# Anomaly Detection in Domestic Animals using Real-Time Veterinary Data

**Charlotte Mary Appleton, BSc (Hons), MSc**

Centre for Health Informatics, Computing and Statistics

Lancaster University

A thesis submitted for the degree of

*Master of Philosophy*

September, 2025

**Anomaly Detection in Domestic Animals using Real-Time Veterinary Data**

Charlotte Mary Appleton, BSc (Hons), MSc. Centre for Health Informatics, Computing and Statistics, Lancaster University

A thesis submitted for the degree of *Master of Philosophy*. September, 2025

# Abstract

This thesis delves into anomaly detection in domestic animals using SAVSNet veterinary consultation data. The main aim is to create an automatic surveillance system that takes the consultation from recorded, through a method of analysis that is effective for the data and displaying the results back to the veterinary surgeons and stakeholders. This thesis explores Gaussian Processes as well as mixed models to decide on an anomaly detection methodology before creating an automatic surveillance system and using Shiny Apps and Tableau to display the results.

The research reveals that using a mixed model to model Spatio-temporal data was more effective than applying just a Gaussian Process as we were able to remove consistent characteristics including seasonality from the data and model departures from these characteristics using a Gaussian Process. One of the key findings is that using the mixed model was successful in detecting outbreak style patterns as it led to the confirmation of a Canine Enteric Coronavirus outbreak in January 2020. These insights contribute to the veterinary surgeons and stakeholders ways of working in that they have direct access to current rates for different Main Presenting Complaints within their county and Nationwide. Having this information accessible aids in the preparation for the veterinary surgeons e.g. having medicines readily available for seasonal periods.

The implications of this research are significant for domestic animal research as it leaves a tool that can be monitored for anomalies and thus can help influence decisions regarding domestic animal health. It is also an offering to the general public and owners of the animals as the dashboard is also created with those with lesser statistical knowledge in mind, alongside an emphasis on people with accessability issues.

In summary, this thesis effectively applies anomaly detection methodologies and through the creation of an automated surveillance system is able to relay results to veterinary surgeons, stakeholders, other researchers and the general public to improve research knowledge in this area.

# Acknowledgements

Firstly, I would like to acknowledge my supervisors, Professor Chris Jewell and Mr Barry Rowlingson. A special thank you to Chris Jewell, who has constantly pushed me to reach my full potential and without him, my knowledge of Bayesian Statistics would not be to this level. I thank him for his constant patience and his belief in me for the past 4 years, and for his encouraging and admirable outlook on research, especially during challenging times like the pandemic.

I would like to thank Dogs Trust for providing the funding for this project. Without this generosity, my means to research my passion, domestic animal health would cease to exist. Further, I would like to thank the members of SAVSNet Agile for their useful knowledge and for the use of their data.

I have thoroughly enjoyed being a student at Lancaster University, and that is purely due to the people there. I am thankful for the students within CHICAS for their constant support, alongside some students in the Mathematics and Statistics department who have now become lifelong friends. A special mention to Jess, Cían, Alin and Aiden, who without their on-going support alongside wine nights, shoulders to cry on, ears to listen to me moaning and the all necessary distractions, the completion of this thesis would've felt a lot more challenging.

On a more personal level, I would like to thank all my close friends, family and my amazing partner Ollie for supporting me throughout this journey. I know it is not an easy journey for many reasons and I appreciate your ongoing support and wouldn't be able to achieve the things I have done in life without you all by my side.

# Declaration

I declare that the work presented in this thesis is my own and has been carried out in accordance with the regulations of Lancaster University. It has not been submitting in this form or similar elsewhere for the award of any other higher degree.

The main body of this thesis contains approximately 28,776 words

# List of Abbreviations

| | |
|---|---|
| GP | Gaussian Process/es |
| HPC | High performance computing |
| HEC | High End Computing cluster at Lancaster University |
| MCMC | Markov Chain Monte Carlo |
| MPC | Main Presenting Complaint |
| NLP | Natural Language Processing |
| NUTS | No U-Turn Sampler |
| NUTS 1 | Nomenclature of Territorial Units for Statistics |
| SAVSNet | Small Animal Veterinary Surveillance Network |

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Small Animal Veterinary Surveillance Network (SAVSNet)

The Small Animal Veterinary Surveillance Network (SAVSNet) is an organisation founded in 2008 by the British Small Animal Veterinary Association and University of Liverpool. Between 2012 and 2017, SAVSNet Ltd was a registered charity and has been the recipient of multiple grants throughout its time, which has allowed it to carry out important research related to animals. Their research priorities are currently surrounding antimicrobial use resistance, climate and environment, and infection and zoonosis [81]. A Dogs Trust funded research group within SAVSNet called SAVSNet AGILE links big and ever-expanding data resources based at SAVSNet to introduce deliverables that enable near real-time actionable health resources. This research group spans multiple universities with projects focusing on dog obesity at the University of Liverpool, stakeholder communication, and management at

Bristol University with this project exploring the narrative for syndromic surveillance using SAVSNet data. This project aims to use existing and newly collected data from SAVSNet-registered veterinary practices and laboratories to develop real-time anomaly detection tools. These tools will be accessible to SAVSNet and veterinary practices, supporting informed decision making.

## 1.2 The Importance of Veterinary Surveillance

Animal health surveillance is the monitoring of disease trends for the purpose of:

- ensuring the safety of the health and welfare of animals through facilitating the control of infection.
- Protecting the food industry and supply chain.
- helping with the prevention of zoonotic diseases.
- helping to inform policy decisions and
- highlighting economic implications for the owner of the animals (be it livestock or domestic). [4]

One could argue the majority of the reasons listed could be relevant to only livestock animals; however, health surveillance research within domestic animals is essential due to the mass amount of domestic animals in the United Kingdom and just how often the general public interacts with them. A true value for population of dogs in the UK is not known, but work by McMillan et al estimated there to be 12.96 million dogs in the UK in 2019 [62] with 31% of households owning dogs

and 26% owning cats as of 2023 [62]. The United Kingdom Surveillance Forum is a government organisation that overlooks the coordination and oversight of activities in the UK and defines surveillance as: "The systematic, continuous or repeated, measurement, collection, collation, analysis, interpretation and timely dissemination of animal health and welfare related data from defined populations. These data are then used to describe the occurrence of health hazards and to contribute to the planning, implementation, and evaluation of risk mitigation actions" [97]. Animal reporting is usually done through main presenting complaints (MPCs). An MPC can be defined as the primary reason an owner might take their animal to veterinary care. The list of MPCs can vary, but some common examples are generalized illnesses, gastroenteric issues, respiratory issues, or pruritus related issues. The systematic categorisation of symptoms to MPCs allows for the easier identification of unusual patterns or spikes through the comparison of standardised and easily comparable data [2].

## 1.3 Implications of MPC for pet health and public health

In the interest of implications for pet health, syndromic surveillance of MPCs allow for earlier disease detection through the analysis of trends, improved veterinary care due to the ability to pre-empt a seasonal change and thus have the medication to resolve symptoms and to help equip researchers information to create better more preventative medications e.g. vaccinations [59, 1].

## 1.4 Zoonosis Diseases and the Spread of Infection

The World Health Organisation defined zoonosis as any infection or disease that is transmissible from nonhuman animals to humans [105]. Zoonotic infections and diseases can be transmitted through contaminated food, direct contact, through water or the environment and can be bacteria, viruses, parasites, or unconventional agents [105]. Due to the close relationship humans have with animals, whether domestic or agricultural, they are considered a major threat to public health. Some examples of zoonotic diseases transmissible from domestic animals are rabies, avian influenza, Salmonellosis and Lyme disease [77]. Domestic animals can transfer infections and diseases through multiple routes, including direct contact, vector-bourne, and airborne. Some direct transmissions are through urine, faeces, skin, and saliva, along with indirect transmission through water bowls, bedding, and food [14]. Vector-bourne diseases can be spread through ticks, mosquitoes or fleas [41]. Finally, some diseases can be airborne and transmitted by respiratory droplets from infected animals. Although some transferable infections and diseases (rabies) can be prevented through vaccination and other methods, it is essential to monitor the health of domestic animals due to the constant threat to public health. Ensuring visibility on domestic animal health and monitoring outbreaks is essential as a preventive measure against the transmission of zoonotic diseases [61]. In addition, it protects the welfare of animals by being aware of seasonal trends, which can help vet surgeries prepare and reduce mortality and suffering rates [42]. From a research perspective, it can also highlight new and emerging diseases that may be caused by new pathogens, which are of concern to both animals and humans [83]. There is also a responsibility for compliance with International Trade Standard to prevent the spread of disease

across borders [79]. Finally, to protect biodiversity as domestic animal infections and diseases can be transferred to wild animals e.g. canine distemper has impacted wild carnivores [22].

## 1.5 Existing Anomaly Detection Methodologies

In this thesis, the term anomaly refers to an observation, or group of observations, that significantly deviates from the expected baseline of disease or health-related events within the monitored population [80]. Anomalies can occur for several reasons, such as random variation, reporting errors, or fundamental changes in the occurrence of the disease. In a statistical framework, anomaly detection methods aim to distinguish background noise from unusual patterns that may require further investigation. An outbreak, on the other hand, is a clinical and epidemiological term. The World Health Organisation defines an outbreak as 'the occurrence of cases of disease in excess of what would normally be expected in a defined community, geographical area, or season' [104]. In veterinary practice, an outbreak is generally recognised when groups of animals show similar symptoms or a confirmed pathogen within a specific time and location, indicating a common source or transmission route [25]. Not every anomaly indicates an actual outbreak, but promptly detecting anomalies is a crucial first step toward identifying and managing outbreaks. This distinction highlights the role of surveillance as both a statistical exercise in pattern detection and a public health tool to prevent and control outbreaks. With the interest of health research, anomaly detection methods are essential to monitor outbreak style patterns and investigate those that arise. It is important in the interest of safety, with results influencing decisions to be taken by government agencies or those alike, to keep prevalence levels

below a certain threshold. It is essential that these methods be robust and can deliver accurate results in a timely manner. This section briefly describes some popular anomaly detection methods.

### 1.5.1 Farrington Method

The Farrington algorithm was developed by Farrington et al. They use a Quasi-Poisson regression to adjust for seasonality and other trends, then is re-weighted to take previous outbreaks into account [30, 46]. The algorithm was first introduced due to the overdispersion of many surveillance data, and used for outbreak detection for applications reported at the Communicable Disease Surveillance Center (CDSC) [109]. One of the major challenges of this algorithm is the lack of information required to train it [107].

### 1.5.2 Stochastic Processes

A stochastic process is a group of random variables which are indexed by some variable $x \in X$. In the equation below, $y$ can be thought of as being a function of $x$ i.e. $y(x) \in \mathbb{R}$ and $X \subset \mathbb{R}^n$.

$$y = \{y(x) : x \in X\}$$

If $X = \mathbb{R}^n$, then $y$ is an infinite dimensional process. However, there only need to be finite-dimensional distributions in order to express the law of $y$

for all

$$x_1, \ldots, x_n$$

and for all

$$n \in \mathbb{NP}(y(x_1) \leq c_1, \ldots, y(x_n) \leq c_n$$

These determine the law of $y$. [102]

### 1.5.3 Gaussian Process methodology

A Gaussian Process is an unsupervised continuous stochastic process used for the prediction of data and provides a thorough method for classification. The mean and covariance functions determine the Gaussian Process [20]. The Gaussian Processes are flexible in that we can have multiple priors all with different sensible distributions and it would still be able to maintain structural integrity [10].

Gaussian Processes work by defining a prior over functions which convert to a posterior in the presence of data. A multivariate normal distribution need only be defined over the functions values at a finite arbitrary set of points, $x_1, ..., x_N$. They assume that $p(f(x_1), ...f(x_N))$ is jointly Gaussian with mean $\mu(x)$ and covariate $\Sigma(x)$ given by $\Sigma_{ij} = k(x_i, x_j)$ where k is a positive definite kernel function [64].

A multivariate normal distribution is a vector of normally distributed variables that describe the finite combination of functions [93].

### 1.5.4 Previous Anomaly Detection using SAVSNet Data

Previous work using anomaly detection performed by Hale et al, uses the SAVSNet veterinary consultation dataset to explore syndromic surveillance over a small period of time for the city of Salford. They modelled the presence of Gastroenteric consultations using a mixed model, with seasonality being a fixed component with the 'unexplained' being modelled stochastically [47]. They calculated the predictive

probability for each premise and day and declared outbreak if the probability exceeded a user specified threshold [47]. As they wanted the system to be ran in near-real-time and for computational reasons, they modelled the data on a moving 9-day window. They chose 9 days as this was long enough to capture temporal correlation [47].

## 1.6 Bayesian Statistics vs Frequentist Statistics

An initial starting point for this project would be to assess the differences between the two inferential statistical approaches. Therefore, we can progress with the type of statistics most suited to the research statement in question. Frequentist statistics works on the premise that a parameter is fixed and unknown. The results are expressed through confidence intervals, hypothesis tests, and p-values [31]. The theory of frequentist statistics is that it can perform without any reliability of any subjective probabilities. The predictions are directly related to the current experiment and do not consider external information about the system being studied [31]. Bayesian statistics, however, works on the premise that our parameters are random variables and are built upon opinions one has about a certain data [31]. A Bayesian approach is more appropriate for this project as I am working with real-time data; it is infinite and at the mercy of real-life biases. The Bayesian framework is especially valuable because of its flexibility. It can handle hierarchical models, complex spatio-temporal structures, and uncertainty propagation in ways that are often impractical in a purely Frequentist framework [34]. Additionally, Bayesian inference provides full posterior distributions rather than just point estimates. For this project, which involves monitoring large-scale, real-time veterinary data streams, Bayesian methods are particularly appealing. They can accommodate the complexity and irregularity

of surveillance data while supporting probabilistic anomaly detection.

## 1.7 Bayesian Statistics

Bayesian Statistics, introduced by Thomas Bayes in the 1770's, is primarily conditional probabilities, which is the probability of an event A happening *given* event B has already happened. Bayes Theorem is widely used amongst medical statistics as the likelihood of seeing false positives or false negatives is greatly increased.

### 1.7.1 Bayes' Theorem

Bayes Theorem is the calculation of the probability of an event occurring based on a similar outcome from a previous event. It is defined by the below formula

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1.1}$$

$P(A)$ is the probability of event A occurring

$P(B)$ is the probability of event B occurring

$P(A|B)$ the probability of event A occurring given event B has occurred

$P(B|A)$ the probability of event B occurring given event A has occurred

### 1.7.2 A Standard Regression Model

A regression model describes the relationship between a dependent variable with one or more independent variables. They are also able to highlight the strength of the relationship between the dependent and independent variables and can be heavily adapted to fit a statistical question where a relationship between variables Y and

X is to be determined. Regression models once ran, are also a very useful tool for prediction. Mathematically, they can be written as below:

$$Y = \alpha + \beta X + u \tag{1.2}$$

where Y is the dependent variable which we are trying to predict or explain, X is the independent variable of interest, $\alpha$ is the y-intercept, $\beta$ is the slope of the explanatory variable and $u$ is the error term or residuals [16].

### 1.7.3 Mixed Effects Models

Mixed effect models are statistical model that include both fixed and random effects. Fixed effects represent overall patterns we are interested in, whereas random effects account for the differences between groups or repeated measurements [87]. The most basic form of a linear mixed effects model is an extension of the linear model with the addition of random effects.

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_j + \epsilon_{ij} \tag{1.3}$$

where

$Y_{ij}$ is the outcome for observation $i$ in group $j$. $\beta_0$ and $\beta_1$ are fixed effects. $b_j \sim N(0, \sigma_b^2)$ is the random effect for group $j$ and $\epsilon_{ij} \sim N(0, \sigma_b^2)$ is the residual error [87].

Mixed effects models are particularly useful for analysing clustered or repeated data, making them a suitable choice for longitudinal studies [87].

### 1.7.4 Prior Distribution

The benefit of using Bayesian Statistics is the ability to assign values or distributions to variables within a model given any knowledge the researcher has surrounding the project [33].

In the context of anomaly detection, there exists much information around that allows us to make sensible decisions around priors and how informative they may be e.g. seasonality or trends.

Using equation 1.1, the prior distribution here on the right hand side of the equation is $P(A)$

### 1.7.5 Likelihoods

Likelihoods are also conditional probabilities and can be seen on the right-hand side of equation 1.1, $P(B|A)$. The likelihood is how plausible each set of parameter values is for generating the data we observed [34].

### 1.7.6 Posterior Distribution

The posterior probability is the resulting conditional probability from Bayes Theorem which updates given new information from the prior and likelihood distributions [56]. Looking at equation 1.1, it is the left hand side of the equation.

## 1.8 Markov Chain Monte Carlo

This section will discuss Markov Chains, Monte Carlo and in turn Markov Chain Monte Carlo with various algorithms within it.

### 1.8.1 Markov Chains

A Markov Chain explains the process of moving from one state to another in coherence with certain probabilistic rules. The main property of a Markov Chain is that state $t+1$ is dependant on current state t only [34].

Markov Chains are stochastic processes, but unlike other stochastic processes, must remain 'memory-less'.

### 1.8.2 Monte Carlo

Monte Carlo are basically simulations which are used primarily to evaluate integrals where analytic solutions are not possible or extremely laborious. Using the case of the expectation:

$$E[X] = \int x f(x) dx \tag{1.4}$$

for random variable X which has no analytical form. By taking a sample from the distribution X we can use the sample mean as an estimate of the theoretical mean. This approach is easy to use even for multi-dimensional distributions [34]

### 1.8.3 Markov Chain Monte Carlo

For most complex probabilistic models, the posterior distribution cannot be calculated exactly due to analytical intractability and computational constraints [73]. Therefore, we use approximation methods such as Markov Chain Monte Carlo (MCMC). MCMC sampling offers an array of algorithms for systematic random sampling from probability distributions [34]. An MCMC algorithm simulates a dependent sample

from the posterior distribution then the summary statistics can be estimated from the sample. There are various different types of MCMC methods for various research questions with popular algorithms explained below.

#### 1.8.3.1 Gibbs Sampler

A Gibbs Sampler is the simplest of the MCMC algorithms and follows the following 4 steps. It is able to obtain samples from multivariate distributions by successively and repeatedly simulating from conditional distributions of each of the components given other components [12]. The Gibbs Sampler is described as below:

**Algorithm** Gibbs Sampler Algorithm

1: Initialise with $\Theta = (\Theta_1^{(0)}, \ldots, \Theta_d^{(0)})$
2: For $i$ = 1,2,...,$n$
   a) Simulate $\Theta_1^{(i)}$ from the conditional $\Theta_1|(\Theta_2^{(i-1)}, \ldots, \Theta_d^{(i-1)})$
   b) Simulate $\Theta_2^{(i)}$ from the conditional $\Theta_2|(\Theta_1^{(i-1)}, \Theta_3^{(i-1)}, \ldots, \Theta_d^{(i-1)})$
   c) ...
   d) Simulate $\Theta_d^{(i)}$ from the conditional $\Theta_d|(\Theta_1^{(i-1)}, \ldots, \Theta_d^{(i-1)} - 1)$
3: Discard the first $k$ iterations and estimate summary statistics of the posterior distribution using $(\Theta_1^{(k+1)}, \ldots, \Theta_d^{(k+1)}), \ldots, (\Theta_1^{(n)}, \ldots, \Theta_d^{(n)})$

Discarding the first $k$ iterations is regularity with MCMC methods and is often called the burn-in period. These values are removed to reduce the effect that they have on the overeall inference [33].

#### 1.8.3.2 Metropolis-Hastings

Another MCMC algorithm is the Metropolis-Hastings algorithm, which is a fundamental method that forms the basis of many other MCMC approaches [33]. Metropolis-Hastings introduces an arbitrary transition probability $q(x, y)$ from which

the simulation is considered straightforward. This transition probability describes the density of moving from $x$ to $y$. The algorithm is defined as below:

**Algorithm** Metropolis-Hastings Algorithm

1: Given the current position, $X_n = x$, generate a value $y^*$ from the proposal distribution $q(x, y)$
2: Calculate the acceptance probability, $\alpha(x, y^*)$, given by:

$$\alpha(x,y) = \begin{cases} \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\} & \text{if } \pi(x)q(x,y) > 0 \\ 1 & \text{if } \pi(x)q(x,y) = 0 \end{cases}$$

3: With probability $\alpha(x, y^*)$ accept the value and set $X_(n+1) = y*$; otherwise reject the value and set $X_(n + 1) = x$
4: Repeat until desired sample size is obtained.

[33]

### 1.8.4 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo is a variation on the standard Markov Chain Monte Carlo that uses gradients to scale to high and correlated dimensions but continuous parameter space. It has become a more popular tool in Bayesian inference with its basis deriving from Hamiltonian dynamics in physics [66]. Hamiltonian Monte Carlo, unlike Metropolis Hastings uses Hamiltonian dynamics to path the chain instead of randomly exploring which may result in folding backwards [66]. Hamiltonian Monte Carlo can only be used to sample from continuous distributions [66]. In basic terminology, there is an introduction of a kinetic energy variable to transmit samples in a parameter space which gives a better sampling property than a Gibbs Sampler for example [5]. The effectiveness of the Hamiltonian Monte Carlo highly depends on the tuning of an approximation path integration method [5] with steps $L$ and step size $\epsilon$. The tuning of these parameters could be troublesome; if $L$ is too large, the

path will begin to retrace its steps, however, if it is too small there will be high auto-correlations between successive path values. Likewise, if $\epsilon$ is too large then there is the expectation of a low acceptance ratio [5]. This method requires trial and error with these parameters to ensure an efficient model, which will in turn take a long time to run the models and assess each of the trace plots [66]. The full algorithm can be seen below

#### 1.8.4.1 Hamiltonian Monte Carlo Equations

**Equations of motion**:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i} \tag{1.5}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i} \tag{1.6}$$

for $i = 1, ..., d$. For any time interval these equations define a mapping, from state at time $t$ to state at time $t + s$ .

**Potential and kinetic energy**:

$$H(q, p) = U(q) + K(p) \tag{1.7}$$

Where U(q) is the potential energy and K(p) is kinetic energy and usually defined as:

$$K(p) = \frac{1}{2} p^T M^{-1} p \tag{1.8}$$

Where M is a symmetric, positive-definite matrix, is often a scalar multiple of the identity matrix, and is often diagonal. Hamiltonian's equations can not be written as

follows for $i = 1, ..., d$ :

$$\frac{dq_i}{dt} = [M^{-1}p]i \tag{1.9}$$

$$\frac{dp_i}{dt} = -\frac{\partial U}{\partial q_i} \tag{1.10}$$

**Algorithm** Hamiltonian Monte Carlo algorithm

1: Initialize with $q = q_0$
2: For each iteration $s \geq 1$ :
a) Draw $p_s \sim N_d(0, M)$.
b) Use $q_{s-1}$ and $p_s$ to simulate the Hamiltonian dynamics and propose ($q^*$,$p^*$).
c) Calculate the acceptance probability, $\alpha(q_{s-1}, q^*)$, given by:
$\alpha(q_{s-1}, q^*) = \min\{\exp\{\mathrm{H}(q_{s-1}, p_s) - H(q^*, p^*)\},1\}$
$= \min\{\exp\{U(q_{s-1}) - U(q^*) + K(p_s) - K(p^*)\},1\}$
d) With probability $\alpha(q_{s-1}, q^*)$, accept the candidate value and set $q_s = q^*$; otherwise reject the candidate value and set $q_s = q_{s-1}$
3: Repeat until a sample of the desired size is obtained.

#### 1.8.4.2 No U-Turn Sampler

The difficulties within the Hamiltonian Monte Carlo, are rectified with the No-U Turn Sampler (NUTS) which is an extension of the Hamiltonian Monte Carlo, where the steps $L$ and step size $\epsilon$ are automatically tuned. NUTS builds a set of likely candidate points for $L$ using a recursive algorithm, which is inclusive of the target distribution. The NUTS sampler is able to recognise when it is beginning to retrace its steps and thus automatically stops [43]. NUTS, can therefore perform just as well as an effective Hamiltonian Monte Carlo without the need to hand tune the parameters thus reducing personal and computational time.

## 1.9 Aims

The aims of this project are to apply anomaly detection methodologies to multiple SAVSNet datasets to spot anomalies. Further, to investigate use of high performance computing for quick analysis of health data. Following this is a method of displaying the results for the veterinary surgeons, stakeholders and the general public with a thorough look into visualisation techniques.

## 1.10 Thesis Structure

The structure of this thesis will follow:

- Chapter 2 is an exploration of the datasets used within this research which compose of consultation, laboratory and natural language processing data.
- Chapter 3 is an exploration of Predictive Anomaly Detection methods with an application onto all datasets explored in Chapter 2. This approach was successful in detecting anomalies within the data defined by main presenting complaint (MPC) and location when the data were plentiful however struggled to show meaningful results when applied to smaller datasets.
- Following the results from the approach in Chapter 3, Chapter 4 takes a look into a different approach to anomaly detection by using mixed models only applied onto consultation data. This approach was also able to detect anomalies regardless of dataset size and was a vast improvement on the previous approach.
- Finally, using these results, chapter 5 of the thesis then moves to building an automatic surveillance system from consultation collection through to displaying

the results using a shiny app.

# Chapter 2

# Data

This chapter describes the various datasets used in this thesis, together with extraction and transformation steps to get them to a useable format for statistical inference models. In the following sections, 4 different datasets containing information on the prevalence of disease in small animals will be described, together with the processing steps required to clean them for analysis.

## 2.1 Datasets

The Small Animal Veterinary Surveillance Network (SAVSNet) is an organisation founded in 2008 by the British Small Animal Veterinary Association and University of Liverpool [81]. Between the years 2012 and 2017, SAVSNet Ltd was a registered charity (number 1149531) and has been the recipient of multiple grants throughout its time which has enabled it to perform important research surrounding in small animal veterinary epidemiology.

Their research priorities currently are; antimicrobial use resistance and developing

better antimicrobial stewardship, climate and environment from the linking of health records with climate, landscape and soci-ecosystem data to predict future risks, and infection and zoonosis by quantifying risk of infection in real-time to anticipate health messages [81].

Within SAVSNet is a Dogs Trust funded project called SAVSNet AGILE, which funded this thesis. The group is formed of 4 Universities being Lancaster University, University of Liverpool, University of Manchester and Bristol University. The aim of this project is to link big and ever-expanding data resources based within SAVSNet to develop analytics that enable near real-time actionable information for veterinary practitioners.

SAVSNet collect data from many of the largest laboratories in the United Kingdom and 15% of veterinary practices in the United Kingdom [82]. For a surgery to be eligable to apply for SAVSNet, they need to use RoboVet (Vetsolutions, Covetrus Software Services) or Teleos Systems Ltd software. The veterinary surgeon is met with a short questionnaire at the end of each consultation, which requests the main reason the animal was presented. An example of what this screen looks like can be seen in Figure 2.1. There is also an opt-out option for the owner if they do not want their information passed through SAVSNet.

For the purposes of this thesis, SAVSNet provided datasets covering syndromic disease data through the main presenting complaint (MPC) at consultation, endemic and exotic diseases through laboratory diagnosis, and natural-language processed topic data from free-text clinical records. All records were anonymised and aggregated at the region and weekly level.


Figure 2.1: The questionnaire presented to the veterinary surgeon at the end of each consultation which asks for the MPC for said consultation.

### 2.1.1 Main Presenting Complaint Dataset

The main dataset used in this research is the full consultation dataset of MPCs. Here, a consultation is defined as an appointment with a veterinary professional in order to obtain medical advice, either in person or telephone. This dates from $17^{th}$ March 2014 to the present. The variables it contains are owner postcode, practice postcode, age of the animal and the main reason that the animal was presented; gastroenteric issues, respiratory issues or pruritus. This thesis will focus on the 3 MPCs which are

gastroenteric, respiratory and pruritus. A full list and descriptions of the variables in this dataset can be seen in table A.2. The dataset has 362 practices over 800 premises. Since data collection began there is the chance for practices to have move to a different premise through the years, as well as larger practices having multiple premises simultaneously.

This full consultation dataset has records for 44 different species of animals, with the main consultations belonging to domestic animals: dogs at 66.4%, cats at 25.8%, and rabbits with 1.4% of the total consults. There are also consultations for chickens, snakes, and tortoises however, the focus for this thesis will be on dogs and cats only. Since data collection through to the end of the research period, $22^{nd}$ February 2023 there have been 8,444,942 records for cat and dogs.

Figure 2.2 shows the total weekly consultations up to February 2023. Between the $17^{th}$ March 2014 and January 2016 is considered the 'recruitment period' where SAVSNet were recruiting more veterinary surgeries hence the large growth in consultations. There is also another large influx of cases in mid-2018 which is the result of another recruitment drive at SAVSNet.

An obvious characteristic in the data is the change in consult frequency in March 2020 due to the Covid-19 pandemic which will be discussed in detail in section 2.2. Another large drop in consults for a week is that in week 30 ($25^{th}$ July - $31^{st}$ July) in 2022, this was due to an error in the SAVSNet data capture system which led to a reporting error.

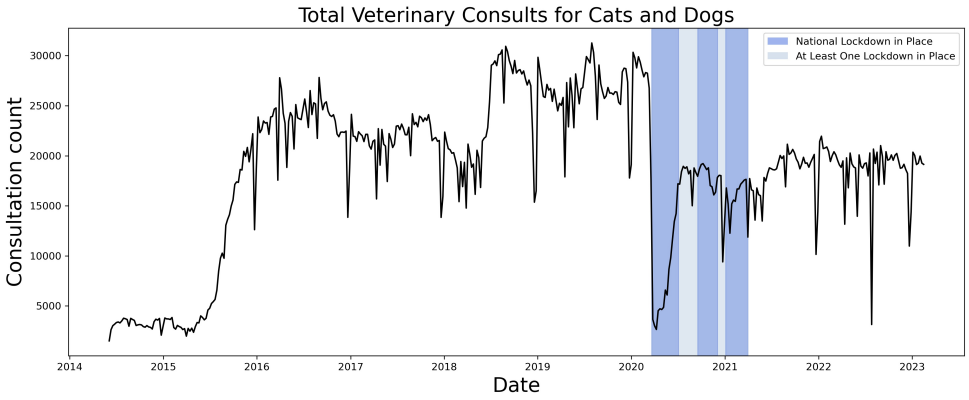



Figure 2.2: Total weekly consultation counts for Cats and Dogs since data origin with additional bars reflecting when the United Kingdom was in a national lockdown and when there was at least one lockdown restriction in place.

Figure 2.3 shows the spatial distribution of cat and dog consultations from veterinary premise locations and the owner postcode across the United Kingdom. shows the spatial distribution of cat and dog consultations from locations of veterinary premises and the owner's postcode throughout the UK. The figure illustrates a wide distribution of the data across the United Kingdom. At the time of writing this thesis (2023), SAVSNet had 362 registered practices across 800 premises. Although specific peer-reviewed sources on the exact number of veterinary practices in the UK are limited, estimates based on the Royal College of Veterinary Surgeons 2022 report suggest approximately 6,249 registered practice premises as of 2022 [100]. This means that SAVSNet has approximately 13% of practice premise data in the UK. Similarly, to practice information, the estimated population of dogs, cats and rabbits in the UK at the time of writing this thesis (2023) was 11 million, 11 million and 1.1 million, respectively (PDSA, 2023). The last accessible dataset that wasn't condensed into MPC consultations and total consultations only was from 2021, and the numbers of unique dogs, cats, and rabbits are 449,869, 290,229 and 34450, respectively. These values aren't necessarily the total animals registered to SAVSNet veterinary surgeries; these are simply animals that have consultation information.

The figure shows the spread of surgeries and owners throughout the whole of the United Kingdom. The consultations are unevenly distributed throughout the different official government regions, with the highest being in the South West (20.08%), then next is the South of England and North West with 13.71% and 10.92 % respectively. Lastly is London with 1.35% of consultations and Scotland and Northern Ireland with 1.96% and 4.45% consultations respectively. A table of the consultation split can be seen in appendix A.3

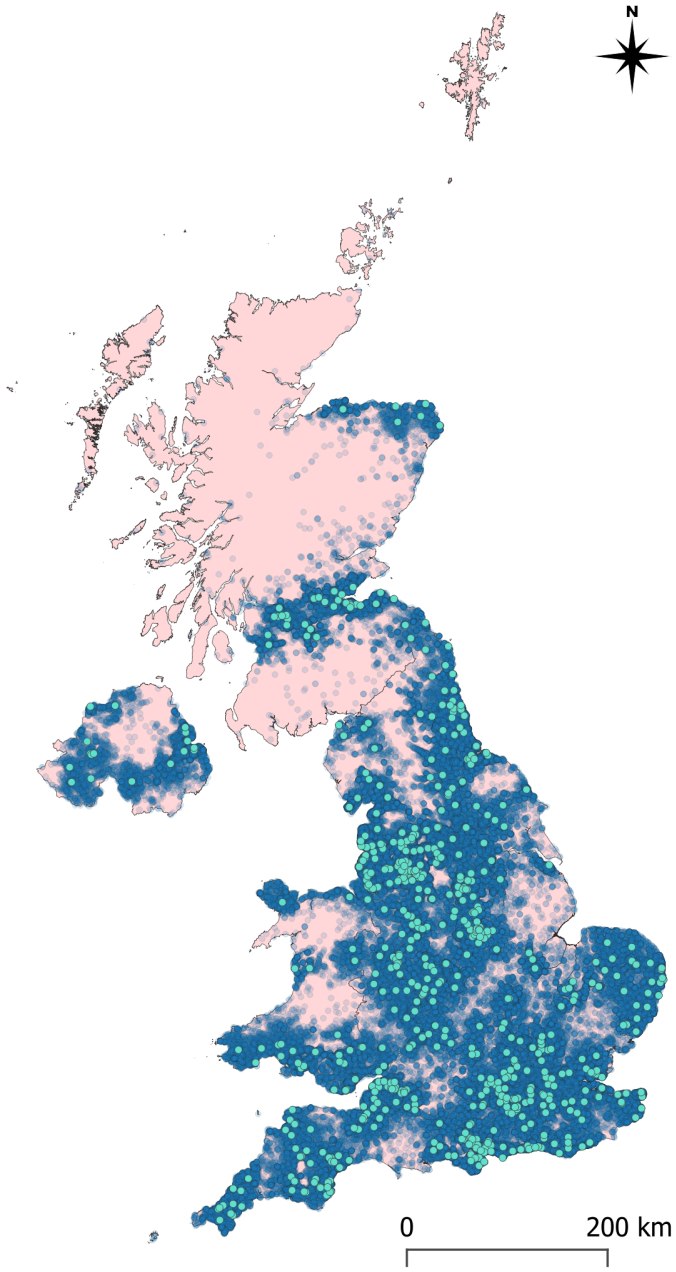



Figure 2.3: Points of the locations of the SAVSNet premises (light blue) and owner postcode (dark blue) across the United Kingdom from data origin.

Figure 2.4 shows plotted prevalence for the MPC's of interest from the data creation date ($17^{th}$ March 2014) where prevalence is defined as the total consultations labelled as said MPC per total consultations for that week. The beginning of each of the timelines for the prevalence is relatively noisy, which shows the previously mentioned recruitment period before it settles around 2017. For each of them, there can be seen to be a slight upward trend throughout the whole timeline, with the exception of a couple of large peaks that are best seen in dog and cat gastroenteric and cat respiratory. Additionally, seasonal trends are observed throughout the years. Which are most prominent in dog gastroenteric and dog pruritus.

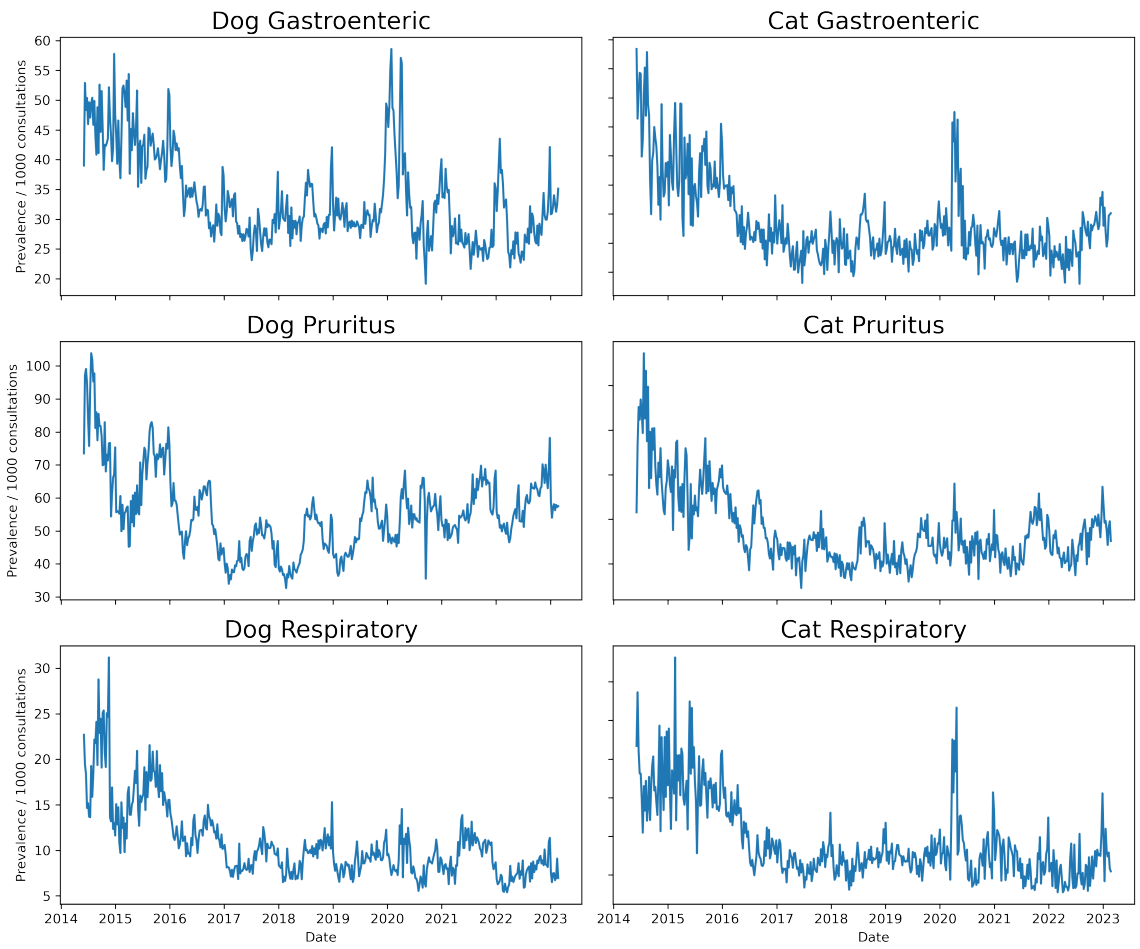



Figure 2.4: Prevalence for each of the MPCs for dog gastroenteric cases (top left), dog pruritus cases (middle left), dog respiratory cases (bottom left), cat gastroenteric cases (top right), cat pruritus cases (middle right) and cat respiratory cases (bottom right)

Given the instability within the data during the recruitment period and noting that this stabilises beginning 2017, the time period for the data use for the thesis will be from January 2017 through to February 2023.

### 2.1.2 Aggregated Main Presenting Complaint Dataset

In the interest of the pet owners' anonymity and data size when moving the dataset to an online storage system, data were aggregated by time and space. Data were aggregated at a local authority level to avoid privacy issues regarding unique identifiers with the counts for the MPC's of interest. With the idea of creating the surveillance system at Nomenclature of territorial units for statistics (NUTS) level 1, which reflects major socio-economic regions throughout Europe [27]. The aggregated dataset contained local authority codes with total counts of gastroenteric, respiratory and pruritus MPCs and a total count of consultations overall. By aggregating the data by local authority code, this allows us to project to various spatial levels if necessary. Aggregation was also done by weekly counts.

### 2.1.3 Natural Language Processed Data

Within the questionnaire the veterinary surgeon fills in, there is also an option for the input of free text. This gives them the ability to write further information about the consultation or the animal which the questionnaire does not cover. This free text may contain valuable information that is otherwise being missed through the MPC questionnaire for the consultation alone.

This thesis uses output from two text mining methodologies from analysis completed by researchers at the University of Liverpool and Durham University, in which topics were identified and extracted from the free-text clinical reports.

#### 2.1.3.1 University of Liverpool Natural Language Processed Dataset

This dataset was results created using an electronic health records annotation using latent Dirichlet allocation topic-modelling to assign different topics to the consultation, given the free text components [67]. Latent Dirichlet allocation (LDA) topic modelling assigns a probability distribution over topics for each consultation, rather than a single topic. This dataset is using the most probable topic per consultation. The dataset columns and descriptions can be found in Appendix A.4. The NLP1 dataset itself contains each consultation with a column for each of the 30 topics, with a binary value whether the consultation fits within that topic or not. Figure 2.5 shows the different topics and the main keywords that are associated with them from the algorithm. Further information about the algorithm can be seen in the paper [67] however, the statistical inference model results will be shown and discussed in Chapter 3.


Figure 2.5: The different topics and keywords associated with them obtained from the latent Dirichlet allocation topic modelling text mining algorithm.

#### 2.1.3.2 Durham University Natural Language Processed Dataset

The second natural language processing method uses PetBERT, which is a language model that has been trained on over 500 million words [29]. The purpose of the NLP2 dataset were to assess the method for identifying the January 2020 Canine Enteric Coronavirus outbreak using free text in a consistent process to the previous NLP1 data. Similarly, to the University of Liverpool Natural Language Processed dataset, this text mining algorithm works on a similar basis in that it assigns the most probable topic to the consultation. The dataset was split by dog, cat and both dog and cat consultations with there being columns added with a binary assignment whether the consultation fits into each of the labels. As it is only the assessment of the January 2020 Canine Enteric Coronavirus outbreak the labels are 'Digestive', 'Digestive and Infectious' and 'Gastroenteric mpc'. The final datasets I received were an aggregation of the results in the same style as the aggregated consultation dataset totalled by day with the dates spanning from $1^{st}$ March 2015 to $11^{th}$ October 2022. Figure 2.6 shows the counts for each of the labels for cats and dogs.

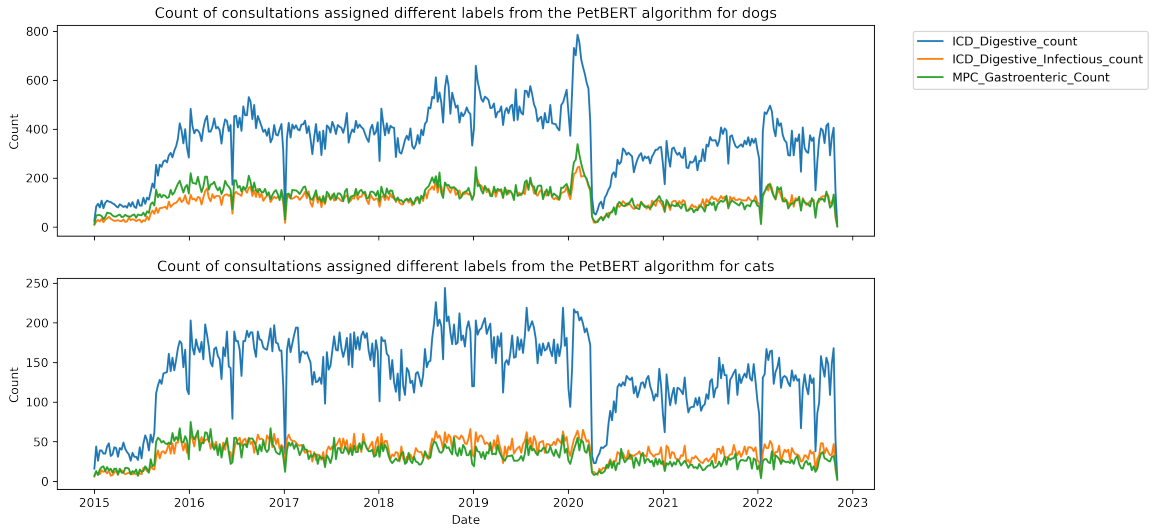



Figure 2.6: Counts of consultations labelled as digestive, digestive and infectious and gastroenteric from the PetBERT algorithm for dogs (top) and cats (bottom).

