Forecasting Lyme Disease Incidence in England with Spatiotemporal Gaussian Processes

Harrison Bo Hua Zhu[1,2,*], Elizaveta Semenova[2,*], Mengyan Zhang[3], Hengguan Huang[1], and Samir Bhatt[1,2]

[1]Department of Public Health, University of Copenhagen, Denmark
[2]School of Public Health, Imperial College London, United Kingdom
[3]Department of Computer Science, University of Oxford, United Kingdom
[*]Equal Contribution

1 Introduction

Lyme borreliosis, the most common tick-borne zoonosis in Europe, is caused by the spirochete *Borrelia burgdorferi* and primarily transmitted by *Ixodes ricinus* ticks [1]. UK case reports have risen over the past two decades, likely due to land-use and climate change, host dynamics, and improved reporting, increasing public-health concern. Untreated infection can progress from mild symptoms to severe neurological, cardiac, and rheumatological disease [2].

Ecological studies show woodland and deer abundance drive tick presence [3], vegetation and altitude shape Scottish exposure [4], and INLA-SPDE captures spatiotemporal autocorrelation in incidence [5]. Yet most models still lack climate and host data integration, robust uncertainty quantification, and interpretability.

We address these gaps with a spatiotemporal Gaussian Process model for Lyme incidence across England. Using high-resolution environmental, demographic, and host layers, it provides accurate, uncertainty-aware forecasts and applies explainability tools to reveal key risk drivers. Our contributions are: (i) a comprehensive curated dataset, (ii) an uncertainty-aware incidence model, and (iii) interpretable insights to guide targeted surveillance and intervention.

2 Methods

2.1 Data

We assembled a 2017–2022 LTLA-level panel of Lyme incidence and drivers across England. Annual cases and populations from UKHSA yielded incidence per 100 000 (Figure 1).

Predictors span geography, climate, ecology, and demography. Geographic terms are LTLA centroid lat/long and year. Elevation is mean ERA5 geopotential height (1 Jan 2020) divided by 9.80665 ms^{-2} and assumed constant [6]. Monthly climate covariates such as precipitation, specific/relative humidity, solar radiation, pressure, and vegetation indices, came from ERA5-Land (9 km), then LTLA-averaged. Woodland density is the share of LTLA area with $\geq 30\%$ canopy from the 2019 Copernicus Tree Cover Fraction product [6]. Tick density was interpolated with a spatiotemporal Gaussian Process from VecDyn occurrence records and aggregated to LTLAs [7]. Host abundance (roe, Chinese water and red deer; wood mice; blackbirds; others) used NBN Atlas occurrences converted to LTLA counts [8]. Population proportions for 5-14 and 50-65 years old are also included, since there are studies showing that they are at-risk populations [9–12]. This multi-layer dataset underpins our incidence modelling. Table 3 shows the description of each covariate.

2.2 Models

Standardized Lyme incidence for each LTLA–year i is modeled with a hierarchical Gaussian-process (GP) regression. Let y_i be the incidence rate and $\mathbf{x}_i$ the covariates. Two likelihoods are compared:

$$y_i \mid \eta_i, \sigma^2 \sim \mathcal{N}\big(\eta_i, \sigma^2\big), \tag{1}$$

$$y_i \mid \eta_i, \phi \sim \mathrm{NegBin}\big(\mu_i, \phi\big), \qquad \mu_i = \exp(\eta_i). \tag{2}$$

Both share the predictor

$$\eta_i = f_{\mathrm{sp}}\big(\mathrm{lon}_i, \mathrm{lat}_i, \mathrm{year}_i\big) + \sum_{j=1}^{p_{\mathrm{static}}} f_j\big(x_{ij}\big) + \sum_{r=1}^{p_{\mathrm{climate}}} f_r\big(z_{ir}\big),$$

where the residual spatiotemporal term follows

$$f_{\mathrm{sp}} \;\sim\; \mathcal{GP}\Big(0,\, k_{\mathrm{space}} \cdot k_{\mathrm{time}}\Big),$$

with k_{space} a Matérn-$3/2$ kernel in Euclidean distance and k_{time} a squared-exponential kernel in year. Each static effect has $f_j \sim \mathcal{GP}(0, k_j)$ with k_j squared-exponential (age covariates use a 2-D version), and each climate effect has $f_r \sim \mathcal{GP}(0, k_r)$ with k_r squared-exponential on the 12-month seasonal profile.

