We present analyses of real-data examples using the I-probit model for a variety of applicaitons, namely binary and multiclass classification, meta-analysis, and spatio-temporal modelling of point processes. Examples in this section have been analysed using in R using the in-development **iprobit** package written by us. Code for replication is provided at http://myphdcode.haziqj.ml. All of these examples had assumed a fixed error precision $\boldsymbol{\Psi} = \mathbf{I}_m$.

0.0.1 Predicting cardiac arrhythmia

Statistical learning tools are being used in the field of medicine as a means to aid medical diagnosis of diseases. In this example, factors determining the presence or absence of heart diseses are studied. Traditionally, cardiologists inspect patients' cardiac activity (ECG data) in order to reach a diagnosis, which remains the "gold standard" method of obtaining diagnoses. The study by Guvenir et al. (1997) aimed to predict cardiac abnormalities by way of machine learning, and minimise the difference between the gold standard and computer-based classifications.

The data set[1] at hand contains a myriad of ECG readings and other patient attributes such as age, height, and weight. Altogether, there are $n = 451$ observations and $p = 279$ predictors. In order for a valid comparison to be made to other studies, we excluded nominal covariates, leaving us with $p = 194$ continuous predictors, which we then standardised. In the original data set, there are 13 distinct classes of cardiac arrhythmia—again, following the lead of other studies, we had combined all forms of cardiac diseases to form a single class, thus reducing the problem to a binary classification task (normal vs. arrhythmia).

Following (5.6), the relationship between patient i's probability of having a form of cardiac arrhthmia p_i and the predictors $x_i \in \mathcal{X} \equiv \mathbb{R}^{194}$ is modelled as

$$\Phi(p_i) = \alpha + f(x_i).$$

Further, assuming $f \in \mathcal{F}$ a suitable RKHS with kernel h_λ, we may assign an I-prior on the (latent) regression function f. We consider three RKHSs: the canonical (linear) RKHS, the fBm-0.5 RKHS and the SE RKHS. The first of these three assumes an underlying linear relationship of the covariates and the probabilities, while the other two assumes a smooth relationship. As all covariates had been standardised, it is sufficient to assign a single scale parameter λ for the I-probit model.

For reference, fitting an I-probit model on the full data set takes about 4 seconds only, with convergence reached in at most 15 iterations. Figure 1 plots the variational lower

[1] Data is made publicly available at https://archive.ics.uci.edu/ml/datasets/arrhythmia.

bound value over time and iterations for the cardiac arrhythmia data set. As expected, the lower bound value increases over time until a convergence criterion is reached.

