A logo for a company

Description automatically generated

University of Auckland

Department of Engineering Science

Part IV Project Report

**Inferring the Detailed History of a Geothermal Field Using Sparse Data and Machine Learning**

*Author*

Michael Terekhin

|  |  |
| --- | --- |
| *Supervisors*  Dr Andreas Kempa-Liehr  Dr Michael Gravatt | *Project Partner*  Daniel Clark |



# Introduction

Geothermal energy plays a crucial role in meeting the energy demands across New Zealand. In 2021, geothermal energy supplied 18% of New Zealand’s electricity with it being expected to range from 15% to 19% in 2035[1]. Currently, 8 geothermal fields are in use with them being Wairakei, Tauhara, Rotokawa, Mokai, Nga Tamariki, Ohaaki, Kawerau, and Ngawha. With each geothermal field contributing heavily to New Zealand’s geothermal and total energy supply, it is of interest to understand the behaviour of these geothermal fields. Understanding how much mass has been taken from the geothermal fields and where the mass has been taken from is important for ensuring the field is ran well and future developments can happen.

Unfortunately, the relevant data describing the field history is very sparse and is typically collected for the entire geothermal field instead of individual wells. As a result, data regarding mass taken from each field and well is very rare leading to a lot of uncertainty. However, total steam flow produced for each field is typically known as it’s collected in regular intervals. By utilizing this regularly collected steam flow data, it’s hoped that the corresponding mass flow data for each field and subsequent well can be determined.

Initially, the simplest version of this problem will be solved where we assume relevant data is known, assume all wells are on, assume constant parameters and so on. Once a reliable and complete model/technique is established, we will move to increasingly complex versions of this problem where we remove known data, add additional parameters, don’t assume all wells are on and so on. In doing so, it’s hoped with increased complexity we will be able to generate a model/technique that can generalize better to different geothermal fields.

This project will have us investigate two different methods of tackling this problem: using a Bayesian approach and physics-based machine learning approach. The Bayesian approach aims to tackle uncertainty by treating parameters as random variables. Using these random variables, likely combinations of parameters are determined by comparing the model results against the existing data. The physics-based machine learning approach instead relies on understanding the possible outcomes through understanding the physical restrictions of a geothermal field. As both approaches require data that accurately reflects a geothermal field, initially simulated data is used to ensure our models aren’t subject to erroneous data issues that real collected data might have.

Using both a Bayesian and physics-based machine learning approach enables us to deal with the large uncertainty associated with sparse geothermal field history data. In combination, these approaches would ideally align at certain parameters and estimates and thus give us confidence in the result.

# Literature Review

Successfully filling in the missing data from the sparse geothermal field history is a crucial process yet is difficult due to lack of quality data. Before a possible solution to this is explored, an overview of existing methods that aim to solve problems like this both in the geothermal space and outside of it are explored.

## Belly Button Bacteria Problem (BBBP)

In 2011, Rob Dunn and his team of ecologists conducted a pilot study where they took swabs of the navels of 60 volunteers in order to understand what bacteria lived in this area. The pilot study resulted in the discovery of 2368 bacterial species, 1458 of which were potentially newly discovered[2]. One year later in 2012, Allen B. Downey came out with his book titled ‘Think Bayes’[3] that applied Bayesian statistical methods to the belly button bacteria problem.

Downey first created a simplified version of this problem using initially just three species: lions, tigers, and bears. To describe the prevalence of each species, he insisted on using a Dirichlet distribution[4] which ensured that the sum of the species prevalence would sum to 1. Once he solved the simplified problem of estimating the prevalence of each species, he went back the other way to try to estimate the total number of species. In order to generate random samples he suggested two methods with them being the marginal beta distribution and using n gamma distributions[5] then normalizing them by dividing by the total, with the latter method being recommended.

With the simpler problem solved, he returned to speaking about the belly button bacteria problem. After replicating what he did in the simpler problem, he generated posterior distributions of prevalence using the marginal beta distribution. By looping through the possible number of occurrences for each species and their associated probabilities he was able to generate a range of prevalence for each species. Using the posterior distributions defined, predictive distributions could be generated. Downey was able to define multiple classes and functions that produced a rarefaction curve[6] which was able to show the number of species against the number of samples. Afterwards, a joint posterior distribution was fit using the simulations to compare the fraction of species seen against the probability.

While this problem was related to an ecological problem and hence the solution developed was geared towards an ecological setting, components of the solution can be applied to our geothermal problem. Utilizing a Dirichlet distribution, as done in the belly button bacteria problem, is very applicable to our problem because we need the total sum of the fraction the wells produce steam or mass at each field to be exactly 1, otherwise our problem is improperly defined. Random sample generation is another thing we can use as we would need to generate plenty of random variates to account for uncertainty. If we use the Dirichlet distribution then using the normalized gamma distributions, as done in the BBBP, for our random variates would be the correct way to go about generating these samples.

