Lyme disease forecasting report

Huw James and Gordon Yong, 30 June 2025

This report outlines the work we have undertaken to forecast reported Lyme disease cases in England in the years 2023 and 2024. The first section gives a summary of our approach, the second section provides geospatial data, and the concluding section gives a summary of covariates considered.

1. Summary of approach

When conducting any forecasting task, the first step is to understand the factors that are likely to affect the variable you are trying to forecast, and how they interact. There are three main factors that affect the reported incidence of Lyme disease, namely:

1. The prevalence of Lyme disease-carrying ticks.
2. The number of people visiting tick-infested areas.
3. General awareness of Lyme disease and how to report it.

The interaction of the first two factors is what affects the incidence, and the interaction of the first two with the third is what affects the level of reporting. Ideally, a forecasting model would include all three of the above factors as inputs. However, this is difficult in practice as:

1. There is no comprehensive data set on the prevalence of ticks across England.
2. There is no comprehensive data set on the number of visits to tick-infested areas.
3. It is difficult to even define a variable to represent awareness of Lyme disease.

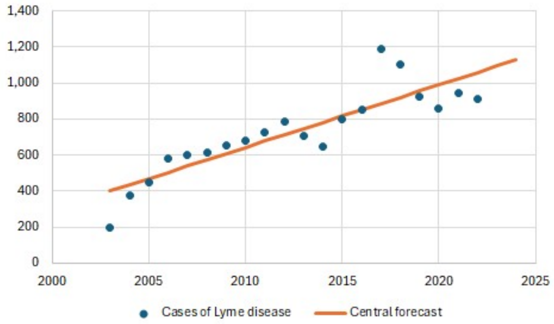
As the three factors above have all been increasing over time and are likely to continue to do so in future, we have proxied them all using the year as the input. Specifically, let y_k denote total the number of Lyme disease cases reported in England in year k. Then we assume that:

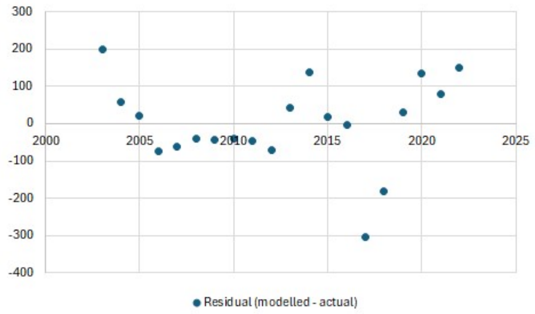
$$y_k = bk + c + e_k.$$

Here, a and b are regression coefficients, and e_k is an error term. The first chart below shows a clear linear relationship between the year and the number of reported Lyme disease cases. We obtained the central forecast for England using the least-squares estimates $\hat{b}$ and $\hat{c}$:

$$\hat{y}_k = \hat{b}k + \hat{c}.$$

Note that in the first chart below, we have included the modelled figures for previous years for illustrative purposes, even though technically only the figures for years 2023 and 2024 represent a forecast. The second chart shows a good scatter of residuals, and the R^2 value is 76%, indicating a good fit to the data.





Our geographical unit of analysis is District or Unitary Authority, depending on the region of England we are looking at. Let y_k^i denote total the number of Lyme disease cases reported in area (District / Unitary Authority) i in year k, and let p^i denote the proportion of cases reported in area i in the historical data, so that:

$$p^i = \sum_k y_k^i \,/\, \sum_k y_k.$$

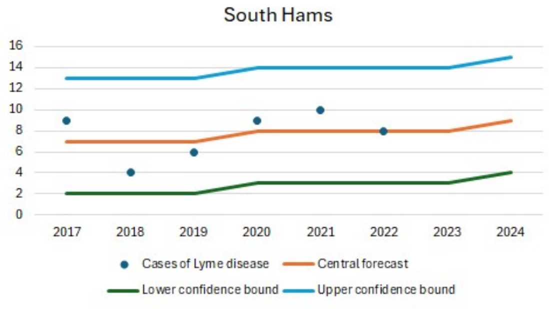
We obtained the central forecast for area i in (future) year k by multiplying the proportion p^i by the central forecast for England in year k, $\hat{y}_k$, and rounding to the nearest integer. In other words, letting $[x]$ denote the integer part of a real number x, we set:

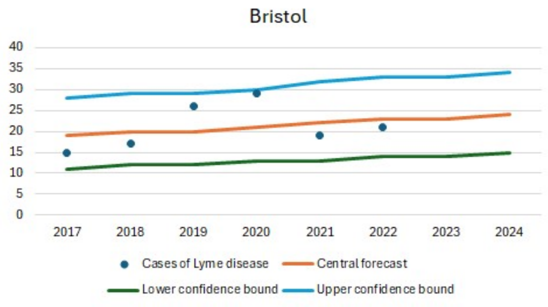
$$\hat{y}_k^i = [p^i \hat{y}_k].$$

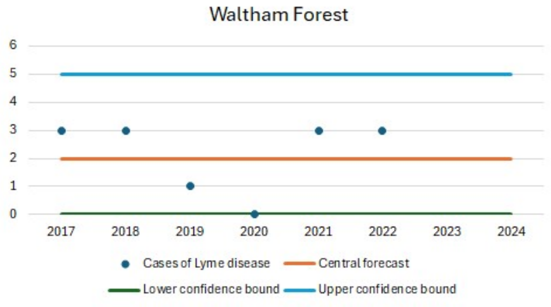
To obtain the lower and upper confidence bounds, we assumed that y_k^i has a Poisson distribution with rate $\hat{y}_k$. The Poisson distribution is a natural choice to model the number of events occurring in a fixed interval of time if these events occur with a known rate. The confidence interval for the forecast for area i in year k is then:

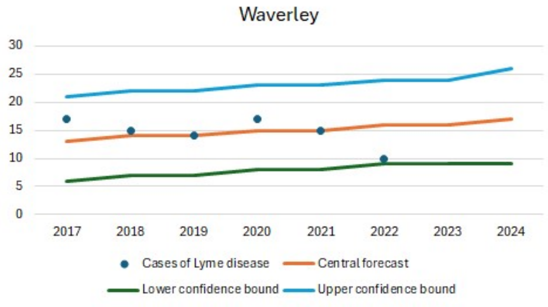
$$(F^{-1}[(1-a)/2 \mid \hat{y}_k^i],\, F^{-1}[1-(1-a)/2 \mid \hat{y}_k^i]).$$

Here, $a = 95\%$ is the confidence level, and $F^{-1}[\cdot \mid \lambda]$ is the inverse of the cumulative distribution function for a Poisson variable with rate λ. The figures below show the modelled central forecasts, and confidence bounds, for four randomly selected areas. Again, we have included the modelled figures for previous years for illustrative purposes.