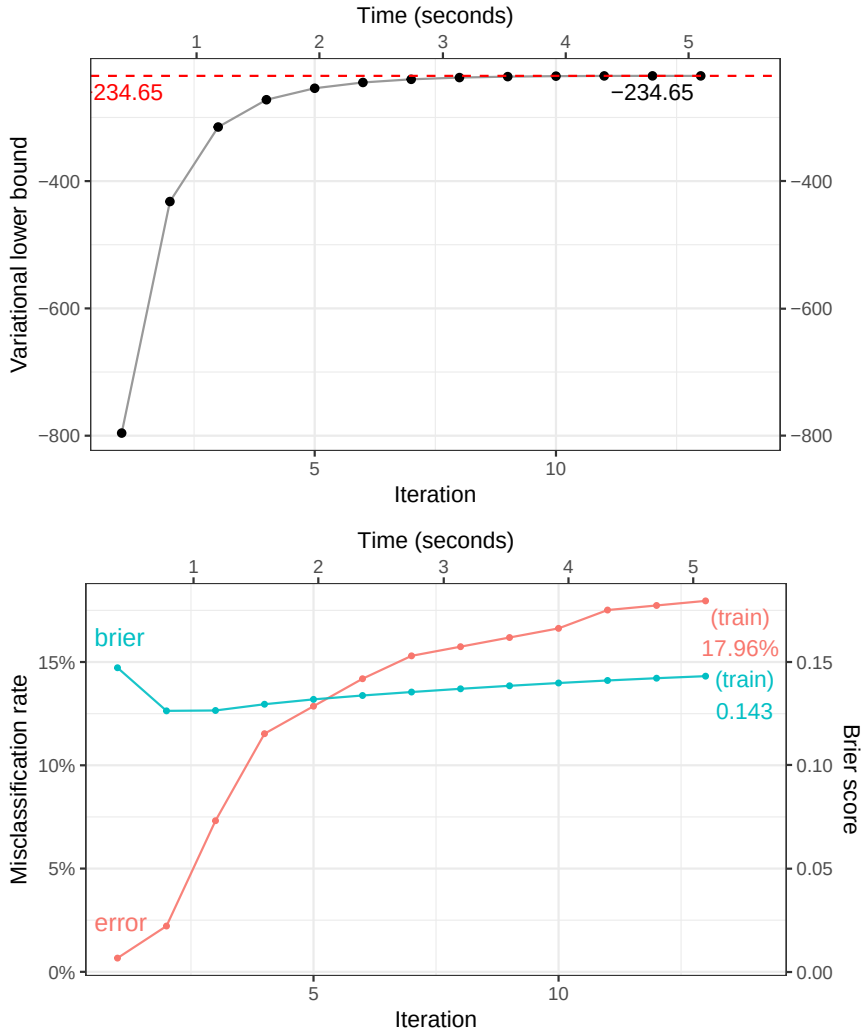


Figure 1: Plot of variational lower bound over time (top), and plot of training error rate and Brier scores over time (bottom).

To measure predictive ability, we fit the I-probit models on a random subset of the data and obtain the out-of-sample test error rates from the remaining held-out observations. We then compare the results against popular machine learning classifiers, namely: 1) linear and quadratic discriminant analysis (LDA/QDA); 2) k-nearest neighbours; 3) support vector machines (SVM) (Steinwart and Christmann, 2008); 4) Gaussian process classification (Rasmussen and Williams, 2006); 5) random forests (Breiman, 2001); 6) nearest shrunken centroids (NSC) (Tibshirani et al., 2002); and 7) L-1 penalised logistic regression (Friedman et al., 2001). The experiment is set up as follows:

1. Form a training set by sub-sampling $s \in \{50, 100, 200\}$ observations.

2. The remaining unsampled data is used as the test set.

3. Fit model on training set, and obtain test error rates defined as

$$\text{test error rate} = \frac{1}{s}\sum_{i=1}^{n}[y_i^{\text{pred}} \neq y_i^{\text{test}}] \times 100\%.$$

4. Repeat steps 1-3 100 times to obtain the *average* test error rates and standard errors.

Results for the methods listed above were extracted from the in-depth study by Cannings and Samworth (2017), who also conducted identical experiments using their random projection (RP) ensemble classification method. These are all tabulated in Table 1.

Table 1: Mean out-of-sample misclassification rates and standard errors in parantheses for 100 runs of various training (s) and test ($451 - s$) sizes for the cardiac arrhythmia binary classification task.

	Misclassification rate (%)		
Method	$s = 50$	$s = 100$	$s = 200$
I-probit			
Linear	35.52 (0.44)	31.35 (0.33)	29.45 (0.38)
Smooth (fBm-0.5)	33.64 (0.66)	28.12 (0.34)	24.33 (0.24)
Smooth (SE-1.0)	48.26 (0.40)	48.32 (0.43)	47.11 (0.37)
Others			
RP-LDA	33.24 (0.42)	30.19 (0.35)	27.49 (0.30)
RP-QDA	30.47 (0.33)	28.28 (0.26)	26.31 (0.28)
RP-k-NN	33.49 (0.40)	30.18 (0.33)	27.09 (0.31)
Random forests	31.65 (0.39)	26.72 (0.29)	22.40 (0.31)
SVM (linear)	36.16 (0.47)	35.61 (0.39)	35.20 (0.35)
SVM (Gaussian)	48.39 (0.49)	47.24 (0.46)	46.85 (0.43)
GP (Gaussian)	37.28 (0.42)	33.80 (0.40)	29.31 (0.35)
NSC	34.98 (0.46)	33.00 (0.40)	31.08 (0.41)
L-1 logistic	34.92 (0.42)	30.48 (0.34)	26.12 (0.27)