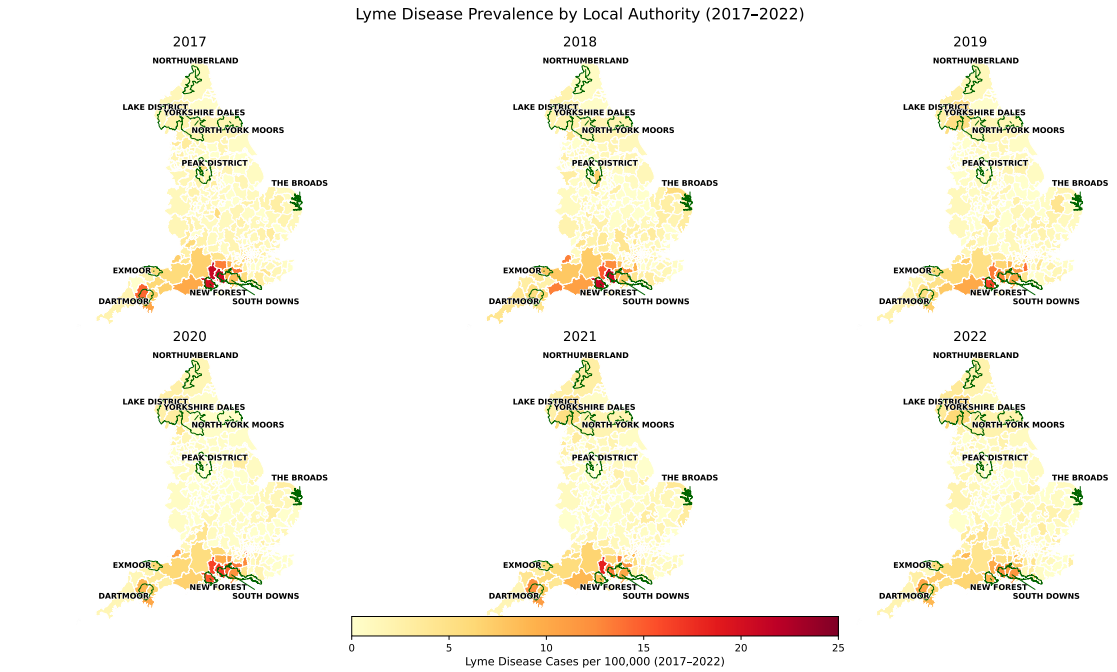

Figure 1: Lyme disease prevalence during 2017-2022. Green bounding polygons indicate key national parks.

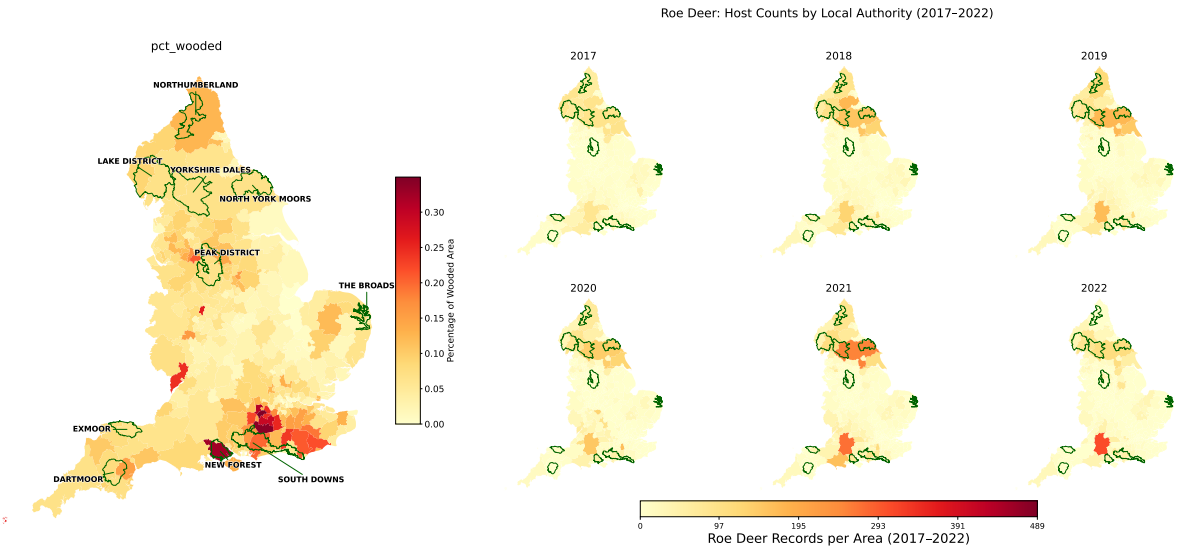

Figure 2: (Left) Percentage woodland for each local authority (Right) Roe deer host counts by local authority from 2017 to 2022.