## Gaussian Process Regression (GPR)

GPR is a powerful non-parametric model and stochastic process that is used in supervised machine learning problems[7]. It can be used to formulate a Bayesian framework for both regression and probabilistic classification problems where many, possibly even infinite, random variables are involved. Initially, a prior distribution is formulated where a distribution is fit across a function without considering additional information about it. This prior distribution uses the mean and covariance function to define a Gaussian process[8]. Using this allows for a function to be fit without any data observations and thus provide some levels of uncertainty before adding information into the problem. Rasmussen[7] demonstrates the prior distribution being fit using a Gaussian Process on three functions. There are no data points and the functions are created using fitted or estimated parameters, as shown in figure 1 below.

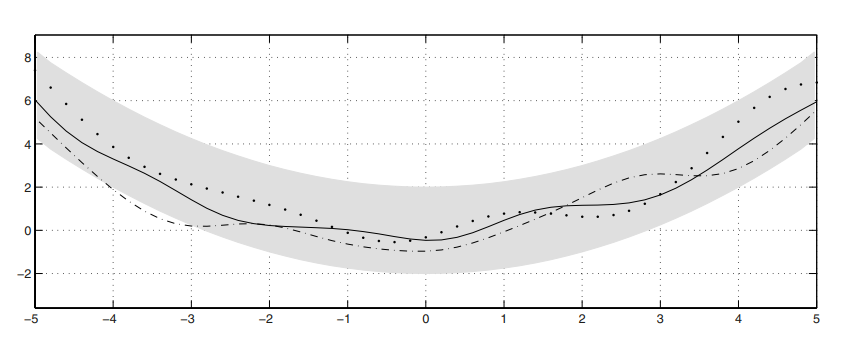


Figure 1: Example of Gaussian Process being applied to 3 random functions creating the prior distribution.

Once the prior distribution is defined, the posterior distribution can be defined which uses both the prior distribution and existing data observations to create a more accurate function. By using a set of equations, the mean and covariance of the posterior distribution can be determined. Variance in the posterior distribution is less than the prior because it takes the prior variance and subtracts a positive term off of it, which depends on the training data[7]. In figure 2 the posterior distribution using the prior distribution in figure 1 as well as training data is shown.

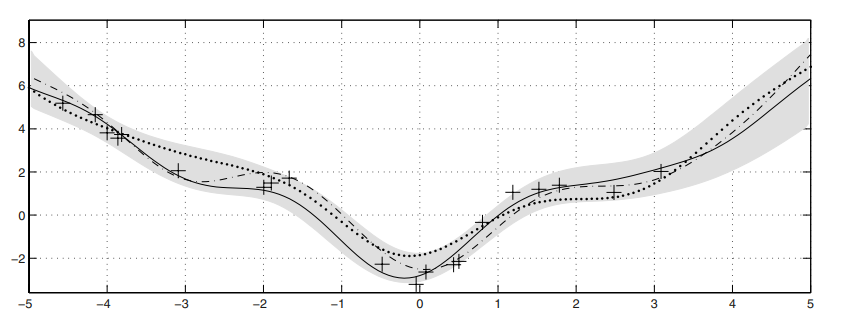


Figure 2: Posterior distribution example.

The posterior distribution is more clearly defined now and there is less noise, as evident through the narrower confidence intervals. This noise level is determined by considering the effect that every function has an extra covariance with itself only, with a magnitude equal to the noise variation[7].

Typically, prior information is vague and inadequate to create a posterior distribution, so the mean and covariance functions are parameterized using hyperparameters. These parameters can be found by taking the log marginal likelihood of the Gaussian distribution.

# References

[1] - <https://www.eeca.govt.nz/insights/energys-role-in-climate-change/renewable-energy/geothermal/#:~:text=Total%20geothermal%20electricity%20capacity%20in,1%2C000%20MW%20of%20electricity%20generation>.

[2] - [What lives in your belly button? A 'rainforest' of species (nationalgeographic.com)](https://www.nationalgeographic.com/science/article/121114-belly-button-bacteria-science-health-dunn#:~:text=Belly%20buttons%2C%20it%20turns%20out%2C%20are%20a%20lot,had%20adopted%20a%20new%20focus%20on%20citizen%20science.)

[3] - [thinkbayes.pdf (greenteapress.com)](https://greenteapress.com/thinkbayes/thinkbayes.pdf)

[4] - [The Dirichlet Distribution: What Is It and Why Is It Useful? | Built In](https://builtin.com/data-science/dirichlet-distribution)

[5] - [Gamma Distribution: Uses, Parameters & Examples - Statistics By Jim](https://statisticsbyjim.com/probability/gamma-distribution/)

[6] - [Rarefaction Curve: A Measure of Species Richness and Diversity - CD Genomics (cd-genomics.com)](https://www.cd-genomics.com/microbioseq/rarefaction-curve-a-measure-of-species-richness-and-diversity.html)

[7] - <https://link.springer.com/chapter/10.1007/978-3-540-28650-9_4>

[8] - <https://www.geeksforgeeks.org/prior-and-posterior-gaussian-process-for-different-kernels-in-scikit-learn/>