Of the three I-probit models, the fBm model performed the best. That it performed better than the canonical linear I-probit model is unsurprising, since an underlying smooth function to model the latent variables is expected to generalise better than a rigid straight line function. The poor performance of the SE I-probit model may be due to the fact that the lengthscale parameter was not estimated (fixed at $l = 1$), but then again, we notice reliable performance of the fBm even with fixed Hurst index ($\gamma = 0.5$). It can be seen that the fBm I-probit model also outperform the more popular machine learning algorithms out there including k-nearest neighbours, support vector machines and Gaussian process classification. It came second only to random forests, an ensemble learning method, which is also generally faster to train than Gaussian process-

like regressions such as I-prior models. The time complexity of a random forest algorithm is $O(pqn\log(n))$ (Louppe, 2014), where p is the number of variables used for training, q is the number of random decision trees, and n is the number of observations.

0.0.2 Meta-analysis of smoking cessation

Conider the smoking cessation data set, as described in Skrondal and Rabe-Hesketh (2004). It contains observations from 27 separate smoking cessation studies in which participants are subjected to either a nicotine gum treatment or a placebo. The interest is to estimate the treatment effect size, and whether it is statistically significant, i.e. whether or not nicotine gum is an effective treatment for quitting smoking. The studies are conducted at different times and due to various reasons such as funding and cultural effects, the results from all of the studies may not be in agreement. The number of effective participants plays a major role in determining the power of the statistical tests performed in individual studies. The question then becomes how do we meaningfully aggregate all the data to come up with one summary measure?

Several methods exist to analyse such data sets. One may consider a fixed-effects model, similar to a classical one-way ANOVA model to establish whether or not the effect size is significant. Because of the study-specific characteristics, it is natural to consider multilevel or random-effects models as a means to estimate the effect size. Regardless of method, the approach of analysing study-level treatment effects instead of patient-level data only is the paradigm for meta-analysis, and our I-prior model takes this approach as well.

A summary of the data is displayed by the box-plot in Figure 2. On the whole, there are a total of 5,908 patients, and they are distributed roughly equally among the control and treatment groups (46.3% and 53.7% respectively, on average). From the box-plots, it is evident that there are more patients who quit smoking in the treatment group as compared to the placebo control group. There are various measures of treatment effect size, such as risk ratio or risk differences, but we shall concentrate on *odds ratios* as defined by

$$\text{odds ratio} = \frac{\text{odds of quitting smoking in } \textit{treatment} \text{ group}}{\text{odds of quitting smoking in } \textit{control} \text{ group}}.$$

The odds of quitting smoking in either group is defined as

$$\text{odds} = \frac{\text{P(quit smoking)}}{1 - \text{P(quit smoking)}},$$

and these probabilities, odds and ultimately odds ratio can be estimated from sample proportions. This raw odds ratio for all study groups is calculated as $1.66 = e^{0.50}$. It is