### 2.1.4 Laboratory Data

In a similar format to the consultation dataset, once an animal has displayed signs or symptoms of a specific disease or pathogen, samples can be sent for testing at a laboratory. The samples can be that of blood, faeces, urine and swabs [68]. SAVSNet obtains the data from seven laboratories across the United Kingdom and has received samples from those points seen in figure 2.7 which reflect owner postcode.

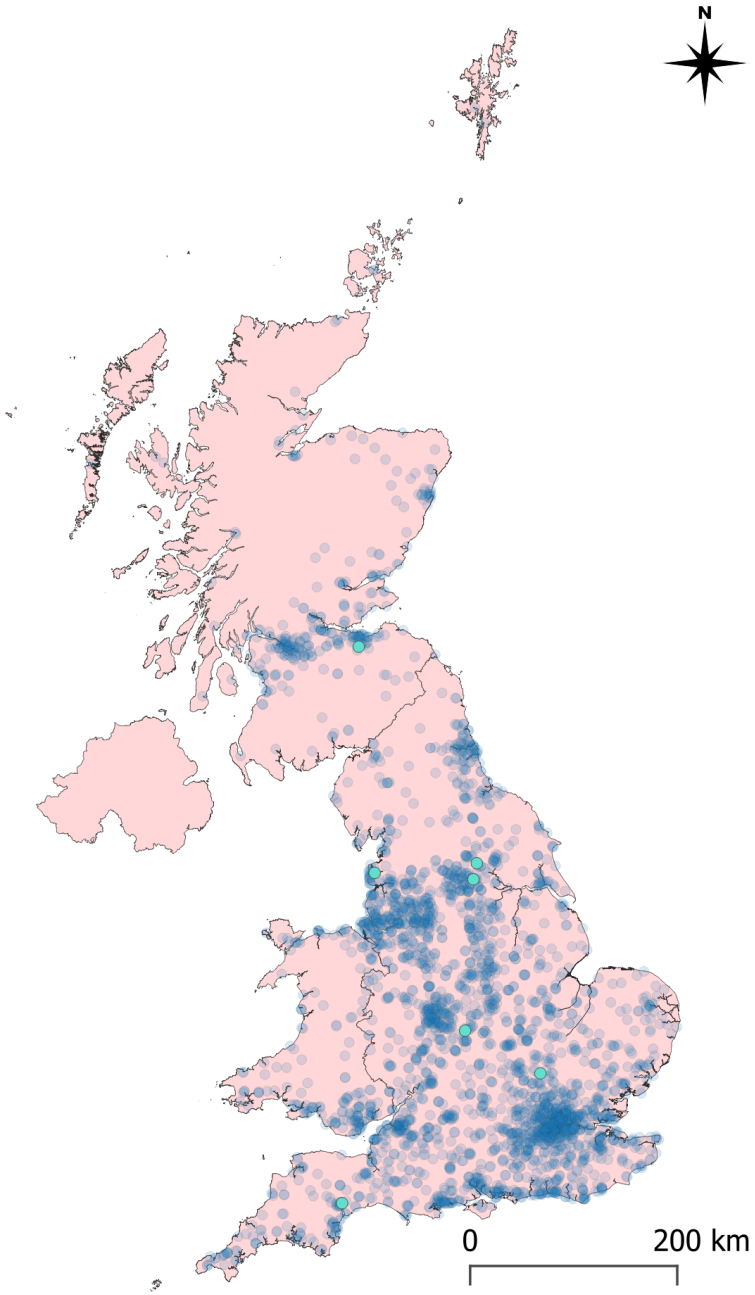



Figure 2.7: Locations of the owner postcodes (dark blue) for samples sent to laboratories (light blue) for testing.

The data, similar to the consultation data has the name of the pathogen/disease

tested and a binary column on whether that pathogen was found within the sample. The species within this dataset are dogs, cats and rabbits with the diseases being tested as follows: Angiostrongylosis (angio), Bordetella Bronchiseptica (Bb), Canine Adenovirus, Canine Distemper Virus, Canine Parvovirus (CPV), Crensoma, Feline Calicivirus, Feline Leukaemia Virus, Feline Herpesvirus, Feline parvovirus / panleukopenia (FPV), Herpes, Lepto, Lungworm, Myxomatosis and Parvovirus. As mentioned earlier, this thesis will focus solely on dogs and cats. The dataset contains both the laboratory postcode and the owner postcode, as well as the date the test was done (which is not necessarily the date that the laboratory received the sample), and the breed and age of the tested animal. A brief overview of the counts for the tested pathogen by species can be seen below in table 2.1.

| Species | Pathogen | Positive Results | Number of Tests |
|---|---|---|---|
| Cat | Angio | 0 | 22 |
| | Bb | 358 | 4174 |
| | CDV | 0 | 4 |
| | CPV | 0 | 5 |
| | FCV | 3801 | 16523 |
| | FELV | 405 | 18023 |
| | FHV | 797 | 17652 |
| | Herpes | 7 | 77 |
| | Lepto | 0 | 24 |
| | Lungworm | 0 | 178 |
| | Parvo | 9 | 842 |
| Dog | Angio | 153 | 5219 |
| | Bb | 414 | 1775 |
| | CAV | 15 | 2701 |
| | CDV | 105 | 9703 |
| | CPD | 661 | 8124 |
| | Crenosoma | 6 | 1836 |
| | FELV | 0 | 1 |
| | FHV | 1 | 66 |
| | FPV | 0 | 2 |
| | Herpes | 1 | 109 |
| | Lepto | 221 | 4327 |
| | Lungworm | 68 | 3198 |
| | Parvo | 116 | 2263 |
| Rabbit | Myxomatosis | 7 | 18 |

Table 2.1: Table of laboratory test results for cat, dog and rabbit for each of the diseases/pathogens with a sum and count of the results.

## 2.2 Covid-19 Pandemic effects

In March 2020, the World Health Organisation declared Covid-19 as a pandemic which means a worldwide spread of the virus [35]. As a way to control the disease, the Government and devolved administration across the United Kingdom introduced non-

pharmaceutical interventions, which are public health measures put into place with the aim of preventing the spread of the virus mainly through the form of lockdown [26]. The dictionary definition of a lockdown is 'a state or period in which movement within or access to an area is restricted in the interests of public safety or health.' [71] and each country took it upon themselves to introduce laws of enforcement for this. The now withdrawn UK Government advice for March 2020 regarding medical appointments for humans stated 'all routine medical and dental appointments should be cancelled while you are staying at home' [95]. As the lockdown and 'stay at home' orders were enforced, this led to cancellations of medical appointments. The British Veterinary Association (BVA), which supports more than 19,000 veterinary surgeons in the United Kingdom (approx 75% of the 25,400 employed veterinary surgeons [63]) created a guideline following that of the government advice given on human health, which was initially in place for 3 weeks starting $23^{rd}$ March 2020 pending review after this date. These guidelines stated that vaccinations, non-essential consultations, neutering and routine reproductive work should be suspended and/or delayed [11]. Furthermore, available assessments over telephone appointments were for mild trauma, skin issues and wounds alongside others. Although the BVA stated these to be suggested guidelines only following government legislation regarding human health, the uptake of these guidelines is reflected in the drop of consultations seen in figure 2.2. Throughout the lifting of restrictions in the United Kingdom, consultation levels increased.

So following this, there needed to be an indication of the COVID-19 pandemic due to the mass loss of consults visible in figure 2.2 for our choice of model to take into account if applicable. This dataset has a weekly binary value for whether that week

was in lockdown (1) or not (0). The dataset was created by myself using information from government policies throughout the pandemic [37, 19]. Here, the definition of a week being in a lockdown state was if there was at least one restriction within the country at any one time. A good example of this is when the English government implemented a tiered system, although only certain areas would be affected by non pharmaceutical interventions (Liverpool at tier 3 and London at tier 4 restrictions), England would be labelled as in a lockdown state. This was done to adjust for any change in public behaviours that were not being enforced by government restrictions. There was also, at times, restrictions in Scotland and Wales when England had loosened theirs and visa versa so the lockdown variable was also split by country. As our datasets used for the analysis are grouped by week, the same is true here, so if the majority of a week was under non pharmaceutical interventions, then the whole week was labelled as so.

## 2.3 Biases and Wrangling

This section looks into the biases within the datasets and any other issues that arise within them with the appropriate data wrangling techniques to eliminate any biases.

### 2.3.1 Human or System Errors and Bias

The common theme throughout all the datasets is that they require human input in some way or other. The consultation questionnaire (excluding information about the animal i.e. personal profile) requires human input in terms of inputting MPC reasoning or typing further information into the free text section. A further issue is

the veterinary surgeons bias. What one veterinary surgeon might feel a noteworthy amount of illness to mention about an animal or considers the main reason the animal was present, another veterinary surgeon might not feel necessary to mention or label differently. These systematic biases are difficult to remove however, due to the volume of consults we can assume that these biases be minimal and have little to no effect on the results.

For the natural language processing algorithm result datasets, the obvious issue for any text mining algorithm is misspellings and abbreviations. The veterinary surgeon could feel fatigue after a volume of consultations and in turn could result in standard human error in terms of misspellings. Further, there are abbreviations which are used in the veterinary industry e.g. PE: physical examination, RR: respiration rate and US: ultrasound. Although there are commonly used acronyms and abbreviations [17], there is no list of standardised acronyms that every veterinary surgeon adhears to so this could show a weakness with the algorithms. Following this, there are also acronyms which are similar, TNF (tumor necrosis factor) and TGF (transforming growth factor), so following a keyboard error from the veterinary surgeon, this could result in a misassignment from the algorithm. Bias of this nature can be taken to be random noise as there is the assumption that misspellings are few.

Although the systems seem robust enough and there is a constant stream of data fed, there is always the ability for the systems to malfunction or break which is evident in figure 2.2 during the week in July 2022 where there was a drop in consultation counts due to a system failure. Unfortunately, there's no way of determining just how many records there should be at this point but our inference model can 'borrow' information from the same time point from different years.

Finally for human biases comes from the owner of the animals themselves. There may be owners that take their animal the vet unnecessarily, or visa versa, contributing to the consultation count but we assume here these two events cancel one another out. Also, as of April 2023 we are still yet to reach pre-lockdown levels which could be reflective of the current cost of living issues [44]. As of 2023, the rise in the cost of living could lead to under reporting as veterinary services are expensive and not offered on any kind of government health service scheme.

### 2.3.2 Missing and Erroneous data

This section looks further into the data with assessments on missing data and other questionable values, with an emphasis on how it was dealt with.

#### 2.3.2.1 Main Presenting Complaint

The investigation into missingness within the MPC data was looked into until $10^{th}$ September 2021 where we then switched to the aggregated dataset which removed the ability to manipulate data this side of the pipeline. The findings can be assumed to hold true for the full, non aggregated data.

The main check here is for missing data within the dataset, as performing analysis on data where some are missing could introduce bias in the results and in turn lead to misrepresentation. An example of this could be if there were missing geolocation data for a specific region, this could lead to under-representation of a MPC and miss outbreak calls. Here, we're defining a missing consultation if there is at least one missing entry within the overall row. If there is a practice that had zero consultations that week then they will not be present in the data for said week. Out of a total

of 7,872,573 cat and dog records from data collection up to $10^th$ September 2021, 10.69% (841,867 rows) contain at least one missing field. Of these records, they all had postcode recorded which was then used as a proxy for precise geolocation data. This is considered to be missing completely at random (MCAR) as there is no systematic differences between those consultations with latitude, longitude and owner postcode and those without [58]. Even though 10.69% missing from a dataset is not overly large, its appropriate to investigate using visualisation methodologies and inputting information where able to.

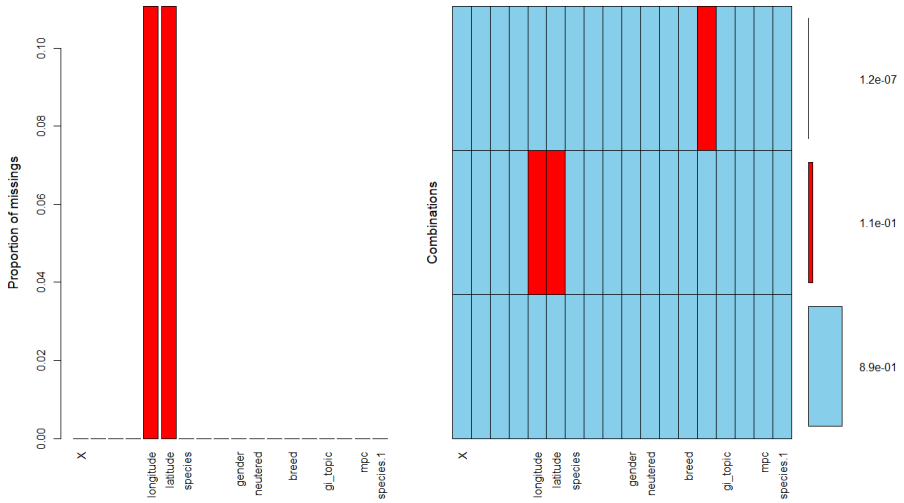



Figure 2.8: Plot of the missingness across the columns within the main consultation dataset using the VIM package. The left part of the figure displays the proportion of missing data whereas the right part of the figure shows combinations of missing data across rows.

Using the VIM package in R [54], we can visualise any other missingness within the dataset which can be seen in figure 2.8. When interpreting this output, the left side of the figure shows the proportion of missingness across the consultations, i.e. which columns had the largest proportion of missing data. The right side of the figure shows

us the combinations of missing data across the rows, i.e. how many rows that had missing geographical information also had missing owner postcode. From the figure it can be seen that for the MPC data between $17^{th}$ March 2014 and $10^{th}$ September 2021, there were 841,472 records with missing geolocation data. Details of this data can be seen below. A second, more minor issue with the missing data is that of owner postcode, where there are 76,621 missing. Any records lacking both postcode and latitude/longitude could not be geolocated and would have to be removed from the dataset, however from the right plot in figure 2.8, we can see that there are no overlapping combinations of these missing data.

The geolocation data was sourced from an open-access dataset obtained from Doogal [8]. This gave region and county information too so the analysis can be tailored and assessed at a more confined level if necessary. Bell, the creator and maintainer of Doogal uses information from Office of National Statistics and Royal Mail to curate a dataset with the full list of variables shown in appendix number A.5. Following the aggregation of the dataset, this removed the need for such a merge as we had all the local authority codes so there was no missing information that needed filling however, this requires a smaller dataset which has geospatial information from local authority codes up to the region.

Upon further investigation, there are also errors and anomalies in the ages of the animals e.g. for dogs the highest age recorded is 2020.64 years old - which is clearly an error. We restrict the age to a maximum of 25 years old, in line with a previous study by Hale et al. using the same dataset [40].

With the necessary merges and deletions, we're left with 99.03% of the data.

#### 2.3.2.2 Natural Language Processed Results Datasets

The NLP1 dataset was 5,655,032 records each representing a single consultation dating from $1^{st}$ January 2016 to $17^{th}$ February 2020. There is some missingness within the dataset. There are 2,328,914 consistently missing records which could imply a model error which yielded no information across any of the topics, including 'topic unknown'. With the removing of these missing values there are 3,027,435 (53.5%) consultations that remain. There are also latitude and longitude columns missing (481,849 rows) however, these have available postcodes to geolocate the columns. This missingness can be said to be MCAR for the same reasons as earlier, and given the remaining information in the dataset unable to make predictions to recover missing entries.

As the NLP2 dataset was received already aggregated, this removed the ability to check for missingness.

### 2.3.3 Laboratory Data

There is missing data within this dataset however, amongst columns that for the purposes of this analysis we have little to no interest in. For example of the 96,748 total samples for cat and dogs, there is 32,044 (33.1%) missing breeds in the data but as we are not looking across breeds we are able to ignore this. Further, there is also 230 (0.2%) missing pathogen test results and as we would be unable to locate these, will have to be removed. Finally, there are 1519 (1.6%) missing postcodes but as we have the full list of NUTS 3 (small regions) codes we are able to use the prior mentioned geospatial dataset to fill in the missing information. Once all of the missing data has been removed we are left with 94,999 sample entries (98.2%). The missingness can be said to be missing at random (MAR) as these missing fields could

be predicted given the other information in the dataset [58].

## 2.4 Spatial Dataset

The second type of spatial datasets used is for plotting. The previous datasets even though they have spatial information, they do not have the information for plotting. This dataset is digital vector boundaries for previously mentioned NUTS 1 territorial units in the United Kingdom obtained from the Open Geography Portal within Office for National Statistics [69]. Having this information allows us to spatially assess the data and results.

## 2.5 Discussion

Once the missingness within the datasets are rectified and aggregated to a weekly timestamp, they are suitable to be input to our methods of choice and analysed however, the question is raised as to whether it is worth analysing data. For example take the MPC data, we need to analyse this to assess weekly levels of syndromes nationwide. The benefit of this dataset is that it is received in real-time and the results reflect what is happening in that week. Now, considering the laboratory data, during busy periods, there may be a lag between the laboratory receiving the sample and the analysis, which doesn't reflect the current weekly rate at which the sample was taken. Samples may be forwarded to an external laboratory, as the veterinary surgery may not have in-house facilities for some illnesses that require advanced testing, or the samples may require the expertise of technicians or researchers [101]. An overwhelming weakness for all but the laboratory data is the reliance on human

input and the bias this introduces. From biases introduced from potential varying opinions on an animal to submitting spelling mistakes in the free text options. Another limitation with the datasets however, is the opportunity for duplicate entries given an animal's condition i.e. if a dog has gastroenteric illness for a couple of weeks and their owner brings them to the vets multiple times. The dataset does not have individual pet ID/Names with the closest unique identifier being owner postcode. Given this is the only unique identifier, it is difficult to say which consultations are different animals within the same postcode and which is the same animal having recurring visits. This also highlights a discussion in what amount of time is needed between consultation visits for an animal for the MPC's to be considered independent. The data is currently stored within the University of Liverpool servers with the need for a manual sending over to us at Lancaster University. Requiring the need for manual movement is time-consuming and can lead to errors occurring. The data should be set up to feed straight through to an online source so that all those required can access it. This will be discussed further in chapter 5.

# Chapter 3

# Predictive Anomaly Detection

The aim of this chapter is to assess some of the available predictive anomaly detection methodologies and choose the one appropriate for the data. The data, as explored in Chapter 2, are 4 datasets which contain main presenting complaint (MPC) data, Laboratory results and two datasets containing the results of different text mining algorithms. This chapter aims to explore the methodology of choice against the datasets with a thorough evaluation of the method.

## 3.1 Anomaly Detection Methodologies

Anomaly detection is important as it can help identify any rare events within different industries and can influence guidelines surrounding these anomalies e.g. constant surveillance during the Covid-19 pandemic and the influencing of the social distancing measures. The data in this research is real-time veterinary data from SAVSNet veterinary surgeries across the United Kingdom and contains counts of the 3 main presenting complaints of why an animal might be at their appointment. These are

gastroenteric, pruritus and respiratory. Anomaly detection in a healthcare setting is vital to prevent excess deaths and to reduce the spread of any illnesses.

Anomaly detection was initially used in the 1980's for the use of intrusion detection [23]. It was this work by Dorothy Denning that highlighted the impact of anomaly detection in real life applications. The next major use of anomaly detection was in the fraud sector with the first application on public health data in 2002 [23]. The different applications through various subjects has lead to many existing methodologies useful for this type of data.

An anomaly is defined as a point within the data that does not conform to the same behavioural characteristics as the other data points [94]. It is important for a methodology for anomaly detection, especially where health is concerned, to be robust and provide accurate results in a timely manner. Farrington et al define the aim of an anomaly detection model to detect any increases on top of expected trends or patterns [30]. My definition of an anomaly is an observation on any week that exceeds a 95% credible interval, whereas my definition of an outbreak is two consecutive weeks above the 99% credible interval. This is due to the probability of there being one week exceeding the 99% credible interval is 0.025 so two consecutive weeks above the credible interval will be calculated as follows:

$$\text{Outbreak Call } = \frac{\frac{1}{(0.025*0.025)}}{52} = 30.769 \text{ years}$$

This section will look through a couple of relevant outbreak detection algorithms which have been either linked to the dataset or will be close to giving the desired outcome of our algorithm following other applications of said methodologies.

### 3.1.1 Farrington Algorithm

The Farrington algorithm was proposed by Farrington et al in 1996 [30]. They highlighted that an outbreak detection algorithm should work in a timely manner with the correct sensitivity that false positives are kept to a minimum. The Farrington algorithm adjusts for trends and seasonality by using a Quasi-Poisson regression and is re-weighted to account for previous outbreaks [46]. One of the applications of the algorithm was on a cluster of cases in February 1995 of *Salmonella agona*. The application of the algorithm found the 'outbreak' flagged as early as mid-December 1994. Even though the algorithm is successful at observing anomalies within a dataset as 40% of the identified excess cases correspond to outbreaks [30], it is not an acceptable method for my research due to the sensitivity weakness as it has potential to learn from past outbreaks if not removed. As we're wanting to take a more automatic approach regarding the creation of a surveillance system, the option for the manual removal of outbreaks is not feasible. This also introduces the question of what an outbreak is and when it starts and ends.

### 3.1.2 Previous Work on SAVSNet Consultation Data

Previous work on SAVSNet consultation data by Hale et al [40] applied a Bayesian framework to predict Gastroenteric MPC for dogs and cats. They used a spatio-temporal mixed effects random model and modelled the presence of Gastroenteric MPC as mutually independent Bernoulli variables with probabilities defined as

$$\Phi^{-1}(p_{j,i,t}) = d_{j,i,t}^{T}\theta + S_{i,t} \tag{3.1}$$

where $j$ is the consultation at the $i^{th}$ premise on day $t$. $\Phi^{-1}\cdot$ is the Probit link. The expression $d^T_{j,i,t}\theta$ indicates a set of exploratory variables and their corresponding regression parameters. While $S_{i,t}$ is a spatio-temporal random effects value for premises and days. The vector S $\sim$ Multivariate Normal$(0, \Omega)$ where $\Omega$ is a spatial covariance matrix. The spatial matrix is constructed using Voronoi Polygons by assessing the distance between practices, once obtained each branch was assigned a distance-decay weight. Each $S_{i,t}$ was conditionally distributed given all other values of S.

Following this, they simulated from the Bayesian predictive distribution using an MCMC algorithm using vague priors so that there was minimal influence on the predictive inferences. The surveillance system created was intended to be ran in near-real-time however, there was the issue of computational power and thus, they decided to run the model on a nine-day moving window. This is a limitation as running it on this time frame, there is no ability to check for any seasonal components in the data.

We want to look at the annual trends in the data thus this method will not be appropriate and can be improved upon in order to spot anomalies on a longer temporal scale.

### 3.1.3 Gaussian Processes

Another popular methodology for both prediction and classification is Gaussian Process (GP) regression. This methodology has grown increasingly popular over the years due to its flexibility and robustness [78]. GP's are non-parametric models, meaning they do not assume a fixed functional form for the data, but instead place assumptions on the structure of the function through the choice of kernel [78]. With

one of the focuses of this work being to capture annual trends, GPs provide predictions based on observed data and quantify uncertainty through their posterior distribution, making them robust learners from past observations [78]. GP applications have increased, with some of the newer applications being on COVID-19 data. Given an overview of the most relevant methodologies for outbreak detection, it seems the GP is the most promising, reliable, and up-and-coming approach in the health sector. Thus this will be our first methodology of choice to analysis the data at hand.

## 3.2 Introduction to Gaussian Processes

We have been presented with a problem in which we have observations $y$ and are wanting to create predictions $\hat{y}$ based on the training data/observations. Given we're using a Bayesian approach, and the problem is 'inductive' [78], we are able to use a function $f$ to predict unobserved values for all possible inputs of $y$. Applying a function rather than working with the data directly allows us to apply any underlying assumptions about the characteristics. Rasmussen and Williams detail two different approaches to this problem and highlight the issues that accompany them. Their first mentioned approach is to restrict the functions used to that of linear functions only for simplicity. However, rarely do real-life data follow a strict linear trend thus the predictions will be inadequate [78]. Their second approach is a theoretical application of a prior probability to *every* possible function however, this could become subject to over-fitting to a training dataset thus performing badly when applied to a test dataset [78]. The second approach has a progressive ideology in theory, but applying separate priors to an infinite amount of functions is time and resource expensive, this is where a GP provides a solution. GP's are stochastic processes which are a

generalisation of the Gaussian probability distribution [78]. They provide a flexible framework given multiple priors each with different sensible distributions whilst still maintaining structural integrity [10]. Following the definition of stochastic processes in the introduction, GP's are a generalisation of multivariate distributions across infinite dimensions [102].

Only the mean and covariance functions of a GP are needed to specify the law below

$$y(x) \sim \text{MVN}(\mathbb{E}(y(x)), \mathbb{C}ov(y(x), y(x')))$$

The mean can be chosen within a rule set mainly through a process of 'trial and error' however, the most common choice for a basic model would be 0 as GP's have the flexibility to model the mean well. The covariance function describes similarities between data points [78]. The GP learns the following from the covariance function: smoothness, differentiability and variance. With a GP, there is also the assumption of stationarity which provides an appropriate level of simplification without excessively reducing the generalisation [21]. A GP function is said to be stationary if:

- m(t) = $\mathbb{E}y_t = \mu$ is independent of time point $t$ and
- $\mathbb{C}ov(y_{t+h}, y_t)$ is independent of $t$ for all $h$. [21]

GP's are flexible robust methodologies which are mathematically tractable. They are also able to give reliable estimates of their own uncertainty which make them an adequate methodology for anomaly detection.

## 3.3 Literature

Brahim-Belhouari and Bermak used a GP for the prediction of non-stationary time series data. They used a respiratory signal dataset which consisted of 725 samples representing a recording of the respiratory rhythm via a thoracic belt [10]. The dataset was split to train and test data to assess a GP model with an exponential covariance function and a non-stationary covariance function. They found that the predictions created by the GP with the non-stationary covariance function match reasonably well with that of the real data patterns whereas the model with the exponential covariance function deviates significantly when applied to the test data [10]. Research performed by Hubin et al [45] applied a Binomial regression model using a latent GP in order to model DNA methylation. They modelled the spatial dependence of methylation probability from within a pool of cells using a Binomial regression model and a Logit link function. They also approached this in a Bayesian framework and included prior specifications in the model design. Their MCMC method of choice was a Mode Jumping algorithm and upon this calculated the individual marginal inclusion probabilities for the covariates. They found the methodology to be successful at identifying methylated dense regions and made use of the marginal posterior probabilities to assess each of the variables fits [45]. Research by Ketu and Mishra, 2020 used a Multi-Task GP regression model for the Covid-19 outbreak. The main purpose of this proposed model was to predict the Covid-19 outbreak worldwide to assist countries in planning their preventative measures. They used data from the World Health Organisation from 213 affected countries and used time series data between 31/12/2019 and 25/06/2020 to create their forecasting model [51]. The result obtained from the GP model was compared to other predictions models to assess the

results using the mean absolute percentage error and root mean squared error [51]. The other models chosen for comparison were Linear Regression, Random Forest Regression, Support Vector Regression and Long Short Term memory with the use of mean absolute percentage error and root mean square error to measure performance. With these values they found their proposed model to outperform the other prediction models and have stated the GP as a stable solution [51]. Another application of GP regression on Covid-19 data was performed by Faricha et al [28] where they assessed the spread of Covid-19 in Indonesia. Their dataset contained three cases, positive cases, recovered cases and death cases. They split each of the cases into test and validation sets. They used 2 models and compared the results. The first model they used was an optimisable support vector machine and an optimisable GP regression. Upon assessment of the root mean squared errors for both methodologies for the 3 separate cases, they found that the GP regression performed substantially more effectively than the support vector machine with an average RMSE of 19.54 for the validation set and 15.85 for the validation training set [28]. This also coincides with the findings of Ketu and Mishra in regard to a GP and support vector comparison. The literature relating to GP's is dominated by applications on time series data surrounding climate predictions, stock pricing and human health. By extension, there is an emerging cause for its application in the study of animal health.

## 3.4 Absorbing the Loss in Consultations Caused by Social Distancing Measures

As previously mentioned in the data chapter, during our period of analysis the presence of social distancing restrictions due to Covid-19 had a marked effect on the apparent prevalence of MPC. The cancellation of routine consults (e.g. vaccinations and health checks) reduced $N_t$ for weeks during which restrictions applied, however emergency consults for different illnesses had still taken place. Upon the first attempt of trying to salvage as much data as possible, we needed to increase the apparent prevalence of MPC during the affected weeks - which initially began with re-evaluating the calculation for prevalence. The usual calculation for prevalence was

$$\text{Prevalence} = \frac{\text{Total count of MPC of interest}}{\text{Total consults}}$$

whereas the new proposed prevalence calculation was

$$\text{Prevalence} = \frac{\text{Total count of MPC of interest}}{\text{Total count of all Unwell MPC consults}}$$

This was investigated within the GP model application for the no lockdown variable state of the model, mainly to test the limitations of our data and how the GP would perform with what would be substantially fewer data points.

## 3.5 Bayesian Model

This section describes the Bayesian logistic latent GP model we use to analyse each of the datasets described in Chapter 2.

A decision within the creation of a model is choosing the appropriate distribution to model our observations against. Our data is discrete at regular time intervals, so an appropriate distribution to model the observations would be either Poisson or Binomial. The pair behave in similar ways however there is a significant difference. The binomial distribution describes binary data from a finite sample, whereas Poisson describes it from an infinite sample. The binomial distribution therefore gives the distribution of getting $x$ events out of $N$ trials where the Poisson distribution describes the probability of obtaining $x$ events in a population. As we have a finite sample of $N$ cases, modelling the observed data $y$ using a binomial distribution is the most appropriate. [34]

### 3.5.1 Consultation Dataset without Lockdown Variable

To model the prevalence of MPC consultations in the United Kingdom, we use a longitudinal latent GP model with Binomial observation process. We assume that we observe $y_t$ MPCs out of a total of $N_t$ consults per week $t = 1, ..., 53$ spanning from 01/01/2019 up to real-time. We model $y_t$ as a Binomial random variable such that

$$y_t \sim \text{Binomial}(N_t, p_t) \tag{3.2}$$

where $p_t$ is the probability of a consult in week $t$ being an MPC, with the log odds of being an MPC in week $t$ modelled as a linear combination of terms as described below.

$$\log\left(\frac{p_t}{1-p_t}\right) = \alpha + \beta t + u_t \tag{3.3}$$

where $\alpha$ is the mean log odds of an MPC consult, $\beta$ represents a linear time trend capturing long-term drift in MPC prevalence (the effect on the log odds of MPC for a 1 week increase in time) and $u_t$ represents a time-varying random effect. We model the linear trend as a fixed effect as we expect a gradual rise in MPC consults over time due to improvements in reporting and the recruitment of more SAVSNet veterinary surgeries, as well as underlying drivers that make outbreaks more common year on year.

The random effect ut allows us to model periodic correlation in our weekly observations, as well as any extra Binomial variation that might contribute to the overall variability of cases from one week to the next. We model the vector $u$ as a GP with mean 0 and covariance matrix $\Sigma^2$ such that;

$$\mathbf{u}_t \sim \text{MultivariateNormal}(0, \Sigma^2) \tag{3.4}$$

The covariance matrix $\Sigma^2$ captures the correlation between two variates $u_t$ and $u_s$ spaces $s - t$ weeks apart, and we assume the correlation follows a periodic function and is seen in equation 3.5.

$$\Sigma^2_{ts} = \begin{cases} \sigma^2 exp\left(-\frac{sin^2\left(\pi|s-t|\frac{1}{52}\right)}{2\phi^2}\right), & \text{if t} \neq \text{s} \\ \sigma^2 + \tau^2, & \text{if t} = \text{s} \end{cases} \tag{3.5}$$

We justify the periodicity following the clear seasonality displayed in the plotted prevalence of the MPC consultations which can be seen in figure 2.4 in section 2.1.

### 3.5.2 Consultation Dataset, Laboratory Dataset, NLP1 and NLP2 with Covid-19 Lockdown Variable

The addition of the Covid-19 lockdown variable alters 4.7 ever so slightly by letting

$$\log\left(\frac{p_t}{1-p_t}\right) = \alpha + \beta_t + \delta z_t + u_t \tag{3.6}$$

where $\delta$ represents an offset in the log odds for MPC for weeks in which social distancing was imposed.

The same model is also used for the topic modelling and laboratory data implementation of the methodology.

### 3.5.3 Models

Model 1 is composed of equations 3.2, 3.3, 3.4 and 3.5 and will be referred to as model 1 when discussing through the results of this model application. Model 2 is composed of equations 3.2, 3.6, 3.4 and 3.5 and will be referred to as model 2 when discussing through the results of this model application.

### 3.5.4 Priors

The priors are relatively uninformative except for $\phi^2$ (our length-scale parameter) which we set equal to 0.32. The choice of the length-scale parameter is important as this describes how smooth a function is. We need a length-scale that is small enough to follow the data closely but one large enough to allow the visibility of outliers. An assessment of different values of $\phi$ on the Dog Gastroenteric MPC data can be seen in section 3.8.1 but the final value decided was 0.32. $\sigma$ and $\tau$ we also restrict to being

strictly positive. Our likelihood for the model including the lockdown variable is in equation 4.11. Noting that $\delta$ is only present in the lockdown variable models.

$$\pi(\alpha, \beta, \sigma, \tau, \delta | y, t) \propto \pi(y | \alpha, \beta, \sigma, \tau, \delta, t)\pi(\alpha)\pi(\beta)\pi(\sigma)\pi(\tau)\pi(\delta) \tag{3.7}$$

$$\pi(\alpha) \sim \text{Normal}(0, 1000) \tag{3.8}$$

$$\pi(\beta) \sim \text{Normal}(0, 100) \tag{3.9}$$

$$\pi(\sigma) \sim \text{HalfNormal}(5) \tag{3.10}$$

$$\pi(\phi) \sim 0.32 \tag{3.11}$$

$$\pi(\tau) \sim \text{HalfNormal}(5) \tag{3.12}$$

$$\pi(\delta) \sim \text{Normal}(0, 100) \tag{3.13}$$

## 3.6 Implementation

The models were ran in Python using package PyMC3 [70] and ran for 6000 MCMC iterations, discarding the first 1000 as a burn-in period. Burn-in periods are used for convergence as it allows the algorithm time to reach its equilibrium distribution. PyMC3 uses the built-in No u-Turn Sampler as its default algorithm. A No U-Turn Sampler is an extension of the Hamiltonian Monte Carlo with its main benefit being that it is able to tune the number of steps L and the step size parameter $\epsilon$.

## 3.7 Prediction

For each iteration, the priors are sampled from each of the distributions listed above and the Gaussian Processes are constructed. These variables are then used, along with the existing data to create posterior predictions from our Binomial distribution. For the creation of the results plots, these posterior predictions are summarised at the 1%, 5%, 50%, 95% and 99% quantiles with the observations colour-coded by tail-probability. Using this colour coding allows the easy visibility of where the observed week value is sitting within or outside the credible intervals.

## 3.8 Results

This section will look into the results of the various applications of the GP methodology on the above mentioned datasets along with comparisons.

### 3.8.1 Length-scale Value

As mentioned in section 3.5.4, there was a trial of different length-scale values on the dog gastroenteric MPC dataset. These values were tested on the data without the Covid-19 lockdown variable (model 2). Figure 3.1 shows the different values of the length-scale being 0.16, 0.32 and 0.64. All of the values show the seasonal trend that follows the pattern of the data with a larger spike around December-January time. The difference between them all is how influenced they are by the smaller spike around Summer. The smaller length-scale value, 0.16, is more influenced by this influx of cases whereas the larger value of 0.64 underestimates the summer cases. 0.32 visually, has taken enough of an influence to alert of a summer spike while not

directly following the pattern of the observed data, therefore 0.32 is the best fit from these tested values.

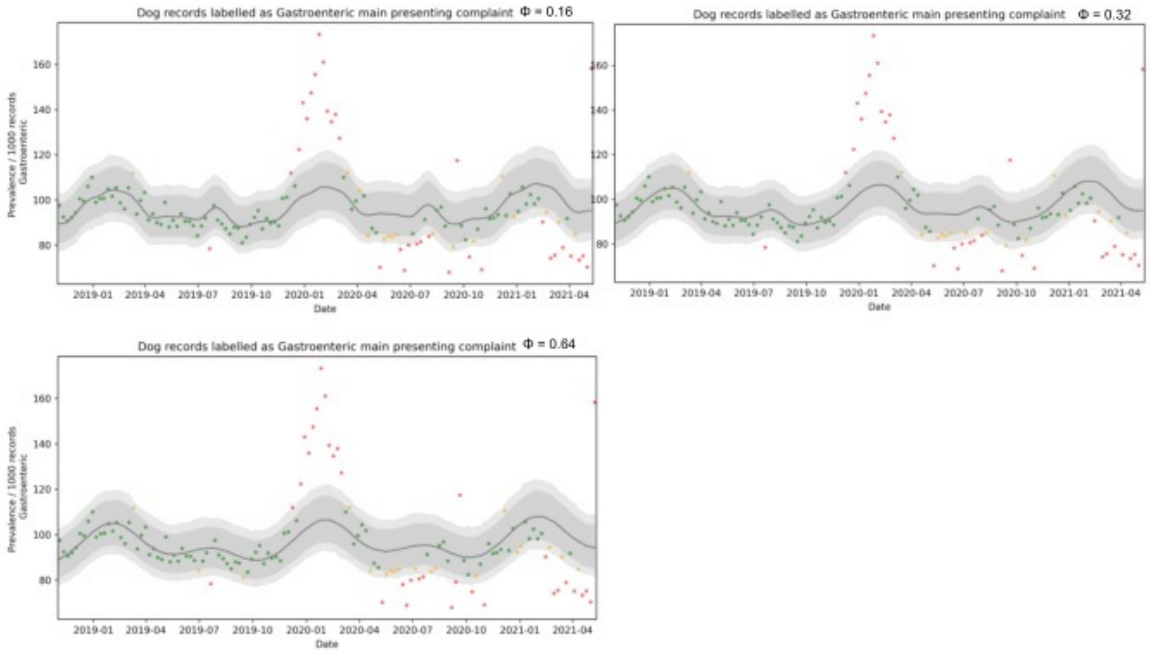



Figure 3.1: GP analysis ran on dog gastroenteric MPC with different values of $\phi$. $\phi$ values of 0.16 (top left), 0.32 (top right) and 0.64 (bottom left).

### 3.8.2 Absorbing Loss of Consults

In order to fully assess the results from the different alterations of the model our use case will be dog Gastroenteric MPC from January $1^{st}$ 2018 up to $14^{th}$ December 2020 as within these dates we have what would be considered 'normal' behaviour up until November 2019 where we have a confirmed outbreak of Canine Enteric Coronavirus until end March where we begin to see the effects from our first lockdown. Figures 3.2, 3.3 and 3.4 show the different applications of the GP on the different prevalence calculations using model 1 and the original prevalence calculation using model 2. In

this case 'no alterations' refers to there being no changes to the data pre-analysis. One thing that is clear from all of them is that they are able to detect the January 2020 outbreak relatively clearly. There is an added uncertainty within model 2 due to the extra variable but it is still able to show an outbreak-style pattern. The second, larger issue is that the second peak (April 2020) just after the first outbreak style pattern, is easily detected by both model 1 applied to the source data and model 2 however, is not being highlighted in figure 3.3. This is due to the use of unwell consultations as the denominator for the prevalence calculation and therefore there being minimal difference between the count of gastroenteric MPC and unwell consultations. There is also a difference in the predicted prevalence after we go through the initial lockdown state. It can be seen from both the unwell consultation prevalence calculation in figure 3.3 and no alterations to the data seen in figure 3.2 that the drop in consults has meant the predicted prevalence is being estimated to be below the 1% credible interval whereas the lockdown variable seems to salvage the lost data and is also able to spot outbreak style patterns, so there is justification for keeping the lockdown variable.

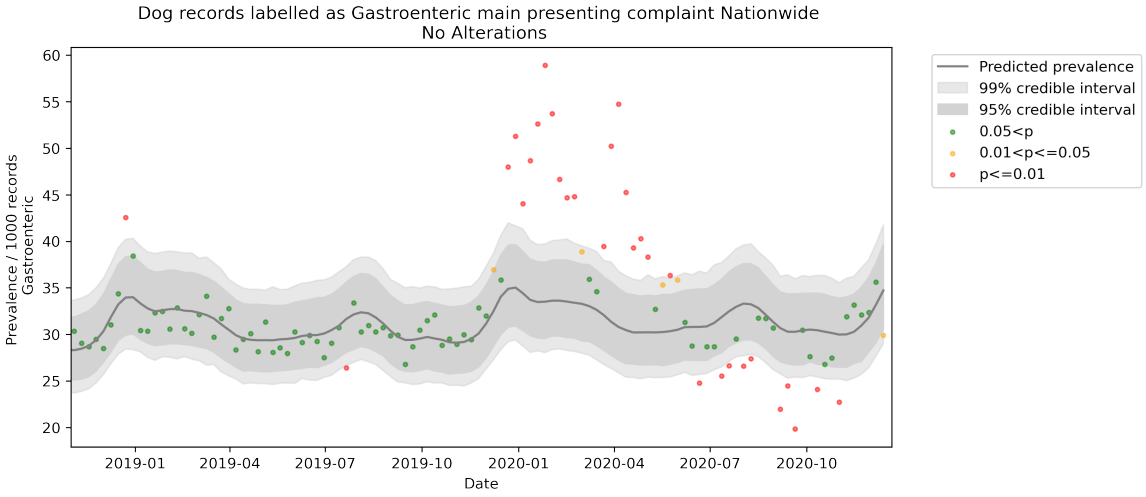



Figure 3.2: GP ran on consultation data for dog gastroenteric MPC nationwide using model 1 using the total consultations as the denominator for the prevalence calculation.

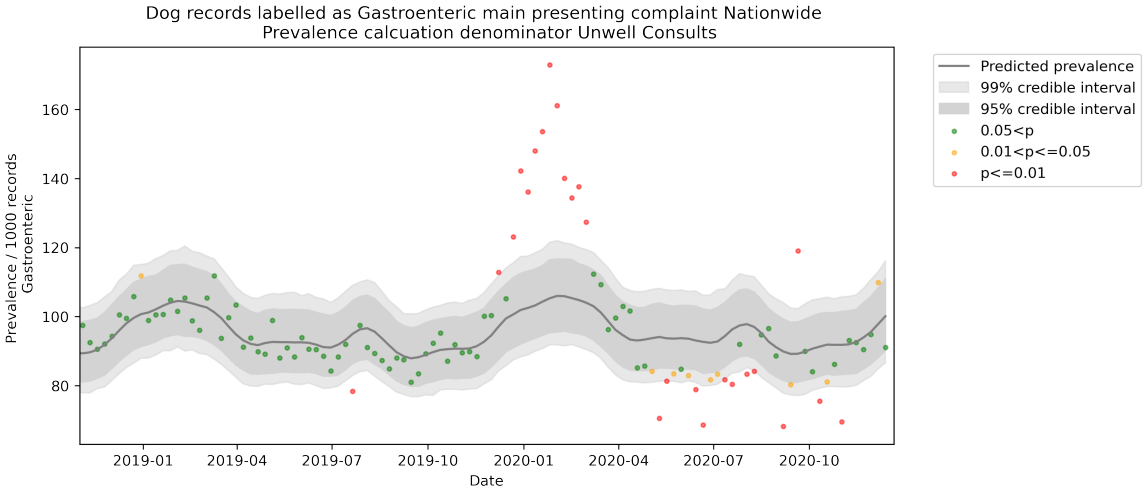



Figure 3.3: GP ran on consultation data for dog gastroenteric MPC nationwide using model 1 with using unwell consultations as the prevalence calculation denominator.