For inference of the second model, we use sparse variational GP (SVGP) of [13], maximizing the evidence lower bound (ELBO) over inducing points and hyperparameters. Baselines include an ℓ_1-penalized Lasso and an XGBoost ensemble trained on the same predictors. All models are fitted on five train-validation splits over data in 2017-2022, and predictive accuracy is evaluated by RMSE on the validation set (Table 1).

2.3 Explainability

To translate predictive performance into actionable public-health insight, we applied two complementary classes of explainability techniques: Shapley-value attribution for *who* drives the prediction, and accumulated local effects (ALE) for *how* a covariate moves risk across its range [14].

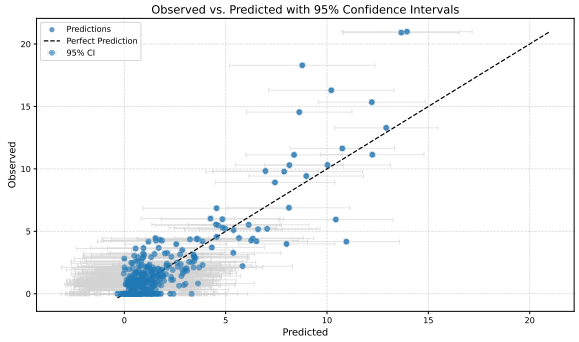

Figure 3: Predicted vs ground truth incidence on validation data for one seed with GPR + Normal model.

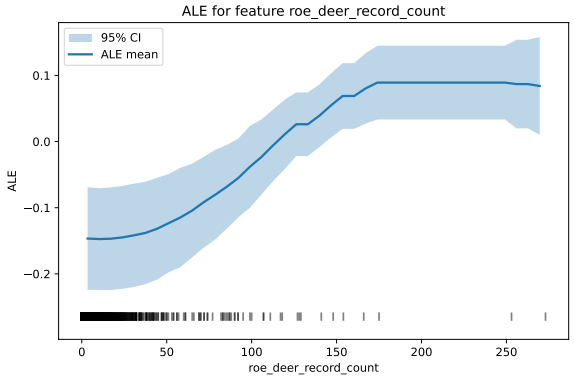

Figure 4: ALE plot for Roe deer host counts for GPR + Normal model. The shaded area represents the 95% confidence interval of the ALE.

3 Results

In this section, we present the results of our study.

3.1 Predictive Performance on 2024 and 2025 data

We report the RMSE for all models computed over 5 train-validation splits. We see that GP regression with Normal likelihood gives the best RMSE out of all the models and has particularly low performance variation across different seeds. We therefore choose this model's prediction as our final prediction in **forecast.csv**. Figure 3 shows the observed incidence vs predicted incidence by this model. We can see that most values are clustered around 0 incidence, since most local authorities record very low levels of incidence. For the high-incidence areas, we still are able to predict values that are reasonably close to the observed incidence.

Table 1: Validation set comparison computed on incidence (number of Lyme disease cases / population $\times 10^5$). The results are repeated across 5 train-validation splits.

Model	RMSE (mean $\pm$ se)
LASSO	2.31 $\pm$0.0761
XGBoost	1.64 $\pm$0.1100
GPR + Normal	$\mathbf{1.46 \pm 0.0450}$
GPR + NegBin + SVGP	1.62 $\pm$0.1350

3.2 Explainability results and plots

Table 2 shows the top 10 features selected by taking the mean absolute ALE for each model. We can see that there is strong agreement in the set of features deemed important, notably pct_wooded, roe_deer_record_count, longitude, latitude, 50_to_65y_fraction and 5_to_14y_fraction. It is expected that longitude and latitude both play an important role in the predictability, since there is strong spatial heterogeneity. For pct_wooded and roe_deer_record_count, it can also be seen from Figure 1 and 2 that their values are higher in places where there are high Lyme disease prevalence. As for 50_to_65y_fraction and 5_to_14y_fraction, it is suggested in multiple past works that they are age groups at risk of contracting Lyme disease.

4 Discussion

First, we develop a GP-based probabilistic model for Lyme incidence using climate, landcover, host-abundance, and at-risk human-population covariates, which outperforms Lasso and XGBoost baselines. Second, we apply explainability methods across all four models, consistently pinpointing woodland cover, roe deer abundance, at-risk human population, and spatial heterogeneity as primary drivers, consistent with prior field and observational studies.