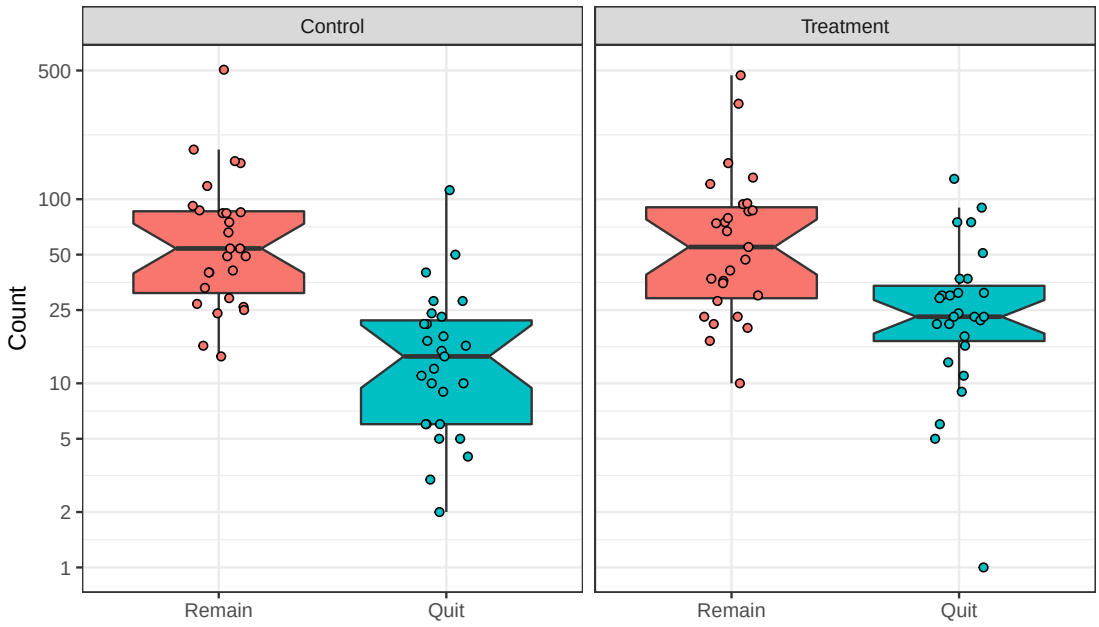


Figure 2: Comparative box-plots of the distribution of patients who successfully quit smoking and those who remained smokers, in the two treatment groups. It is evident that there are more successful patients quitting smoking in the treatment group than in the control group. The raw odds ratio of quitting smoking (treatment vs. control) is 1.66.

also common for the odds ratio to be reported on the log scale (usually as a remnant of logistic models). A value greater than one for the odds ratio (or equivalently, greater than zero for the log odds ratio) indicates a significant treatment effect.

A random-effects analysis using a multilevel logistic model has been considered by Agresti and Hartzel (2000). Let $i = 1, \ldots, n_k$ index the patients in study group $k \in \{1, \ldots, 27\}$. For patient i in study j, p_{ik} denotes the probability that the patient has successfully quit smoking. Additionally, x_{ik} is the centred dummy variable indicating patient i's treatment group in study k. These take on two values: 0.5 for treated patients and -0.5 for control patients. The logistic random-effects model is

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}x_{ij}$$

with

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}\right)$$

Agresti and Hartzel (2000) also made the additional assumption $\sigma_{01} = 0$, so that, coupled with the contrast coding used for x_{ik}, the total variance $\mathrm{Var}(\beta_{0k} + \beta_{1j}x_{ik})$ would be

constant in both treatment groups. The overall log odds ratio is represented by β_1, and this is estimated as $0.57 \approx \log 1.76$.

In an I-prior model, the Bernoulli probabilities p_{ik} are regressed against the treatment group indicators x_{ik} and also the patients' study group k via the regression function f and a probit link:

$$\begin{aligned}\Phi^{-1}(p_{ik}) &= f(x_{ik}, k) \\ &= f_1(x_{ik}) + f_2(k) + f_{12}(x_{ik}, j).\end{aligned}$$

We have decomposed our function f into three parts: f_1 represents the treatment effect, f_2 represents the effect of the study groups, and f_{12} represents the interaction effect between the treatment and study group on the modelled probabilities. As both x_{ik} and k are nominal variables, the functions f_1 and f_2 both lie in the Pearson RKHS of functions $\mathcal{F}_1$ and $\mathcal{F}_2$, each with RKHS scale parameters λ_1 and λ_2. As such, it does not matter how the x_{ik} variables are coded (dummy coding 0, 1 vs. centred coding -0.5, 0.5) as the scaling of the function is determined by the RKHS scale parameters. The interaction effect f_{12} lies in the RKHS tensor product $\mathcal{F}_1 \otimes \mathcal{F}_2$. In the I-probit model, there are only two parameters to estimate, while in the standard logistic random-effects model, there are six. The results of the I-prior fit are summarised in the table below.

Table 2: Results of the I-probit model fit for three models.