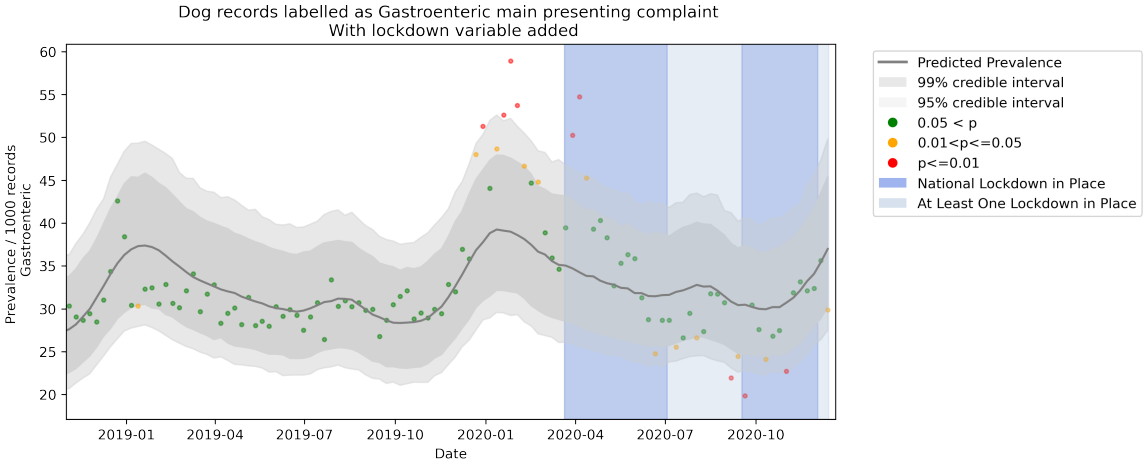



Figure 3.4: GP ran on consultation data for dog gastroenteric MPC nationwide using model 2 with the prevalence calculation using total consultations as denominator.

### 3.8.3 Consultation Dataset

This section will assess the results of GP on the consultation dataset, initially split by MPC and species and looked at nationwide, with a further split by region.

#### 3.8.3.1 National Levels for Each of the MPC and Species

The plots in this section are the GP results for each of the MPCs for dogs and cats at a national level and can be seen in figure 3.5. From a visual inspection, it is clear to see that they have performed well. There appears to be seasonality within all the MPCs for each species. The strongest seasonality is Pruritus which affects both species mainly around the summertime.

The gastroenteric MPC has a stronger seasonality within dogs than cats which is surrounded by two peaks a year in summer and winter. These peaks coincide with the veterinary surgeon's beliefs that dogs are more likely to have access to food that will make them ill during these periods e.g. barbeque food and raw meats in summer and chocolate at Christmas. [86, 15]

In terms of outbreak style patterns, the most notable is the dog gastroenteric MPC in January 2020, which was later confirmed from samples from the dogs to be a Canine Enteric Coronavirus [76]. However, other notable outbreak-style patterns are in dog respiratory between April and October 2021 and cat gastroenteric and pruritus both having outbreak-style patterns in March/April 2020. There is currently no evidence from samples surrounding a pathogenic or viral outbreak for these, however, it could be a rush of appointments at veterinary surgeries following the news of the pandemic and a reflection of that.

Another more severe mention of these results from a visual perspective applied

to cat gastroenteric, cat respiratory and dog pruritus, in that during the addition of the lockdown variable at these times there seems to be a sharp shift in the predicted prevalence. This could be due to a smaller amount of data in these criteria than the others, as we can see with dog gastroenteric we do not have this issue so visually. This begs the question of how well the GP in its current state is able to deal with data that is significantly smaller in size.

Looking through the trace plots in appendix section A2 it is clear that each of the prior predictive distributions in each application has performed well and looks to have fully explored the distribution assigned.

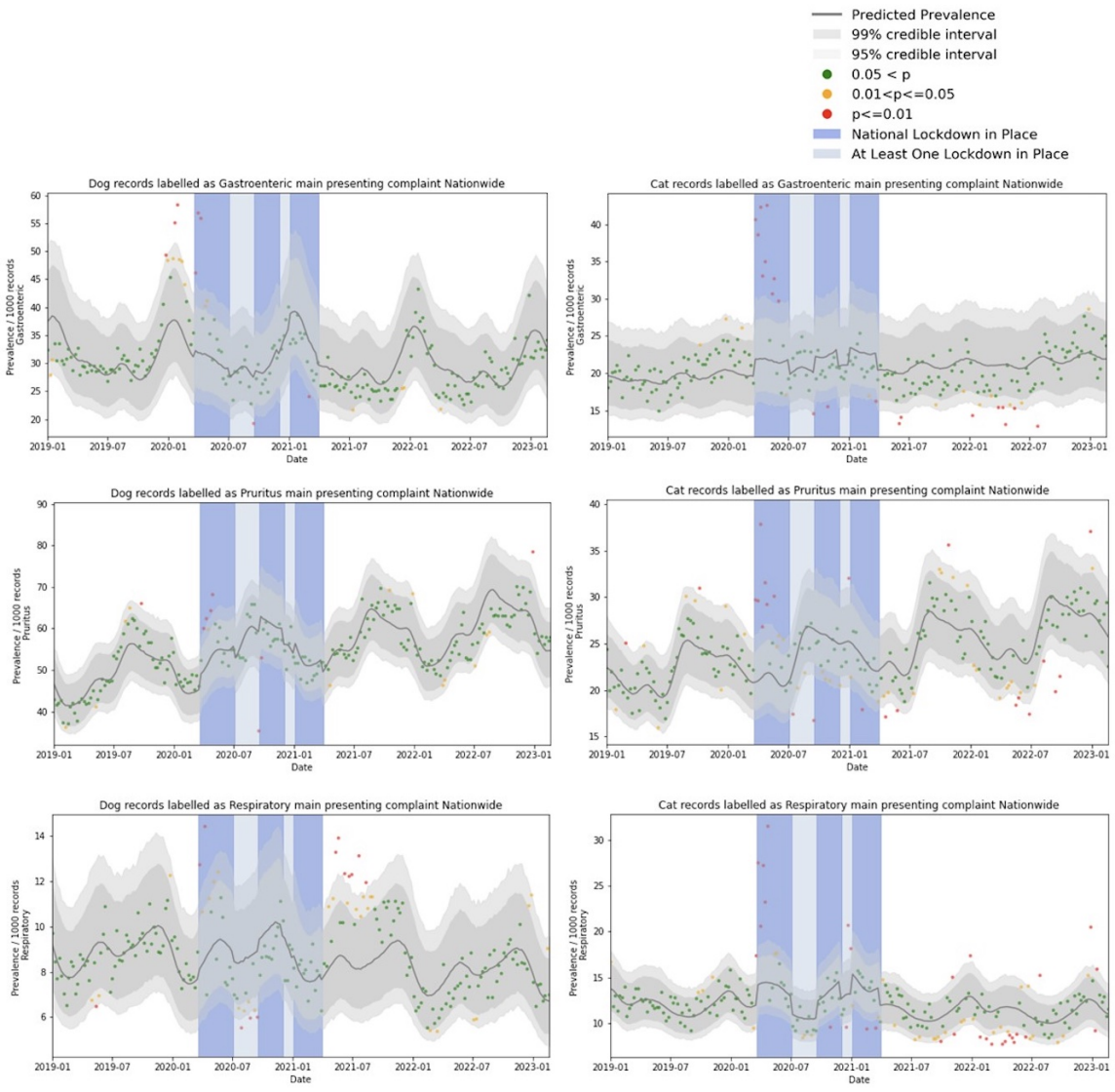



Figure 3.5: Gaussian Process results for dog (left column) and cats (right column) with gastroenteric MPC (top), Pruritus MPC (middle) and Respiratory MPC (bottom)

#### 3.8.3.2 Lower Level Applications by MPC, Species and Region

This section looks into the GP methodology at a lower level application by MPC, species and region. This aims to highlight any underlying issues with the methodology when there are fewer consultations. Figure 3.6 shows successful applications of the methodology onto the data as there is a clear seasonal trend within them, with the ability to witness outbreak style patterns and deviations from the credible intervals.

The left column in figure 3.6 show dog gastroenteric MPC for the regions of North West of England and Yorkshire. Within both of the plots, the January 2020 outbreak style pattern that was first noticed at a national level is also within these. Further, in the Yorkshire figure, there are two high points in January 2022, this reflects what veterinary surgeons were highlighting as there was a link between gastrointestinal symptoms and dogs visiting the beach in the North East England. SAVSNet found this outbreak to be a new variant of Canine Enteric Coronavirus which had been seen in January 2020 [32].

The previous 3 visualisations of the methodology are promising and offer some insightful information, however how does the method work with fewer consultations? Examples of this are shown in the right column of figure 3.6. Another obvious issue within not just the poor visual plots is the strong immediate jumps in the lockdown periods where the model seems to be highly influenced by the lockdown indicator.

The plot showing cat respiratory records in the South West, the credible intervals seem incredibly tight and therefore a lot of the observed weekly prevalence values are outside of the regions. This example here is not a single case and is reflective of more of the smaller regions for the different MPC's.

These plots show the inability to confidently make judgements regarding outbreak

calls as when we're defining an outbreak to be two consecutive weeks above the 99% credible interval, on multiple occasions there are many consecutive weeks above the credible intervals as the model is over-fitting the Binomial noise.

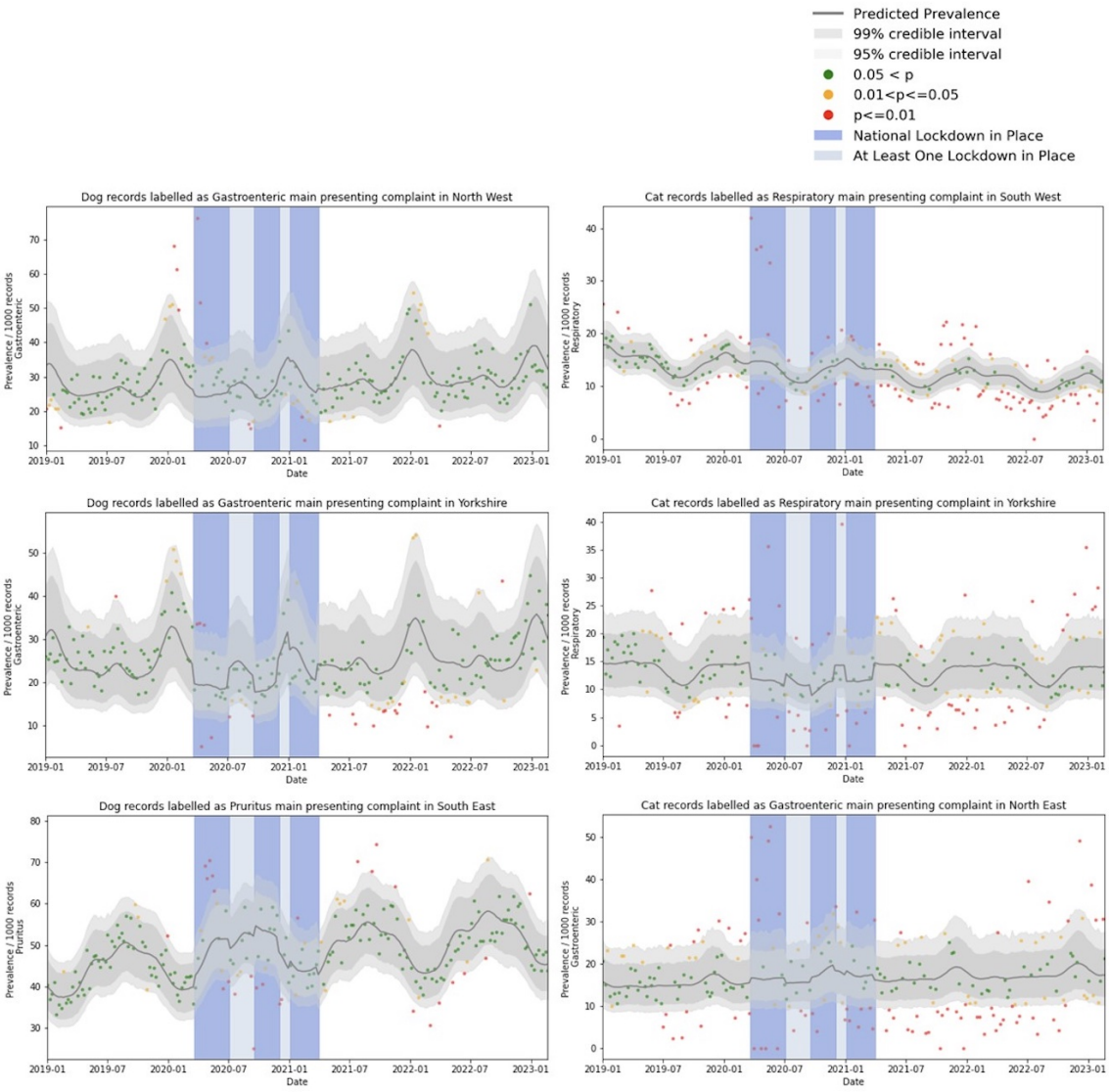



Figure 3.6: Plots which display insightful (left column) and uninsightful (right column) results when Model 2 is applied to a lower spatial level. Insightful plots are dog gastroenteric MPC in North West (top left), dog gastroenteric in Yorkshire (middle left) and dog pruritus in South East (bottom left). Uninsightful plots are cat respiratory in South West (top right), cat respiratory in Yorkshire (middle left) and cat gastroenteric in North East (bottom left).

### 3.8.4 Natural Language Processed Datasets

The main aim from the topic modelling analysis was to see if any of the assigned topics from the algorithm align with that of the Canine Enteric Coronavirus outbreak in January 2020. The next topic modelling projects aim was to also see if any of them would also follow the same pattern as both the first method and the Canine Enteric Coronavirus. The people who performed the text mining algorithms were interested in the results for their own reasons e.g. whether the NLP2 algorithm was able to spot the outbreak earlier than the NLP1 algorithm, however we will be using these results solely to test the validity of the methodology and see whether it is able to consistently spot outbreak style patterns from different variations of the data source.

#### 3.8.4.1 NLP1

From the 30 different topics seen in figure 2.5, below are the ones which include keywords which would reflect our main presenting complaints. Figure 3.7 contains Topic 5 that contains keywords that relate to respiratory issues e.g. coughing, cough, chest and heart. Further, it shows Topic 16 which reflects keywords associated with gastroenteric issues, food, eating and diarrhoea. Even though there may be other keywords relating to the MPC's in the other topics, from the diagram these topics seem to be the main umbrella for the terms. Upon being sent the results of the topic modelling, the major request was to see if any of the topics reflect the January 2020 Canine Enteric Coronavirus outbreak. From Topic 16, there is a clear seasonality around January-February with a large outbreak style pattern that mirrors that of the prior mentioned outbreak. The GP application to the results indicates that this natural language processed model was able to find and assign keywords relating to

the January 2020 outbreak.

Topic 21 in the figure reflects words that one would associate with pruritus e.g. skin, itching and itchy. The credible intervals, similarly to what we seen earlier from the cat respiratory in the South West shown in figure 3.6, are small. Although there are a few occurrences of multiple consecutive weeks above the credible interval there is still quite a noticeable pattern in the data. There is also quite clearly a very strong seasonal trend which peaks around September every year, and as expected this coincides with the patterns from the nationwide pruritus plots shown in figure 3.5. Interestingly however, is the apparent slight upward trend from the end of 2018 onwards in where the observed points have an upward trend but the GP does not follow. This could be due to the lack of historical data in the model and therefore the GP is unable to pick up these trends.

Finally, Topic 5 in figure 3.7 shows the topic that contains the words one would associate with respiratory issues such as coughing, chest and heart. Looking into the plots there seems to be, similarly to all the other MPC's across the species a seasonal trend which peaks around September. The GP appears to fit the data and trend well until September 2018 where there is a large jump in the observed cases which the model does not follow or trend to.

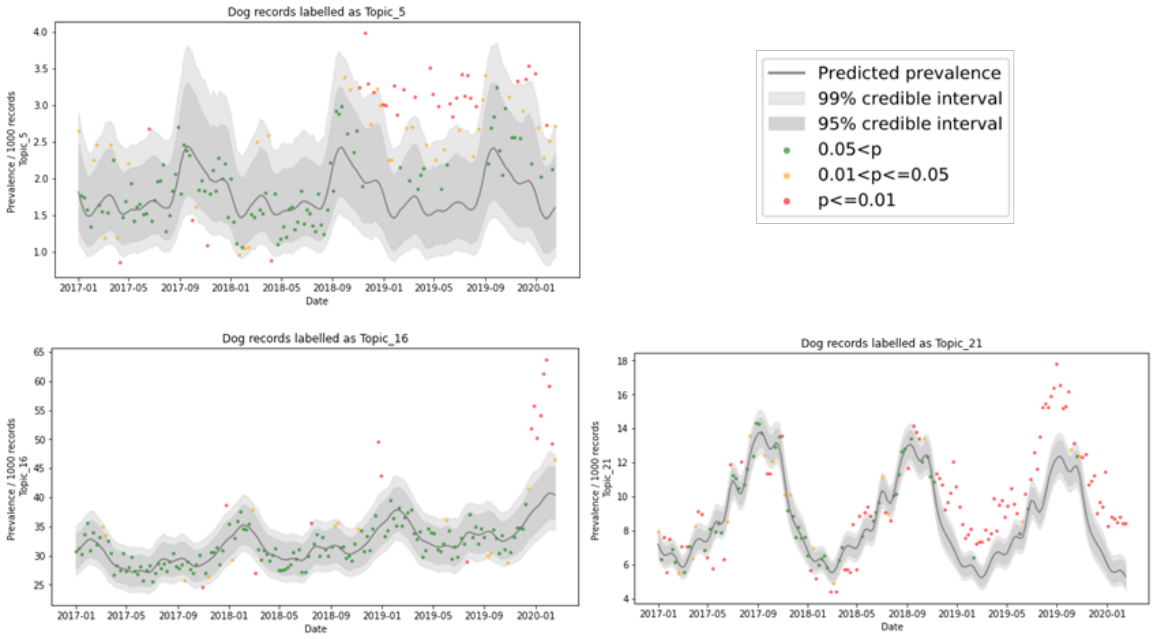



Figure 3.7: GP results for dog topic 5 which reflects free text including coughing, chest and heart (top left), GP results for dog topic 16 which reflects the free text that includes the terms food, diarrhoea, eating and blood (bottom left) and GP results for dog topic 21 which reflects the free text that includes terms such as skin, itchy and allergy (bottom right)

#### 3.8.4.2 NLP2

As mentioned in the Data section, the wanting to run the results of this NLP method through the GP was for an easy comparison between the two NLP methodologies. There was also a wanting to see if any of the categories had an outbreak style pattern the same time as the January 2020 outbreak. Figure 3.8, shows each of the categories for both cat (left column) and dog (right column). The obvious pattern displayed in all 6 of the plots is the outbreak style pattern when lockdown began in March 2020.

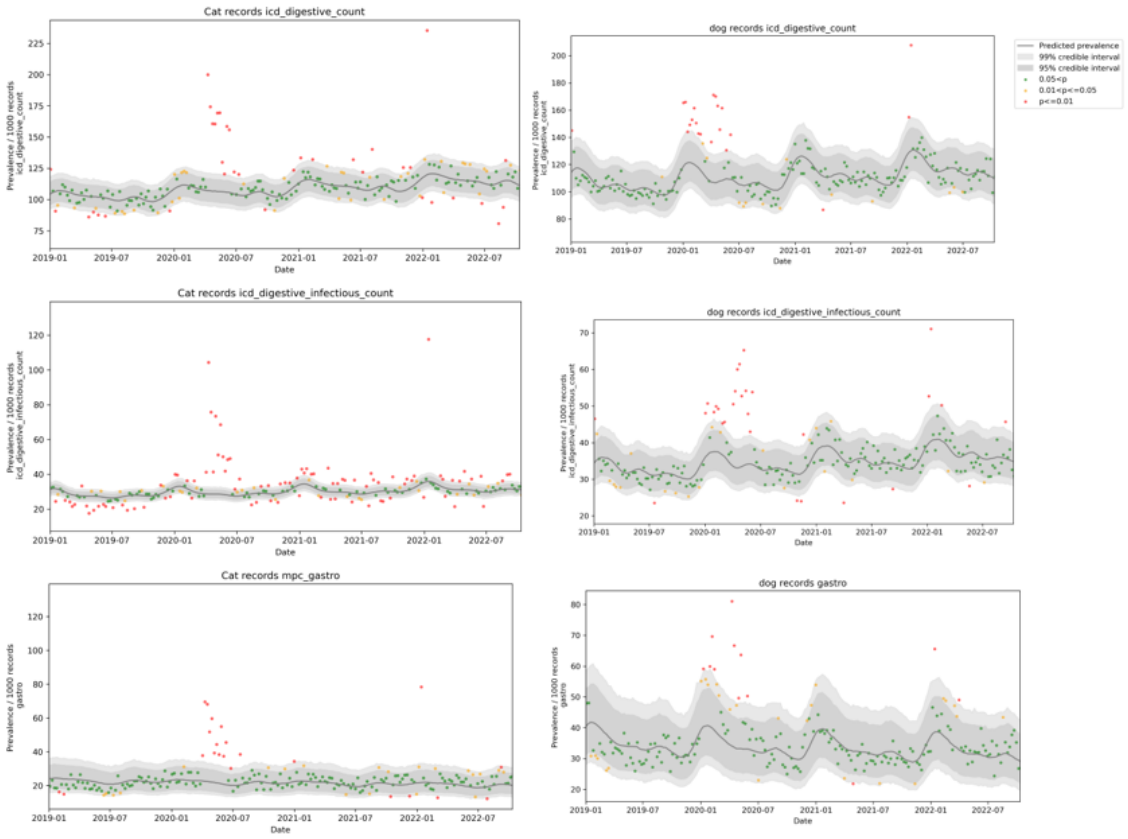



Figure 3.8: GP results for second Natural Language Processed methodology. Cats digestive count (top left), cat infectious digestive count (middle left), cat gastroenteric (bottom left), dog digestive count (top right), dog digestive infectious count (middle right) and dog gastroenteric records (bottom right).

Another obvious characteristic more prevalent in the dog plots (right column) than the cat plots (left column) is seasonality. The data displaying much stronger seasonal trends, showing an influx of expected cases during Christmas time which, as previously mentioned, is usual behaviour. The seasonality is less strong within the cat plots, perhaps due to the small amount of data it had to learn from.

### 3.8.5 Laboratory Dataset

As previously mentioned, there will be some applications of the method to the laboratory dataset. The work performed at Bristol University in the SAVSNet Agile group identified from a group of veterinary surgeons the main pathogens in canines they would want to hear about if there was a positive case [91]. These are Parvovirus, Leptospirosis, Lungworm, Distemper and Cutaneous and Renal Glomerular Vasculopathy. The temporally modelled results are seen below in figure 3.9 except for Renal Glomerular Vasculopathy as this is not something measured by SAVSNet. This data was run using model 1 as there was no need to alter the data for the introduction of lockdown as these samples came from animals that were allowed to attend their veterinary appointments.

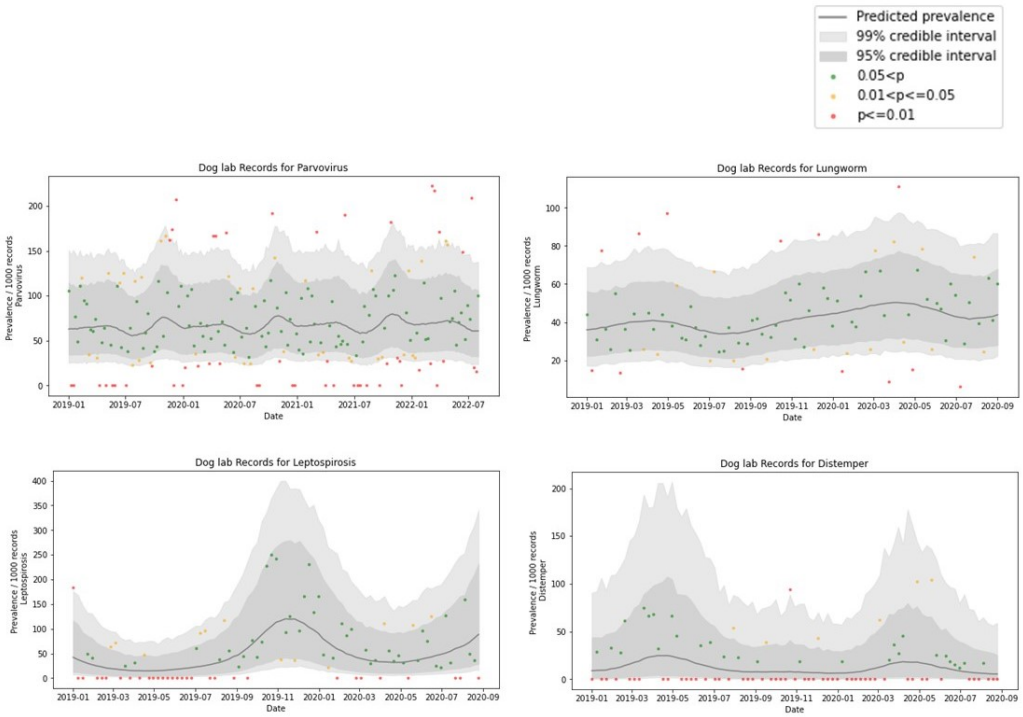



Figure 3.9: GP results for the laboratory data showing the main canine endemic diseases that a veterinary surgeon would want to know about following research performed by [91]. These are Parvovirus (top left), Lungworm (top right), bottom left (Leptospirosis) and bottom right (Distemper)

Initially looking at the results for all the pathogens there is apparant seasonality within them, some have stronger seasonal aspects e.g. parvovirus (top left) and leptospirosis (bottom right), whereas lungworm and distemper have weaker seasonal components. The main observation here however is how similar the plots are to that of the results from the fewer data points for the main presenting complaint which was discussed in section 3.7.2.2. Having this issue and the model overestimating binomial noise leaves the results hard to interpret and incapable of providing a confident call of an outbreak.

The first issue highlighted within the discussion of the Data section, is that there are little to no weekly counts for a vast majority of the pathogens across the different species, only the main pathogens (parvovirus, leptospirosis, etc.) have enough data to provide semi-plausible results. There is also the unknown lag within the laboratory results which affects the representation of the true weekly counts. Finally, there was the end-goal thinking in terms of the surveillance system, are the lab results something that also need relaying back to the public or is this something that the veterinary surgeons can enquire about as and when their concerns rise? Following this thinking, the use and analysis of the lab results was abandoned as they are unable to give real weekly estimates of prevalence for important pathogens.

## 3.9 Discussion

We have shown that a latent GP can be used to detect outlying case numbers in time-series data. This section will assess the GP so far and other aspects of what has been touched on during this analysis.

An initial point of discussion for this model is that the initial implementation of the model was run with a single chain, and trace and posterior plots were assessed visually. This represents a methodological limitation. Relying on one chain prevents a robust assessment of convergence, as there is no basis for comparison across independent runs from different starting values. A single chain may appear to mix well and, when assessed visually, appear to produce stable estimates; however, there is no quantitative assessment to ensure that the parameters haven't failed to fully explore the parameter space. For future development of this model, there will be more than a single chain run and formal convergence diagnostics, such as the Gelman-Rubin statistic, will be

calculated to ensure the posterior inference is trustworthy.

Another point of discussion with this model and a limitation with how its currently designed is that the model was initially trained on all available data each week, meaning that when generating predictions for a given week, the 'training' set inadvertently included data from the same period. This approach introduces a limitation, as it effectively allows the model to have future insight and results in the posterior predictive distributions that are overly optimistic to what would be achieveable in practice. In a real-world setting, only historical data would be available and the model would need to be retrained each week using information up to that point intime before generating predictions for the subsequent week. Consequently, the results presented here should be interpreted with the caveat in mind, and future implementations should adapt a rolling retraining framework to ensure that evaluation more closely reflects out-of-sample predictive performance.

Further, another aspect of the chapter assessed different values of the length-scale value $\phi$. To evaluate the model fit across the different values of the length-scale, we compared the joint log posterior. This diagnostic metric was chosen in this case as it gives a good summary of the fit of the model, with the model with the highest log joint posterior to be the favoured [108]. Both the mean and maximum values indicated that the model with $\phi = 0.32$ had the best fit, followed by $\phi = 0.16$, and then $\phi = 0.64$. The max and mean joint log posterior results were as follows: $\phi = 0.32$ (-1632, -32295), $\phi = 0.16$ (-1666, -41942) and $\phi = 0.64$ (-1710, -47166). The consistency across both the maximum and mean summaries suggests $\phi$ =0.32 to be the best chosen length-scale for this data.

Similarly to the value of $\phi$, the addition of the lockdown variable into the model

was a crucial part of the chapter. Regarding the addition of the lockdown variable, model diagnostics were reported for both models with and without the lockdown variable. The diagnostic method chosen was Leave One Out Cross Validation (LOO). This method assesses the statistical performance of a model by repeatedly withholding a single observation from the dataset and fitting the model to the remaining data. Once it has been fit, it then evaluates how well the model predicts the withheld point [99]. This process is repeated for every observation, with results combined to estimate the model's predictive accuracy [99]. When using the LOO method, the model with the highest value indicates better prediction [99]. The GP model, when run without the lockdown variable, achieved a LOO value of -1,585,917 and the model with the lockdown variable achieved -1,413,460, concluding that accounting for the lockdown effect provides a better fit to the observed cases.

We also have the regional information for the laboratory dataset however because the counts are low it is a similar situation to the earlier mentioned effect on the lower data in that they can sometimes look relatively uninformative. There is also an issue with the laboratory data in general in that there is a delay between the veterinary consultation, the veterinary surgery sending off the needed sample, and the laboratory receiving it and then testing it, loosely speaking the sample once assessed may be irrelevant to the week it was obtained. The identified pathogens that surgeons would want to know about that were derived from my colleagues are the more serious issues that would be reported and made aware of regardless if this analysis was done or not. Another issue regarding the laboratory results is the amount of bias within them, as there is bias introduced through the veterinary surgeon choosing to send an animals sample to a laboratory, then the bias of the owner accepting a large payment to

perform the sample and have it sent for testing. Finally, veterinary surgeries also have in house lateral flow tests for different pathogens which reduces the power of this data as some results of specific pathogens won't be displayed throughout it and thus the data is not showing the full scope of the real weekly figures [52].

There is also an expense of computational and time resources as the GP takes a long time to run. As it currently stands the GP's take around 2 hours and 36 minutes to run each one. When splitting out into the different MPCs, species and regions, we will be running 78 altogether which is not feasible in an outbreak detection method due to how an immediate response is needed in outbreak situations. Chapter 5 discusses and explored computing optimisation techniques.

Having the ability to apply this type of methodology to multiple layers of geospatial data allows us to assess any outbreak style patterns nationally, with the ability to drill down to regional levels to see where is most affected by them. Ideally, we would be able to have lower layer analysis to local authority level however, the lack of data would leave these lower level applications uninformative.

Regarding the 'un-insightful' results in 3.6 (the right hand side plots) and the comments regarding the smaller variance estimates, this could either be due to a mixing problem with the MCMC algorithm or simply due to there being limited information available to estimate covariance parameters. Upon initial visual investigation of the trace and posterior plots, the parameters did not suggest major convergence issues; however, in future development of this model, there would need to be more quantitative and formal diagnostics to conclude the reason for the small variance estimates confidently.

With the GP model as it is, likewise with any regression style model, there is the

issue of learning from all the past data. The model receiving and creating predictive probabilities based on data that is high in periodicity and containing past outbreaks in turn reduces the sensitivity of the model and after enough time and training may not confidently be able to call out an outbreak due to being desensitised to it. This model as it currently stands then requires the manual removal of these 'outbreaks' and seasonal components which, when the aim of this method is to produce results in a timely manner, is not a feasible option. This also brings us up the question about what is considered an outbreak and at what time point the outbreak begins and ends. Likewise where there are seasonal trends, how to determine where the seasonal aspects begin and end and how to rectify them.

## 3.10 Conclusion

In conclusion, the Gaussian Process methodology using model 2 was able to absorb the loss of consultations and was able to identify outbreak style patterns consistently from the various data sources. However when applied to lower level spatial units with less data, there are issues with the fitting of the model. Now, this isn't to say that the Gaussian Process methodology is not a good fit for this data as it was able to show anomalies, the issue is that it was over-fitting the Binomial noise which was caused by the lack of data. Further, with the idea of creating an automatic surveillance system, it is important to eliminate the need for human intervention with the data while also generating longevity within the surveillance system. Therefore a methodology applied to the input data that removes any patterns whether seasonal or outbreak style patterns is necessary.

# Chapter 4

# Model Based Anomaly Detection

In the previous chapter, we developed a method for detecting anomalous behaviours in time series of clinical metrics in small animal clinical records. This was based on using a Gaussian process to learn seasonal patterns of case incidence, and therefore construct a predictive distribution of incidence for each week $w = 1, \ldots, 53$ of the year. Anomalous incidence was the defined as those weeks in which the observed incidence exceeded the $95^{th}$ percentile of the predictive distribution.

Whilst this previous method was shown to be effective for detecting the January 2020 outbreak of Canine Enteric Coronavirus, it suffers from the same drawback that other predictive methodologies have in that they create predictions based on all the past data input. This chapter explores a different approach to this problem in which we model seasonality using fixed effects and model departures from mean seasonality using a Gaussian Process. Intuitively, the fixed effects can be thought of as de-trending and de-seasonalising the time series, whilst the GP models any left-over local non-stationary behaviour such as an outbreak. Importantly, the GP correlated residuals in time, mitigating the effect of the occasional outlier or outbreak detected.

## 4.1 De-sensing

This section discusses and demonstrates the de-sensing phenomenon.

De-sensing in this context refers to the modelling and removing of any underlying recurring patterns and noise from time-series data.

Time series data have trends, and a trend is defined as a continued increase or decrease over time [106]. Trend information from a dataset can impact methods and models in a negative way however, there are also useful implications to knowledge of trends in the data. We can use the information to improve model performance by taking this prior knowledge into account [106]. Another way it can be useful is it can influence methods and model selections which in turn can make evaluation more efficient [106]. A final use to knowing the trend is that it can simply be removed using methods/adaptation of modelling which again cam improve model performance.

When thinking about trends within data, it can be a relatively subjective process and further once a trend has been identified there needs to be the decision whether it is a Global or Local trend. Global trends are relevant to the whole time series e.g. seasonality whereas Local trends are isolated instances [106].

To demonstrate the idea of de-sensing, below are figures showing a time series with yearly seasonality and an overall upward trend. The removal of each bias will be shown in stages to highlight the effect of each one.

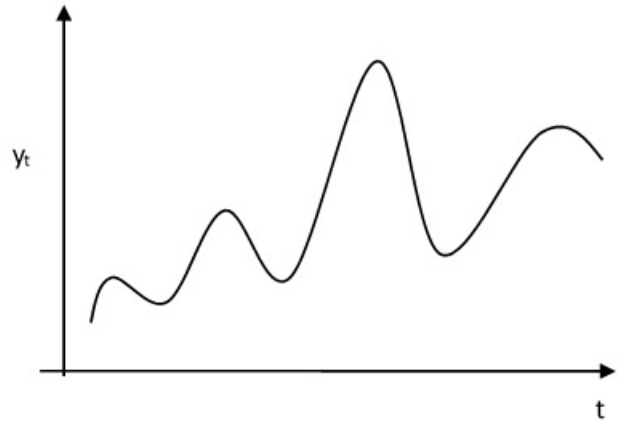



Figure 4.1: Example plot of a time series with seasonal, upward trend and outbreak style pattern components.

Figure 4.1 shows time series $y_t$ plotted against time, the point of this is to show a gradual upward trend, seasonality and an outbreak style pattern. Now, say from here we notice the upward trend, we can add this into the model as a fixed effect, as below:

$\frac{y_t}{n_t} - \alpha + \beta(t - \bar{t})$

This example model takes us from figure 4.1 to 4.2. The difference between the time series now is that the levels of $y_t$ have been reduced to be at 0, i.e. these are the residuals of the model.

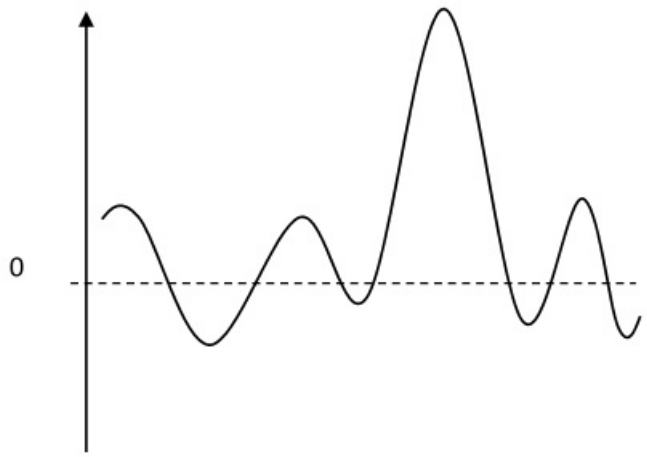



Figure 4.2: Example plot of a time series with seasonal and outbreak style pattern components.

Notice now from figure 4.2, we have a seasonal component. This is also something we can fix and model for. Thus our model above can be modified to include a seasonal component as such

$\frac{y_t}{n_t} - \alpha + \beta(t - \bar{t}) - \Theta_1 cos(.) - \Theta_2 sin(.)$

Which will leave the residuals as seen in figure 4.3. At this stage, we are now able to investigate any remaining effects of the data from the residuals $u_t$ and have more confident conclusions regarding outbreak calls.

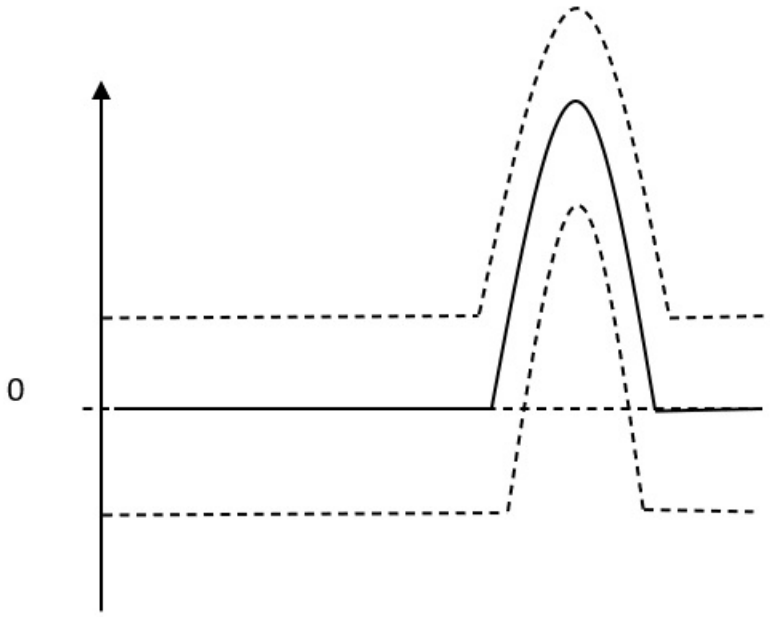



Figure 4.3: Example plot of a de-trended time series with an outbreak style pattern. The solid line represents a de-trended time series with one visible outbreak and the dotted lines are credible intervals.

## 4.2 Mixed-Effects Model Based Anomaly Detection

Based on the de-sensing description and demonstration in section 4.1, and the exploration and analysis of the data explored in Chapters 2 and 3, it was determined there was an upward overall upward trend, seasonality and an effect when social distancing measures were in place. Therefore we are able to use a mixed-effects model to model anomalies where the fixed effect components are for the three components mentioned. We are then able to model the residuals to assess outbreak style patterns.

Data points that do not conform to the 'normal' behaviour of a dataset are

describes as anomalies. Algorithms and models are created to capture these anomalies and make decisions based on them. In the case of this project, we are attempting to detect outbreak style patterns.

There are many methods for outbreak detection including the Farrington algorithm. These methods can be useful and have mass supporting literature, but if you have prior knowledge of the data, a model based approach might be more appropriate.

### 4.2.1 Previous Research

The initial consideration of modelling repeated measure data fundamentally as signals that can be explained into fixed and random effects were Laird and Ware [55]. This can be expressed as

$$Y_{ij} = a_i^T \alpha + x_{ij}^T \beta + d_{ij}^T U_i + \sigma Z_{ij} \text{for} j = 1, \ldots, n_i, i = 1, \ldots, m \tag{4.1}$$

Where $a_i^T \alpha + x_{ij}^T \beta$ are the fixed effect terms and $d_{ij}^T U_i$ are random effects with $\sigma Z_{ij}$ representing an error term, $n$ is number of measurements for $i^th$ subject and $m$ is the number of subjects [55]. A popular application of this model is a 'random-intercept and random slope' where a subject's random effect represents a linear function of time and works well for smaller numbers of repeated measures per subject [6]. However, for longer sequences of data, the idea that individual random-effect trajectories can be represented by straight lines becomes unrealistic because these trajectories exhibit non-linearities [6].

Following this, Diggle [24] proposed to introduce a stationary stochastic process to model a time-varying random-effect term. This alters equation 4.1 to be:

$$Y_{ij} = a_i^T \alpha + x_{ij}^T \beta + d_{ij}^T U_i + W_i(t_{ij}) + \sigma Z_{ij} \text{for} j = 1, \ldots, n_i, i = 1, \ldots, m \qquad (4.2)$$

where $W_i(t)$ are independent zero-mean Gaussian Processes with covariance function $\gamma(t,t) = (cov)W_i(t), W_i(t)$ [24].

Hale et al [40] looked at a model based approach to model anomalies using the SAVSNet data. They justify adjusting for known effects by using measured exploratory variables. Following this adjustment, they treat the remaining data as stochastic and include this as a latent variable [40]. Although Hale et al have this approach on the same dataset, one major limitation in this research is that, for reasons of computational performance, it is only performed over a 9 day period, so there is no capability to account for seasonality.

Following supporting literature for this model based approach, and following the characteristics of the MPC data defined in chapter 2; we have an upward trend, social distancing measures and clear seasonality. Adapting the model in Equation 3.6 we can add harmonic terms to model seasonality. The difference here between the two approaches is that the approach in chapter 3 assumed the seasonality within the Covariance matrix of the Gaussian Process, whereas this approach models the seasonality as a fixed effect.

There are many ways to model seasonality, but by definition of periodicity it can be reduced down to the use of a combination of trigonometric terms. A popular method for modelling periodicity in time series analysis is the Fourier series.

## 4.3 Mixed-effects model for SAVSNet Clinical Data

The routinely collected data that we will be using for this analysis is veterinary consultation data for dogs and cats for the three main presenting reasons the animal might have been at the consultation; gastroenteric related illness, respiratory related illness and pruritus related illness. The data is representative of the United Kingdom and analysis will be reported nationally and regionally. A more in-depth description and exploration surrounding the data source seen in Chapter 2.

For this dataset, we are aware and have previously determined seasonality, an upward trend and also the effects of multiple lockdowns which we can treat as fixed effects within the model to remove these effects from the data. There is also knowledge regarding over-dispersion (highlighted in Chapter 3 Results and Discussion) and occurrence of outbreaks which we can make adaptations to the model for.

### 4.3.1 Fixed-Effect Model

A Fourier transform decomposes a time series into an infinite series of progressively higher order sin and cosine terms. Modelling an infinite series is unachievable for many reasons, therefore we have to truncate it as the $p^th$ order. Section 4.7.1 looks at choosing the $p$ order whereas this section continues to explain the concept and its advantages. To begin a look at the Fourier series analysis we first begin by defining a periodic function on a time series as

$$\text{logit}(p_t) = \alpha + \xi_1 \cos(\frac{2\pi t}{365} + \phi) + \xi_2 \cos(\frac{4\pi t}{365} + \phi) + \dots \tag{4.3}$$

and using trigonometry identities we know that

$$\cos(\theta + \phi) = \cos\theta\cos\phi - \sin\theta\sin\phi \tag{4.4}$$

Which then reduces to

$$\theta_1 \sin(\theta) + \theta_2 \cos(\theta) \tag{4.5}$$

Now we have a simple periodic function defined, the theory initialised by Joseph Fourier, states that any periodic function can be expressed by a combination of periodic components [38].