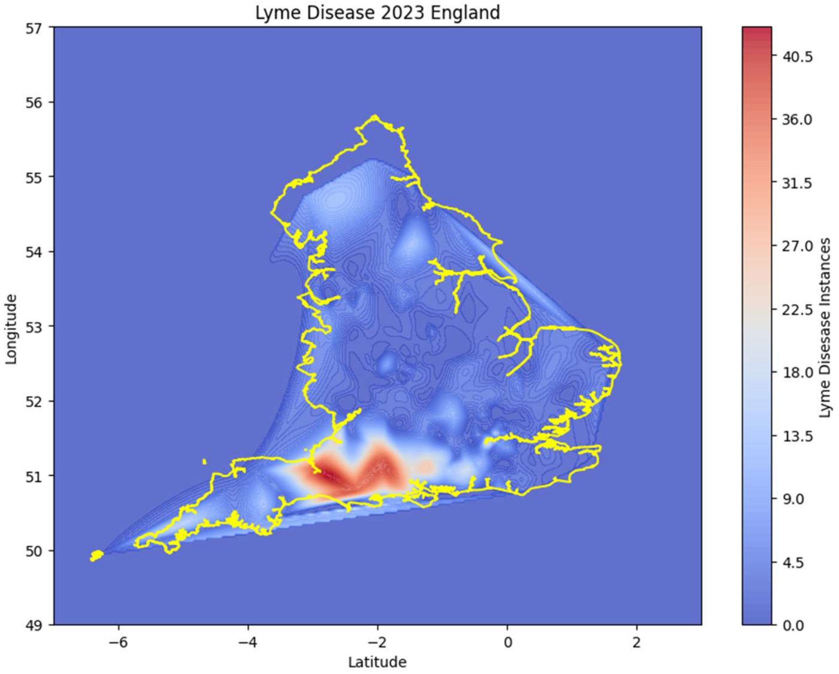


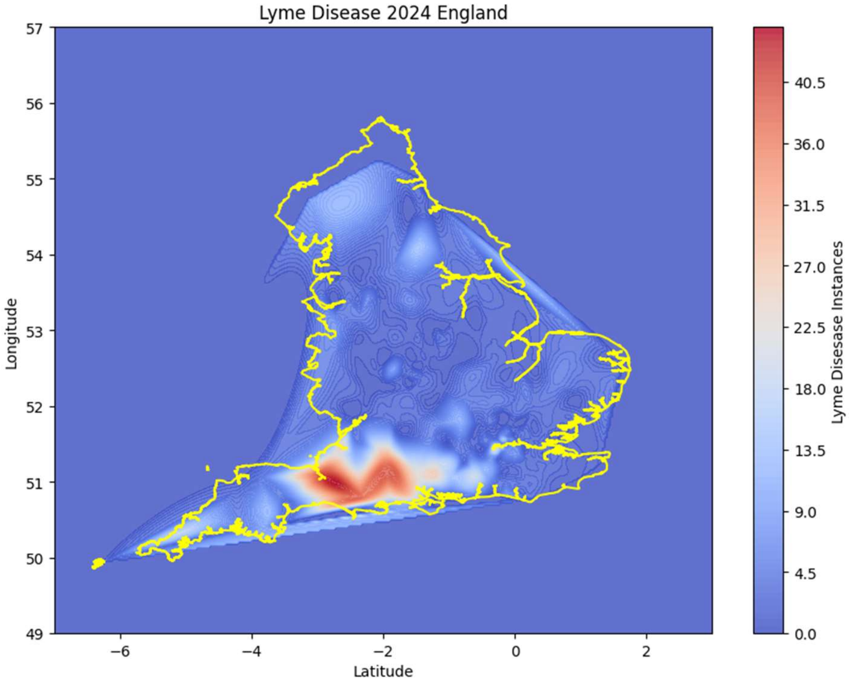


2. Geospatial data

To understand how Lyme disease changes across the UK, we used the forecast per district as epicentres for a contour plot, using the central coordinates as reference points.

We then applied a cubic interpolation using the SciPy Python library to produce estimates in areas between the epicentres. This allowed us to generate a contour plot, which we overlaid onto a map of England to create a geospatial estimate. The figures below give a visual representation of how the incidence rate is likely to vary geographically.





3. Covariates

When building the forecasting model, we considered including two additional variables:

- Average air temperature, to capture environmental conditions affecting tick activity.
- Public awareness, account for changes in reporting behaviour.

Average temperature is a proxy variable for both tick activity and human exposure. Studies suggest warmer weather tends to increase tick activity and suggest people spend more time outdoors, often with less protective clothing. However, we decided not to include temperature in the model for two reasons:

- There was not enough variation in annual temperatures over the forecasting period to make it a useful predictor.
- While temperature helped explain regional trends (e.g. higher incidence in the south), it did not align well with patterns at a more local level. For example, Eastern England has elevated temperatures but low incidence.

We considered public awareness as a covariate because it could help explain year-to-year changes in reported cases, especially spikes like the one in 2017. In the end, we excluded this variable because it is difficult to measure awareness consistently, and there is no reliable dataset capturing the scale or timing of awareness campaigns.