Model	ELBO	Error rate (%)	Brier score	No. of parameters
f_1	-3210.76	23.65	0.179	1
$f_1 + f_2$	-3142.24	29.30	0.206	2
$f_1 + f_2 + f_{12}$	-3091.20	23.48	0.168	2

The approximated marginal log-likelihood value for the I-prior model (i.e. variational lower bound), the Brier score for each model and the number of RKHS scale parameters estimated in the model are reported in Table 2. Three models were fitted: 1) a model with only the treatment effect; 2) a model with a treatment effect and a study group effect; and 3) Model 2 with the additional assumption that treatment effect varies across study groups. Model 1 disregards the study group effects, while Model 2 assumes that the effectiveness of the nicotine gum treatment does not vary across study groups (akin to a varying-intercept model). A model comparison using the evidence lower bound indicates that Model 3 has the highest value, and the difference is significant from a Bayes factor standpoint—$\text{BF}(M_3, M_1)$ and $\text{BF}(M_3, M_2)$ are both greater than 150. The misclassification rate and Brier score indicates good predictive performance of the models, and there is not much to distinguish between the three although Model 3 is the best out of the three models.

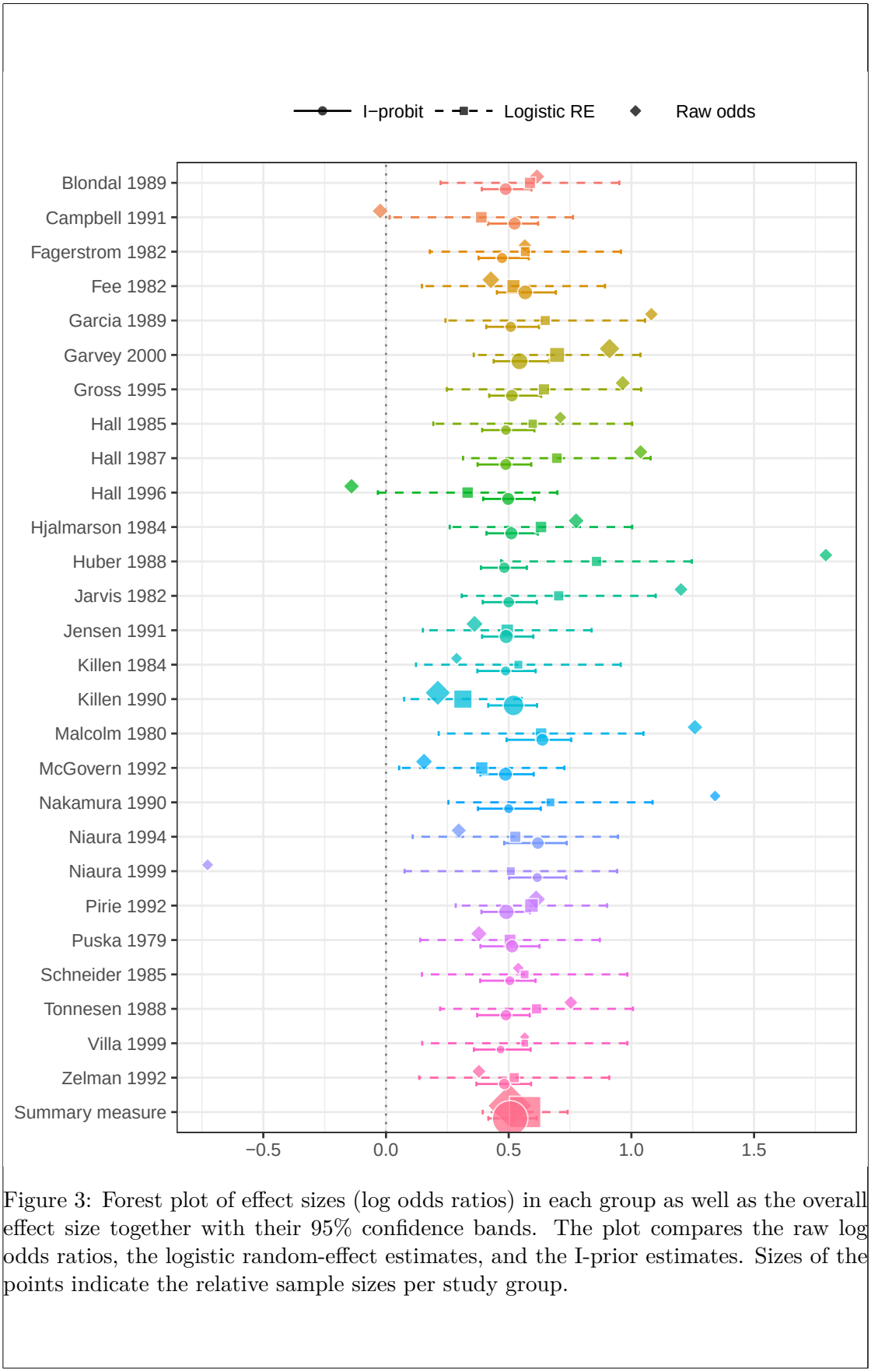


Figure 3: Forest plot of effect sizes (log odds ratios) in each group as well as the overall effect size together with their 95% confidence bands. The plot compares the raw log odds ratios, the logistic random-effect estimates, and the I-prior estimates. Sizes of the points indicate the relative sample sizes per study group.

Unlike in the logistic random-effects model, where the log odds ratio can be read off directly from the coefficients, with an I-prior probit model the log odds ratio needs to be calculated manually from the fitted probabilities. The probabilities of interest are the probabilities of quitting smoking under each treatment group for each study group k—call these $p_k(\text{treatment})$ and $p_k(\text{control})$. That is,

$$p_k(\text{treatment}) = \Phi\big(\hat{\nu}(\text{treatment}, k)\big)$$
$$p_k(\text{control}) = \Phi\big(\hat{\nu}(\text{control}, k)\big),$$

where $\hat{\nu}$ represents the standardised posterior mean estimate for the regression functions which are distributed according to

$$f(x_{ik}, k)|\mathbf{y}, \hat{\theta} \sim \text{N}\left(\hat{\mu}(x_{ik}, k), \hat{\sigma}^2(x_{ij}, k)\right),$$