The harmonic regression approach has many advantages. First of which is that it allows and can be adjusted for any length of seasonality [48]. Secondly, for data with more than one seasonal period, there can be an adjustment made for different frequencies of Fourier terms. A final advantage is that the researcher can control the smoothness of a seasonal pattern [48].

Using the periodic function defined in equation 4.4, and the known effects described in section 4.3, we can assign prior distributions to these effects highlighted given knowledge we already have and model these as fixed effects, leaving only the residuals to model anomalies.

### 4.3.2 Modelling of Anomalies

Once the fixed effect part of the model is applied to the dataset and it has been desensed, we will be able to use a methodology to model the correlated residuals. This regression of choice is a Gaussian Process. For the purposes of this chapter I will give a brief overview of the Gaussian Process methodology, but a more in-depth description

can be found in section 3.2. Gaussian Processes are an unsupervised stochastic process which provide a solution to the premise that a dataset can effectively have infinite functions to fit it and Gaussian Process assign probabilities to each of these functions [36].

## 4.4 Model

In chapter 3, we explored the different model possibilities given the introduction of nationwide social distancing measures. The consensus on this was that the model with the added variable to simulate values for when the area was in a Covid-19 lockdown state was the most appropriate for the drop in consultations. Updating this model to now add in the Harmonic regression, it becomes as below.

To model the prevalence of MPC consultations in the United Kingdom, we apply a Harmonic regression and use a longitudinal latent GP model with Binomial observation process. We assume that we observe $y_t$ MPCs out of a total of $N_t$ consults per week $t = 1, ..., 53$ spanning from 01/01/2019 up to real-time. We model $y_t$ as a Binomial random variable such that

$$y_t \sim \text{Binomial}(N_t, p_t) \tag{4.6}$$

where $p_t$ is the probability of a consult in week $t$ being an MPC, with the log odds of being an MPC in week $t$ modelled as a linear combination of terms as described below.

$$\log\left(\frac{p_t}{1-p_t}\right) = \alpha + \beta(t-\bar{t}) + \delta z_t + h_t^T\beta + u_t \tag{4.7}$$

where $\alpha$ is the overall mean prevalence of the MPC, $\beta$ captures a longitudinal linear trend. $\delta$ is representing an offset of log odds for weeks in which social distancing measures were enforced. H, the harmonic series matrix for which $h_t$ represent a row, is a T $X$ 10 matrix with elements:

$$h_{tk} = \begin{cases} cos(\frac{2\pi kt}{365}) & \text{for} k = 1, \ldots, 5 \\ sin(\frac{2\pi kt}{365}) & \text{for} k = 1, \ldots, 5 \end{cases} \tag{4.8}$$

Finally, $u_t$ represents a time-varying random effect.

As mentioned in Chapter 3, the random effect $u_t$ allows us to model both serial and periodic correlation in our weekly observations, as well as any extra Binomial variation that might contribute to the overall variability of cases from one week to the next. We model the vector $u$ as a GP with mean 0 and covariance matrix $\Sigma^2$ such that;

$$u_t \sim \text{MultivariateNormal}(0, \Sigma^2) \tag{4.9}$$

The covariance matrix $\Sigma^2$ captures the covariance between two variates $u_t$ and $u_{\bar{t}}$ spaces $t - \bar{t}$ weeks apart, and we assume the correlation is stationary, given we have removed seasonality with the harmonic fixed-effects model.

$$\Sigma^2_{t,\bar{t}} = \begin{cases} \sigma^2 \left(1 + \frac{\sqrt{3(t-\bar{t})^2}}{\phi}\right) \exp\left[-\frac{\sqrt{3(t-\bar{t})^2}}{\phi}\right], & \text{if t} \neq \bar{t} \\ \tau^2 & \text{if t} = \bar{t} \end{cases} \tag{4.10}$$

## 4.5 Priors

Our priors on this model are similar to the priors of the previous model, described in section 3.5.4, except for this model we no longer have a fixed length-scale parameter. This is due to the harmonic regression before running the GP methodology on the residuals. Due to the limited noise and seasonality, we have the computer capacity to sample the length-scale parameter whereas we were unable to do that in a timely manner using the approach in Chapter 3.

Our likelihood for this model is

$$\pi(\alpha, \beta, \sigma, \tau, \delta | y, t) \propto \pi(y | \alpha, \beta, \sigma, \tau, \delta, t)\pi(\alpha)\pi(\beta)\pi(\sigma)\pi(\tau)\pi(\delta) \tag{4.11}$$

$$\text{coefficients} \sim \text{Normal}(0, 10) \tag{4.12}$$

$$\pi(\sigma) \sim \text{HalfNormal}(1) \tag{4.13}$$

$$\pi(\phi) \sim \text{HalfNormal}(10) \tag{4.14}$$

$$\pi(\tau) \sim \text{HalfNormal}(10) \tag{4.15}$$

$$\pi(\delta) \sim \text{Normal}(0, 100) \tag{4.16}$$

## 4.6 Implementation

Similarly to the models in Chapter 3, this model was ran in Python using PyMC3 [70] and ran for 6000 MCMC iterations, discarding the first 1000 as a burn-in period. Burn-in periods are used for convergence as the MCMC algorithm as this

reduces the dependence on the initial (usually user input) start conditions. The MCMC algorithm used here is the No U-Turn Sampler which is an extension of the Hamiltonian Monte Carlo with its main benefit being that it is able to tune the number of steps L and the step size parameter $\epsilon$. The code can be found here https://github.com/seeemmayy/Harmonic-Regression.

## 4.7 Prediction

The process of the prediction is that the model is built to remove the upward trend and the bias from the sampled prior values of $\alpha$ and $\beta$. Once we have removed these characteristics, the social distancing variable adjusts the data accordingly then, the Fourier terms removes the seasonality from the data. We model these residuals using a GP to detect anomalies. These sampled priors are combined with our data to calculate our predicted prevalence values and similarly to the outputs in chapter 3, plot credible intervals of the posterior predictive distribution.

Figure 4.4 shows an example of the residuals obtained from the model shown in equation 4.7 that are yet to be modelled by the GP. The use case here is the dog gastroenteric MPC nationwide. The purpose of this plot is to show the effect of the fixed effects part of the model where anomalies are now more obvious.

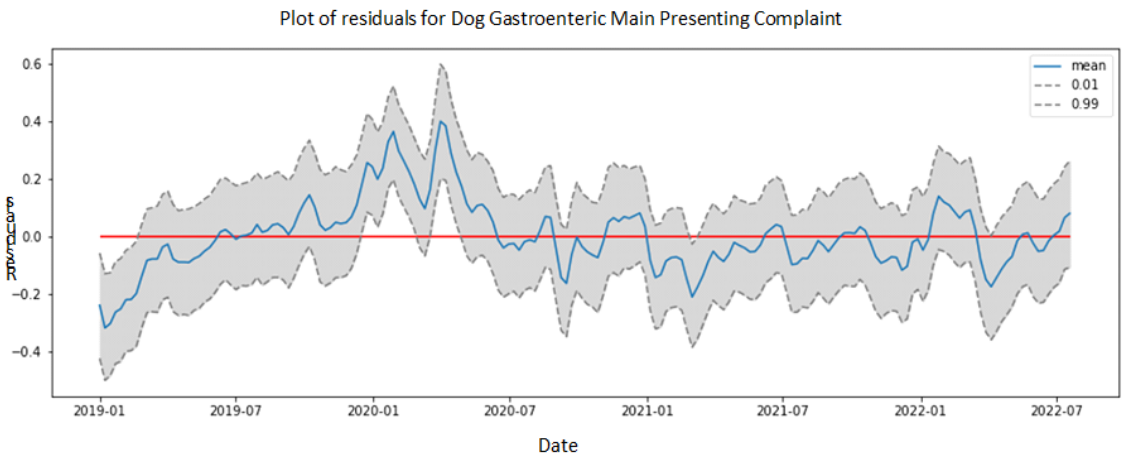



Figure 4.4: Plot of the residuals for Gastroenteric MPC in dogs to be modelled using a Gaussian Process

The plots then follow an identical structure to the plots in section 3.8, where the prediction intervals are quoted at the 1%, 5%, 95% and 99% level. The observed prevalence is then plotted on top following the same colouring system as in section 3.8; red for below 1% or above the 99% intervals, orange for those points between the 1% and 5% and 95% and 99% where the other points are green.

### 4.7.1 How to Choose Amount of Harmonics

Following the description that Fourier series can be technically infinite, how do we pick how many Harmonics to use to best describe the seasonality? A way to check this is by running the model for multiple choices of Fourier terms and assessing the Deviance Information Critera (DIC) for each model and picking the amount of harmonics that resonate with the smallest DIC. This is a generic method for model selection, however as we have a continuous time series we can use a Fourier Transform. Once the time-series data has been transformed to a frequency, we are able to smooth the frequency

by averaging years of data over each time point t, from here we can deduce how much of the data each harmonic explains and choose $n$ Fourier terms based on this. Figure 4.5 shows these results. From here it's visibly clear we will get the most information out of our dataset using 5 Fourier terms as the visual decline after the 5th harmonic suggests that any additional terms provide little explanatory value. A limitation of this however, is that at the time of writing this thesis, there was no quantitative comparison metrics to formally support this choice. For future analysis and development of this model, it would be a requirement to include some formal model comparison metrics (e.g. DIC or LOOCV).

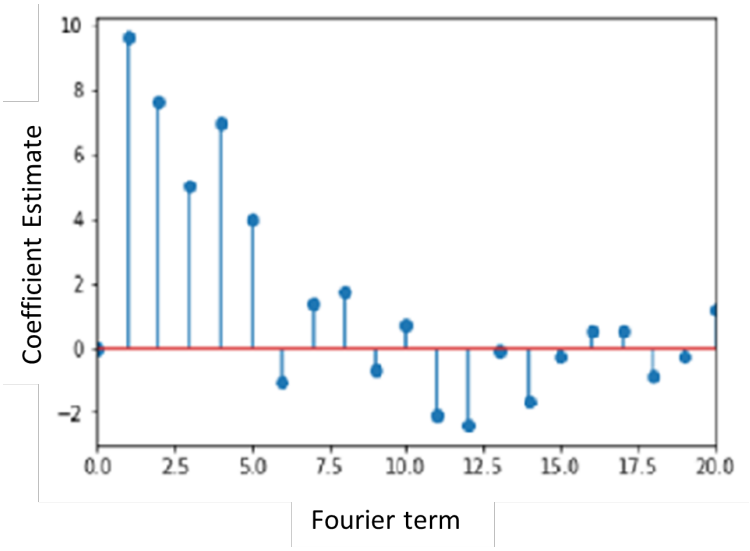



Figure 4.5: Figure of different amounts of Fourier terms modelled on a Fourier transformed time series of the MPC consultation data.

## 4.8 Results

This section looks at results from this methodology. Using the same examples as Chapter 3, section 3.8.3, we will be able to assess visually, how well the model fits

and any visual differences between the two methods.

This model was ran on the consultation dataset, described in detail in chapter 2, which is the main dataset used throughout the thesis and contains consultations for the MPC information for both cats and dogs. The three MPC's are gastroenteric, respiratory and pruritus. Traceplots for each of the results shown are in appendix .3.

### 4.8.1 National Levels for each of the MPC and Species

Figure 4.6 shows the results of the model on the data for each of the main presenting complaints nationwide for both our species.

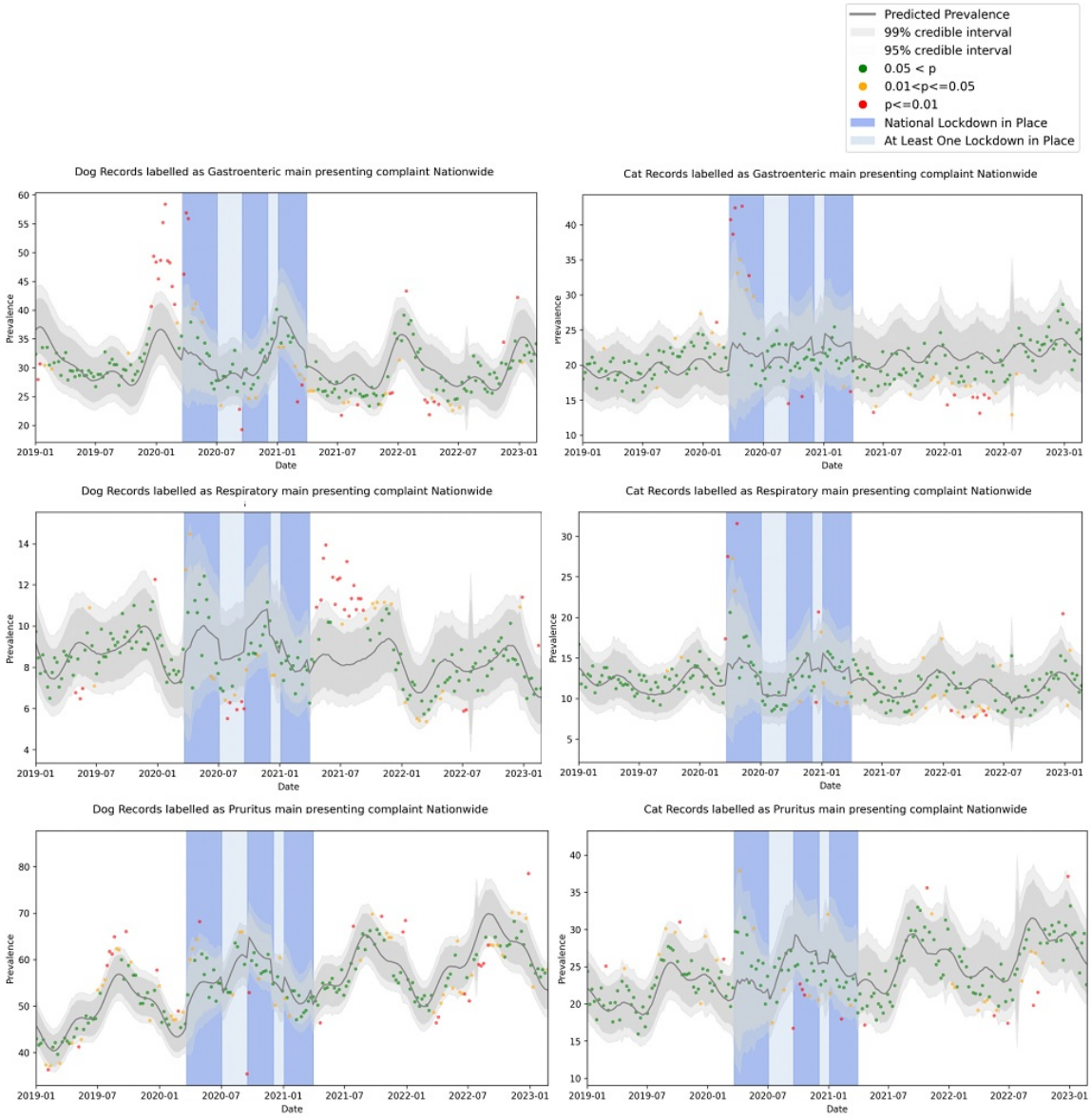



Figure 4.6: Harmonic Regression then Gaussian Process on the Main Presenting Complaint consultation data. Results for Dog Gastroenteric nationwide (left top), Dog respiratory nationwide (left middle), Dog pruritus nationwide (bottom left), Cat gastroenteric nationwide (top right), Cat respiratory (right middle) and cat pruritus (bottom left).

In figure 4.6, the variance is smaller than the results from the first method which

can be seen in figure 3.5. The anomalies in the data have more of an assertion and that could be due to the smaller variance within the prediction intervals. Note that these prediction intervals are the same as in the figure 3.5, which are at the 95% and 99% intervals. They are also displaying the same seasonal component that we expected to see and that is also visible in the previous methodology.

All of the results are also being influenced by the lockdown variable as when there is the added sampled lockdown variable there is a slight increase in prevalence, where as we then come out of a lockdown state there is a small decrease.

Another notable visual is that the Canine Enteric Coronavirus outbreak in January 2020 is highly visible within the dog gastroenteric MPC as expected, however within dog respiratory there is an outbreak style pattern that lasts a couple months. Some of the points at the tail ends of this outbreak style pattern are shown in red here whereas in figure 3.5, they are only shown as orange which we would only consider a cautionary prevalence level.

A further comment on the figures is the spike in variance around September 2022, this was due to an outage of the software for the surgeries for a day which resulted in that week missing consultations.

### 4.8.2 Lower Level Applications by MPC, Species and Region

These results will mimic the categories of the results in figure 3.6 for a visual comparison of the two methods.

When the previous model was broken down further geographically, it became apparent that from a lack of data, the previous model was over-estimating the Binomial noise which in turn led to relatively uninterpretable results which lacked

confident and credible conclusions.

Figure 4.7 shows what we identified from the previous method as readable usable plots and uninformative plots. As expected, the readable plots are consistent across the different methodologies being able to display clear seasonality around Summer and Winter along with the outbreak of Canine Enteric Coronavirus in January 2020 to a regional level (top left and middle left plots).

They also have the spike in September 2022 where there was a system outage which resulted in missing data.

A comment to say, however, that even though these plots were considered usable and informative, there are two consecutive weeks above the prediction interval for dog gastroenteric MPC in Yorkshire (left middle plot) approximately February 2022. A reminder that what we consider an outbreak is two consecutive weeks above the 99% interval. This was later confirmed to be an outbreak of a gastroenteric related illness [32]. Referring back to figure 3.6, both of these prevalence counts are within the 95% and 99% credible intervals, so by our own definition we would just consider this as a caution.

The main difference between the results from the two methods is shown in the 'uninformative' plots. Using this method, we are able to see rare occasions where the prevalence count is above the credible intervals and with that are able to make more confident judgements surrounding anomaly calls within these lower level MPC categories. The model is unable to capture any seasonality however, this could again be due to the low amounts of data across these regional levels. The prediction intervals are less smooth compared to those in figure 3.6, however this is due to the estimating of the length-scale as opposed to fixing it.

Another behaviour to mention in figure 4.7 for cat records labelled as Respiratory in Yorkshire (right middle) and cat records labelled as gastroenteric in North East (right bottom), around March 2020 there are a couple of weeks with zero prevalence even with the addition of the lockdown effect. At this time, veterinary surgeries closed to the public for all non-emergency appointments so there may of been no consultations of respiratory or gastroenteric in cats for these weeks. It is already known that cats are less likely to be taken the veterinary surgery [9] and following closures it makes sense there being no consultations in these distinctive areas at this time due to there being low prevalence levels when it is business as usual.

This methodology has been able to capture more useful information in results that we even considered to be already informative, and has allowed for the ability to make decisions and statements regarding prevalence levels for the lower data areas which we were unable to make with the previous methodology.

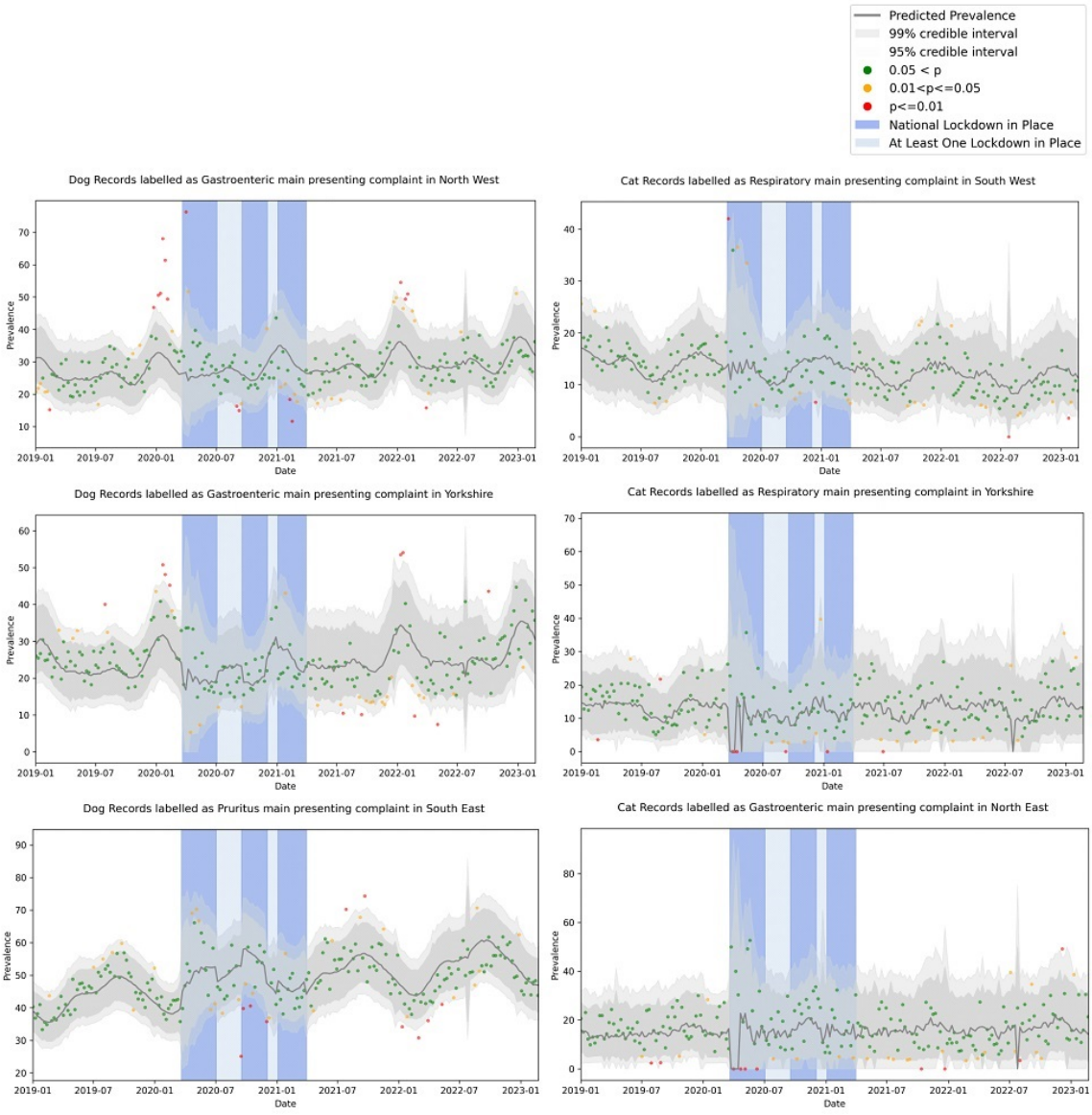



Figure 4.7: Harmonic Regression then Gaussian Process on the Main Presenting Complaint consultation data. Results for 'good' and 'bad' applications identified with the previous methodology. The 'good' applications here are dog gastroenteric in North West (top left), Dog Gastroenteric in Yorkshire (left middle), dog pruritus in South East (left bottom). The 'bad' applications, cat respiratory in South West (top right), cat respiratory in Yorkshire (middle right) and cat gastroenteric North East (bottom right).

## 4.9 Threshold Plotting Alterations

A debate within health related research and results is the idea of thresholds and at which one to report anomalies. The most common thresholds, or intervals, for statisticians are usually the 95% and the 99% intervals but could it be made more informative? Hale et al used user defined thresholds for their surveillance system in that the viewer could choose their own. The issue with this, is that allowing the full freedom of threshold defining can lead to meaningless results. If the intervals are made too small that there are fewer data points outside the intervals we could miss a vital outbreak. Likewise, if the intervals are too large and there's too many points outside the interval we could also miss important information.

It begs the discussion around whether thresholds should be flexible and not the standard, and who should make those decisions? A member of SAVSNet Agile's project focuses on just this. They conducted interviews with veterinary surgeons to gather their opinions on when they themselves would like to hear of a weekly prevalence level above a certain threshold [90]. They found that for gastroenteric and pruritus MPC the veterinary surgeons wanted to be alerted at the 95% and 99% credible intervals, whereas for respiratory MPC they wanted to know at the 90% and 95% credible intervals [90]. These thresholds were used for the final displaying of results.

## 4.10 Discussion

Overall, the approach used within this chapter was successful at identifying previously identified outbreaks along with other outbreak style patterns which the approach in

Chapter 3 failed to identify. This approach was also more successful at retaining meaningful information for lower level applications for categories with minimal data e.g. those figures presented in 4.7.

Further, as the new model and approach was able to successfully detect outbreak style patterns with the use of mixed effects, this increases the longevity of the approach as it shows that a lot of noise was a result of seasonality and overall upward trends within the data. Having these removed through the use of fixed effects increases the longevity as these characteristics are having little emphasis on the modelling of anomalies. Going into this chapter, creating a model that introduced longevity within the analysis was crucial due to the emphasis of an automatic surveillance system with no want for intervention to manually remove certain characteristics. After time, the model as it currently stands will begin to learn from outbreak data as it uses all the previous data, therefore further work from this is to introduce a sliding window in which the predictions will be created.

Threshold alterations was an important investigation and implementation from the approach in Chapter 3 to the approach within this chapter. As these results would be displayed and for the use of veterinary surgeons and stakeholders mainly, it is important to factor their opinion when deciding when to declare a weekly prevalence point as above the credible interval. Having these opinions have influence on the reporting credible intervals however, it is kept within reason so that the results are meaningful. Following a project by another SAVSNet Agile member where they looked at communicating information back to veterinary surgeons and stakeholders [91], it influenced the reporting of credible intervals to be 95% and 99% for gastroenteric and pruritus MPC and 90% and 95% for respiratory.

## 4.11 Conclusion

In conclusion, this new approach was successful at identifying anomalies within the consultation MPC datasets alongside identifying other anomalies and outbreak style patterns that the approach in Chapter 3 failed to identify.

This is not to say that this is the most efficient methodology for this dataset as with removing the noise the model becomes very sensitive to deviations from 'normal' behaviour and thus can greatly increase variability. For the purpose of this project however, this model and approach is more than capable to distinguish outbreak style patterns and anomalies within our dataset.

Having this method in place now, we are able to progress through to the creation of the automatic surveillance system and the development and creation of a dashboard to display the results.

# Chapter 5

# Design and Implementation of an Automatic Surveillance System

Highlighted in section 3.1.1 was the discussion around needing to get life changing health-related results in a timely manner due to the decisions that might need to be made from them. High Performance Computing is necessary for this in order to perform complex and important calculations at great speed and in turn create visualisations to be delivered back to those who want it. This chapter explores taking code running from a local machine and turning it into an automatic surveillance system. It first looks into what high performance computing is and how the High End Computing system based at Lancaster University can be used to do this. This chapter will also discuss a viable data stream that keeps the integrity of any personal and unique identifiers of our animals and their owners.

Following this it delves into the data visualisation side of statistics and the relaying of information back to those who desire it, who in my case are the veterinary surgeons, the stakeholders of the surgeries and the public. It explores data visualisation in a

deep accessability context with a literature review of some popular dashboards and surveillance tools that were created throughout the Covid-19 pandemic. We then move onto the visualisation of the results from the mixed effects model, which can be found in chapter 4 and the creation of a useful tool for this information to be relayed back to those who need it.

## 5.1 What is High Performance Computing?

High performance computing (HPC) is the collecting of computer power to process and perform complicated tasks at speed [3]. HPC is used across many disciplines including science, engineering and business. It is very useful in the context of health research due to the need to get highly influential results in a timely manner so that decisions can be made e.g the need for social distancing measures for a lockdown of a city/town. To put the power of HPC into perspective, an average personal computer at the time of writing can perform billions of calculations a second, whereas a supercomputer cluster (computers with high performance capability) can perform one quadrillion ($10^15$) calculations a second [3].

After originating in the 1960's to aid Government and academic research, HPC began to gain popularity and was used by companies in industry from the 1970's to help create products in pharmaceutical firms, aerospace and in financial services [3].

HPC works by using clusters of powerful processes in order to read and analyse multi-dimensional datasets e.g. electronic health records or data regarding the weather worldwide.

A paper by Ko et al 2022 [53], describes the power of high performance computing in a medical statistics setting. They analysed the onset of type-2 diabetes from 200,000

patients and approximately 500,000 single nucleotide polymorphisms using HPC [53]. They found that fitting the half-million-variate model took less than 45 minutes on the HPC, further proving the power of the HPC and how important computer power is in having to make life-altering decisions in a timely manner.

## 5.2 The Lancaster High End Computing Cluster

The High End Computing (HEC) Cluster is a service ran at Lancaster University to aid researchers who need high end computing services and high throughput computing. Currently, as of 2024, the facility offers 13,000 cores, 24 Tesla V100 GPUs, 59TB of aggregate memory and 230TB of GPFS-based file storage [72].

The HEC is comprised of 3 components: a login node, the compute nodes and dedicated file systems [72]. Users connect to the HEC login node via SSH. From there they can submit jobs to a queue for execution on the commit nodes. Users have separate private file store within the HEC.

## 5.3 Data Engineering

Data engineering is the concept of designing and implementing of systems for the collection, organisation and processing of data. A system like this in place allows for a researcher to utilise the data to find meaningful results. At the beginning of the project an infrastructure like this was not in place and is a necessity for the timely and automatic analysis of our data.

### 5.3.1 Pre-existing Infrastructure

Information regarding the main dataset used for this project is examined in-depth in section 2.1, but for the purposes of this chapter I will provide a summary. Once a consultation is finished, the veterinary surgeon is presented with a questionnaire in which they have to select the main reason for the consultation. Following this completion, the consultation data is sent to University of Liverpool. The challenge here is to create a system in place which smoothly moves data cross institutionally without jeopardising GDPR, automatically analysing the data and then relaying the results back to University of Liverpool/the veterinary surgeons and stakeholders via a method of display. An ideal pipeline of data and results flow for the project can be seen in figure 5.1.

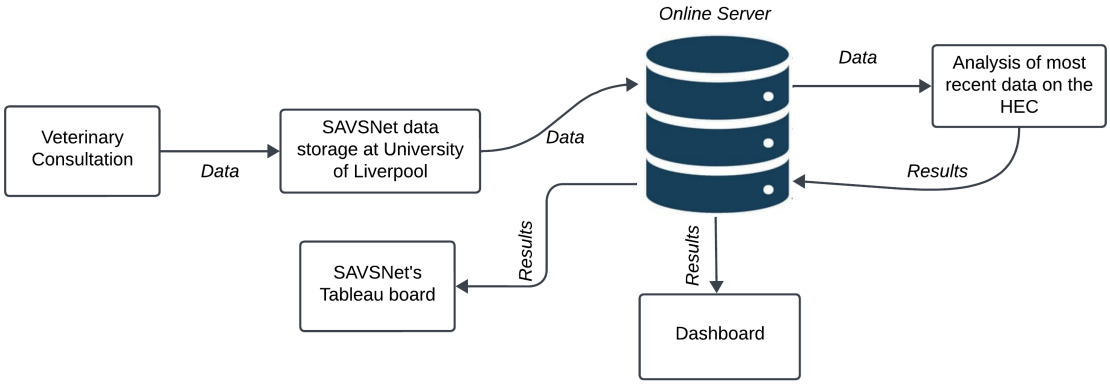



Figure 5.1: Data flow diagram illustrating the SAVSNet pipeline, from the collection of veterinary consultation data through storage, analysis and finally visualisation of results on dashboards and reporting tools.

### 5.3.2 Initial New Data Infrastructure

At the beginning of this thesis with no data engineering infrastructure in place, this data was then being sent via Microsoft Teams to Lancaster University where analysis was done and figures of results were sent back over to Liverpool over Microsoft Teams. This setup, albeit not ideal was unavoidable due to the rapid response needed to the January 2020 outbreak in Canine Enteric Coronavirus. At this point, the model was still being developed and ran on my personal machine which lengthened the time to relay results due to the inability to run jobs in parallel.

Once the outbreak had concluded, work began to shift towards creating a system for a smooth data flow for the consultation data through to the results being relayed. This meant that I could also focus on moving analysis to the HEC to allow for quicker analysis. At this point, I was manually adding veterinary consultation files, running the analysis and saving the plots locally to email to University of Liverpool.

Although the code had been mounted to the HEC which reduced the analysis timing due to the ability to run models parallel, it is still inefficient due to the manual aspects. Further, there is still the issue of holding unique identifiers in an unsecure location. It was at this point we began working with the aggregated dataset described in section 2.1.2, which had a spatial component, date of the week end, counts for each of the MPC's and total consultations for this week. Moving to the aggregated data eliminated any GDPR issues as there was no unique identifiers within the dataset and also solved an issue with disk space due to the file size being vastly smaller.

### 5.3.3 Revised Improved Infrastructure

The pipeline between the veterinary surgery and University of Liverpool had been set up prior to the starting of this project, so the pipeline needed to be between University of Liverpool and Lancaster. Once the analysis was mounted onto the HEC and the security risk of having unique identifiers in the dataset was resolved, I began to look at developing a streamlined pipe to transfer data from consultation through to the results.

There are many options for data sharing cross-institutionally including networked file sharing, data repositories, remote shared database access' and object storage servers for secure file transfers. Although data repositories e.g. dataverse, Redcap, CKAN; networked file sharing e.g. Office 365 or Dropbox and remote shared database access e.g. MSSQL or PostgreSQL are well known and well used, they lack easy API access for services to use and fine access control. Object storage servers advantages are the other mentioned disadvantages in that they have fine-grained access control for data transfer, user and service authorisation and standard API methods compatible with the AWS S3 storage system. A type of object storage server is MinIO which is an open-source server. This was the choice for a platform to house the data and results as it is AWS compatible, is on a free subscription base, offers local hosting and is open source.

Now we have chosen an object storage server, we are able to link this to the code using the python package S3FS. University of Liverpool were able to create a stream their side so that the weekly aggregated veterinary consultation data would also be saved onto this server. The idea was also that the result data and plots would also be stored on this server so that University of Liverpool also had access to it, alongside

being able to create a dashboard that can also read the results. At this time in the creation of the pipeline, I was using Crontabs on the HEC to schedule analysis at weekly intervals in which it would obtain the most recent data on the server and run the python code automatically. It would then write the results and plots back to the online server. At this point we now had this pipeline in place, as an easy way for University of Liverpool to see the results a simple web server seen in figure 5.2 was created to show each of the results while the dashboard was being created fit for the purpose of the veterinary surgeons, stakeholders and the public.

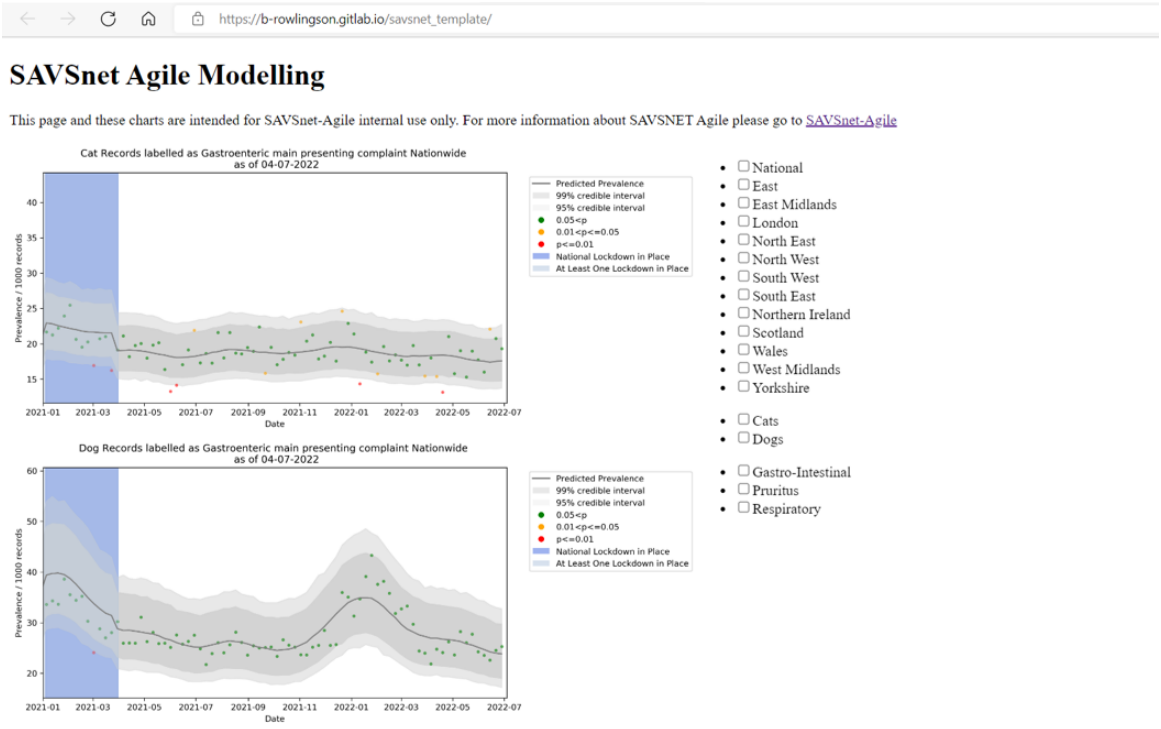



Figure 5.2: Screenshot of simple web server to easily display the results to University of Liverpool as an interim as the dashboard was being developed.

Now we have the basics decided in regards to the data stream and important conversations and discussions about transferring sensitive data cross institutions, the

rest of the chapter dives into the data visualisation and the dashboard creation.

## 5.4 Data Visualisation

This section will explore the basics of data visualisation including different ways it can be communicated to people of both statistical and non statistical backgrounds with focus on the various outputs people can access the information they are wanting. There is also an emphasis on SARS-Cov-19 and how the constant exposure to data has changed the public's view of statistics.

### 5.4.1 What is Data Visualisation?

Data visualisation is the displaying of data by the means of visual representation e.g. graphs, charts and maps. The displaying of data is vital in regards to spotting and understanding any underlying trends or stories that the data holds [88].

Data visualisation is an important tool of relaying information back to those who would benefit from seeing it, for example in our case veterinary surgeons, stakeholders and domestic animal owners. Thankfully the aspect of data visualisation is broad and there is the ability to display results to a very wide demographic of people in various forms of media. The data visualisation can be either through an interactive tool or a still image of a graph or map, this would be at the choosing of the researcher given what they are wanting to relay. There is the addition within an interactive tool to cater to every demographic within the same space, this could be through the add on of informational hovers, alongside extra pages or 'tabs' displaying information regarding methodologies or definitions.

An effective tool from this research will be able to aid veterinary surgeons and stakeholders to make informed decisions regarding their surgeries e.g. buying the right amount of medicine when it is known there will be an expected peak of gastroenteric illness around Christmas and Summertime. There is also the wanting to relay information back to the public, as of February 2023, 34% of households in the United Kingdom own a dog, while 28% own a cat [60].

When there is results in the interest of the public and with public and animal safety in mind, it is important to allow those that could be affected an avenue to look without being overwhelmed by statistical results.

### 5.4.2 Examples of Data Being Communicated

Relaying health related results back to the public is crucial to ensure there is a trust and confidence from the public to the researcher. Health research is likely to effect if not everyone, a vast majority of people whether it be human or animal health. Reasons for the importance of the research were touched on in the introductory chapter.

Over many years, the presentation of data has become a major tool in Epidemiology. Dating back to 1854, Dr. John Snow used mapping to identify the source of the cholera outbreak in London by mapping residents deaths and identifying the water pump that was contaminated [84]. Following this was a tool created and used by Florence Nightingale known as the polar area diagram. She displayed data on sanitation in India which impacted soldiers lives massively by reducing mortality from 6.9 % to 1.8% [57]. The next arguably largest notable public health related issue that required relaying information back to the public was that of Spanish Flu in 1918 - 1920. The death toll of Spanish flu in the UK was an estimated 228,000 people [7],

therefore information needed relaying back to the public in order to help aid policies to eliminate the illness. Upon assessment the display methods of choice were mainly bar charts of deaths and worldwide maps showing the spreads of the waves [7].

More topical for this research project is the illnesses and diseases spread within livestock. Namely the Swine Flu outbreak in 2009 and Foot and Mouth outbreak in the United Kingdom 2001.

The most recent public health emergency however and with the highest access to data visualisation tools is that of Covid-19. This will be explored in the next section.

### 5.4.3 How Covid-19 Changed the Public's View

For the general public, especially an older generation, there is an assumption they have not had access to plots and in depth visualisations unless they truly seek them. The 2020 pandemic was devastating for many reasons, however in the United Kingdom it give the general public the exposure to visualisations that they would not have had. Everyday during the course of the pandemic there was a 'daily briefing' in which the Prime Minister and the Health Secretary would present the daily infection rates and hospitalisations. Gloomy as this was, for a portion of the population it was a recap in how to interpret plots and regular, forced interaction with data visualisation tools.

The pandemic give the experience of allowing the public to be force fed statistics and visualisations in many media streams, e.g. television, newspapers and the internet. There was also the outlet of many dashboards created to display data for the users local area and worldwide information.

## 5.5 Dashboard comparisons

The starting point for creating a dashboard to relay health related information back to users is to review those already available. The importance of Covid-19 and the need to relay information back to the public to help enforce new social distancing measures in turn meant the creation of different dashboards by different governments, academic institutions and from people's personal research. This section will look into the different dashboards created, note what information they're relaying back to the public, how much literature there is in regards to informing about the results/how to read the results and links to other pages for more information and how easy the dashboards are to navigate. There will also be an assessment of another non-Covid-19 related dashboard created by researchers at Lancaster University. This section is not to assess the values of Covid-19 and other health conditions displayed, it is to assess the dashboards.

The first, and perhaps most relevant dashboard for those who wanted to see figures and numbers for England during Covid-19 is the Government ran dashboard. Seen in figure 5.3, this dashboard up until December 21st 2023 showed the weekly and daily new case count, weekly and daily fatality count and hospitalisations along with vaccination counts and testing figures. This data has since changed and the dashboard is used by the UK Health Security Agency to display all respiratory virus' which includes Covid-19. The first page is a summary of all the information they have shown primarily with time series plots with the choice to search down to your postcode for results in your area or look at interactive maps of England for cases and vaccinations. The colour scheme is a basic white and blue and the figures are basic in that they just show counts, no model results. This is helpful for those with accessibility issues and

little knowledge in the statistics area. When the user clicks into one of the areas of interest e.g. case count or death count, that's where there are some more interesting plots including a heat map of case count split by age which is the first demographic split we see. Further down to the end of the page there is also the cases split by Covid-19 variant. In summary, these pages that delve into each of the main areas of interest with Covid-19 begin by showing a top-line overview of that section which then goes further into the splitting of data given demographics and variants for those that have an interest in that. Data visualisation aside, there are also pages to inform the reader of any new updates, an option to download the data, a developers guide and a full page of descriptions and explanations of the different aspects of the dashboard, data collection and specific definitions the data adheres to [96].