Limitations:

- **Tick data quality:** VecDyn Explorer records are sparse, so we interpolated LTLA-level tick prevalence. Although three models flag tick_pred as influential, this finding warrants caution. Future work could incorporate higher-resolution estimates (e.g. [3]).
- **Under-reporting:** UKHSA case counts omit asymptomatic or unreported infections, so our incidence predictions likely underestimate true prevalence, despite remaining useful for policy and driver analysis.

Acknowledgements

S.B. acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research

Table 2: Top 10 features for each model ranked by Mean Absolute ALE. Mean absolute SHAP is also attached for LASSO and XGBoost.

model	feature	Mean Absolute ALE	Mean Absolute SHAP
GPR + Normal	tick_pred	0.791	—
GPR + Normal	**pct_wooded**	0.738	—
GPR + Normal	**roe_deer_record_count**	0.678	—
GPR + Normal	**longitude**	0.421	—
GPR + Normal	lai_lv_3	0.357	—
GPR + Normal	**50_to_65y_fraction**	0.333	—
GPR + Normal	**latitude**	0.251	—
GPR + Normal	lai_lv_9	0.224	—
GPR + Normal	lai_lv_2	0.222	—
GPR + Normal	lai_lv_10	0.202	—
SVGP + NegBin	**longitude**	1.310	—
SVGP + NegBin	**roe_deer_record_count**	1.102	—
SVGP + NegBin	**latitude**	0.694	—
SVGP + NegBin	blackbird_record_count	0.427	—
SVGP + NegBin	**50_to_65y_fraction**	0.412	—
SVGP + NegBin	**pct_wooded**	0.334	—
SVGP + NegBin	chinese_water_deer_record_count	0.283	—
SVGP + NegBin	**5_to_14y_fraction**	0.251	—
SVGP + NegBin	src_8	0.243	—
SVGP + NegBin	wood_mice_record_count	0.238	—
LASSO	**roe_deer_record_count**	0.554	0.264
LASSO	tick_pred	0.545	0.531
LASSO	**latitude**	0.522	0.501
LASSO	blackbird_record_count	0.380	0.030
LASSO	**longitude**	0.326	0.310
LASSO	**pct_wooded**	0.294	0.272
LASSO	**5_to_14y_fraction**	0.197	0.182
LASSO	spec_humidity_3	0.072	0.069
LASSO	**50_to_65y_fraction**	0.064	0.058
LASSO	rel_humidity_3	0.051	0.048
XGBoost	**longitude**	0.818	0.326
XGBoost	**latitude**	0.806	0.984
XGBoost	tick_pred	0.342	0.257
XGBoost	**roe_deer_record_count**	0.317	0.385
XGBoost	lai_lv_6	0.244	0.046
XGBoost	lai_hv_2	0.215	0.105
XGBoost	**50_to_65y_fraction**	0.199	0.082
XGBoost	**5_to_14y_fraction**	0.173	0.135
XGBoost	lai_hv_4	0.168	0.040
XGBoost	blackbird_record_count	0.167	0.013

Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking. S.B. acknowledges support from the Danish National Research Foundation via a chair grant (DNRF160). S.B. acknowledges support from The Eric and Wendy Schmidt Fund For Strategic Innovation via the Schmidt Polymath Award (G-22-63345) which also supports H.B.H.Z. S.B. acknowledges support from the Novo Nordisk Foundation via The Novo Nordisk Young Investigator Award (NNF20OC0059309). ES acknowledges being supported in part by the AI2050 program at Schmidt Sciences (Grant [G-22-64476]). ES acknowledges HPRU HAM.

Lastly, ChatGPT was used to refine the grammar and for redacting the text.

References

[1] Rizzoli, A. *et al.* Lyme borreliosis in europe. *Eurosurveillance* **16**, 19906 (2011).

[2] Nguyen, A., Mahaffy, J. & Vaidya, N. K. Modeling transmission dynamics of lyme disease: Mul-

tiple vectors, seasonality, and vector mobility. *Infectious Disease Modelling* **4**, 28–43 (2019).

[3] Burdon, M. G. *et al.* Modelling the distribution of the tick ixodes ricinus in england and wales using passive surveillance data from citizen science reports (2025).

[4] Lou, Y. & Wu, J. Modeling lyme disease transmission. *Infectious Disease Modelling* **2**, 229–243 (2017).

