e, is there any evidence that the growth rates *r* and basic reproduction numbers R_0 differ between the two districts?

Part 2. Deterministic model for dog rabies dynamics

We model the spread of rabies in a dog population with four compartments:

S = number of susceptible dogs
E = number of latently infected dogs
I = number of infectious (rabid) dogs
D = cumulative number of rabies death.

using this set of differential equations:

$$\frac{dS}{dt} = b - \frac{\beta S I}{N} - m\,S$$

$$\frac{dE}{dt} = \frac{\beta S I}{N} - \alpha\,E - m\,E$$

$$\frac{dI}{dt} = \alpha\,E - \omega\,I - m\,I$$

$$\frac{dD}{dt} = \omega\,I$$

where $N(t) = S(t) + E(t) + I(t)$ is the number of dogs alive at time t. **In the following, we use day as the time unit**.

Task 2.1. Model definition and parametrisation

a. Write a brief definition for every parameter in the model (no coding required). Pay particular attention to parameter *b*.

b. First, consider the population dynamics in the absence of rabies. Estimate the values of parameters *b* and *m* for a stable population of 1,000 dogs with an average lifespan of 5 years.
Hint: recall the monomolecular model from 1A Mathematical Biology (Block C).

c. Use your results from Part 1 to estimate the values of parameters α and ω.

d. Explain why the basic reproductive number is equal to $R_0 = \frac{\beta}{\omega+m}$, and calculate values of β for each district.

e. Using Practical 2 from Block C as a template, implement the system of differential equations in Python:

- Write a function seid_model(t, var, par) which returns an array with the numerical values of the four derivatives of S, L, I and D, given the current time t, an array var containing the current values of the 4 variables and an array par containing the values of the 5 parameters $b, m, \alpha, \beta, \omega$.
- Use the solve_ivp function from the scipy.integrate library to solve the system of differential equations for each district, using the initial conditions $S(0) = 1000$ and $I(0) = Y_0$ as estimated in Task 1.2, so that the start t=0 corresponds to 1/1/2002 in the dataset.

Note: you should use the option method = “LSODA” in solve_ivp() to override the default ODE integration method with a more robust one.

- Plot the solution for each district over a period of 2 years.

In order to compare the model output with the data, we need to calculate the monthly cases predicted by the model. Given the short infectious period, we are going to use the number of rabid deaths, represented by the variable D(t), as a proxy for the cumulative number of cases up to time t.

f. Write a function seid_deaths(t, var_init, par) that takes an array t of time points (in days), an array of initial conditions var_init and an array of parameter values as arguments, computes the SEID model solution using solve_ivp, and returns an array containing the tally of rabies deaths over each period [t[i], t[i+1]].

   Hint: use the argument t_eval in solve_ivp() to specify an array of time points at which the solution should be returned. You may also find the Numpy function diff() useful.

g. Plot the recorded monthly case numbers in each district with the prediction from the SEID model and the prediction from the Poisson regression model from Part 1.

Task 2.2. Model exploration

In this section, we will use $R_0 = 1.17$, $I(0) = 0.5$ and $S(0) = 1{,}000$ as default values.

a. Plot and describe the model dynamics over a 6-year period.

   Investigate the effects of varying, in turn, the values of R_0 between 1.1 and 1.2, $I(0)$ between 0.1 and 1, and $S(0)$ between 1,000 and 10,000, on each of the following outputs:

b. Cumulative case numbers in the first two years,
c. Epidemic peak,
d. Asymptotic steady state.

e. What happens if *b*=*m*=0? Consider the initial outbreak dynamics and the longer term dynamics.

   Note: the solver may fail (or become very inaccurate) when running the model for more than 5 years.

f. Discuss the differences with a standard SEIR model.

Part 3. Stochastic model

a. Using the code from Block C Practical 2 Part 5 as a template, write a function **`SEID_sim(t_max, init, par)`** which implements the Gillespie algorithm for the SEID model with initial conditions **init**, parameter values **par** and for **t_max** days.

b. Run 100 simulations for 2 years and plot the number of rabies deaths against time. Add the curve of D(t) from the deterministic model to the same graph.

c. Plot the distribution of the total number of rabies deaths after 2 years. Compare with the output from the deterministic model (code and discussion).

d. Calculate the probability of early extinction (Lecture C5), and write some code to calculate the proportion of simulations that end "early" (without causing a large outbreak). Justify your choice of definition.

e. Considering the variations in outbreak sizes generated by demographic stochasticity, discuss the implications for parameter estimation based on a single observed outbreak. (No coding required)