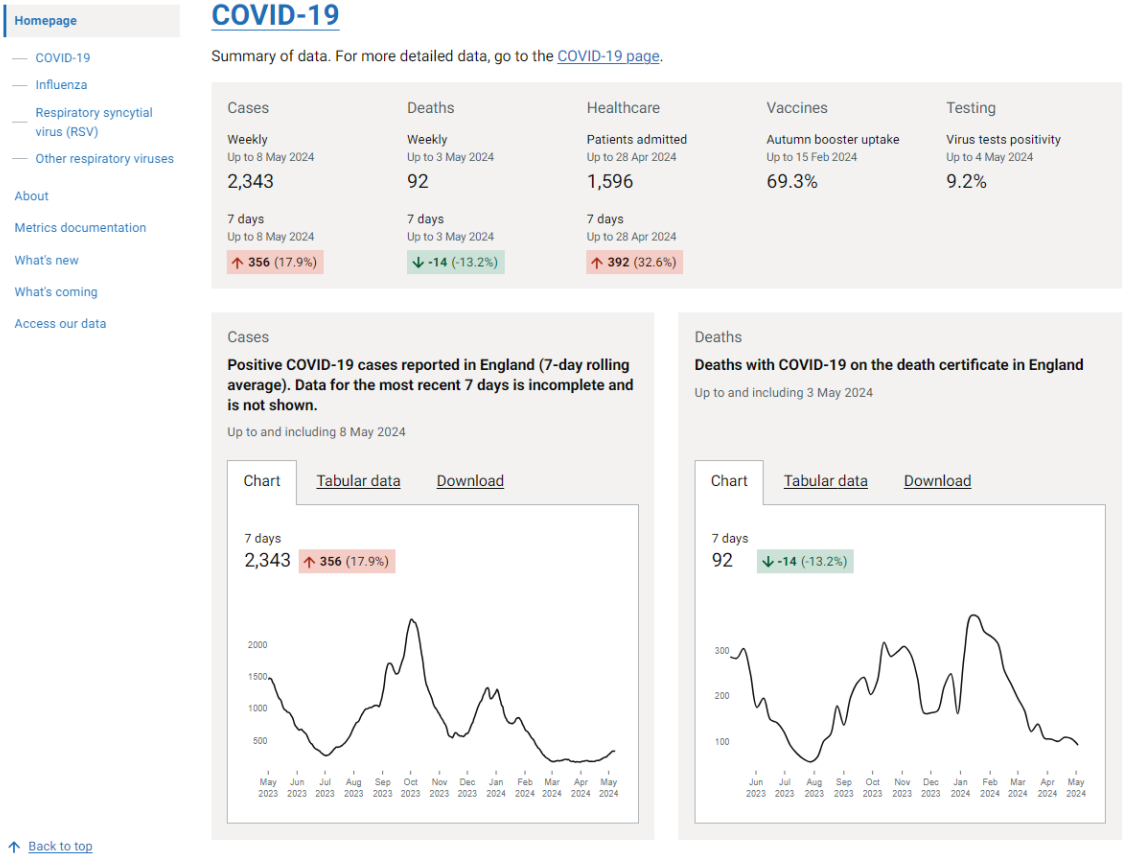



Figure 5.3: Screenshot of the UK Governments dashboard for Covid-19 [96].

The second arguably most notable dashboard during the pandemic that isn't the government dashboard was one produced by John Hopkins University seen in figure 5.4. Even though they stopped updating the tool on 3rd October 2023, during the pandemic it was a staple to see an overview of not just of a specific county but worldwide. The initial page shows an interactive world map where you can filter by country, along side it has figures of cases and deaths split by country and a global figure. They also have a vaccination figure too. The colour scheme is black/grey, green, red and white. The use of green and white could prove difficult for those with accessibility issues however the figures they are showing are not needing to differentiate

different results via colour, the figures are simple and show a single measure e.g. cases over time or deaths over time. The dashboard itself is just an overview with extra pages that show FAQ's about the data and Covid-19 in general, information about the data and a link to a post showing how they modelled Covid-19 and access to the raw data [49].

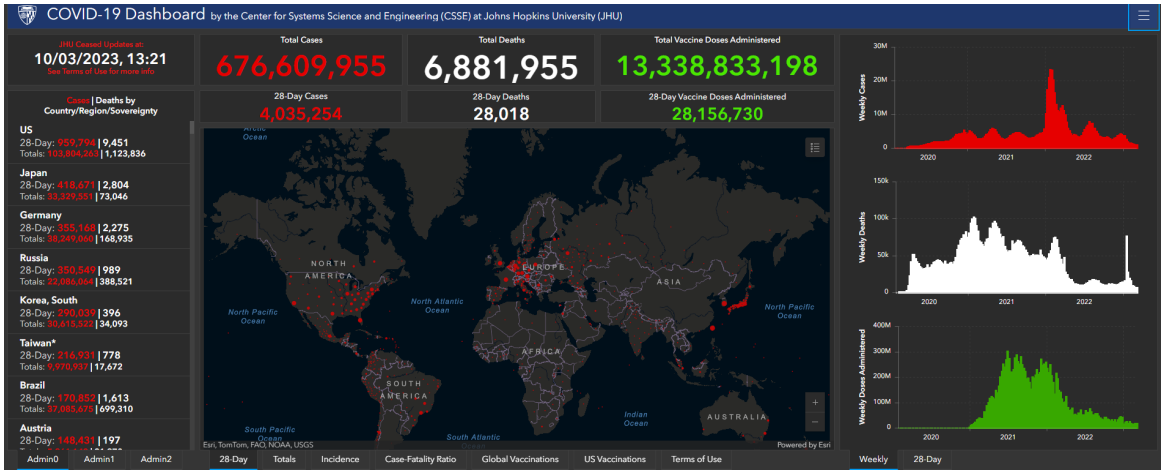



Figure 5.4: Screenshot of John Hopkins Covid-19 dashboard [49].

The final Covid-19 dashboard that will be discussed is the one created by the World Health Organisation. This dashboard contains several pages each dedicated to a single measure, these are cases, deaths, vaccinations and variants. There is also an option to download the data and an 'About' page which describes different visualisations and has any update information regarding the dashboard. The main visualisation on each of the pages is a world map, with a different sized point to illustrate the amount of data for that country. The continents are also split by colour. There is the option to look at the data on a 7 day window, a 28 day window or a cumulative value. Further down on the pages is a look at the data on a time series plot showing the weekly count of the different metric. Following this figure is a stacked time series plot split

by continent. Overall a useful and easy to use dashboard showing basic information through simple plots with no statistical modelling results [103].

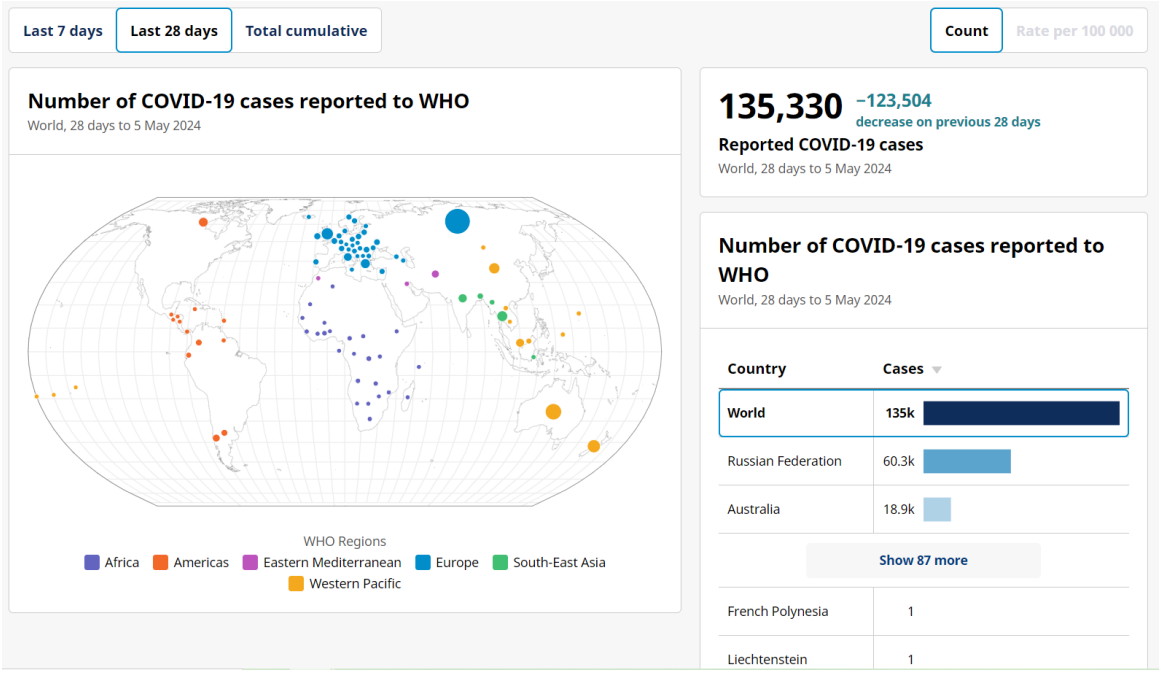



Figure 5.5: Screenshot of The World Health Organisation's dashboard for Covid-19 [103].

Finding a dashboard from an individual researcher was a difficulty as there is usually not a want to have more in-depth information open to the public as not to create issues. The final dashboard I will review is the Dynamic Health Atlas created by Hale and Diggle [39] seen in figure 5.6. The dashboard shows two separate health scenarios across different pages, one being prevalence of chronic kidney disease in Salford using NHS records, and the other being Dengue cases in Sri Lanka. The dashboard has a map component and a line plot when a specific area is selected. The colour scheme for this dashboard is blue-scale with a red line plot, which demonstrates the need for an accessible colour scheme. This dashboard also displays statistical

inference information with the aim to create an impact for researchers and those interested giving them a tool which shows a health concern in a specific area [39]. As the dashboard is built for researchers and medical professionals alike, there is a simplicity with it being there are only two graphics on the app which filter as the user makes selections. A simple app is easier for readabilities sake and not overwhelming to the viewer.

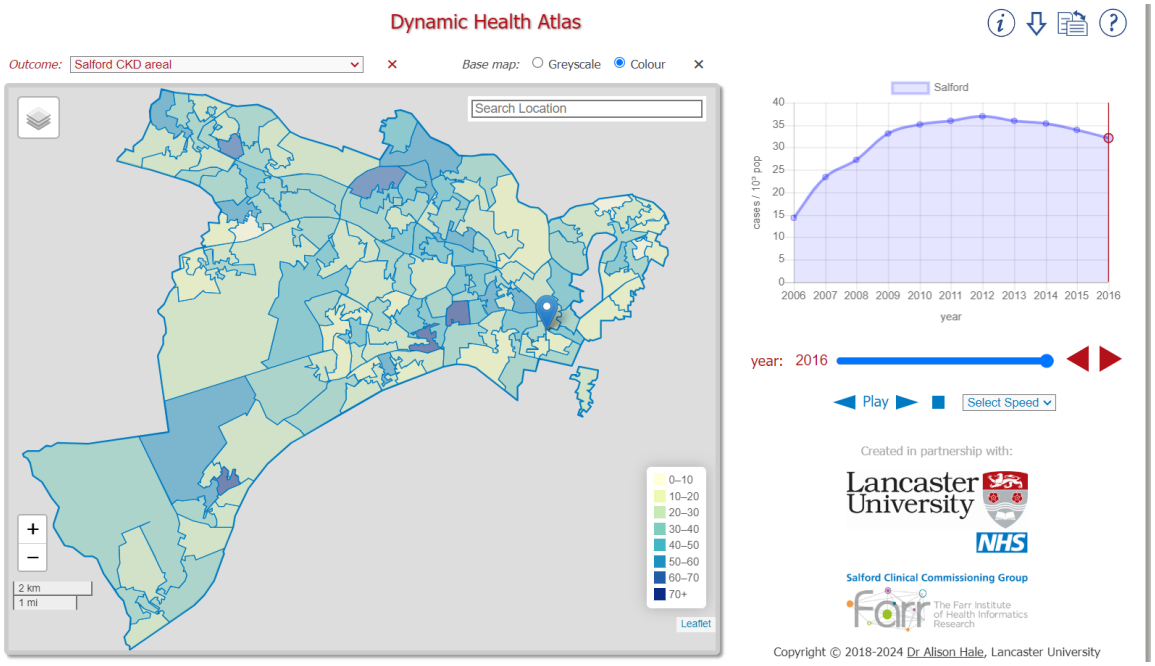



Figure 5.6: Screenshot of the Health Atlas dashboard created by Hale and Diggle used to show multiple health scenarios worldwide [39].

The common theme between the first three dashboards, perhaps for the simplicity of the user is that they're using many figures to each display one time series and not showing any statistical modelling results, simply just count data with a weekly percentage change. The aim for the dashboard I will create is to show the statistical modelling results as an interactive time series plot split by species and region, thus

introducing the need of colour to differentiate between the meaning of different values on the same plot. This will require the use of a specific colour scheme to accommodate for those people with visual accessibility issues. The choice not to include simple count data plots is due to the fact the data can be somewhat meaningless without any context, as from an initial glance of the count data there could appear to be a larger number of cases of gastroenteric main presenting complaint in December, however this is expected due to the seasonal component of the syndrome but this might not relay in a count data plot.

## 5.6 Accessibility Highlights

Accessibility is a part of data visualisation which is an essential to have the most people be able to read the results. This section looks into the main accessibility issues and explains how to help accommodate for them.

### 5.6.1 Colour Blindness

With the idea of displaying results in a traffic light pattern, one could consider colour blindness to be one of the most important and influential factors when considering data visualisation. Colour blindness means that a person sees colours differently than other people, and although there are tools out there to try combat colour blindness, there is no cure for it [65].

Colour blind awareness state that approximately 8% (1 in 12) men and 1 in 200 women in the world suffer from colour blindness of some kind [18]. So as colour blindness affect 4% of the male British population alone it's important to tailor to

this statistic.

In total there are 7 types of colour blindness, 4 of which are red-green colour blindness, 2 are blue-yellow colour blindness and also complete colour blindness [65].

The different types of red-green colour blindness are explained here. Deuteranomaly is the most common type of red-green color blindness and makes green look more red whereas Protanomaly makes red look more green. Both of these are usually mild cases and doesn't get in the way of day to day activities. Further, there is Protanpia and deuteranopia which make a person unable to tell the difference between red and green.

The different Blue-Yellow colour blindness types are described here. Firstly there is Tritanomaly which makes it hard to tell the difference between blue and green, and between yellow and red. Secondly, Tritanopia makes colours less bright and also unable to tell the difference between blue and green, purple and red and yellow and pink.

Finally there is complete colour blindness which means a person only sees in black and white, however this is quite uncommon [65].

The below figure 5.7, shows the different known colour blindness variants and how each one sees different colours.

### 5.6.2 Blindness/Visually Impaired

Users with strong visual impairments and total blindness will need extra aid in order to receive the results. Usually this would come in the form as a text to speech operator. This can be an assessment into the different dashboard software to see which options are available and attempt to build this into the final dashboard.

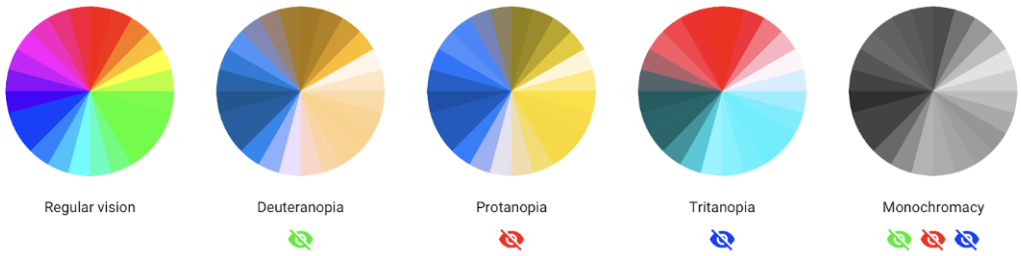



Figure 5.7: Examples of how different colour blindness's see various colours. [74]

### 5.6.3 Cultural Consideration

Although this work is referring to the United Kingdom only, there may also be people outside of the UK that look at it where culturally there are significant meanings for each of the colours. e.g. in the Western world, red symbolises negativity and poor performance whereas in China red is seen as a lucky, powerful colour [85]. I will be keeping things culturally appropriate with where the data is set and also in line with other health dashboards so using the colour red to indicate danger. There will also be any complications removed with the meanings of the colours by use of a descriptive legend.

### 5.6.4 Neurodivergent Considerations

There are various levels of understanding when it comes to statistics and interpreting visualisations, so the creation of the dashboard has to be tailored to all different levels of understandings. The dashboards that were earlier reviewed did this by posting simple count data plots, however as previously discussed this is not a viable option for our results. The way we can combat that is to include a page within the dashboard that discusses and described the statistical methodologies to different

levels. To include all the information without boring or overwhelming the reader, I will use drop down boxes which contain further information in so it is there for those that want to read it. The first box will be always visible and will describe the methodology in a couple of sentences, with the next box going into further detail about the methodologies and the final box describing the model applied to the data. The use of colours will also make the plots easier to read, aiming to use a traffic light colouring system will allow the reader to quickly see what statistical level the weekly prevalence count is at. There is also the choice of language when aiming to not alienate a group of people for example, there may be a demographic of people that have never crossed the word prevalence and therefore do not understand its use, an easy way round this is to change this to 'cases per 1000 consults'. The pandemic helped in many ways in terms of exposure of data visualisations but it's incorrect to assume that every user is familiar with the same terminology. Another way around using specific terminology is to introduce a glossary with word definitions, but the aim of this tool is to for people to get a quick look into results and not have to look further than they have to to find the definition of a word.

### 5.6.5 Motor Impairment

There are also people who have motor impairments, in that they might be unable to use their mouse correctly and thus require to use a keyboard to navigate around dashboards. This will have to be a consideration when assessing the different dashboard software's and whether they have this when the dashboard is in browser mode. [85]

## 5.7 Building Based on Criteria

This section, following what was learnt in the previous will go through the different ways the accessibility's can be catered for in the creation of the dashboard.

### 5.7.1 Colour Blindness Consideration

So given what we know about colour blindness and just how many people in the United Kingdom alone it affects, this would be a main motivator to the creation of the dashboard. In a standard traffic light format that has been implemented throughout all of the statistical analysis visualisations within this thesis, we are able to process these through a colour blind visualiser to see how they would appear to people with all the different variants listed above. This can be seen in figure 5.8, where it shows the default Tableau colours for green, red and orange (top left), this same map which had a green colour blindness filter applied to it (top right) and the map again with a red colour blindness filter applied (bottom left). There is an immediate issue with using the standard traffic light colouring system to identify areas of concern as for those with colour blindness, this is difficult to decipher from the maps.

Between the two figures displaying the results of the Gaussian Process (the map showing this weeks colour and the time series plot), it could be argued that the time series plots colours need not altering as there is also the benefit of a visual representation of where the weekly points are sitting within the credible intervals. The maps however, as these have little other context or information except for the colour of each region these will need to be altered and in the interest of consistencies sake, it is also worthwhile to alter the time series plots to reflect the same colourways.

So just how to choose a colourway? There was a highlight in one of the

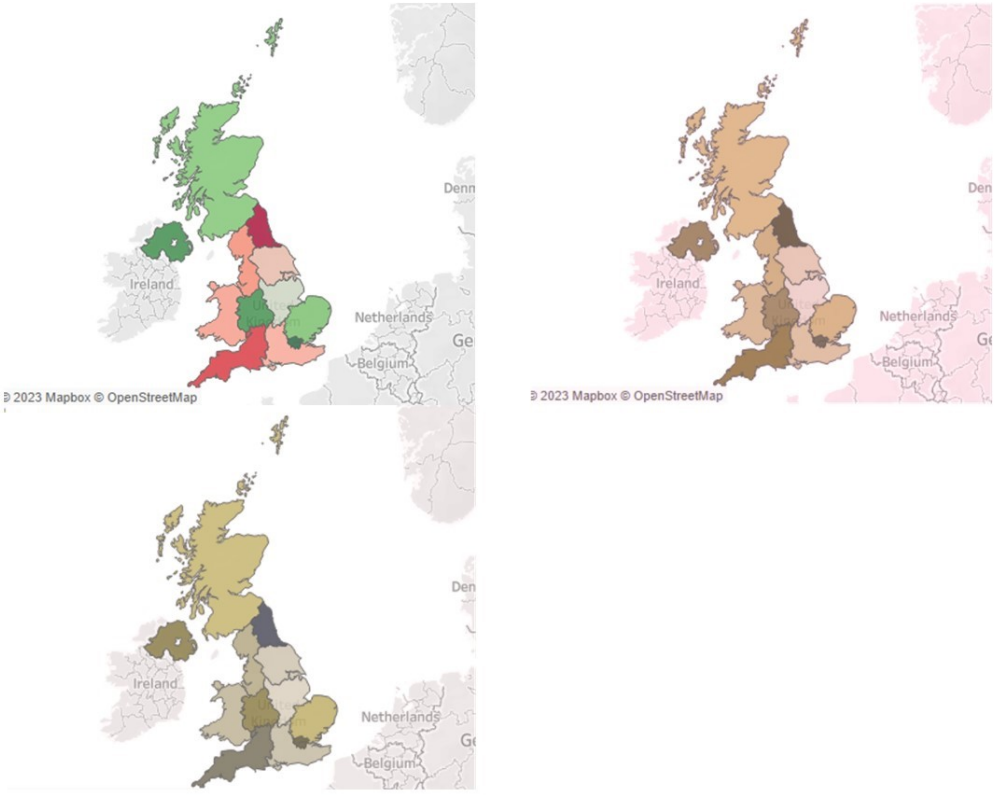



Figure 5.8: Maps of the UK for Dog Gastroenteric MPC consultations coloured by where the prevalence sits on the credible intervals. The top left is the standard green, red and orange that Tableau produces, the top right is the map put through a green colour blindness filter and the bottom left is through a red colour blindness filter.

presentations by Hannah Thomas [92] at the Communicating Mathematics to The Public workshop ran by Isaac Newton centre at University of Cambridge in January 2023 which explained a recommended categorical colourway that can be used that is suitable for all known colour blindness variants.

In order to keep with a standard idea of a traffic light system which is universally known the appropriate colours to choose would be Turquoise (hex code #28A197) to represent points within the credible interval, Orange (hex code #F46A25) for those

| Colour | Hex Code | Example Of Colour |
|---|---|---|
| Dark Blue | #12436D | |
| Turquoise | #28A197 | |
| Dark Pink | #801650 | |
| Orange | #F46A25 | |
| Dark Grey | #3D3D3D | |
| Light Purple | #A285D1 | |

Table 5.1: Table of the colour scheme suitable for people with certain colour blindness, with the Hex codes and an example of the colour from the talk by [92].

points between the credible intervals and Dark Pink (hex code #801650) for those that are outside of the upper credible interval range. The choosing of these colours along with an appropriate key allow for the easy interpretation into the plots.

### 5.7.2 Choice of Language and Words

In the creation of the app, it needs be in the forefront of the mind regarding the words and type of language used. The assumption is that some users of this app (the public) do not have mathematical knowledge past GCSE level i.e. they have not done mathematics since the age of 16. Thankfully, in the light of recent health emergencies and the constant relaying of statistics through various media platforms, the hope is that the public have more of an understanding of what the plots are showing and what the statistics mean in relation to themselves and their area. That being said however, it deems appropriate to simplify the language so that there is no confusion and to eliminate any opportunity for elitism, for example changing something so simple as 'prevalence' to 'cases' can eliminate the need for any external research into definitions.

## 5.8 Dashboard Software Decisions

This section will look into the various software to display results at an interactive level and debate them. The most appropriate software will then be chosen.

### 5.8.1 Dash, Shiny or Tableau

Dash is an open-source framework used for building apps within Python [13]. Following its release in 2017, it has evolved to include implementations for R, F and Julia [13]. Dash is used for building visualisation interfaces and helps the developer create usable dashboards and apps without them having an intensive web development background [13]. Dash uses libraries within Python that already incorporate JavaScript thus knowledge of this is not necessary. It is built upon the Flask, Plotly and React libraries which are the web development framework for Python, and provides a range of visualisation tools, drop-downs and sliding components [75]. Its simple use of the Python language to create a useful and insightful data visualisation tool is what makes it an attractive option to creating the dashboard.

Similarly to Dash is Shiny, which is an app building tool within the R language. This is also an open-source framework and has the ability to create powerful web applications for the aim of data visualisation. Shiny creates a reactive environment where the creator can choose to allow filters on the dataset which the user can use and the current data visualisations will automatically show the filtered data. An issue with the Shiny app is in regards to the publishing of it. There are multiple options to publish however, the majority of them work on a subscription based service and those that are free offer a limited experience. The Social Science Computing Cooperative is

a hosting platform for publishing and running Shiny apps on a server. Deploying the shiny app is an option here, but they are limited only to users with a paid account [98]. They expressed this means the app can be slow with more than one user, especially when there is statistical analysis being performed. One advantage that we have here, is that the analysis is being performed by an external HPC and we are simply plotting the results in our app, not performing the analysis within it. This means that we have a bit of a buffer in choosing a publisher as we're using the app as a display of the results and not to perform the statistical analysis.

Tableau was founded in 2003 off the back of a computer science research project whose main aim was to make data visualisations more accessible. Since it's origin, the founders have invested heavily into the development of Tableau to ensure that users get visualisation results faster and help visualise unexpected results [89]. Tableau is a subscription based service and currently have over one million active members. The software is able to show data in all various formats including maps and different plots and tables. There is also the option to make these interactive for the user through use of hover boxes and subsetting options to show the user more concise information, be it data reflecting an area of the UK or a specific illness. Tableau also has other applications that allow for easy manipulation of datasets ready to be input to the Tableau visualisation tool.

In the interest of the grant and from a charity funders standpoint, the idea of creating a dashboard in an open source format will be more of an appeal to a subscription-based platform like Tableau. Those dashboards being hosted at Liverpool will be hosted from the University of Liverpool's account so a tableau dashboard is adequate in this circumstance however from the funders perspective, a dashboard

created in an open-source language is more sustainable.

A Tableau page displaying these results to be hosted on the SAVSNet website could be seen as a simple duplication, however there will be extra information within the Shiny App that is not on the Tableau dashboard. The argument also for creating the Tableau dashboard is that for some of the veterinary surgeons and stakeholders that already know where to see these dashboards, they'll be able to go to a familiar place to view them and they'll all be under one basic URL. Whereas the idea for the app is to be hosted at Lancaster University containing more in depth information.

As we're performing the analysis on a separate basis and saving those results elsewhere for the app to retrieve automatically, Shiny in R seems to be the better choice due to its advantage in data visualisation over Python given the strength of visualisation libraries in R e.g. ggplot and plotly. These libraries are able to integrate seamlessly with a shiny app [50]. Another advantage is the active community within Shiny as there are thousands of peer reviewed applications and dashboards across domains [50]. A final reason is that Shiny has greater accessibility and collaborated in that I allows non R users to interact with analysis and visualisations ( [50].

## 5.9 The Final Shiny App

This section shows the final dashboard created with commentary regarding decisions implemented given the research into the topic.

### 5.9.1 Home Page

When creating the Home Page seen in figure 5.9, the main criteria would be that any user would be able to, at a glance, get a indication of the position of their area for the species and syndromic illness they're looking for. The page also contains information regarding how to read the maps, what the colours mean and where to find further information within the SAVSNet website.

Following the investigation regarding different visual impairments people may have, the colours still follow a traffic light system but are suitable for near all impairments to see. For the people that have complete colour blindness, there is also a hover box added to each of the regions that show that weekly count. The example hover here is shown on Scotland for Dog Gastroenteric. The stakeholders and veterinary surgeons are able to see this page and observe current rates for their areas and then can choose whether to look further into how many week consecutively there has been cause for concern, say.

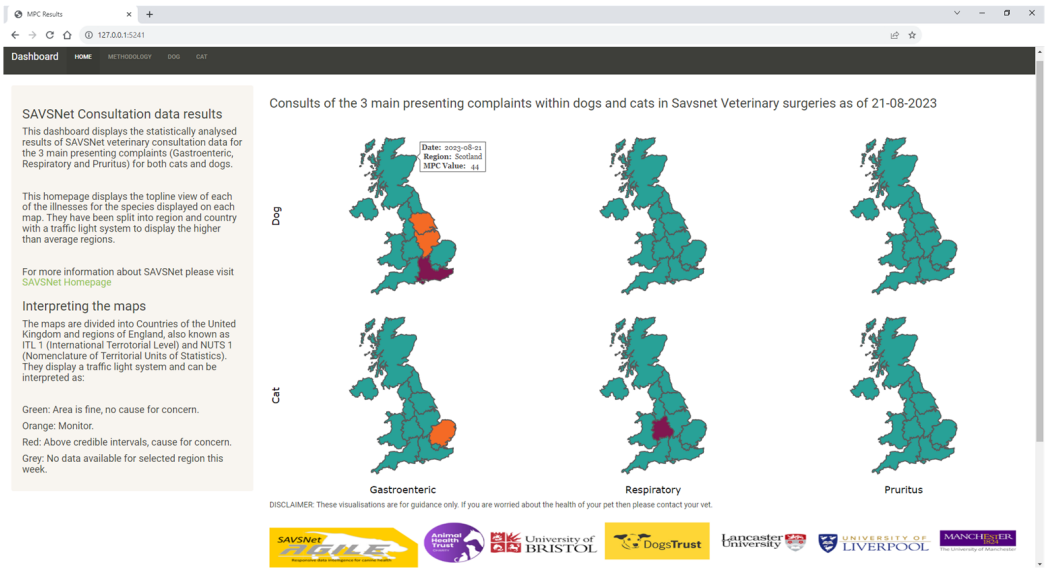



Figure 5.9: Screenshot of the Home Page of the app with a panel on the left describing the showings and how to interpret the maps. The main panel shows maps of the UK split by regions coloured by that weekly prevalence value split by species and MPC.

An emphasis of this chapter and the creation of this surveillance system was on accessibility issues and thus this influenced the colour scheme of the maps. Figure 5.10 shows the homepage of the surveillance system through the same filters as 5.8 and there is a vast difference in choosing the colour scheme. The top right homepage in the figure shows what users with green colour blindness would see, where the bottom left homepage shows red colour blindness. The change was obviously necessary as instead of a few different shades of grey for all the red and green areas in figure 5.8, there is 3 distinct colours in the maps which will allow for easier interpretation by

those with this accessibility issue.

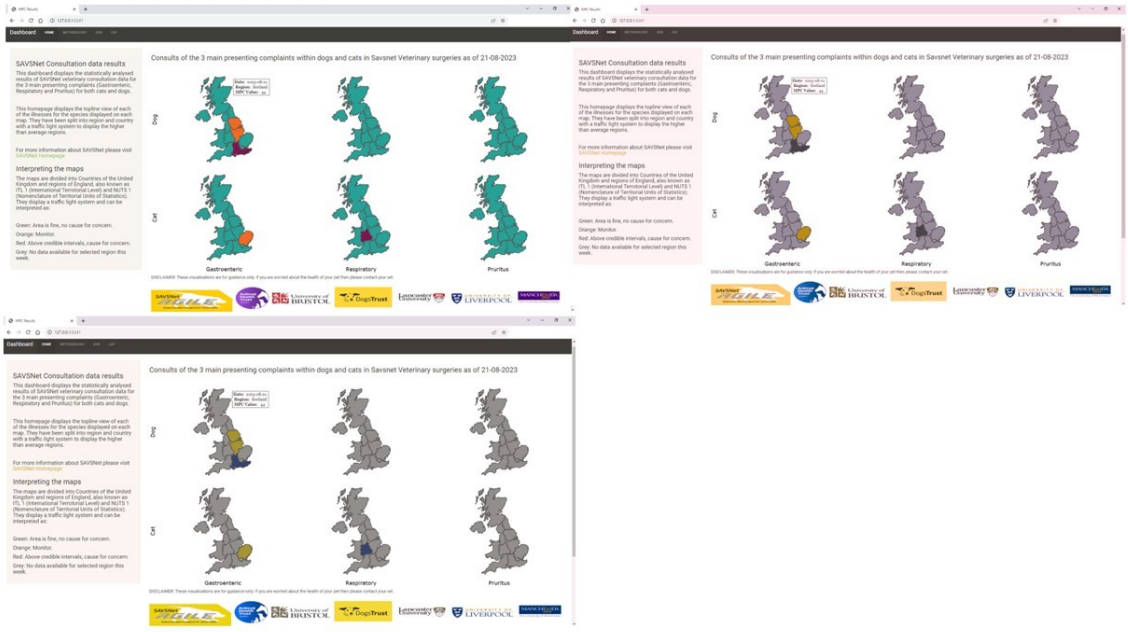



Figure 5.10: Screenshot of the Home Page of the app with the addition of the Home Page having a green colour blindness filter (top left) applied to it and a red colour blindness filter (bottom left) added to it.

### 5.9.2 Methodology Page

The creation of this page is more where the learnt knowledge surrounding language and speaking styles came to play. The idea was not to inundate the viewer with a lot of statistics from probabilistic theory all the way through to creating a model, so this was tackled by using optional drop down boxes. The page has a basic paragraph in which it simply describes what the methodologies are and a top line view of what they're doing with optional drop downs containing more information if the reader requests.

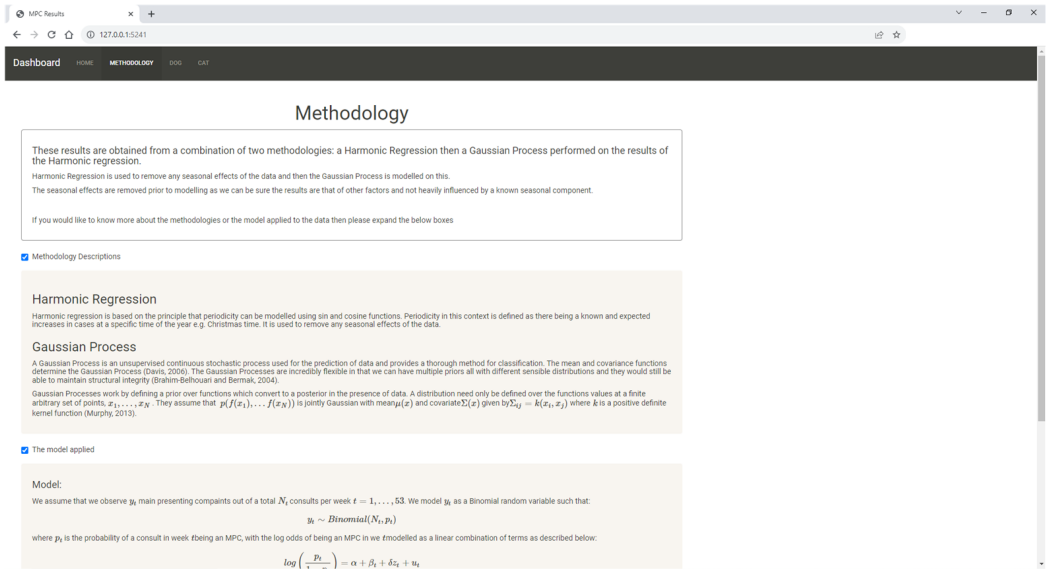



Figure 5.11: Screenshot of the Methodology Page of the app displaying different depths of the methodologies and statistics through some drop down boxes.

It was also an important requirement that the information regarding the statistics be on a separate tab from any of the other information. This again, is to not inundate and overwhelm the reader with information they do not want and/or not seeking. The presence of this methodology page and the different depths is for the reader to use if they would like.

### 5.9.3 Dog and Cat Pages

So following an exploration by the user into the initial page and assessing the levels, they then have the option to look further into the rates for both species across the

different MPC levels by region. These are split into separate pages for the species however, the design of them is the same. These can be seen in figures 5.12 and 5.13.

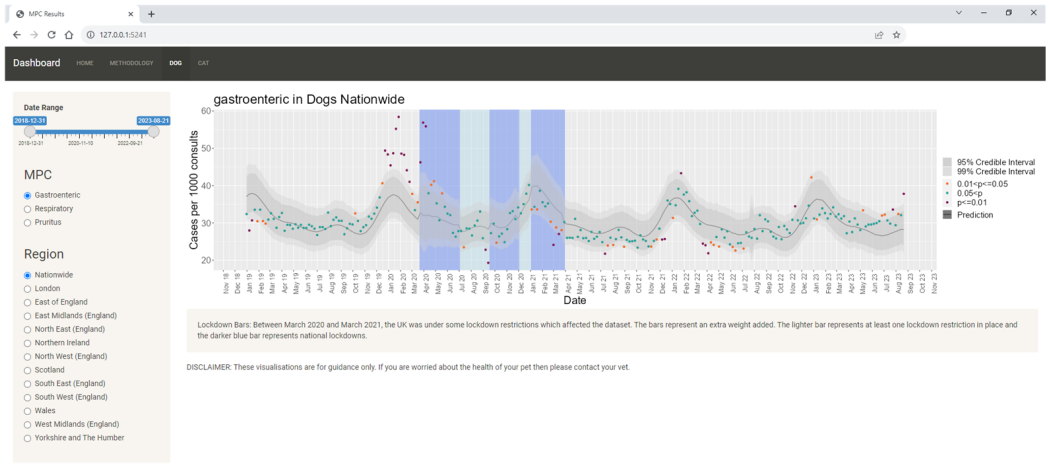



Figure 5.12: Screenshot of the Dog Page of the app. The left panel contains filters including the region and the MPC and a sliding window for the timescale. The main panel shows the time series plot for the results from the Gaussian Process back to January 2019.

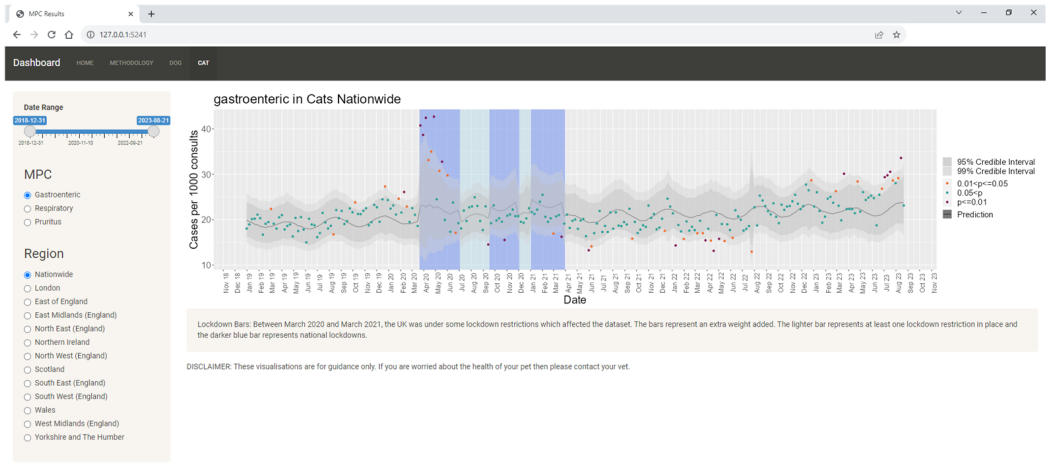



Figure 5.13: Screenshot of the Cat Page of the app. The left panel contains filters including the region and the MPC and a sliding window for the timescale. The main panel shows the time series plot for the results from the Gaussian Process back to January 2019.

The display for these data is also different, instead of the display of this weeks value on a map of the United Kingdom, it is the whole time series. The colour scheme from the maps follows here for accessibility and consistency reasons. The left hand panel has radio-buttons to select a singular MPC and region, this is for the prevention of adding multiple selections which is uninformative. These pages also offer information at a whole national level which the initial page doesn't. Also on the panel is a time slider in which the user can look to refine the plot by week or focus on a desired time period.

There is also the information regarding the lockdown bars and what they represent which can be an important indicator for users looking specifically for lockdown affected periods.

## 5.10 Discussion

Although for this research project the timing of the analysis using the HEC is sufficient enough, there are always ways that efficiency can be improved, especially within the Python language. This analysis as it stands was performed using the PyMC3 package, however there are other high computation performance packages that would also be able to perform the analysis in a more timely manner e.g. Tensorflow and PyTorch.

Another important decision required for this chapter was the design and implementation of the data pipeline from veterinary consultation through to the results. The main consideration here was using a system or software that was capable of being easily written to while maintaining sufficient security. It was also essential that both the institutions involved in this portion of the project had access to it o that data could be uploaded to it and likewise the results from us at Lancaster. MinIO was decided on due to its many advantages including that user access had to be granted by the owner of the bucket and access could be revoked easily. It is also secure as the registered users of that specific project require their own username and key to log on. The online server linked nicely with the code on the HEC through the use of the S3FS package and made collecting the recent data and writing the analysed results to it doable. The Shiny App is also able to easily obtain the recent results and create the dashboard based on them.

This final shiny app has a unique take on showing statistical results and

information without being too cluttered and overwhelming for the user. The strategic use of multiple pages in the dashboard allows the user to access not only a top line view of how the region is doing for that week, but also lets them view historical results for the MPC. There is also the separate page describing the methodologies to different levels and having this information within its own tab reduces the statistical language but also giving it a place for those users that have an interest in it.

The emphasis on making this dashboard accessible to a larger group of people was essential as to not alienate a specific group that might have issues with standard visualisations. The addition of the hovers for the map page showing the values and regions helps relay more information about that week to the user, and altering the colour scheme throughout allows those with colour blindness to be able to use the dashboard without difficulties. The dashboard, once viewed in a browser also has the capability of being controlled with the keyboard, which helps those users with motor impairments.

The link to the SAVSNet page is also an essential as there are more information and more data visualisations through that website that the user might also be interested in and allows the reader to gain more information or definitions that may not be on this dashboard. Saying that, the simplification of the language throughout the dashboard removes the need for a definitions page as the aim was to make this tool as simple and user friendly as possible.

Following the completion of the Surveillance tool that was also created with the veterinary surgeons and stakeholders in mind, I gathered feedback from Dr Carmen Tamayo Cuartero, whos PhD thesis focused on the communication of results back to veterinary stakeholders. Early on in her research project she conducted interviews

with veterinary surgeons to assess their requirements. Her feedback on the surveillance tool is as follows:

"This dashboard is a useful tool for veterinary surgeons. It gives a quick overview of common conditions across the country. In interviews for this research, veterinary surgeons mentioned that a brief summary of consultation levels for different MPCs in their area would be beneficial. The time series plots for each MPC, categorised by species, also help in preparing for possible seasonal peaks. Overall, the dashboard lets veterinary surgeons and stakeholders quickly see important information through colour coded maps. It also provides the option to dive into detailed time series data when necessary."

## 5.11 Conclusion

This automatic surveillance tool is essential to this research as it is the first of its kind with the dataset and allows for the constant surveillance into the 3 main presenting complaints for cats and dogs across the different regions and thus giving valuable insight to people who require it to make life-changing decisions regarding their pets or the patients health.

# Chapter 6

# Discussion

The aim of the thesis was to take routinely collected syndromic veterinary data from veterinary surgeries in the United Kingdom, apply anomaly detection methodologies to them with the interest to relay information and results back to the veterinary surgeons, stakeholders and the general public. This thesis successfully explores approaches to outbreak detection methodologies and the design and creation of an automatic surveillance system.

The thesis builds on research by Hale et al [40] but for a longer period of time (years as opposed to days in Hale et al's research) to introduce a seasonal element to the research while also focussing on a larger spatial area to model anomalies for the UK.

At the start of the thesis there were multiple data sources; veterinary consultation data, laboratory data and two natural language processed datasets. Further into the thesis and as a reaction to the introduction of Covid-19 social distancing measures, a lockdown indicator dataset was created. As the veterinary consultation data is the result of a pre-made questionnaire and not free text, it was clean and needed

little manipulation and imputation of missing values. The discussion arose about condensing this dataset to remove information with no value to the research and to remove personal identifiers. This was done for two reasons; to reduce file size for storage of weekly files and to remove GDPR concerns for the pet owners if there was to be a data leak. The condensed dataset contained a spatial indicator, date of the end of the week collected, total counts for each of the main presenting complaints (MPC) and total consultation count for that week.

The laboratory data contained pathogen test results from samples sent by the veterinary surgeries. It also included spatial components so we could assess spatially any increased prevalence of specific pathogens. It was later decided to not use the laboratory data due to the unknown lag between the laboratory receiving the sample and the test being done. Further, there are PCR tests that the veterinary surgeon can perform at the consultation which reduces the power of the dataset.

Both the natural language processed datasets were used as a proof of concept for research performed outside the scope of this thesis. These datasets were included within this thesis as a different use case to show the versatility of the methodology explored in Chapter 3.

Finally, the lockdown indicator dataset was a necessity when modelling within the time frame due to the introduction of social distancing measures as a result of Covid-19. This was a binary indicator of whether parts of the country were in lockdown or not. This was done at country level as it was an immediate reaction to new guidelines being introduced weekly using the government websites for England, Wales, Scotland and Northern Ireland. This lockdown indicator was able to salvage data that would have been lost due to the closure of veterinary surgeries for routine appointments

during the first lockdown period ($23^{rd}$ March 2020 to $11^{th}$ May 2020).