[5] Neupane, N., Goldbloom-Helzner, A. & Arab, A. Spatio-temporal modeling for confirmed cases of lyme disease in virginia. *Ticks and Tick-borne Diseases* **12**, 101822 (2021).

[6] Copernicus Climate Change Service. ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (2023). URL `https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels`. Accessed: 08 June 2025.

[7] Rund, S. S. *et al.* VecDyn Explorer. University of Notre Dame, VectorByte Initiative (2023). URL `https://vectorbyte.crc.nd.edu/vecdyn-detail/244`. Accessed: 07 May 2025.

[8] NBN Trust. NBN Atlas occurrence download at `https://nbnatlas.org`. Accessed: 07 May 2025 (2025).

[9] Steere, A. C., Coburn, J. & Glickstein, L. The emergence of Lyme disease. *Journal of Clinical Investigation* **113**, 1093–1101 (2004).

[10] Stanek, G., Wormser, G. P., Gray, J. & Strle, F. Lyme borreliosis. *The Lancet* **379**, 461–473 (2012).

[11] Hahn, M., Fingerle, V. & Wilske, B. Epidemiology of Lyme borreliosis in Germany, 2013–2017. *Eurosurveillance* **23**, 170–0824 (2018).

[12] Kugelman, J. R., Shapiro, E. D., Brenner, N. & Huber, M. Age-specific incidence of Lyme disease in Massachusetts: 1993–2015. *American Journal of Tropical Medicine and Hygiene* **99**, 416–423 (2018).

[13] Hensman, J., Fusi, N. & Lawrence, N. D. Gaussian processes for big data. *arXiv preprint arXiv:1309.6835* (2013).

[14] Christoph, M. Interpretable machine learning: A guide for making black box models explainable (2020).

A Description of covariates

Table 3: Description of covariates used in the model.

Covariate	Variable name	Description	Data source	Preprocessing
Latitude	lat	Centroid latitude of LTLA	LTLA shapefiles	Computed as mean latitude per LTLA
Longitude	lon	Centroid longitude of LTLA	LTLA shapefiles	Computed as mean longitude per LTLA
Year	year	Calendar year (2017–2022)	UKHSA	Used directly from reported incidence year
Elevation	elevation	Mean elevation (m)	ERA5 single-level reanalysis	Geopotential height converted to meters, averaged over LTLA
Air temperature	t2m	2-m air temperature	ERA5-Land	Monthly mean, aggregated by LTLA and year
Total precipitation	tp	Total precipitation	ERA5-Land	Monthly mean, aggregated by LTLA and year
Specific humidity	spec_humidity	Specific humidity	ERA5-Land	Monthly mean, aggregated by LTLA and year
Relative humidity	rel_humidity	Relative humidity	ERA5-Land	Monthly mean, aggregated by LTLA and year
Shortwave radiation	src	Surface shortwave radiation	ERA5-Land	Monthly mean, aggregated by LTLA and year
Surface pressure	sp	Surface atmospheric pressure	ERA5-Land	Monthly mean, aggregated by LTLA and year
Leaf area index (high)	lai_hv	High vegetation leaf area index	ERA5-Land	Monthly mean, aggregated by LTLA and year
Leaf area index (low)	lai_lv	Low vegetation leaf area index	ERA5-Land	Monthly mean, aggregated by LTLA and year
Woodland cover	pct_wooded	% area with tree cover $\geq$30%	Copernicus NRT Tree Cover	Clipped to LTLA boundary; percentage of pixels meeting canopy threshold
Tick density	tick_pred	Predicted tick occurrence	VecDyn Explorer (dataset 244)	GP-smoothed interpolation of occurrence records, aggregated to LTLA
Roe deer records	roe_deer_record_count	Estimated abundance	NBN Atlas	Occurrence data aggregated to LTLA
Chinese water deer records	chinese_water_deer_record_count	Estimated abundance	NBN Atlas	Occurrence data aggregated to LTLA
Red deer records	red_deer_record_count	Estimated abundance	NBN Atlas	Occurrence data aggregated to LTLA
Wood mice records	wood_mice_record_count	Estimated abundance	NBN Atlas	Occurrence data aggregated to LTLA
Blackbird records	blackbird_record_count	Estimated abundance	NBN Atlas	Occurrence data aggregated to LTLA
Age 5–14 population	5_to_14y_fraction	% population aged 5–14	ONS	Computed as fraction per LTLA
Age 50–65 population	50_to_65y_fraction	% population aged 50–65	ONS	Computed as fraction per LTLA