Problem Solving and Modelling Task

Alexander Arthur

January 15, 2022

Contents

1 Introduction

The given task is to predict the time period after which a population of bacteria will completely cover the surface of the water in a dam (Figure 1). In order to do this, the area of the lake must first be calculated from the diagram provided. Once this area is known, the growth of the bacteria can then be projected and the point at which the lake will be covered subsequently found.

2 Assumptions and observations

2.1 Observations

When the diagram of the dam is oriented with the straight wall horizontally, the rest of the outline passes the vertical line test, meaning that it can be modelled with some sort of mathematical function.

2.2 Assumptions

In order to simplify the problem, certain assumptions must be made either due to a lack of information provided or to bring the scope of the problem to a reasonable scale.

Should the water level of the dam change, this would change the water's surface area, since it is unlikely the walls of the dam are perfectly vertical. Changing the surface area of the dam would in turn affect the point at which the bacteria would cover the entire lake, affecting the result of the investigation.

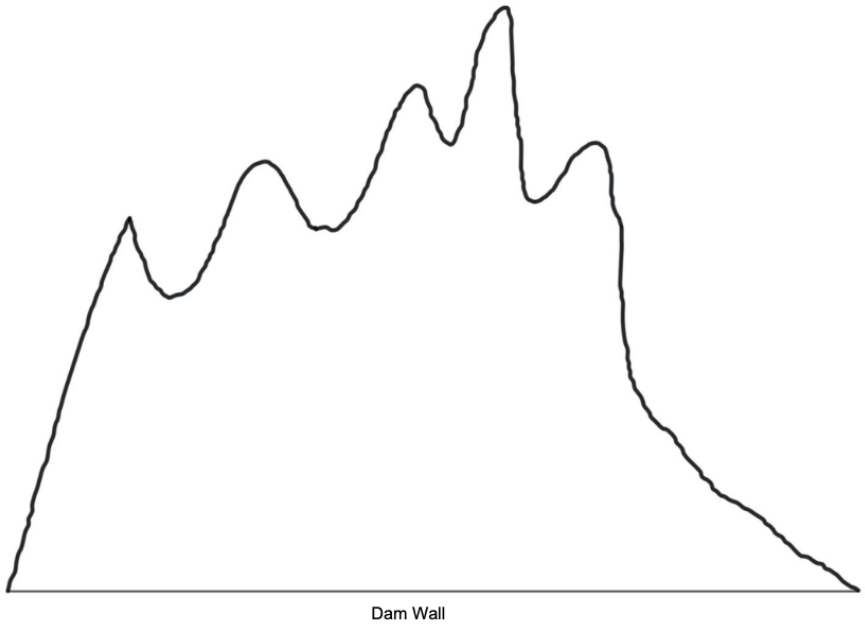

Figure 1: Diagram of outline of dam

Since no data is provided on the water level, it is reasonable to assume that it is static, and will not have any effect on the surface area of of the lake.

In a similar vein, there are many factors which can influence the growth rate of a population of bacteria, including access to a food source, temperature and competition for space. Similarly to the dam's water level, no data is provided around this, meaning that it must be assumed that the none of these factors will affect the bacteria as the population grows.

3 Translation of mathematical concepts

As per the requirements of the task, a mathematical approach must be used to calculate the area of the dam. To satisfy this requirement, the definite integral of a piecewise function will be used to calculate the area of dam.

In statistical modelling, a model can be overfitted to a dataset such that it is highly accurate within the domain of the data, but is wildly inaccurate outside this domain, meaning that it cannot be used to predict values outside the domain of the data and therefore is not useful. However, in this task, for any given segment of the piecewise function the only domain which is applicable is that of that segment, meaning that the concept of the function being over-fitted does not apply. In addition, the Weierstrass approximation theorem states that for any continuous function f over the real interval $[a, b]$ and any $\epsilon > 0$ there

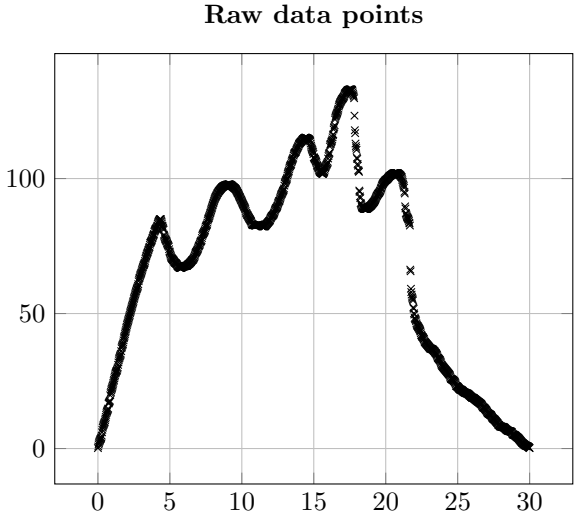

Figure 2: Raw data from Python script

exists a polynomial p such that $|f(x) - p(x)| < \epsilon$, where $a \leq x \leq b$.

This means that since the only key metric used to select a function is its accuracy within the specified domain, regression was used to select a function within each defined section, and the regression models used were strictly polynomial; the degree of the polynomial was simply increased until the function fit the data to a sufficient degree of accuracy ($R^2 > 0.97$).

To generate the points to be used for these regressions, a Python script was written to list the coordinates of the topmost black pixel in each column of pixels. This generated a list of 1262 pixel coordinates, which were scaled to the dimensions of the dam such that one unit on the x and y axes were each equal to 1 meter and subsequently used to generate regression models (Figure 2).

In an ideal scenario, a single function could be used to model the entire length of the dam's outline. However, due to the computational limits of the available software a polynomial regression of sufficiently high degree is not possible. This means that the data was instead divided into several sections (Figure 3, each of which were then used to run a polynomial regression. Since these sections were selected to have more simple shapes than the entire dataset, polynomials of much lower degree (< 6) were sufficient to replicate their shape to a sufficient degree of accuracy.

Let us define a variable b for the boundaries between the sections, such that $b_0 = 0$, $b_1 = 4.3 \ldots b_8 = 30$.

$$A = \sum_{r=1}^{8} \int_{x_{r-1}}^{x_r} f_r(x)\, dx \tag{1}$$

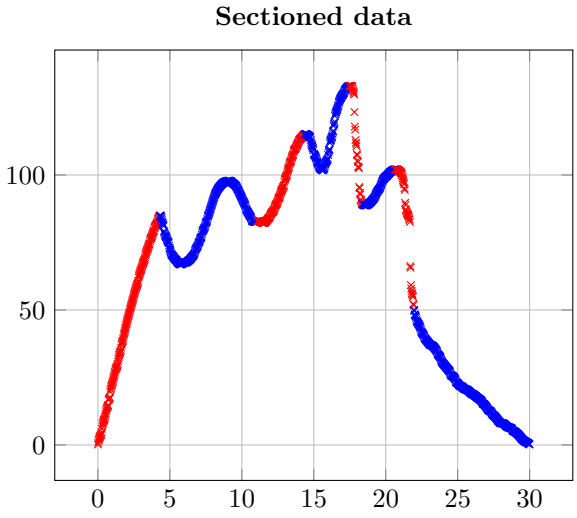

Figure 3: Data shown in coloured sections