### 6.0.1 Predictive Anomaly Detection

This chapter of the thesis covered Gaussian Processes (GP) as a methodology for the use of anomaly detection. It was found that the GP were useful at detecting anomalies for each of the MPC's for data at a higher spatial resolution i.e. more data. We also found that the GP was able to absorb the mass loss of consultations as a result of social distancing measures introduced in the United Kingdom during 2020 and 2021. Having these consultations accounted for by adapting the model allowed us to make confident conclusions regarding any anomalies.

Summarised are the main disadvantages of using a GP, which led the thesis onto using a different approach. The first and obvious issue was that as the data were fewer, the models were over-estimating the Binomial noise. Having this aspect of the model made it impossible to confidently make decisions and to report conclusions back to the veterinary surgeons and the public, which meant we could only make secure conclusions mainly at a National level.

The second disadvantage to this approach was computational time, where before the use of Lancaster Universities High Performance Computer and running the analysis on my PC, just one of the categories took approximately 2 and a half hours. The long computational time contradicts the idea that responses to outbreak style patterns should be performed in a timely manner. It was concluded from here that there would need to be an updated way to run the analysis and to involve more of an automated approach using the least amount of data.

The final, and arguably the most important is that the model as it is currently

built is learning and creating predictions from all the previous data input to it. With the end aim of the thesis to have created an automatic surveillance system, manually removing these outbreak style patterns was not a feasible option. Continually learning from past data reduces the sensitivity of the analysis and therefore removes the ability to confidently report outbreak style patterns. A sliding window time-frame would be an option to help eradicate this issue, however there is then the issue of deciding the length of the time frame to ensure outbreaks are swiftly removed while also giving the analysis enough data to learn from.

### 6.0.2 Model Based Anomaly Detection

This chapter explored a new approach to the research question which was using mixed effect models to detect anomalies in the consultation dataset only. It was successful at spotting anomalies, even down to lower geospatial levels and for MPC's with fewer consultations. This approach not only captured outbreaks that we had witnessed in the predictive anomaly detection in Chapter 3, it was able to also find outbreaks that the initial methodology had missed.

Another positive with this approach was the reduced computational time to generate results. This fits with the ideology that health research and anomaly detection should be performed quickly in order to influence decisions made by those with power. The reduced computational time also has environmental benefits as the High Performance Computer had lower running times. Although the chapter 4 model includes an additional harmonic regression step to remove some seasonal trends prior to fitting the Gaussian Process, this does not add a computational burden. The chapter 3 model takes on average 4 hours and 21 minutes to run a single application

(e.g. dog, gastroenteric, North West). Whereas the chapter 4 model takes on average 2 hours and 46 minutes to run single application (e.g. dog, gastroenteric, North West).

The ability to build known narratives of the data into a model was a useful component for the point of longevity within the planned surveillance system. The idea for the surveillance system was for it to be fully automated, so the more bias and noise that can be removed from the analysis before modelling the anomalies increases the longevity of the surveillance system, and further removed the need for human interaction to remove past outbreaks and seasonality. Over time the model will begin to learn from the past outbreaks, but at a slower pace than the model developed in Chapter 3. This is due to the GP in chapter 3 being trained directly on the full dataset, allowing it to capture both seasonal patterns and anomalies immediately. However, with the model in Chapter 4 (Harmonic Regression and GP), it only models the residual deviations from seasonality. Since these residuals are smaller and noisier the model receives a weaker learning signal from each past outbreak. Consequently, although the model in Chapter 4 does learn from past outbreaks over time it is at a slower pace than the Chapter 3 model using the full dataset.

The assessment and discussion regarding thresholds is important due to the different options that can be given. It is to either give the reader of the plots the freedom to choose their own thresholds, which could result in loss of power and information for the weekly prevalence levels, or to give the standard credible intervals at 99% and 95%. This project is a meeting ground for those concepts where the opinions of professionals who will use these results regularly are taken into account on the definition of thresholds for the different MPC's but kept within reason to still be informative.

### 6.0.3 Automatic Surveillance System

The solidification of an approach that works for longevity of a surveillance system while also allowing us to confidently call out outbreak style patterns means we're now able to complete the thesis by creating an automatic surveillance system. The aim of this dashboard is to relay results back to the veterinary surgeons, stakeholders and to the public to give them visibility on prevalence levels in their counties.

The first discussion needed was creating a path between veterinary consultation questionnaire being submitted, to the analysis being performed automatically to the results then being stored in a place where both the dashboard and colleagues at the University of Liverpool could obtain them. The suggested, and finalised pipeline was shown in figure 5.1. The idea was to store all data and results on an online server using MinIO which both Liverpool and more importantly, the high performance computer at Lancaster could have access to. This ensured the ability to use Crontabs to schedule the analysis at a specific time of the week. MinIO was used due to its comparability to Amazon AWS but without a subscription fee.

The choice of using a Shiny App for the display of the results was a relatively easy decision to make given how renowned it is for such a job, with more literature and real life examples than its Python equivalent. There was also the decision to choose Shiny App over Tableau as the use of an open source software was essential to reducing costs. SAVSNet have a subscription to Tableau at the University of Liverpool which is integrated into their website, so allowing them access to the results to update the figures on that was necessary. Although SAVSNet have a tableau page which displays the results, it is only at a top line level and does not go into depth regarding the methods and historical results as the Shiny App does.

An important element I wanted to build into the Shiny App was accessability. With the idea that this would be displayed back to the public there needed to be some considerations regarding language choice, visualisations that included universal colours with hovers and not an overwhelming amount of statistics. The manipulation of tabs was ideal to relay back all the information I wanted with the reader having the ability to find out more about the methods/results if they wanted to by including a methodology tab. The main page seen in figure 5.9, provides a top line view of how each MPC for dog and cat is reporting in the credible intervals by region with historical results being in the separate dog and cat tabs. All of these visualisations were created using a suitable palette for those with colourblindness.

## 6.1 Conclusion

The aims of this thesis were to explore statistical techniques for anomaly detection using veterinary consultations and through the use of the high performance computing system at Lancaster University create an automatic surveillance system that relays results to the veterinary surgeons, stakeholders and the public. These aims were met and through the use of SAVSNet data and the High Efficiency Computer at Lancaster, I was able to create visualisations and plot results across different platforms (Tableau and Shiny App) to suit different audiences.

This thesis was able to successfully create meaningful conclusions regarding MPC consultations in the United Kingdom and the methodologies were able to reveal outbreak style patterns, one of which was the Canine Enteric Coronavirus outbreak in January 2020. A further outbreak style pattern that was identified was gastroenteric MPC in dogs in Yorkshire which related to another county wide spread of illness.

These outbreak calls demonstrates the success of the surveillance system and it's use to the veterinary surgeons, stakeholders and the general public in a home setting.

There was also a dashboard created alongside this thesis to display the results which had a large emphasis on different accessibility's which many dashboards released during the Covid-19 pandemic failed to incorporate. The dashboard is also an avenue for people to assess domestic animal health rates in their county, an area of research which is, to my knowledge and at time of writing this thesis, vastly under researched and difficult to locate.

Another novel innovation is that the analysis of this SAVSNet data has yet to be done to this level of detail, this allows the veterinary surgeons especially, a point of call for not only current levels for their counties or nationwide, but the ability to assess past behaviour and plan for future potential outbreaks. Research by Hale et al [47], which has been heavily commented throughout this thesis, used a small portion of SAVSNet data as a use case, but this thesis adds to the knowledge by not only providing an updated review of the registered SAVSNet veterinary surgeons but also a broader and more in-depth analysis and reporting system.

## 6.2 Further Work

Although the thesis successfully achieved its aims, there is always further work that could be done, especially on such an interesting and highly research-able area. One concept that was loosely explored before settling on the mixed effects model detailed in Chapter 4 was the Markov Switching Model. This method calculates the probability that you're in an outbreak state given the previous weeks probability of being in an outbreak state. The premise behind thinking of this methodology was to eliminate

the prior discussed issue of the Predictive Anomaly Detection method in Chapter 3 where the model would learn from years worth of information.

In regards to efficiency with code and analysis, there could be some work around the coding and exploring different ways to make it more efficient. The language of choice was Python with MCMC iterations ran using package PyMC3, however there are more packages and coding concepts that could increase the efficiency e.g. TensorFlow and PyTorch. This was not a huge consideration for this thesis as once the analysis was mounted onto the HPC, using parallel programming was efficient enough to reach the aims of the thesis.

The paper discussed throughout this project by Hale et al [40] where they used spatio-temporal methods on the same dataset for the Salford area over a 9 day window had assessed spatial correlation from neighbouring veterinary surgeries. Hale et al are able to perform this analysis as they are using a location matrix of an area of a city, thus reducing computation time and storing of the matrix. Adding a spatial correlation into the existing model designed in Chapter 4 wouldn't be impossible, but as the project has looked at the United Kingdom as a whole, there would need to be careful consideration in the computational time and also where the location matrix for each of the veterinary surgeries would be stored.

Another area that could be improved and re-evaluated is the lockdown variable. As described in Chapter 2, this was a binary variable created as an immediate response to Government decisions for each of the 4 countries as to whether that region had social distancing measures assigned. Off the back of the Canine Enteric Coronavirus outbreak in January 2020, it was imperative that the loss of data due to these measures was salvaged as to monitor the syndromes, most importantly dog

gastroenteric. The approach took in the thesis did work well and was able to increase prevalence levels, however how would it differ to taking a non-binary approach? This non-binary approach would be described as a sudden drop when a country was entered into social distancing measures and from there a gradual recovery as these measures eased. Having this gradual effect would help with human biases when coming out of the lockdown periods with owners not wanting to take their pets to the vet.

The models in each of the chapters could be more developed and fine-tuned, given further development into them i.e. assess different numbers of chains and obtain quantitative metrics to thoroughly investigate the model parameters.

A large expansion and extension of this thesis can be thought of by having access to and using the veterinary free text dataset and performing some Natural Language Processing methodologies for the purpose of detecting any new combinations of symptoms which could contribute to a new illness or disease. This thought comes from the presence of Alabama Rot and how it was defined as a disease by the combination of different symptoms.

Finally, another member of the SAVSNet Agile team completed their project surrounding stakeholder management and how to relay information back. As highlighted in chapter 4, there were some of the decisions from the veterinary surgeons and stakeholders input into the results from that chapter but now that project is finished, further work could go into amending the way deliverables are communicated. From their research, an emphasis was made on threshold reporting and when stakeholders and veterinary surgeries would want to be alerted of an increase of cases [90].

# Chapter 7

# Appendices

## .1 Additional Tables

| Column Title | Description |
|---|---|
| Consultation Date | Date of the Consultation |
| Practice ID | The ID of the practice the consultation took place at |
| Premise ID | The ID of the premise the consultation took place at |
| Longitude | The longitude of the |
| Latitude | The latitude of the |
| Species | The species of animal being consulted |
| GI combo regex | |
| GI mpc selected | Binary column of whether the animal was there for a GI related issue. |
| Gender | The gender of the animal |
| Neutered | Binary column whether the animal was neutered at the time of consultation |
| Age | Age of the animal at the consultation |
| Breed | Breed of the animal |
| Cleaned Postcode | Postcode of the owner of the animal |
| Unwell mpc (main presenting complaint) | A binary column whether the consultation was regarding if the animal was unwell |
| MPC (main presenting complaint) | The main reason the animal was there. Has 108 different inputs ranging from Vaccination, post operation check up, other healthy |

Table 1: Raw Dataset

Table 2: Full main presenting complaint dataset features with a description.

| **Column Name** | **Description** |
|---|---|
| Consult date | Date of consultation |
| Practice ID | ID of practice |
| Premise ID | ID of premise |
| MPC (Main presenting complaint) | The main reason the animal was there. Has 108 different inputs ranging from Vaccination, post operation check up, other healthy |
| GI mpc selected | A binary column whether the consultation was related to GI |
| Respiratory mpc selected | A binary column whether the consultation was related to respiratory |
| Gender | Gender of the animal |
| Species | Species of animal |
| Age | Age of animal at consult |
| Person Postcode | Postcode of the owner |
| Bank Holiday or Weekend | Whether the consultation date was on a bank holiday or weekend |
| Pure or Cross | Whether the animal was purebreed or crossbreed. |
| Premise postcode | Postcode of the premise |
| Premise Latitude | Latitude of the premise |

| **Column Name** | **Description** |
|---|---|
| Premise Longitude | Longitude of the premise |
| Premise easting | Easting of the premise |
| Premise Northing | Northing of the premise |
| Premise postcode area | The inital characters of the alphanumeric postcode |
| Premise Region | Region of the premise |
| Premise District | District of the premise |
| Country | Country of the premise |
| Owner postcode | Postcode of the Owner |
| Owner Latitude | Latitude of the Owner |
| Owner Longitude | Longitude of the Owner |
| Owner easting | Easting of the Owner |
| Owner Northing | Northing of the Owner |
| Owner postcode area | The inital characters of the alphanumeric postcode |
| Owner Region | Region of the Owner |
| Owner District | District of the Owner |
| Person IMD | Index of multiple deprivation score for each of the owners |
| Year | Year of the consultation |
| Month | Month of the consultation |

| **Column Name** | **Description** |
|---|---|
| Day | Day of consultation |
| Week number | Week number of the consultation |

| Region | Total Count | Percentage |
|---|---|---|
| South East | 1529859 | 20.08% |
| South West | 1044865 | 13.71% |
| North West | 831739 | 10.92% |
| East of England | 809251 | 10.62% |
| Yorkshire | 743770 | 9.76% |
| North East | 550661 | 7.23% |
| West Midlands | 520543 | 6.83% |
| East Midlands | 488043 | 6.41% |
| Wales | 366743 | 4.81% |
| Scotland | 338872 | 4.45% |
| Northern Ireland | 149396 | 1.96% |
| London | 102957 | 1.35% |

Table 3: Distribution of dog and cat consultations split by region

| Column Name | Description |
|---|---|
| Consult date | Date of consultation |
| Practice ID | ID of practice |
| Premise ID | ID of premise |
| GI mpc selected | A binary column whether the consultation was related to GI |
| Gender | Gender of the animal |
| Species | Species of animal |
| Age | Age of animal at consult |
| Pure or Cross | Whether the animal was purebreed or crossbreed. |
| Topic columns from 0 to 29 | Binary column whether the consult fit into the column given the text mining determined topics |
| Topic unknown | Binary columns whether the topic of the column is unknown from the text mining algorithm. |
| Year | Year of the consultation |
| Month | Month of the consultation |
| Day | Day of consultation |
| Week number | Week number of the consultation |

Table 4: Topic Modelling Dataset

Table 5: Spatial dataset feature names and descriptions used for geolocation for fields with missing data.

| **Column Name** | **Description** |
|---|---|
| postcode | the postcode providing the information for the row |
| in use? | A yes/no indicator whether the postcode is currently in use |
| latitude | the latitude of the postcode |
| longitude | the longitude of the postcode |
| easting | the easting of the postcode |
| northing | the northing of the postcode |
| Grid Ref | Ordanance survey grid reference |
| County | Name of the county the postcode is in |
| District | Name of the district the postcode is in |
| Ward | Name of the ward the postcode is in |
| District code | Code of the district the postcode is in |
| Ward code | Code of the ward the postcode is in |
| Country | Name of the country the postcode is in |
| County Code | Code of the County this postcode is in |
| Constituency | Name of the Parliamentary Constituency this postcode is in |
| Introduced | Date that the postcode was introduced |
| Terminated | Date that the postcode was not used anymore (if N/A left empty) |
| Parish | Name of the Parish the postcode sits in if applicable |
| National Park | Name of National Park postcode is in if applicable |
| Population | Population at postcode given from the 2011 census |

| **Column Name** | **Description** |
|---|---|
| Households | Number of households in the postcode area given from the 2011 census |
| Built up area | Name of the built up area of the postcode |
| Built up subdivision | Name of the built up area of the subdivision |
| Lower layer super output area | Name of the lower layer super output area (LSOA) that this postcode is in |
| Rural/Urban | Description of the area the postcode is in |
| Region | Name of the region the postcode is in |
| Altitude | Height of postcode above sea level in metres |
| London zone | Transport for London travel zone indicator (if applicable) |
| LSOA code | Code of the LSOA the postcode is in |
| Local Authority | County Council area that the postcode is in |
| MSOA Code | Cosfe for the middle layer super output area that the postcode is in |
| Middle layer super output area | Name of the middle layer super output area that the postcode is in |
| Parish Code | Code of the Parish that the postcode is in |
| Census output area | Code for the census output area |
| Constituency Code | Code of the parliamentary Constituency the postcode is in |

## .2 Traceplots from Gaussian Process application in Chapter 3

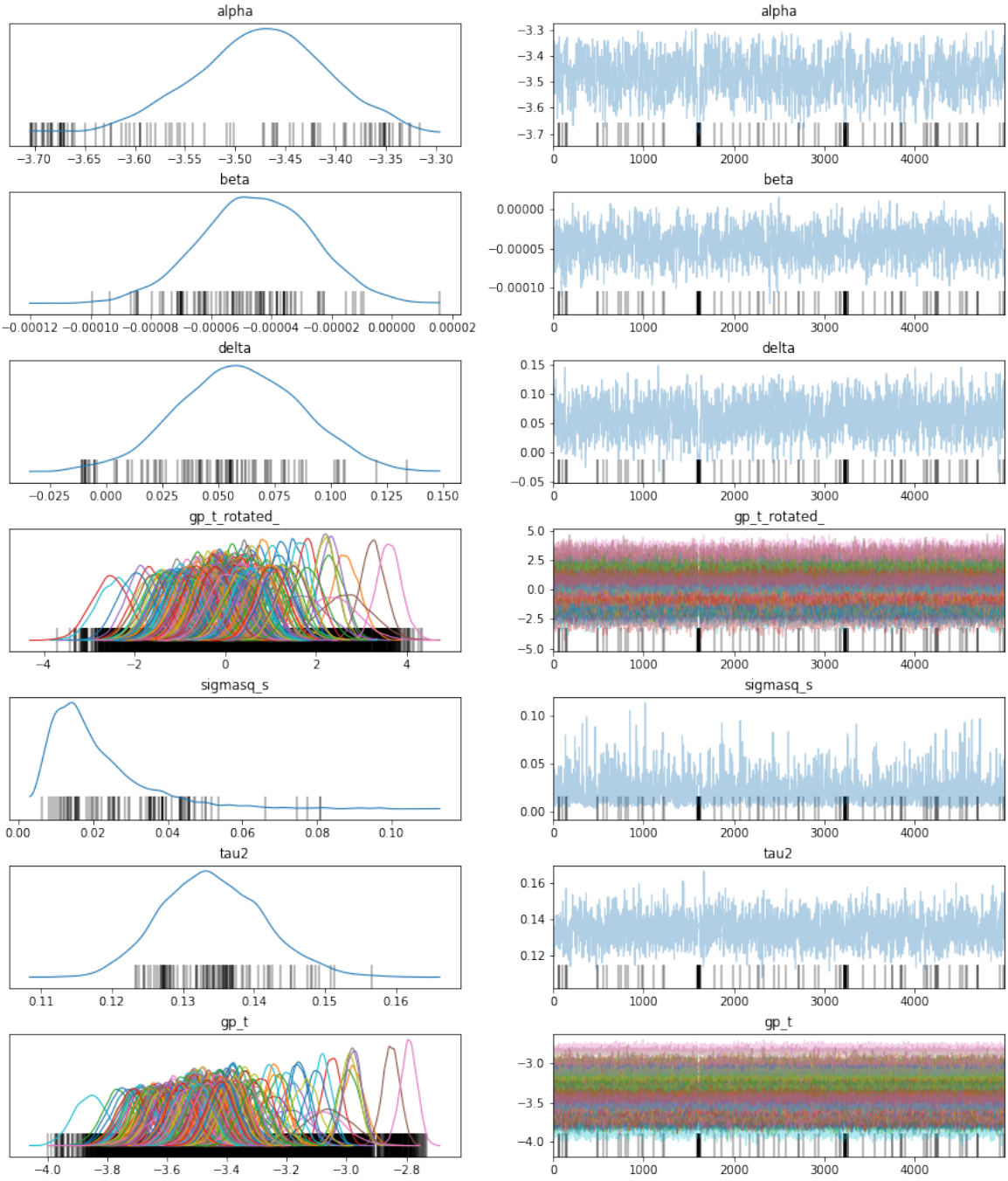



Figure 1: Traceplot for Gastroenteric MPC for dogs Nationwide using the basic Gaussian Process application seen in figure 3.5

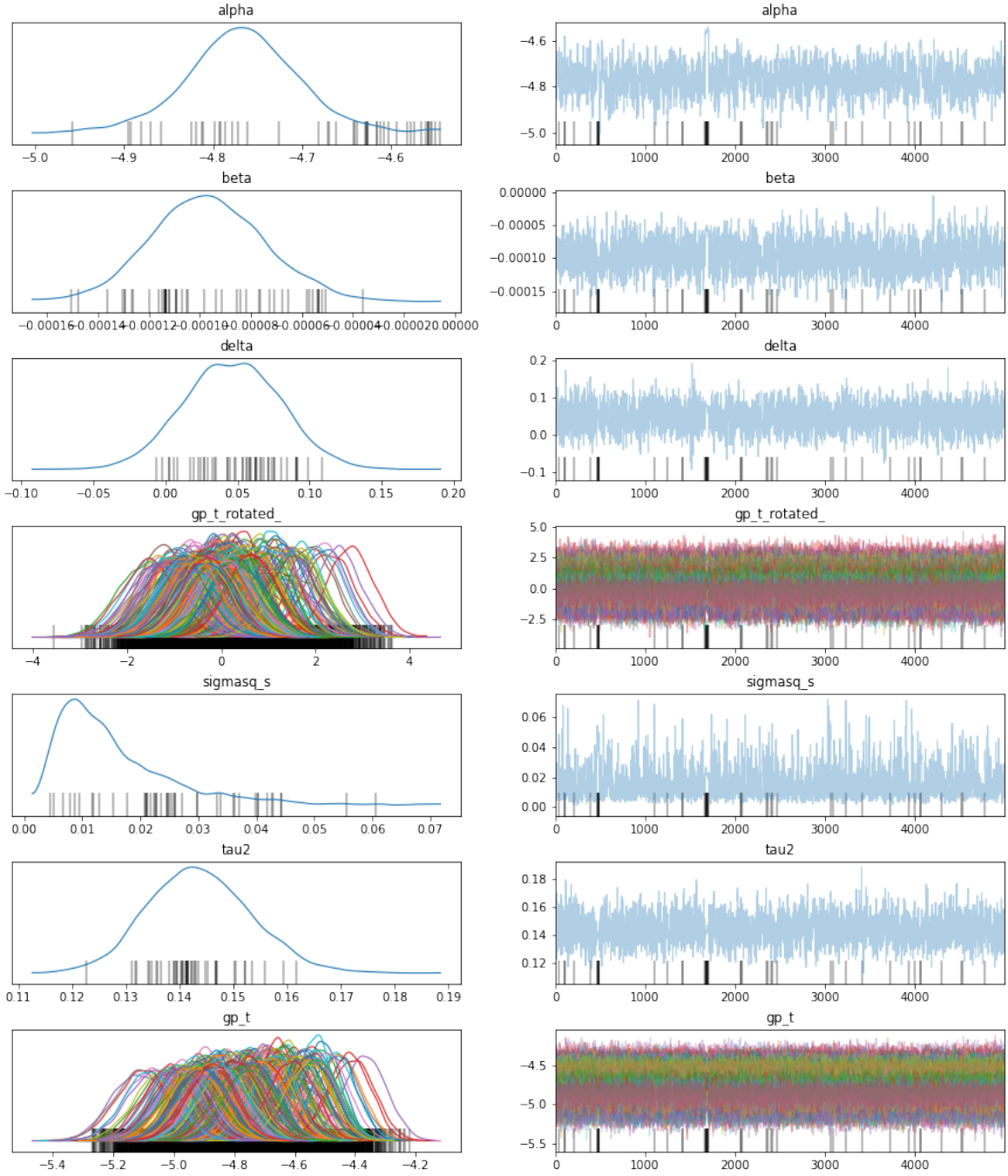



Figure 2: Traceplot for Respiratory MPC for dogs Nationwide using the basic Gaussian Process application seen in figure 3.5.

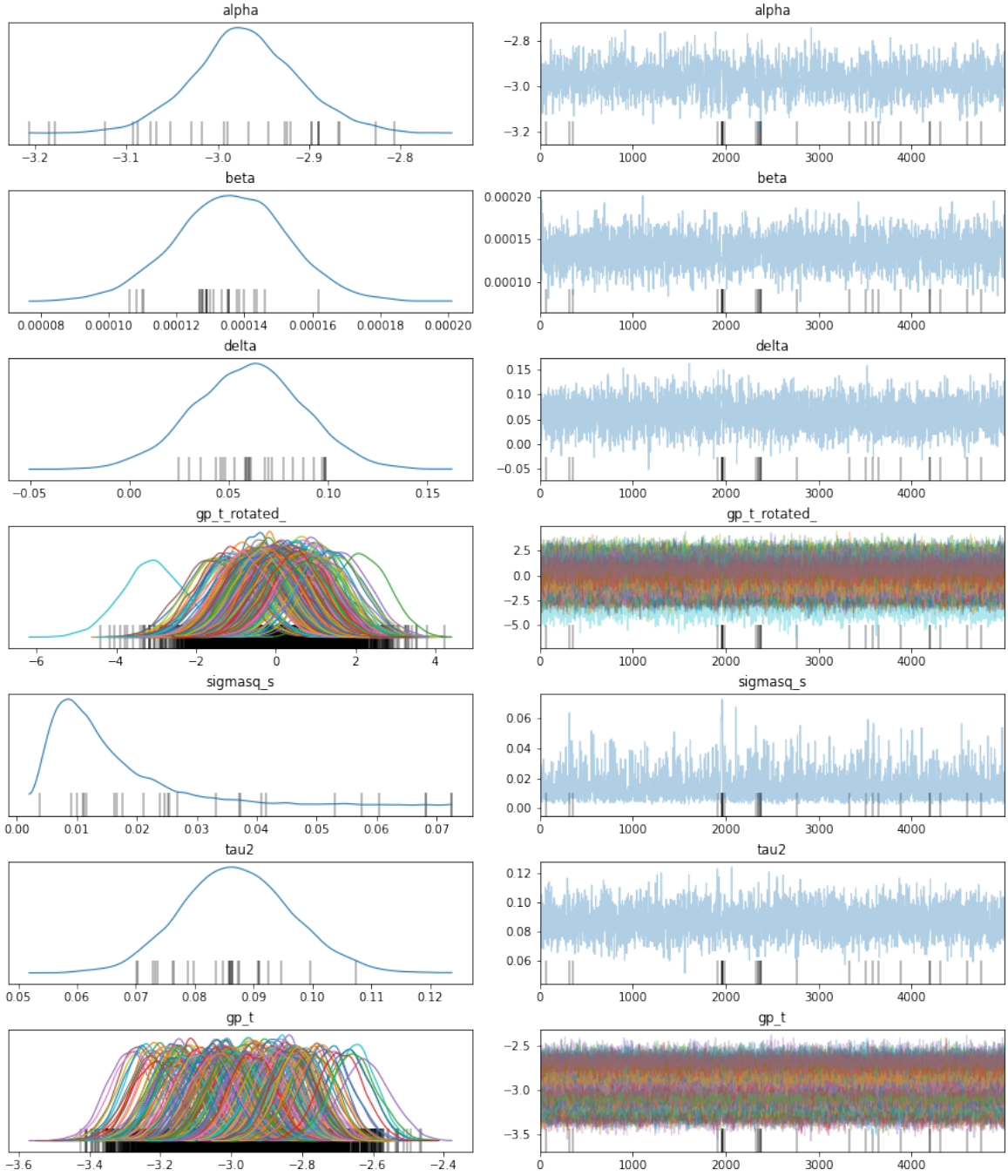



Figure 3: Traceplot for Pruritus MPC for dogs Nationwide using the basic Gaussian Process application seen in figure 3.5.

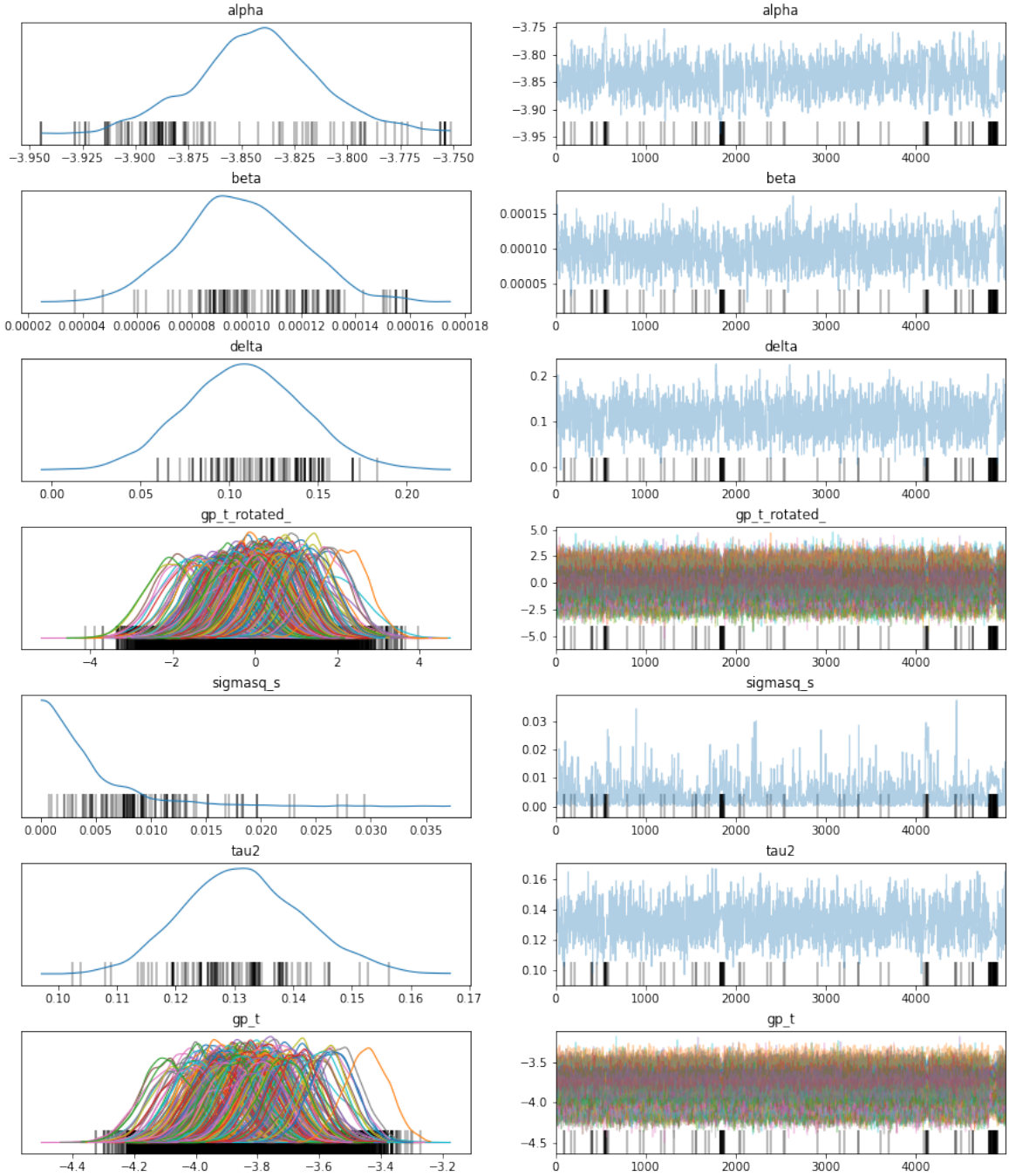



Figure 4: Traceplot for Gastroenteric MPC for cats at a national level using the basic Gaussian Process application seen in figure 3.5.

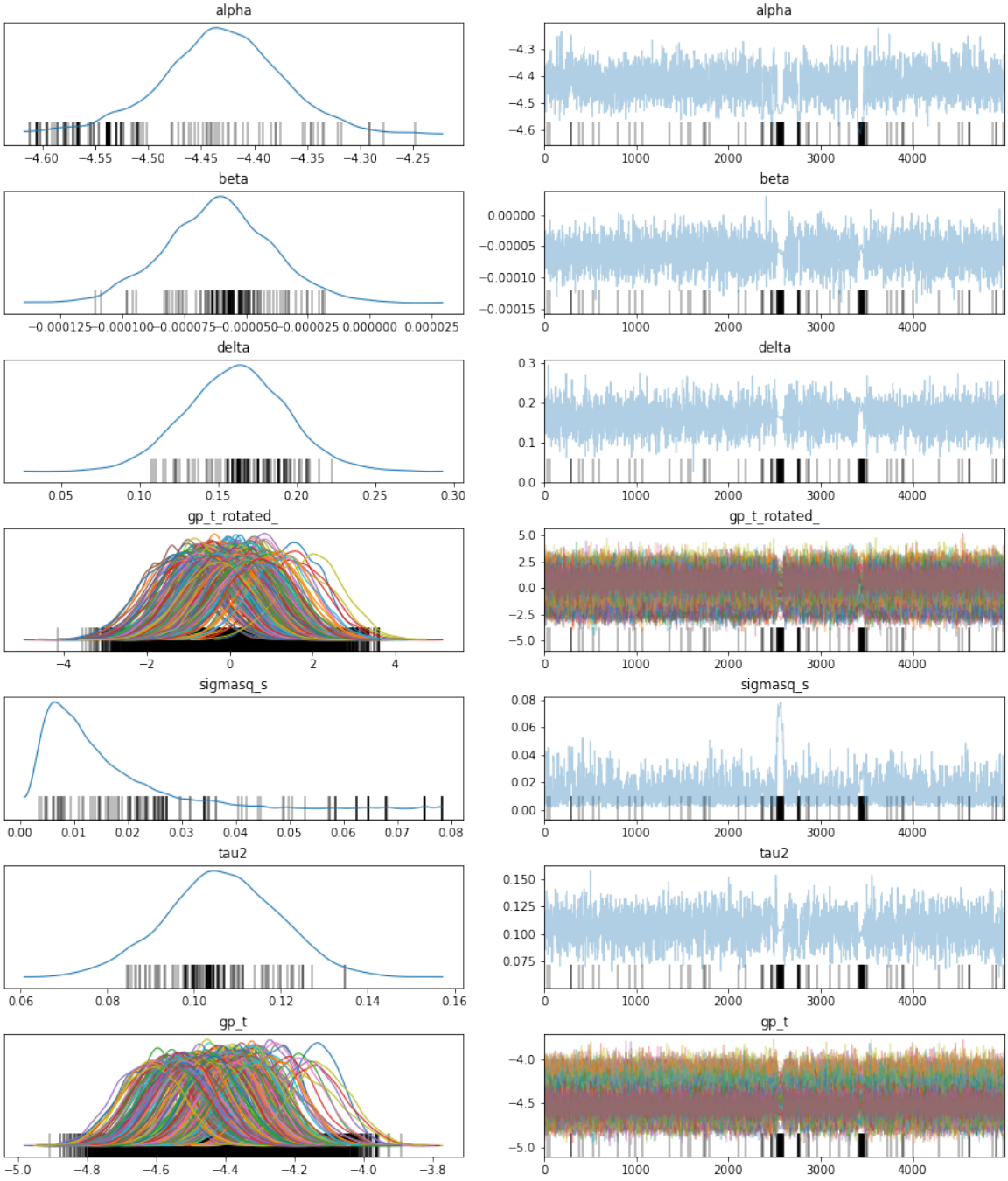



Figure 5: Traceplot for Respiratory MPC for cats at a national level using the basic Gaussian Process application seen in figure 3.5.

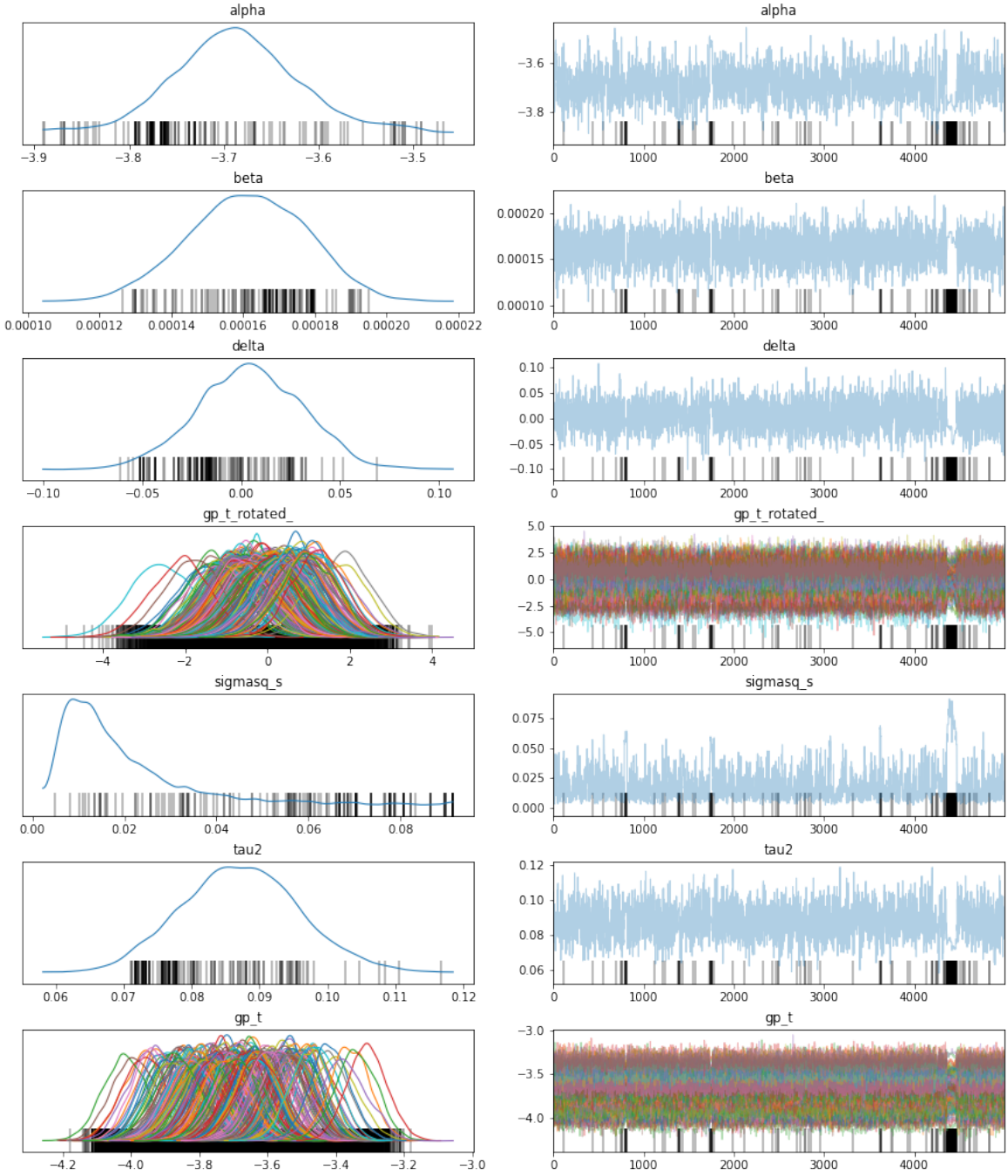



Figure 6: Traceplot for Pruritus MPC for cats at a national level using the basic Gaussian Process application seen in figure 3.5.

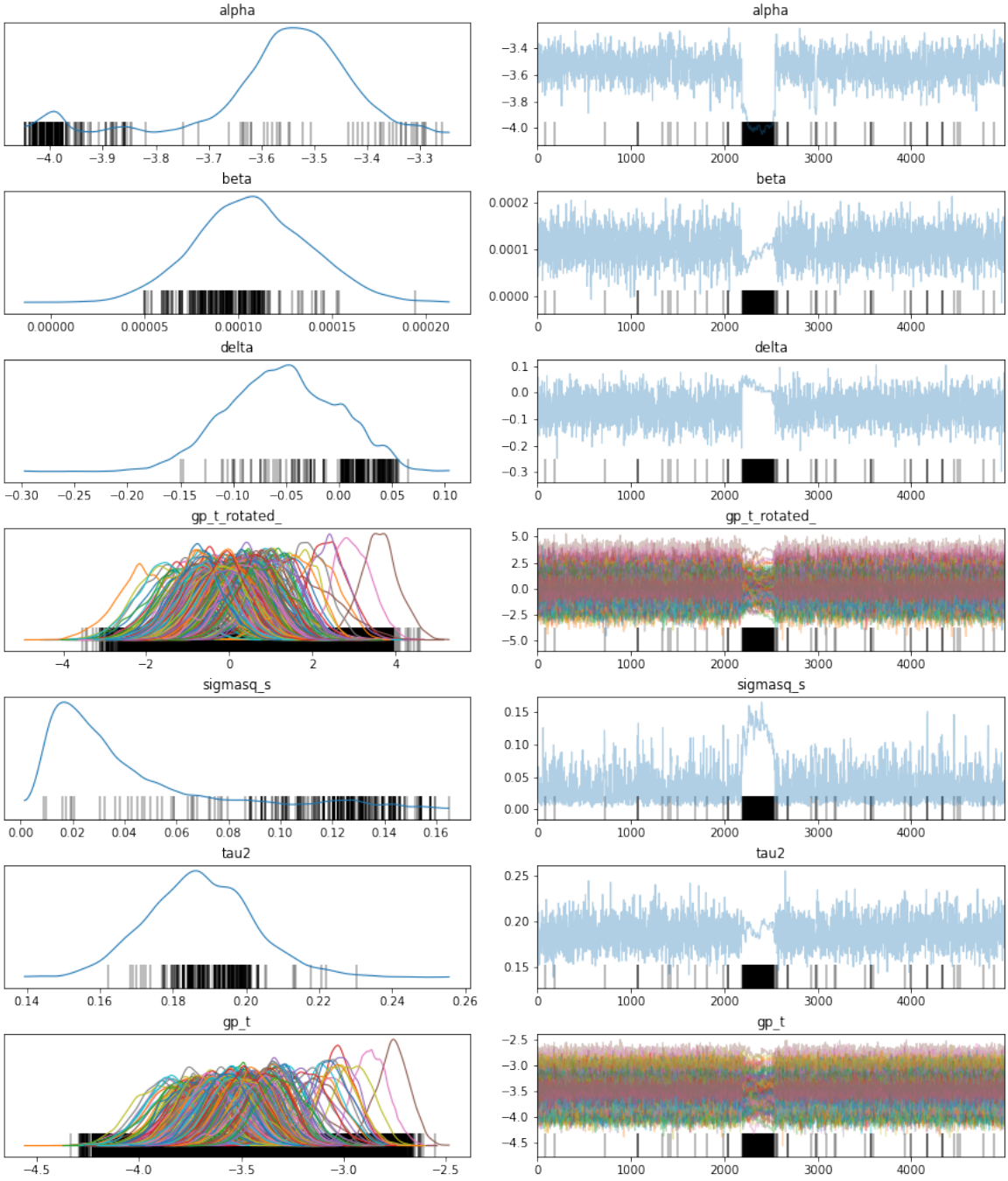



Figure 7: Traceplot for Gastroenteric MPC for dogs in the North West of England using the basic Gaussian Process application seen in figure 3.6.

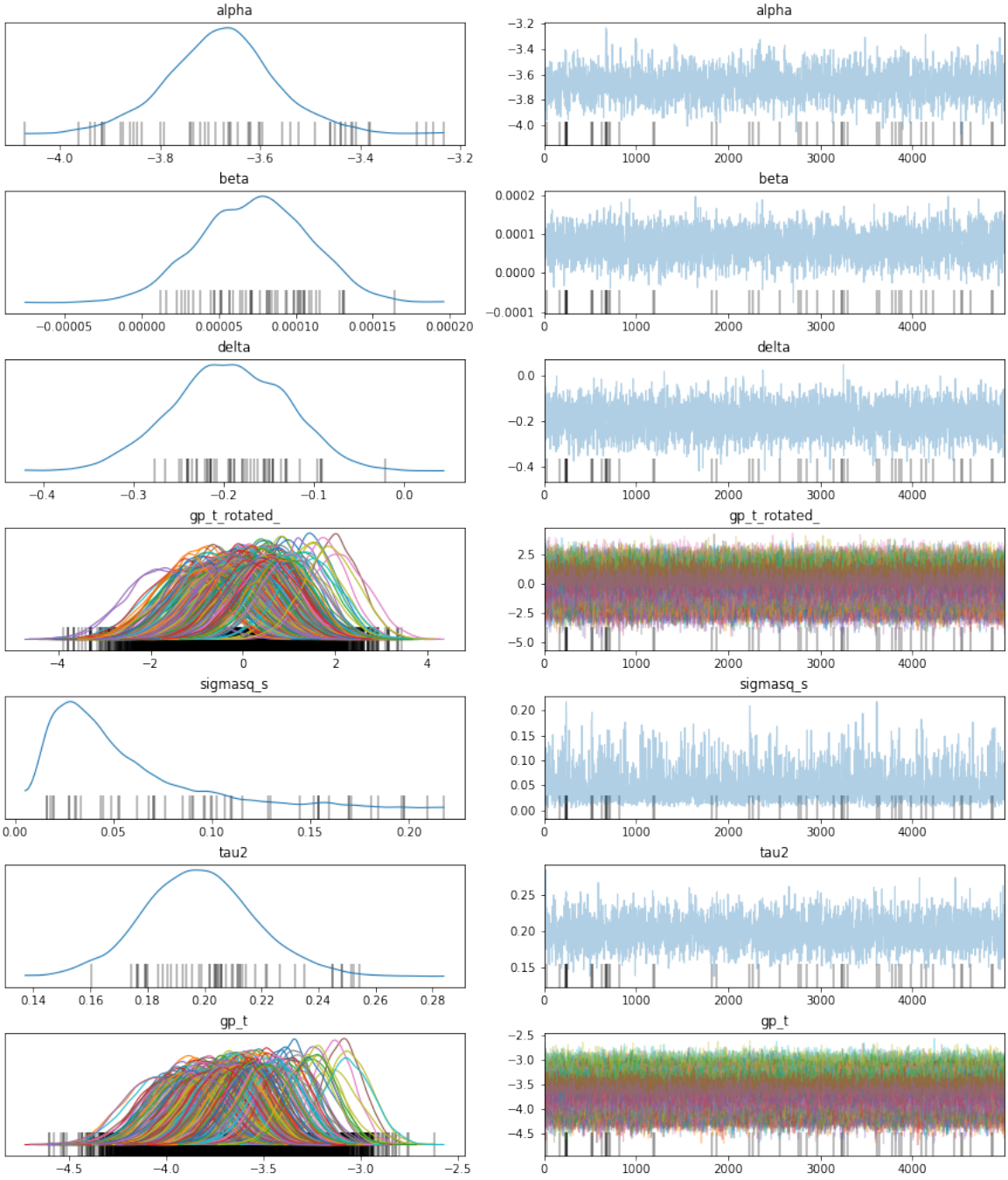



Figure 8: Traceplot for Gastroenteric MPC for dogs in Yorkshire using the basic Gaussian Process application seen in figure 3.6.

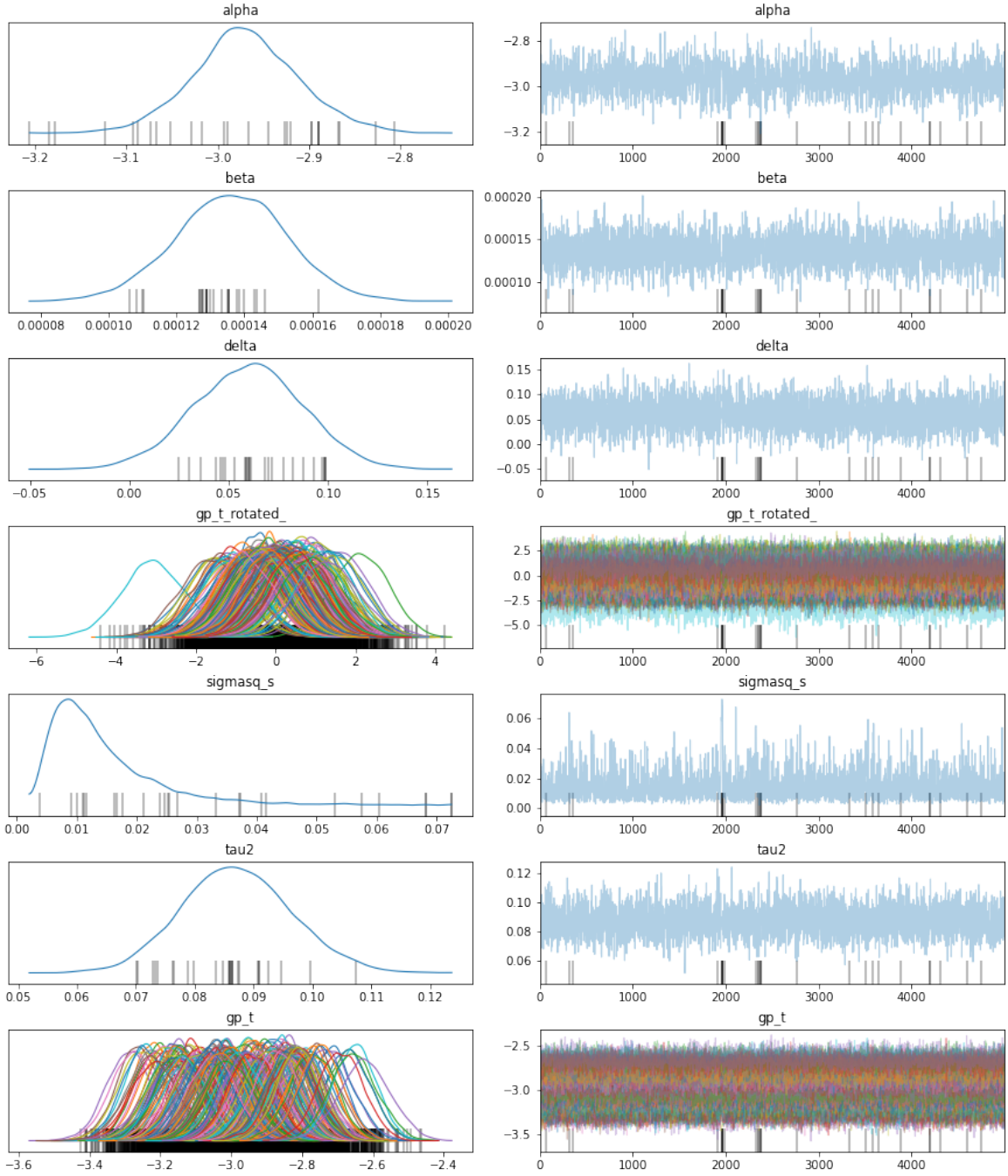



Figure 9: Traceplot for Pruritus MPC for dogs in the South East of England using the basic Gaussian Process application seen in figure 3.6.

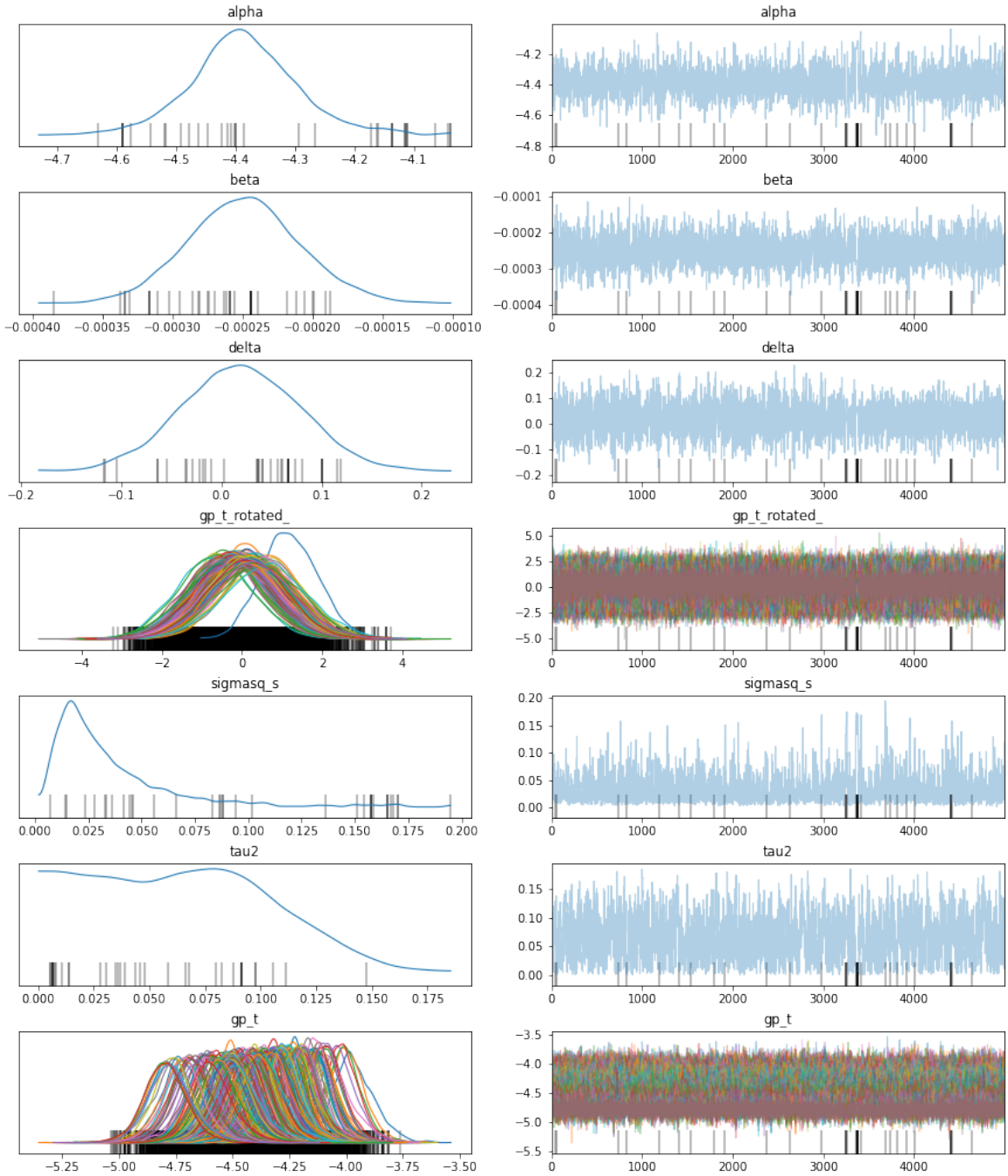



Figure 10: Traceplot for Respiratory MPC for cats in the South West of England using the basic Gaussian Process application seen in figure 3.6.

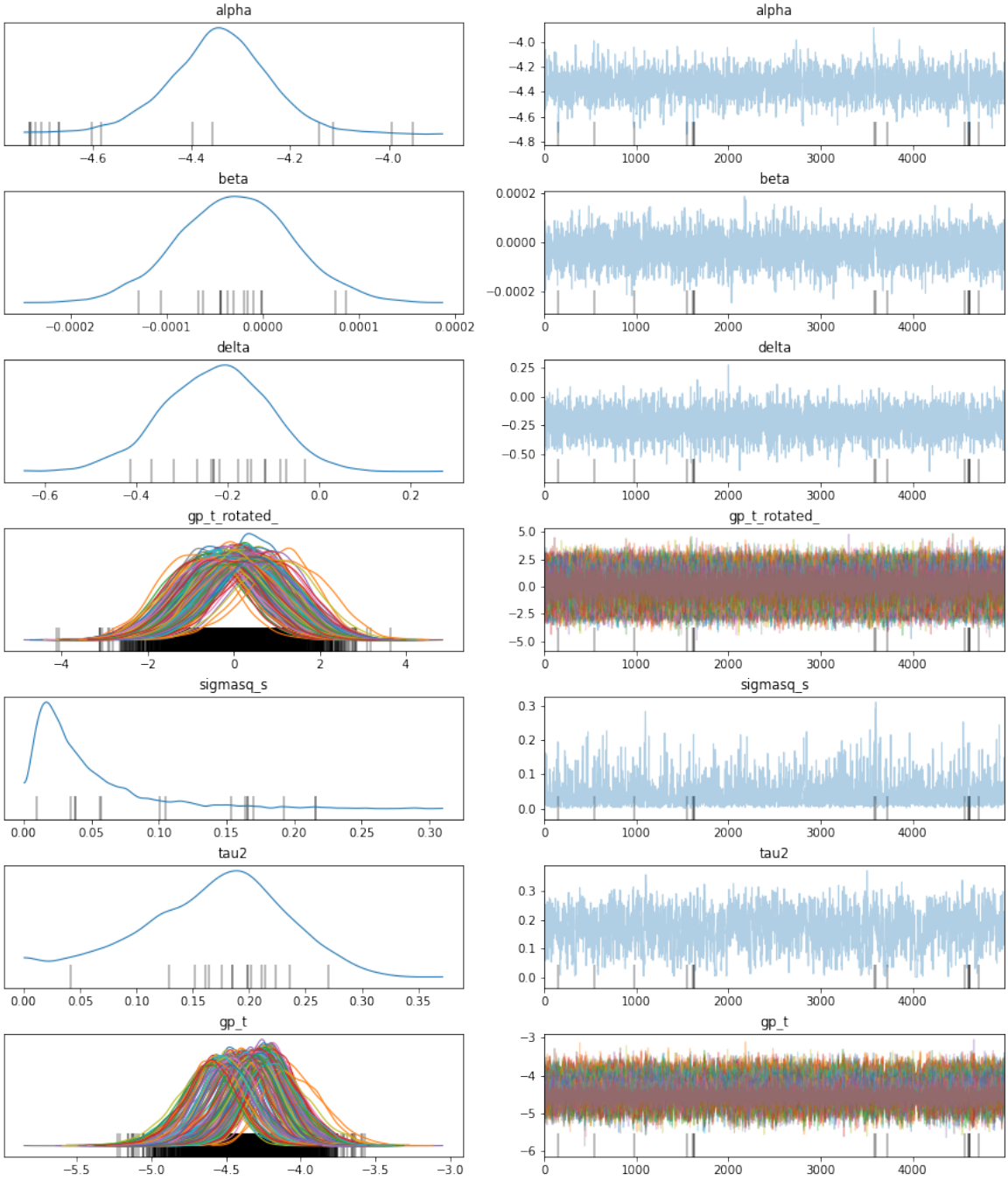



Figure 11: Traceplot for Respiratory MPC for cats in the Yorkshire using the basic Gaussian Process application seen in figure 3.6.

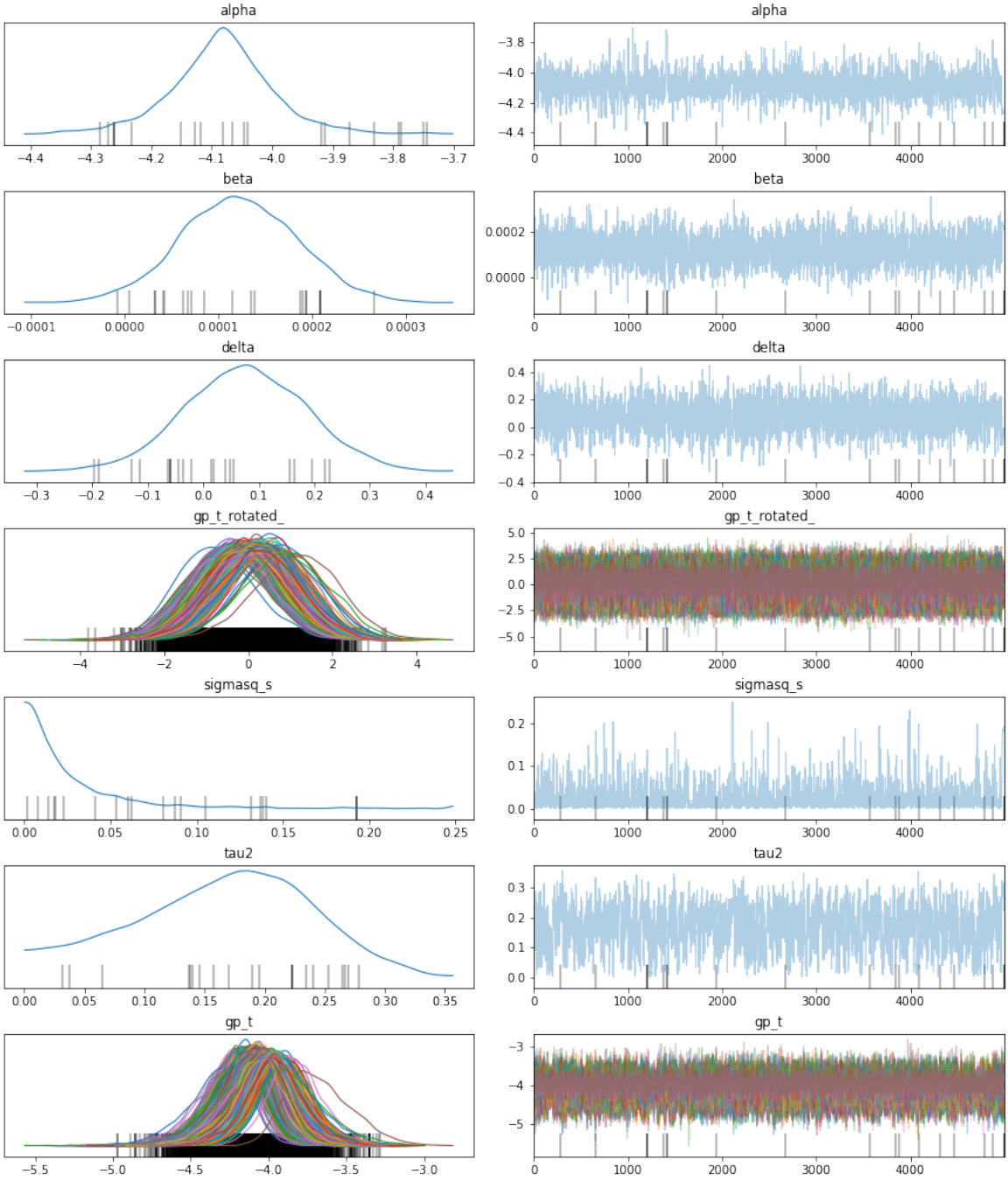
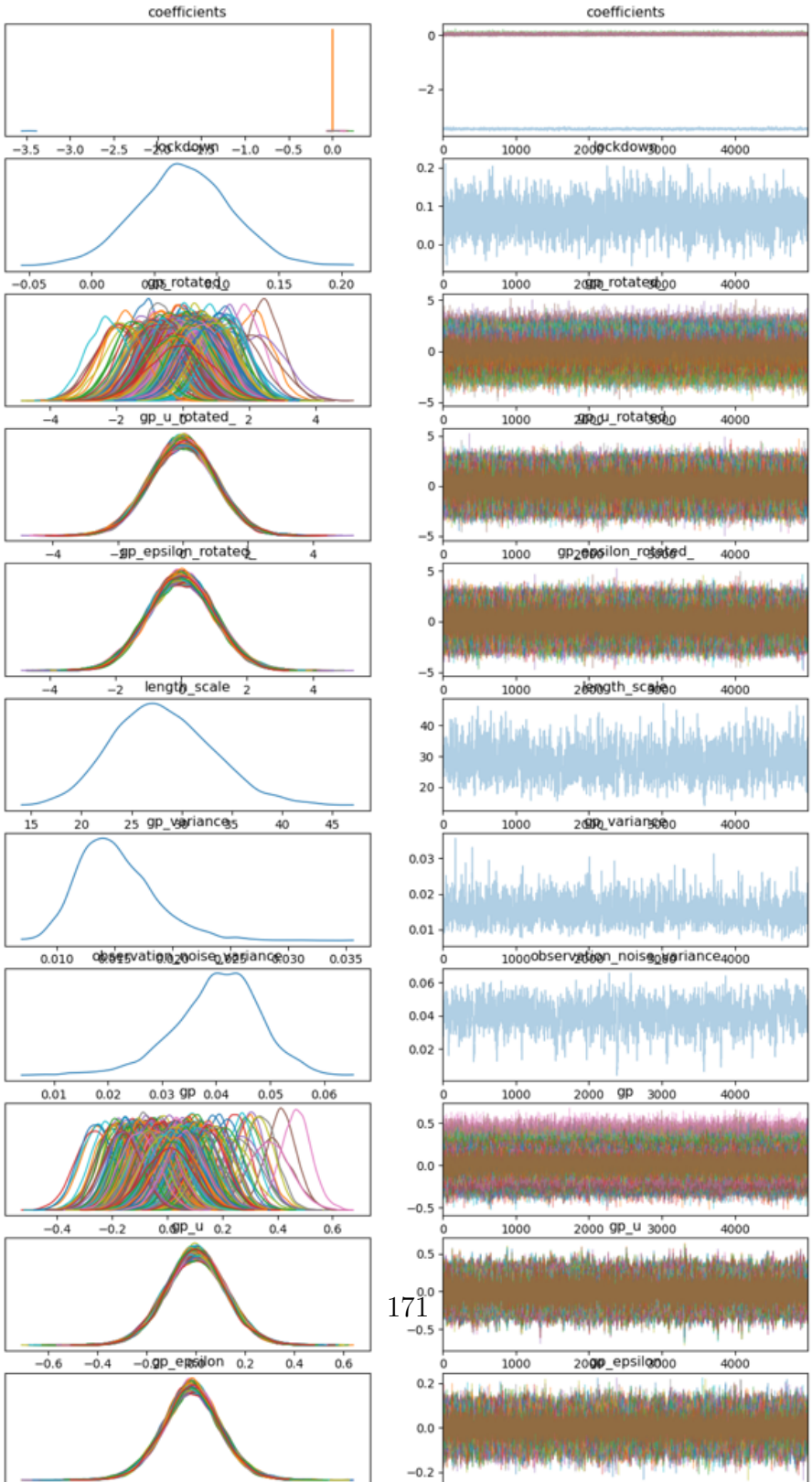



Figure 12: Traceplot for Gastroenteric MPC for cats in the North East of England using the basic Gaussian Process application seen in figure 3.6.

## .3 Traceplots from Mixed Effects model in Chapter 4



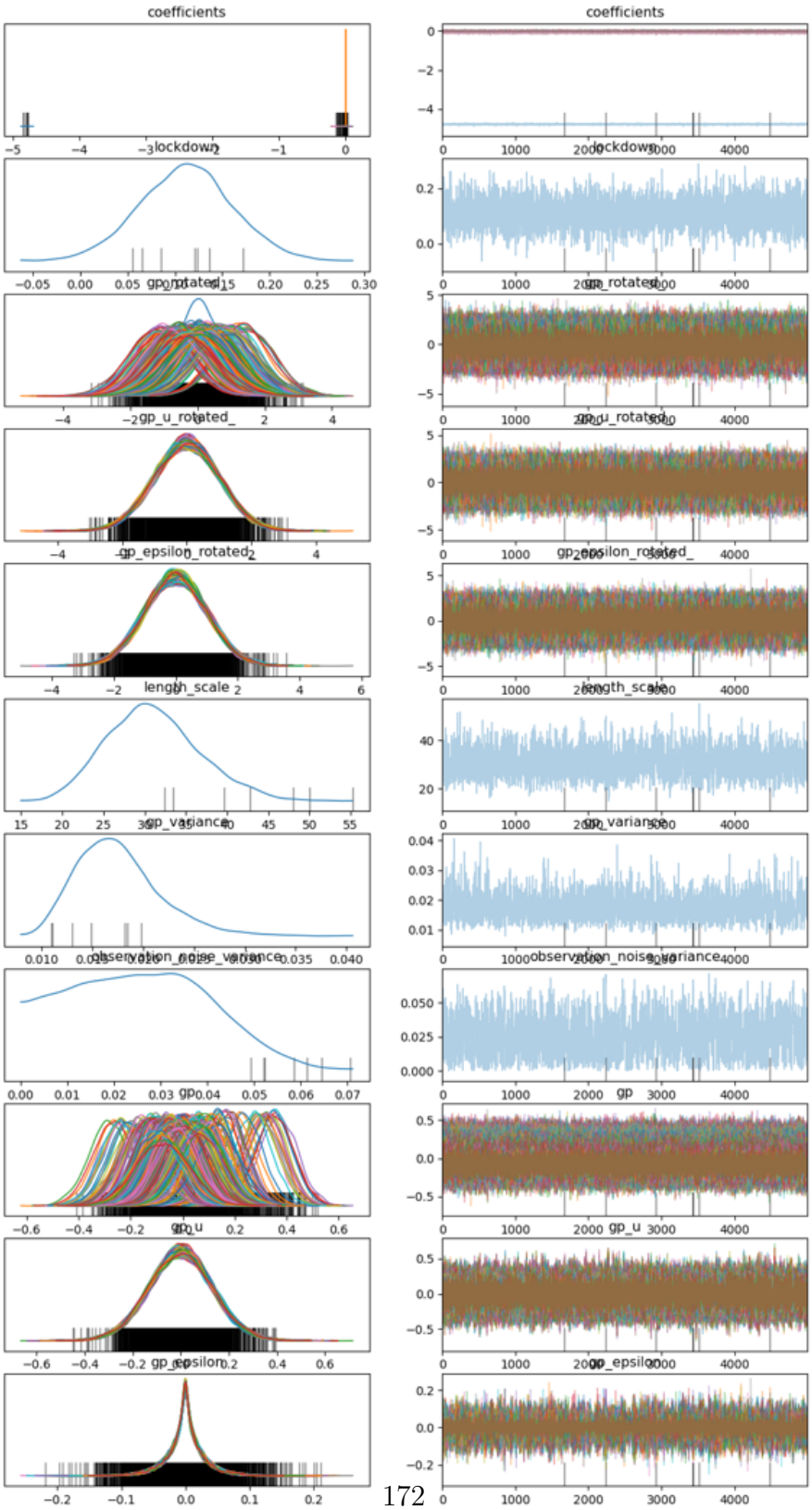





Figure 14: Traceplot for Respiratory MPC for dogs Nationwide using the mixed effect model seen in figure 4.6.

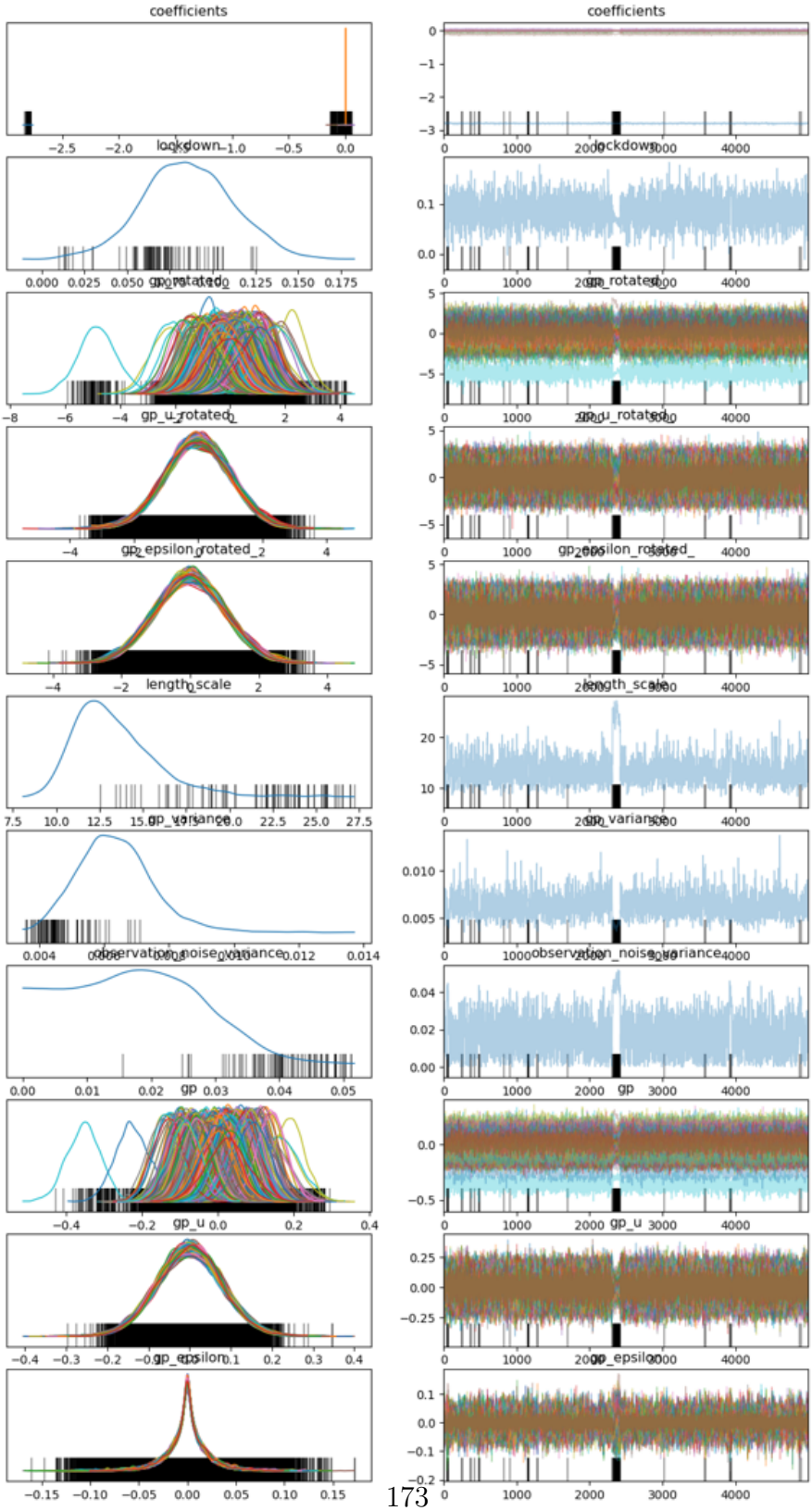





Figure 15: Traceplot for Pruritus MPC for dogs Nationwide using the mixed effect model seen in figure 4.6.

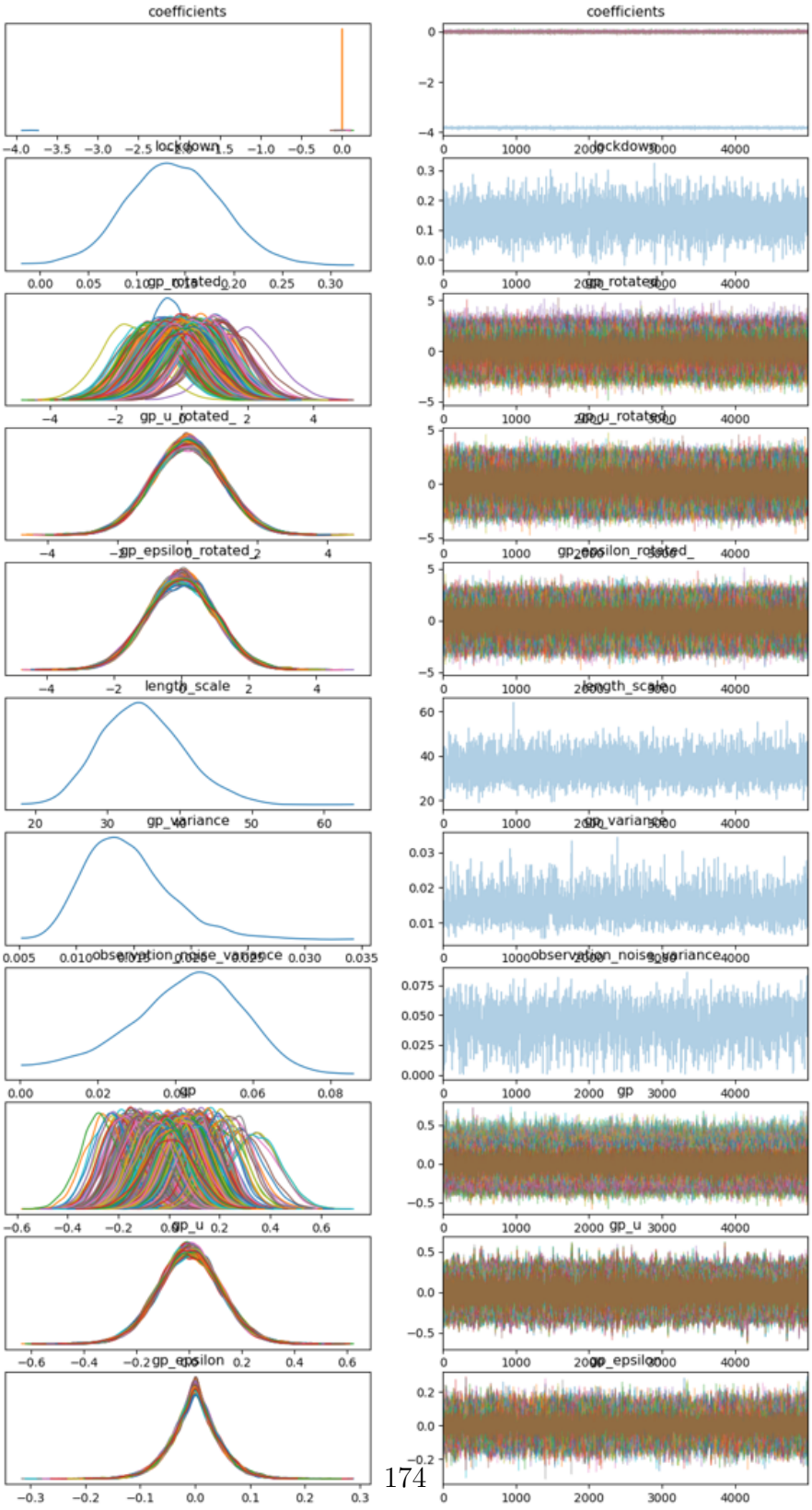



Figure 16: Traceplot for Gastroenteric MPC for cats at a national level using the mixed effect model seen in figure 4.6.

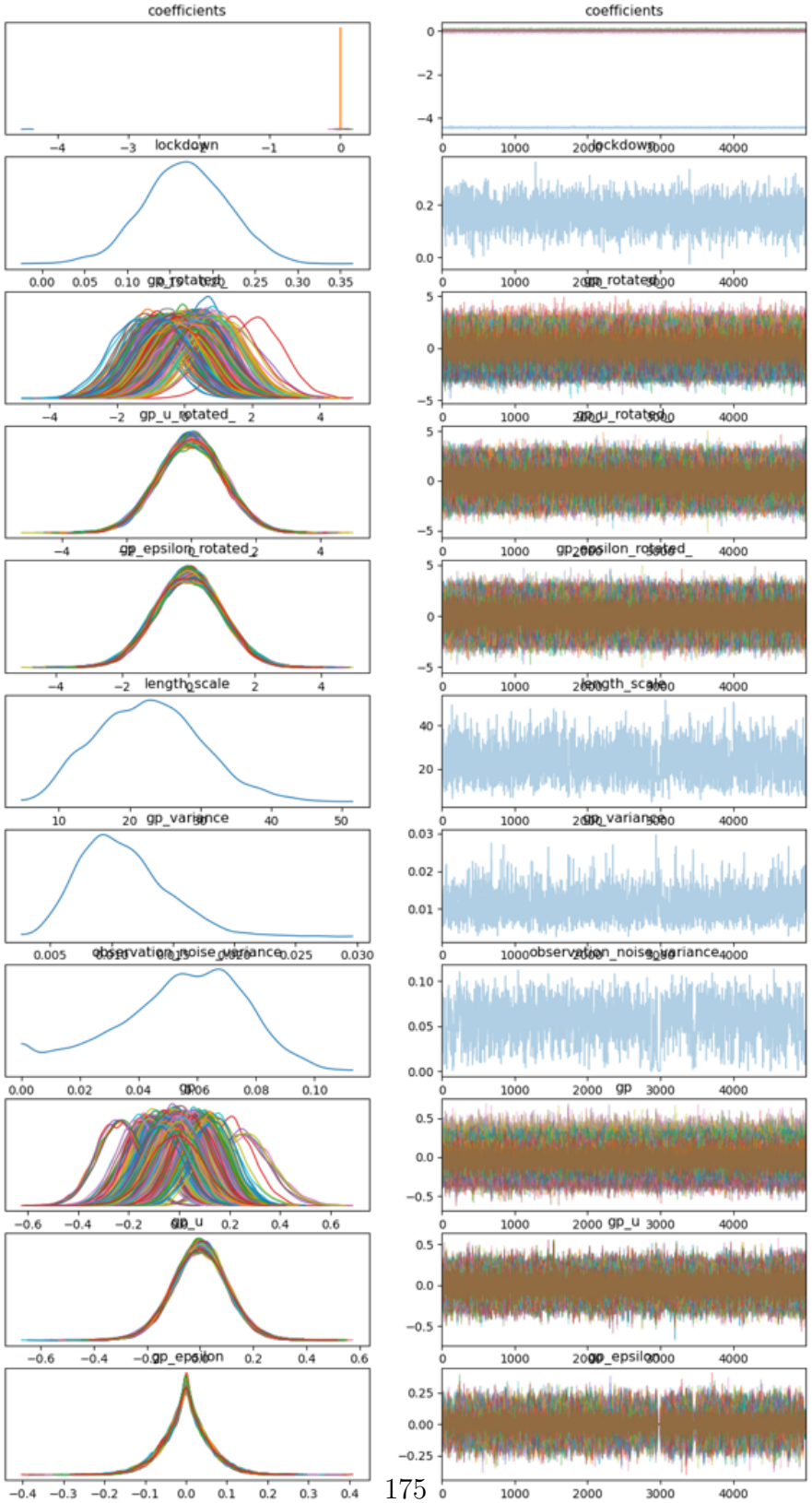



Figure 17: Traceplot for Respiratory MPC for cats at a national level using the mixed effect model seen in figure 4.6.

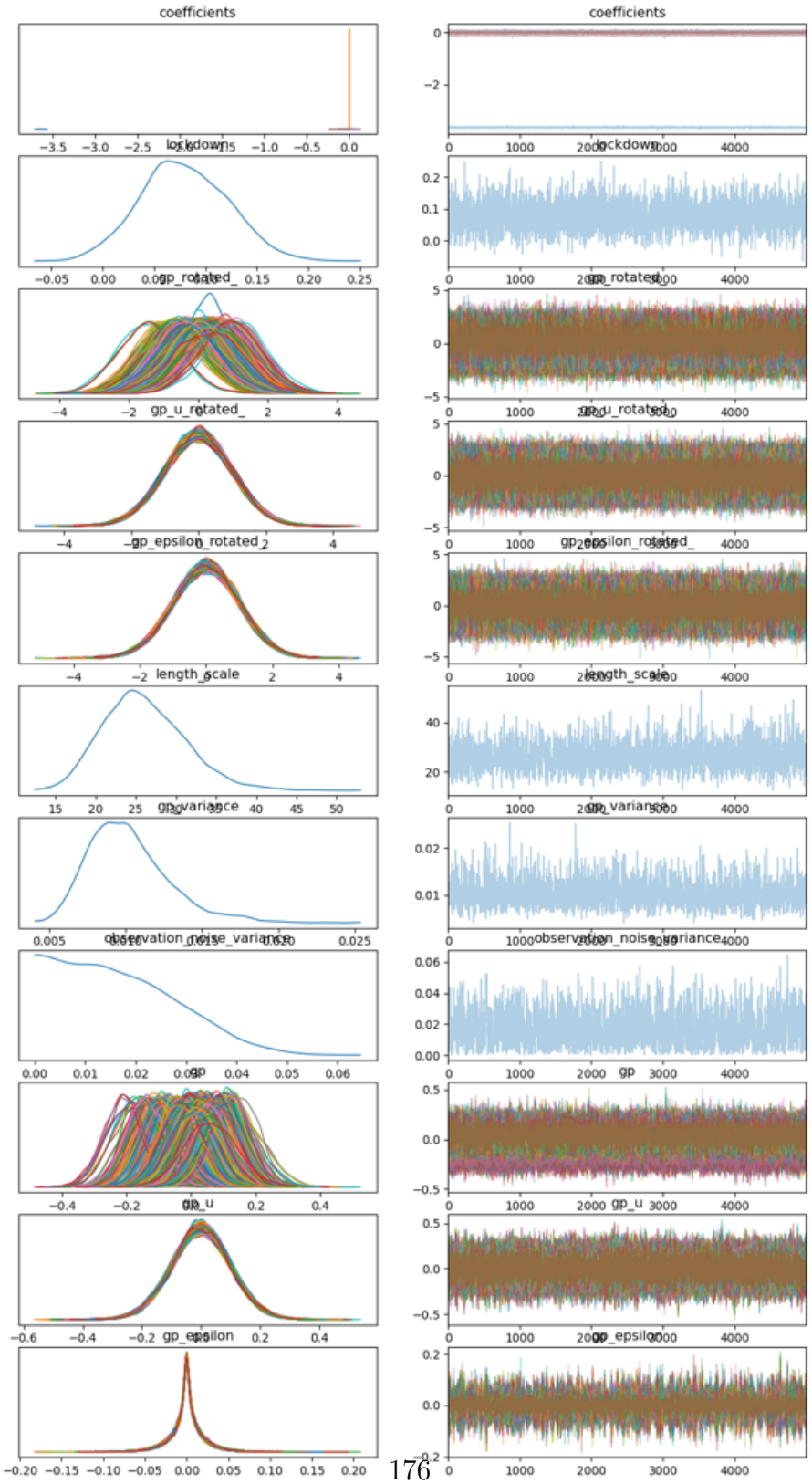





Figure 18: Traceplot for Pruritus MPC for cats at a national level using the mixed effect model seen in figure 4.6.

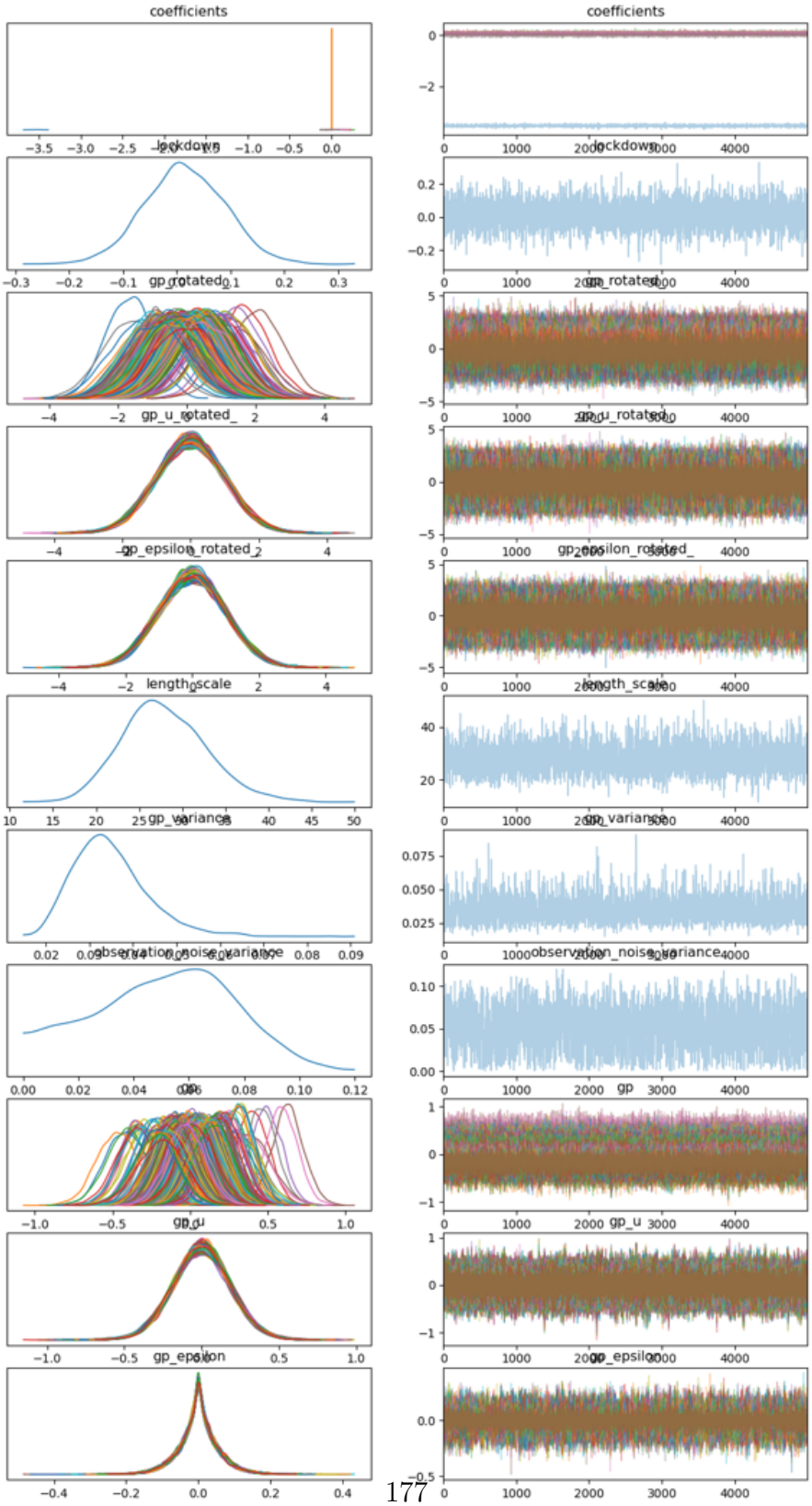





Figure 19: Traceplot for Gastroenteric MPC for dogs in the North West of England using the mixed effect model seen in figure 4.7.

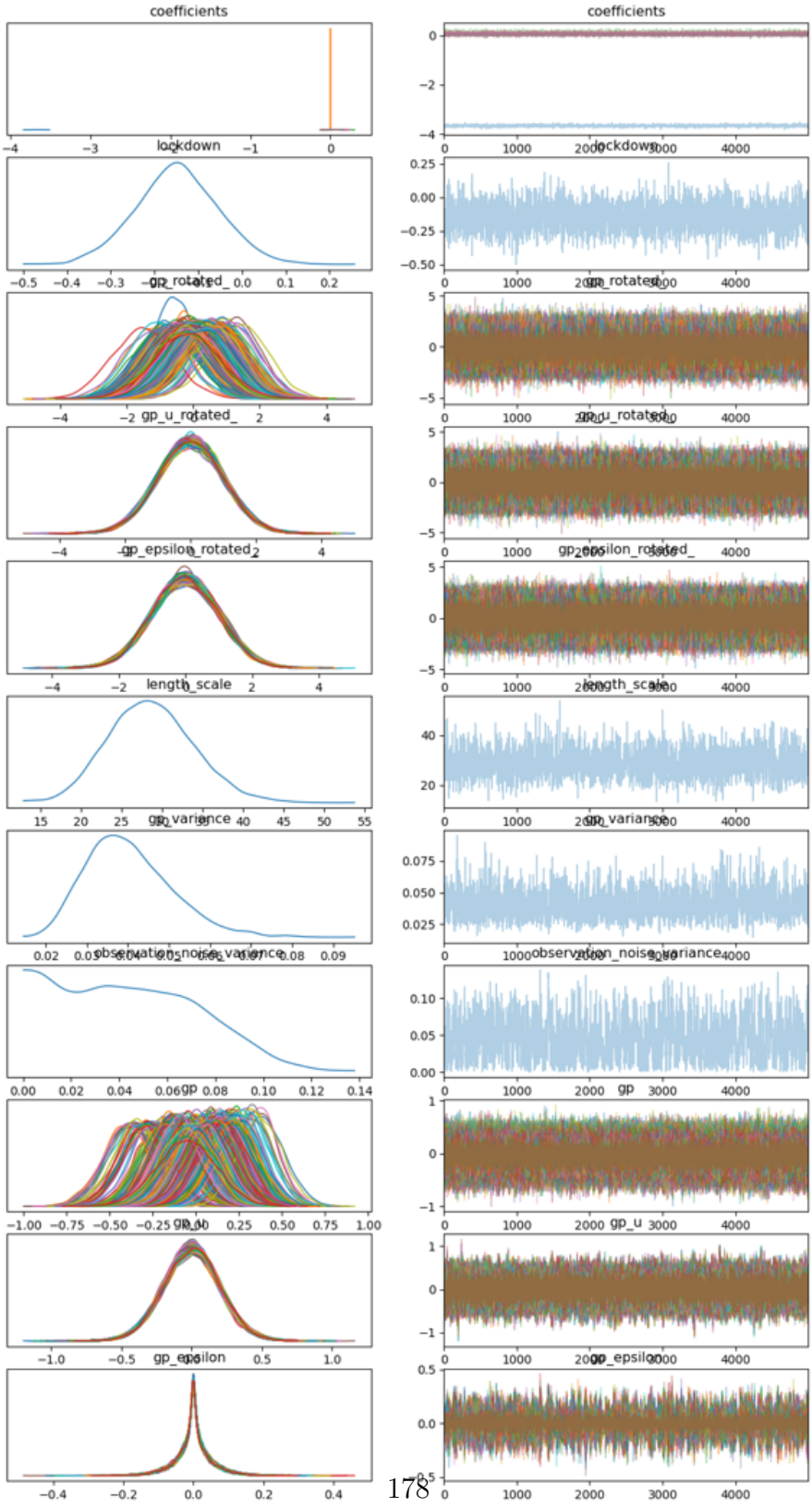





Figure 20: Traceplot for Gastroenteric MPC for dogs in Yorkshire using the mixed effect model seen in figure 4.7.

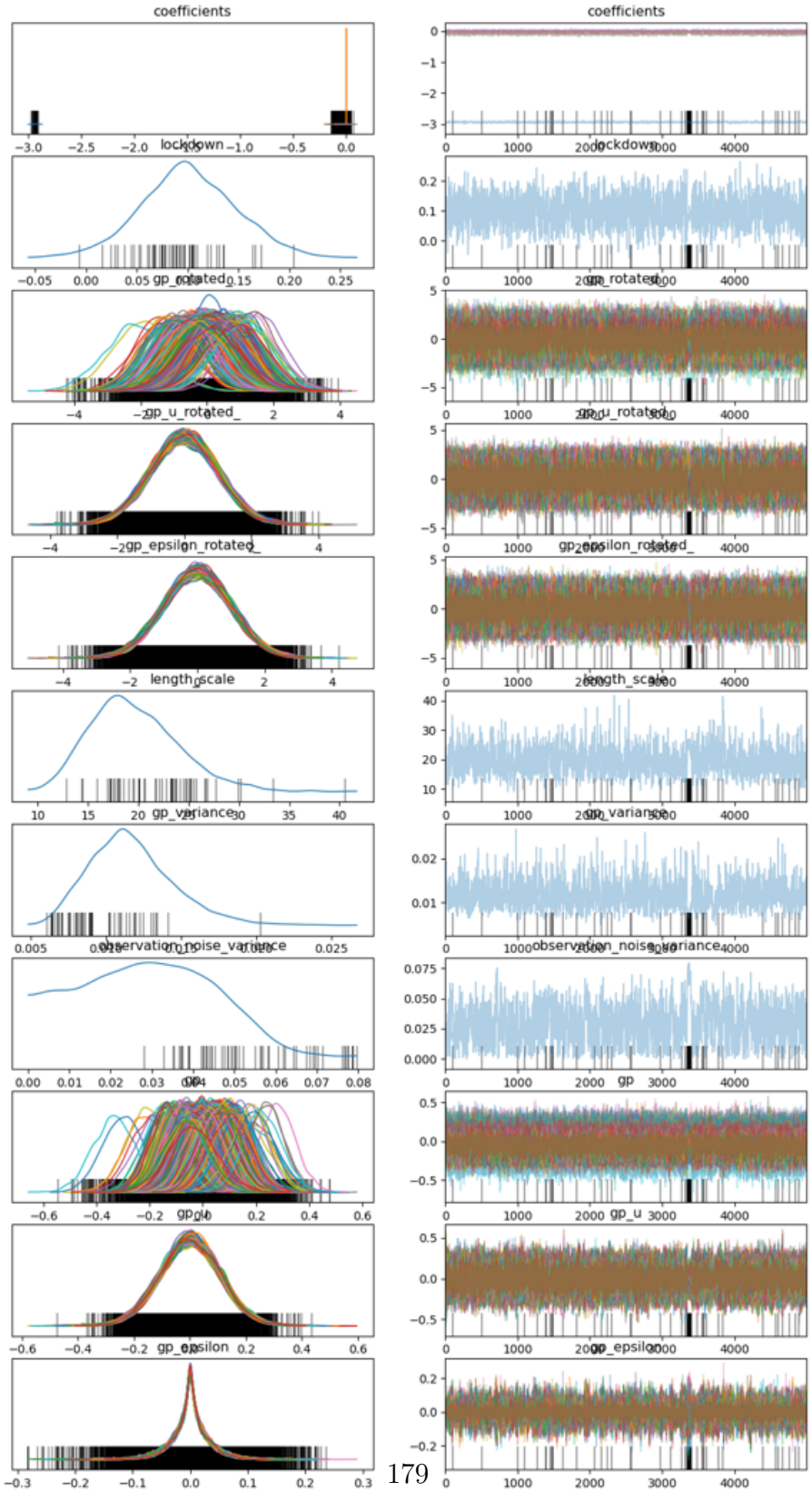



Figure 21: Traceplot for Pruritus MPC for dogs in the South East of England using the mixed effect model seen in figure 4.7.

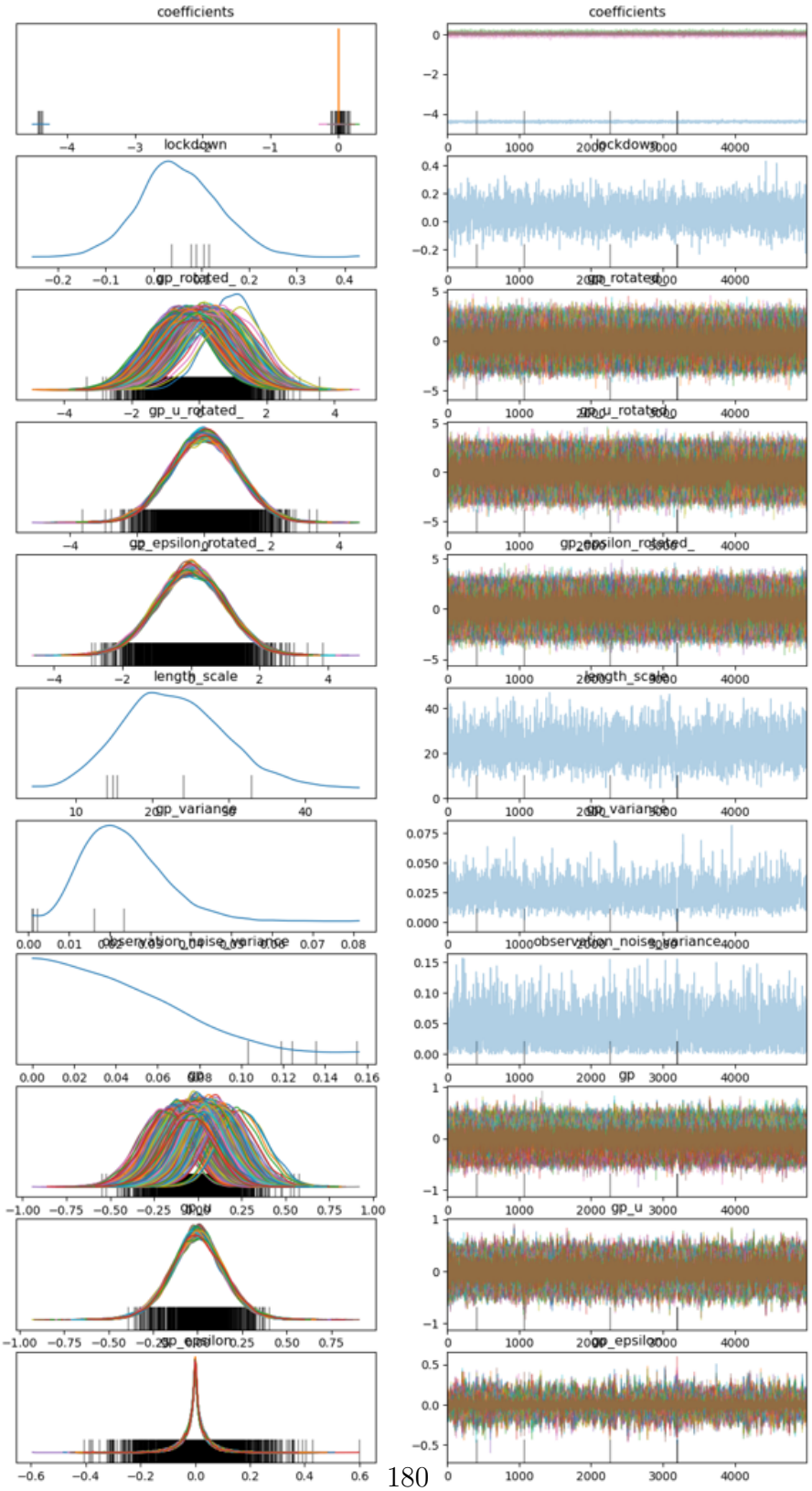



Figure 22: Traceplot for Respiratory MPC for cats in the South West of England using the mixed effect model seen in figure 4.7.

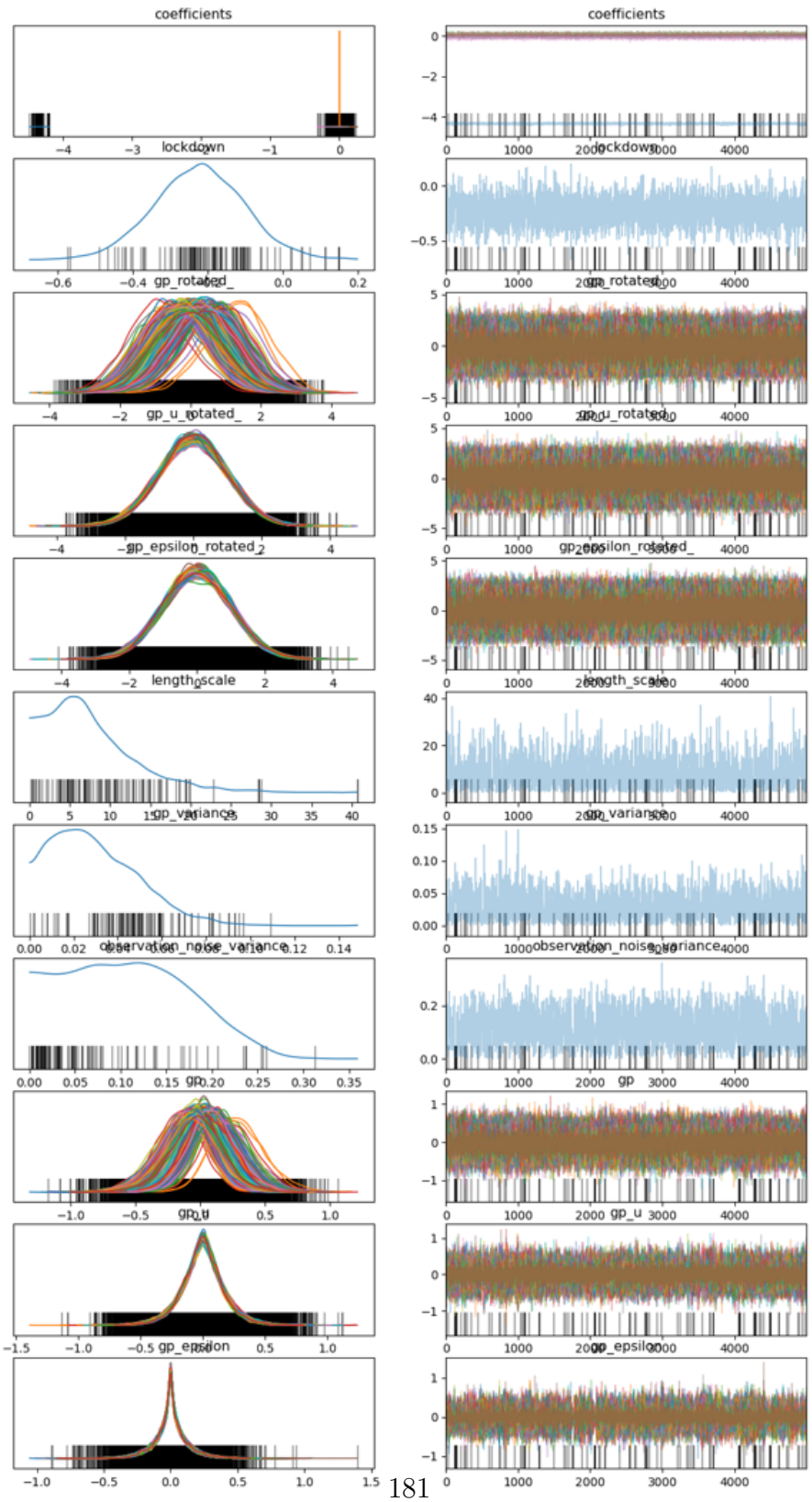





Figure 23: Traceplot for Respiratory MPC for cats in the Yorkshire of England using the mixed effect model seen in figure 4.7.

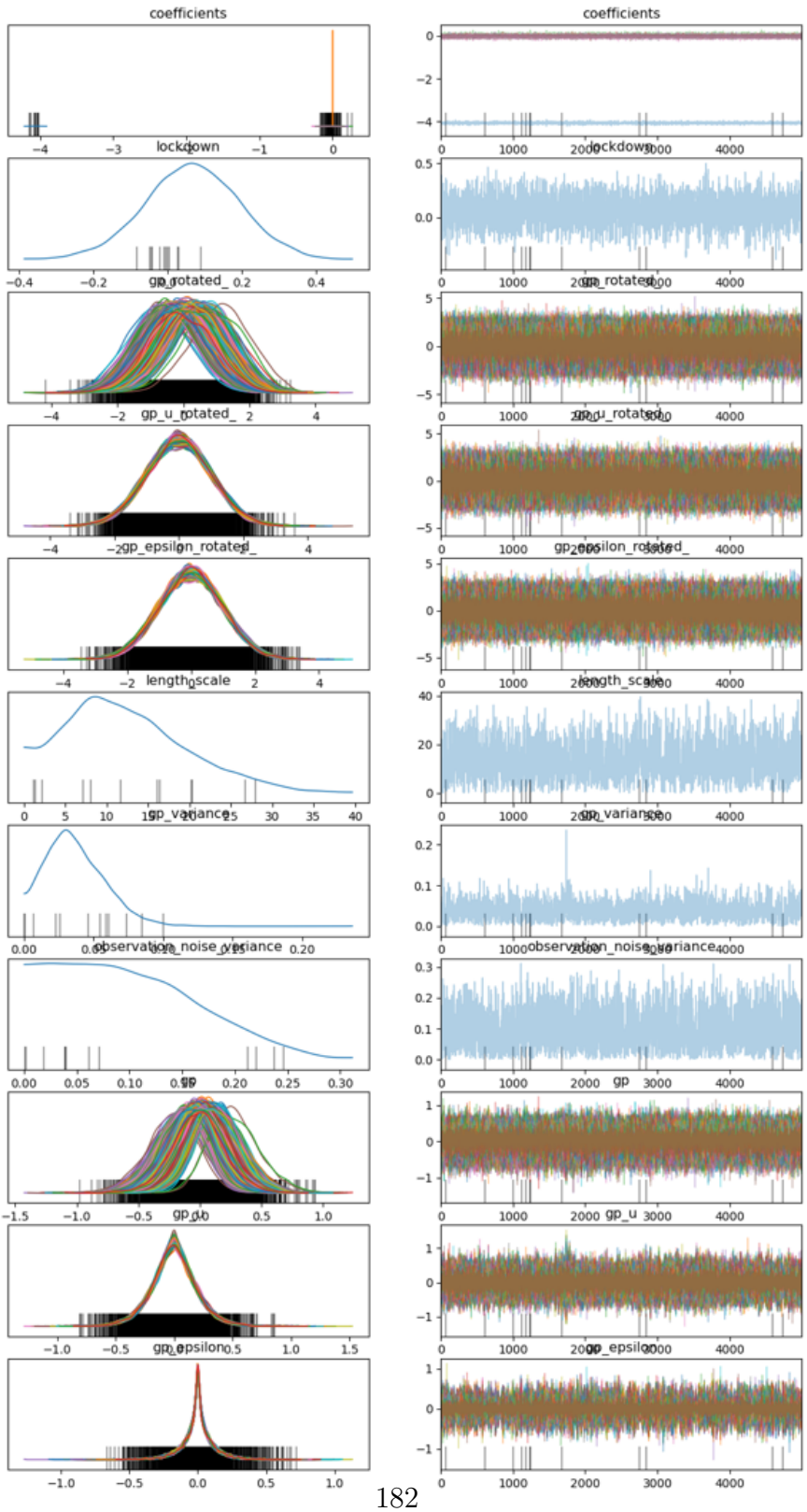





Figure 24: Traceplot for Gastroenteric MPC for cats in the North East of England using the mixed effect model seen in figure 4.7.

# Bibliography


[1] American Animal Hospital Association (AAHA). *AAHA Preventive Healthcare Guidelines for Dogs and Cats*. Accessed: 2025-01-29. 2025. URL: https://www.aaha.org/wp-content/uploads/globalassets/02-guidelines/preventive-healthcare/aaha-preventive-healthcare-guidelines-for-dogs-and-cats.pdf.

[2] Canadian Animal Health Surveillance System (CAHSS). *A Practical Framework for Developing Case Definitions for Animal Diseases*. Accessed: 2025-01-27. 2022. URL: https://www.cahss.ca/CAHSS/Assets/SharedDocuments/CaseDefinitionFramework_220105.pdf.

[3] SÉRGIO ALMEIDA. "An introduction to high performance computing". In: *International Journal of Modern Physics A* 28.22n23 (2013), p. 1340021. DOI: 10.1142/s0217751x13400216.

[4] World Organisation for Animal Health (OIE). *Terrestrial Animal Health Code*. Paris, France: World Organisation for Animal Health (OIE), 2023.

[5] Aisaku Arakawa et al. "Hamiltonian Monte Carlo method for estimating variance components". In: *Animal Science Journal* 92.1 (2021). DOI: 10.1111/asj.13575.

[6] Özgür Asar et al. "Linear Mixed Effects Models for Non-Gaussian Continuous Repeated Measurement Data". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 69.5 (Sept. 2020), pp. 1015–1065. ISSN: 0035-9254. DOI: 10.1111/rssc.12405. eprint: https://academic.oup.com/jrsssc/article-pdf/69/5/1015/49342154/jrsssc_69_5_1015.pdf. URL: https://doi.org/10.1111/rssc.12405.

[7] John M Barry. "The site of origin of the 1918 influenza pandemic and its public health implications". In: *Journal of Translational Medicine* 2.1 (2004). DOI: 10.1186/1479-5876-2-3.

[8] Chris Bell. *UK Postcodes*. Accessed: 2020. 2021.

[9] Caroline Bologna. *Cat owners aren't taking them to the vet enough. that's a problem.* Accessed 2024. 2023. URL: https://www.huffingtonpost.co.uk/entry/cats-veterinarian-undermedicalization_l_63c1ea21e4b0fe267cba4d51.

[10] Sofiane Brahim-Belhouari and Amine Bermak. "Gaussian process for nonstationary time series prediction". In: *Computational Statistics and Data Analysis* 47.4 (2004), pp. 705–712. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2004.02.006. URL: https://www.sciencedirect.com/science/article/pii/S0167947304000301.

[11] British Veterinary Association. *Guidance for veterinary practices in assessing emergency and urgent care during the Covid-19 pandemic*. Accessed: 2022. 2020. URL: https://www.bva.co.uk/media/3399/bva-guidance-for-veterinary-practices-on-covid19-march-2020.pdf.

[12] George Casella and Edward I. George. "Explaining the gibbs sampler". In: *The American Statistician* 46.3 (1992), p. 167. DOI: 10.2307/2685208.

[13] Dylan Castillo. *Develop data visualization interfaces in python with dash*. Accessed: 2024. 2023. URL: https://realpython.com/python-dash/.

[14] Center for Food Security and Public Health. *Zoonotic Diseases of Companion Animals: By Routes of Transmission*. Accessed: 2025-04-08. 2013. URL: https://www.cfsph.iastate.edu/Zoonoses_Textbook/Assets/zoonotic_diseases_by_routes_of_transmission_CA.pdf.

[15] Clarendon Street Veterinary Surgery. *Five common health conditions we treat in winter*. Accessed: 2023. 2023. URL: https://www.clarendonstreetvets.co.uk/article/five-common-health-conditions-we-treat-in-winter/.

[16] Ton J. Cleophas and Aeilko H. Zwinderman. *Regression Analysis in medical research*. Springer International Publishing, 2018.

[17] Jennifer Coates. *Dictionary of veterinary terms: Vet-speak deciphered for the non-veterinarian*. s.n., 2019.

[18] Colour Blind Awareness. *What is Colour Blindness?* Accessed: 2024. 2022. URL: https://www.colourblindawareness.org/#:~:text=is%20Colour%20Blindness%3F-,What%20is%20colour%20blindness%3F,most%20of%20whom%20are%20male..

[19] House of Commons Library. 2021. URL: https://commonslibrary.parliament.uk/research-briefings/cbp-9068/.

[20] Richard Davis. "Gaussian Process". In: Sept. 2006. DOI: 10.1002/9780470057339.vag002.

[21] Richard A. Davis. "Gaussian Process". In: *Encyclopedia of Environmetrics* (2006). DOI: 10.1002/9780470057339.vag002.

[22] S. L. Deem, W. B. Karesh, and W. Weisman. "Putting Theory into Practice: Wildlife Health in Conservation". In: *Conservation Biology* 15.3 (2001), pp. 657–660. DOI: 10.1046/j.1523-1739.2001.015003657.x.

[23] Dorothy E Denning. "An Intrusion-Detection Model". In: *IEEE TRANSACTIONS ON SOFTWARE ENGINEERING* SE-13.2 (1987), pp. 222–232.

[24] Peter J. Diggle. "An Approach to the Analysis of Repeated Measurements". In: *Biometrics* 44.4 (1988), pp. 959–971. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2531727 (visited on 06/04/2024).

[25] Taylor Eisenstein et al. "Enteric disease outbreaks associated with animal contact — animal contact outbreak surveillance system, United States, 2009–2021". In: *MMWR. Surveillance Summaries* 74.3 (2025), pp. 1–12. DOI: 10.15585/mmwr.ss7403a1.

[26] European Centre for Disease Prevention and Control. *Guidelines for the implementation of non-pharmaceutical interventions against COVID-19*. Accessed: 2023. 2020. URL: https://www.ecdc.europa.eu/en/publications-data/covid-19-guidelines-non-pharmaceutical-interventions.

[27] Eurostat. *NUTS - NOMENCLATURE OF TERRITORIAL UNITS FOR STATISTICS*. Accessed: 2021. 2021. URL: https://ec.europa.eu/eurostat/web/nuts/background.

[28] Anifatul Faricha et al. "The Comparative Study for Predicting Disease Outbreak". In: *Journal of Computer Electronic and Telecommunications* 1.1 (2020), pp. 53–59. DOI: 10.52435/complete.v3i2.

[29] Sean Farrell et al. "Petbert: Automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records". In: *Research Square* (2023). DOI: 10.21203/rs.3.rs-3084076/v1.

[30] C. P. Farrington et al. "A statistical algorithm for the early detection of outbreaks of infectious disease". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159.3 (1996), p. 547. DOI: 10.2307/2983331.

[31] Isabella Fornacon-Wood et al. "Understanding the differences between Bayesian and frequentist statistics". In: *International Journal of Radiation Oncology*Biology*Physics* 112.5 (2022), pp. 1076–1082. DOI: 10.1016/j.ijrobp.2021.12.011.

[32] Nicola Frost. *Researchers confirm dog sickness outbreak in Yorkshire - University of Liverpool News*. Accessed: 2023. 2022. URL: https://news.liverpool.ac.uk/2022/01/28/researchers-confirm-dog-sickness-outbreak-in-yorkshire/.

[33] Dani Gamerman and Hedibert Freitas Lopes. *Markov chain Monte Carlo: Stochastic simulation for bayesian inference*. Taylor Francis, 2006.

[34] Andrew Gelman et al. *Bayesian Data Analysis*. CRC Press, Taylor Francis Group, 2015.

[35] Tedros Adhanom Ghebreyesus. *Who director-general's opening remarks at the media briefing on COVID-19 - 11 march 2020*. Accessed: 2023. 2020. URL:

https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

[36] Jochen Görtler, Rebecca Kehlbeck, and Oliver Deussen. *A visual exploration of Gaussian processes*. Accessed 2024. 2021. DOI: 10.23915/distill.00017. URL: https://distill.pub/2019/visual-exploration-gaussian-processes/.

[37] Institute for Government. 2021. URL: https://www.instituteforgovernment.org.uk/sites/default/files/2022-12/timeline-coronavirus-lockdown-december-2021.pdf.

[38] Loukas Grafakos. *Fundamentals of fourier analysis*. Springer, 2024.

[39] Alison Hale and Peter Diggle. *Lancaster Medical School - Chicas - The Dynamic Health Atlas*. Accessed: 2024. 2017. URL: https://chicas.lancaster-university.uk/projects/dynamic_health_data.html.

[40] Alison Hale et al. "A real-time spatio-temporal syndromic surveillance system with application to small companion animals". English. In: *Scientific Reports* 9 (2019). ISSN: 2045-2322. DOI: 10.1038/s41598-019-53352-6.

[41] Healthline. "Disease Transmission: Direct Contact vs. Indirect Contact". In: (2016). Accessed: 2025-04-08. URL: https://www.healthline.com/health/disease-transmission.

[42] J. A. Hernandez, R. Heller, and L. H. Kahn. "Detecting Emerging Diseases in Farm Animals through Clinical Observations". In: *Emerging Infectious Diseases* 12.3 (2006), pp. 440–445. DOI: 10.3201/eid1203.050917.

[43] Matthew D. Hoffman and Andrew Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Accessed: 2023. 2011. arXiv: 1111.4246 [stat.CO].

[44] Peter Hourston. *Cost of living crisis*. Accessed: 2022. 2022. URL: https://www.instituteforgovernment.org.uk/explainer/cost-living-crisis.

[45] Aliaksandr Hubin et al. "A bayesian binomial regression model with latent gaussian processes for modelling DNA methylation". In: *Austrian Journal of Statistics* 49.4 (2020), pp. 46–56. DOI: 10.17713/ajs.v49i4.1124.

[46] A Hulth et al. "Practical usage of computer-supported outbreak detection in five European countries". In: *Eurosurveillance* 15.36 (2010). DOI: 10.2807/ese.15.36.19658-en.

[47] Rob Hyndman and G. Athanasopoulos. *Forecasting: Principles and Practice*. English. 3rd. Australia: OTexts, 2021.

[48] Rob J Hyndman and George Athanasopoulos. *Forecasting: Principles and practice (2nd ed)*. Accessed 2024. 2018. URL: https://otexts.com/fpp2/dhr.html.

[49] John Hopkins University. Accessed: 2023. 2020. URL: https://coronavirus.jhu.edu/map.html.

[50] Peter Kasprzak et al. "Six Years of Shiny in Research - Collaborative Development of Web Tools in R". In: *The R Journal* 12 (2 2021). https://rjournal.github.io/, pp. 20–42. ISSN: 2073-4859.

[51] Shwet Ketu and Pramod Kumar Mishra. "Enhanced Gaussian process regression-based forecasting model for COVID-19 outbreak and significance of IoT for its detection". In: *Applied Intelligence* 1.21 (2020). DOI: 10.1007/s10489-020-01889-9.

[52] Rowland King. *Rapid Labs launch new range of veterinary diagnostic tests*. Accessed: 2024. 2023. URL: https://www.rapidlabs.co.uk/news/rapid-labs-launch-new-range-of-veterinary-diagnostic-tests/.

[53] Seyoon Ko et al. "High-performance statistical computing in the computing environments of the 2020s". In: *Statistical Science* 37.4 (2022). DOI: 10.1214/21-sts835.

[54] Alexander Kowarik and Matthias Templ. "Imputation with the R Package VIM". In: *Journal of Statistical Software* 74.7 (2016), pp. 1–16. DOI: 10.18637/jss.v074.i07.

[55] Nan M. Laird and James H. Ware. "Random-Effects Models for Longitudinal Data". In: *Biometrics* 38.4 (1982), pp. 963–974. ISSN: 0006341X, 15410420. URL: http://www.jstor.org/stable/2529876 (visited on 06/04/2024).

[56] Ben Lambert. *A student's Guide to Bayesian statistics*. SAGE, 2018.

[57] Graham R. Law and Shane W. Pascoe. "4 Dealing With the Numbers". In: *Statistical Epidemiology*. CABI, 2013, pp. 125–125.

[58] Christina Mack, Daniel Weistreich, and Zhaohui Su. *Managing missing data in patient registries: Addendum to registries for evaluating patient outcomes: A user's guide*. Agency for Healthcare Research and Quality (US), 2018.

[59] MSD Veterinary Manual. *Preventative Health Care for Small Animals*. Accessed: 2025-01-29. 2025. URL: https://www.msdvetmanual.com/management-and-nutrition/preventative-health-care-and-husbandry-in-small-animals/preventative-health-care-for-small-animals.

[60] Monika Martyn. *Pet ownership statistics UK – what interesting facts can tell us!* Accessed: 2024. 2023. URL: https://worldanimalfoundation.org/advocate/pet-ownership-statistics-uk/#:~:text=13%20Million%20Pets%20in%20UK%20Households%20Are%20Dogs%20(PFMA)&text=Ten%20million%20or%2034%25%20of,8.2%20million)%20owning%20a%20cat.

[61] J. D. Mayer, H. S. J. Zald, and D. H. Pletscher. "Wildlife disease monitoring and surveillance". In: *Journal of Wildlife Diseases* 49.4 (2013), pp. 751–758. DOI: 10.7589/2012-11-305.

[62] Kirsten M McMillan et al. "Estimation of the size, density, and demographic distribution of the UK pet dog population". In: (2024). DOI: 10.21203/rs.3.rs-3772889/v1.

[63] Frédéric Michas. *Number of veterinarians in the UK 2021 - sourced from ONS Survey data*. Accessed: 2022. 2022. URL: https://www.statista.com/statistics/318888/numbers-of-veterinarians-in-the-uk/?fbclid=IwAR2S34RwoB9pbrmez_Uizwm7NmCVMi9VpWz1NRLekCjFQXTGoIsmkjo10U0#:~:text=In%202021%2C%20the%20number%20of,growing%20number%20of%20veterinarians%20employed.

[64] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN: 9780262044660. URL: https://www.

amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8%5C&qid=1336857747%5C&sr=8-2.

[65] National Eye Institution. *Color blindness*. Accessed: 2024. 2019. URL: https://www.nei.nih.gov/learn-about-eye-health/eye-conditions-and-diseases/color-blindness#:~:text=What%20is%20color%20blindness%3F,and%20contact%20lenses%20can%20help..

[66] Radford M. Neal. "MCMC Using Hamiltonian Dynamics". In: *Handbook of Markov Chain Monte Carlo* 54 (2010), pp. 113–162.

[67] Peter JM Noble et al. "Using topic modelling for unsupervised annotation of electronic health records to identify an outbreak of disease in UK dogs." In: *PLoS ONE* 16.12 (2021). DOI: 10.1371/journal.pone.0260402.

[68] Office for National Statistics. Dataset. Accessed: 2021. 2020. URL: https://geoportal.statistics.gov.uk/search?collection=Dataset&sort=name&tags=all(BDY_LAD%5C%2CDEC_2021).

[69] Office for National Statistics. *NUTS Level 1 (January 2018) Full Clipped Boundaries in the United Kingdom*. Accessed: 2020. 2018.

[70] Abril-Pla Oriol et al. "PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python". In: *PeerJ Computer Science* 9 (2023), e1516. DOI: 10.7717/peerj-cs.1516.

[71] Oxford Learners Dictionaries. *Lockdown*. Accessed: 2023. 2023. URL: https://www.oxfordlearnersdictionaries.com/definition/english/lockdown.

[72] Mike Pacey. *HEC 3.0 User Guide: Introduction*. Accessed: 2024. 2023. URL: https://lancaster-hec.readthedocs.io/en/latest/intro.html.

[73] Jaewoo Park and Murali Haran. *Bayesian Inference in the Presence of Intractable Normalizing Functions*. 2018. arXiv: 1701.06619 [stat.CO]. URL: https://arxiv.org/abs/1701.06619.

[74] Laura Physick. *How to build an Accessible Data dashboard*. Accessed: 2024. 2022. URL: https://home.vizlib.com/accessible-data-dashboard-designing/.

[75] Plotly Europe Ltd. *Dash in 20 minutes*. URL: https://dash.plotly.com/tutorial.

[76] Alan D. Radford et al. "Outbreak of Severe Vomiting in Dogs Associated with a Canine Enteric Coronavirus United Kingdom". In: *Emerging Infectious Diseases* 27.2 (2021), pp. 517–528. DOI: https://doi.org/10.3201/eid2702.202452.

[77] Md Tanvir Rahman et al. "Zoonotic Diseases: Etiology, Impact, and Control". In: *Microorganisms* 8.9 (Sept. 2020), p. 1405. DOI: 10.3390/microorganisms8091405. URL: https://www.mdpi.com/2076-2607/8/9/1405.

[78] Carl Edward Rasmussen and Christopher K I Williams. *Gaussian processes for machine learning*. Cambridge, Mass. Mit Press, 2008. ISBN: 9780262182539.

[79] J. Rushton, R. Viscarra, and J. Otte. "The Economic Impact of Foot and Mouth Disease in the United Kingdom 2007". In: *The Veterinary Journal* 179.1 (2009), pp. 1–14. DOI: 10.1016/j.tvjl.2007.10.019.

[80] Durgesh Samariya et al. "Detection and explanation of anomalies in healthcare data". In: *Health Information Science and Systems* 11.1 (2023). DOI: 10.1007/s13755-023-00221-2.

[81] Small Animal Veterinary Surveillance Network. https://www.liverpool.ac.uk/savsnet/about/. Accessed: 2023. 2020.

[82] Small Animal Veterinary Surveillance Network. *Real Time Data - Small Animal Veterinary Surveillance Network (SAVSNET) - University of Liverpool — liverpool.ac.uk*. https://www.liverpool.ac.uk/savsnet/real-time-data/. Accessed 2023. 2018.

[83] K. F. Smith et al. "Zoonotic Viruses Associated with Illegally Imported Wildlife". In: *PLoS ONE* 6.9 (2011), e17694. DOI: 10.1371/journal.pone.0017694.

[84] J. Snow. *On the Mode of Communication of Cholera*. John Churchill, 1849. URL: https://books.google.co.uk/books?id=-_dZAAAAcAAJ.

[85] Arjun Srinivasan et al. "Azimuth: Designing accessible dashboards for Screen Reader Users". In: *The 25th International ACM SIGACCESS Conference on Computers and Accessibility* (2023), pp. 1–16. DOI: 10.1145/3597638.3608405.

[86] Star Vets. *Beware of five common summer dog diseases*. Accessed: 2023. 2021. URL: https://www.starvetclinic.co.uk/article/beware-of-five-common-summer-dog-diseases/.

[87] Walter W. Stroup, Marina Ptukhina, and Julie Garai. *Generalized linear mixed models: Modern concepts, methods and applications*. CRC Press, 2024.

[88] Tableau. Accessed: 2024. 2023. URL: https://www.tableau.com/en-gb/learn/articles/data-visualization#:~:text=Data%20visualisation%20is%20the%20graphical,outliers%20and%20patterns%20in%20data.

[89] Tableau. *What is tableau?* Accessed: 2024. 2023. URL: https://www.tableau.com/en-gb/why-tableau/what-is-tableau.

[90] Carmen Tamayo Cuartero et al. "Setting clinically relevant thresholds for the notification of canine disease outbreaks to veterinary practitioners: An exploratory qualitative interview study". In: *Frontiers in Veterinary Science* 11 (2024). DOI: 10.3389/fvets.2024.1259021.

[91] Carmen Tamayo Cuartero et al. "Stakeholder opinion-led study to identify canine priority diseases for surveillance and control in the UK". In: *Veterinary Record* 193.9 (2023). DOI: 10.1002/vetr.3167.

[92] Hannah Thomas. "Data Visualisation and Digital Accessbility: What We Can Do to Help". Communicating Mathematics for the Public. 2023. URL: https://gateway.newton.ac.uk/sites/default/files/asset/doc/2301/Hannah-Thomas.slides.pdf.

[93] Y. L. Tong. *Multivariate normal distribution Y. L. Tong*. Springer New York, 2012.

[94] Maximilian E. Tschuchnig and Michael Gadermayr. "Anomaly detection in Medical Imaging - A Mini Review". In: *Data Science – Analytics and Applications* (2022), pp. 33–38. DOI: 10.1007/978-3-658-36295-9_5.

[95] UK Health Security Agency. 2020. URL: https://www.gov.uk/government/publications/covid-19-stay-at-home-guidance/stay-at-home-guidance-for-households-with-possible-coronavirus-covid-19-infection.

[96] UK Health Security Agency. Accessed: 2023. 2020. URL: https://coronavirus.data.gov.uk/.

[97] UK Surveillance Forum. *UK surveillance forum (UKSF)*. Accessed:2023. 2023. URL: https://www.gov.uk/government/groups/uk-surveillance-forum-uksf.

[98] University of Wisconsin–Madison. *8 publishing shiny apps online: Creating shiny apps at the SSCC*. Accessed: 2024. 2021. URL: https://sscc.wisc.edu/sscc/pubs/shiny/publishing-shiny-apps-online.html.

[99] Aki Vehtari, Andrew Gelman, and Jonah Gabry. *Practical bayesian model evaluation using leave-one-out cross-validation and WAIC*. 2016. URL: https://arxiv.org/abs/1507.04544.

[100] Royal College of Veterinary Surgeons. *RCVS Facts 2022*. Accessed: 2025-01-14. 2022. URL: https://www.rcvs.org.uk/news-and-views/publications/?filter-keyword=&filter-type=18&filter-month=&filter-year=.

[101] Trevor J Whitbread. *Laboratory tests routinely performed in veterinary medicine - special pet topics*. 2019. URL: https://www.msdvetmanual.com/special-pet-topics/diagnostic-tests-and-imaging/laboratory-tests-routinely-performed-in-veterinary-medicine.

[102] Richard Wilkinson. "An Introduction to Gaussian Processes". In: University of Nottingham. Gaussian Process Summer School, 2020.

[103] World Health Organisation. *Covid-19 cases — WHO COVID-19 Dashboard*. Accessed: 2023. 2020. URL: https://data.who.int/dashboards/covid19/cases?n=c.

[104] World Health Organization. *Disease Outbreaks*. Accessed: 2025-04-08. 2025. URL: https://www.emro.who.int/health-topics/disease-outbreaks/index.html.

[105] World Health Organization. *Zoonoses*. Accessed: 2025-04-08. 2020. URL: https://www.who.int/news-room/fact-sheets/detail/zoonoses.

[106] Zhaohua Wu et al. "On the trend, detrending, and variability of nonlinear and nonstationary time series". In: *Proceedings of the National Academy of Sciences* 104.38 (2007), pp. 14889–14894. DOI: 10.1073/pnas.0701020104.

[107] Daisuke Yoneoka et al. "Geographically weighted generalized Farrington Algorithm for rapid outbreak detection over short data accumulation periods". In: *Statistics in Medicine* 40.28 (2021), pp. 6277–6294. DOI: 10.1002/sim.9182.

[108] Ying Yuan and Valen E. Johnson. "Goodness-of-Fit Diagnostics for Bayesian hierarchical models". In: *Biometrics* 68.1 (2011), pp. 156–164. DOI: 10.1111/j.1541-0420.2011.01668.x.

[109] Bushra Zareie et al. "Outbreak detection algorithms based on Generalized Linear Model: A review with new practical examples". In: *BMC Medical Research Methodology* 23.1 (2023). DOI: 10.1186/s12874-023-02050-z.