with $x_{ik} \in \{\text{treatment}, \text{control}\}$ and $k \in \{1, \ldots, 27\}$ (see details in Section 5.5). The log odds ratio for each study group can then be calculated as usual. For the overall log odds ratio, the probabilities that are used are the averaged probabilities weighted according to the sample sizes in each group. This has been calculated as $0.51 \approx \log 1.66$, slightly lower than both the raw log odds ratio and the log odds ratio estimated by the logistic random-effects model. This can perhaps be attributed to some shrinkage of the estimated probabilities due to placing a prior with zero mean on the regression functions.

The credibility intervals for the log odds ratios in the forest plot of Figure 3 are also noticeably narrower under an I-prior compared to the fitted multilevel model. One explanation is that empirical Bayes estimates, such as the I-probit estimates under a variational EM framework, tend to underestimate the variability in the estimates because the variability in the parameters are ignored when point estimates are used, compared to distributions in a true Bayesian estimation framework.

0.0.3 Multiclass classification: Vowel recognition data set

We illustrate multiclass classification using I-priors on a speech recognition data set[2] with $m = 11$ classes to be predicted from digitized low pass filtered signals generated from voice recordings. Each class corresponds to a vowel sound made when pronouncing a specific word. The words that make up the vowel sounds are shown in Table 3. Each word was uttered once by multiple speakers, and the data are split into a training and a test set. Four males and four female speakers contributed to the training set, while four male and three female speakers contributed to the test set. The recordings were manipulated using speech processing techniques, such that each speaker yielded six frames of speech from the eleven vowels, each with a corresponding 10-dimensional numerical input vector (the predictors). This means that the size of the training set is $8 \times 6 \times 11 = 528$, while

$7 \times 6 \times 11 = 462$ data points are available for testing the predictive performance of the models. This data set is also known as Deterding's vowel recognition data (after the original collector, Deterding, 1990). Machine learning methods such as neural networks and nearest neighbour methods were analysed by Robinson (1989).

Table 3: The eleven words that make up the classes of vowels.

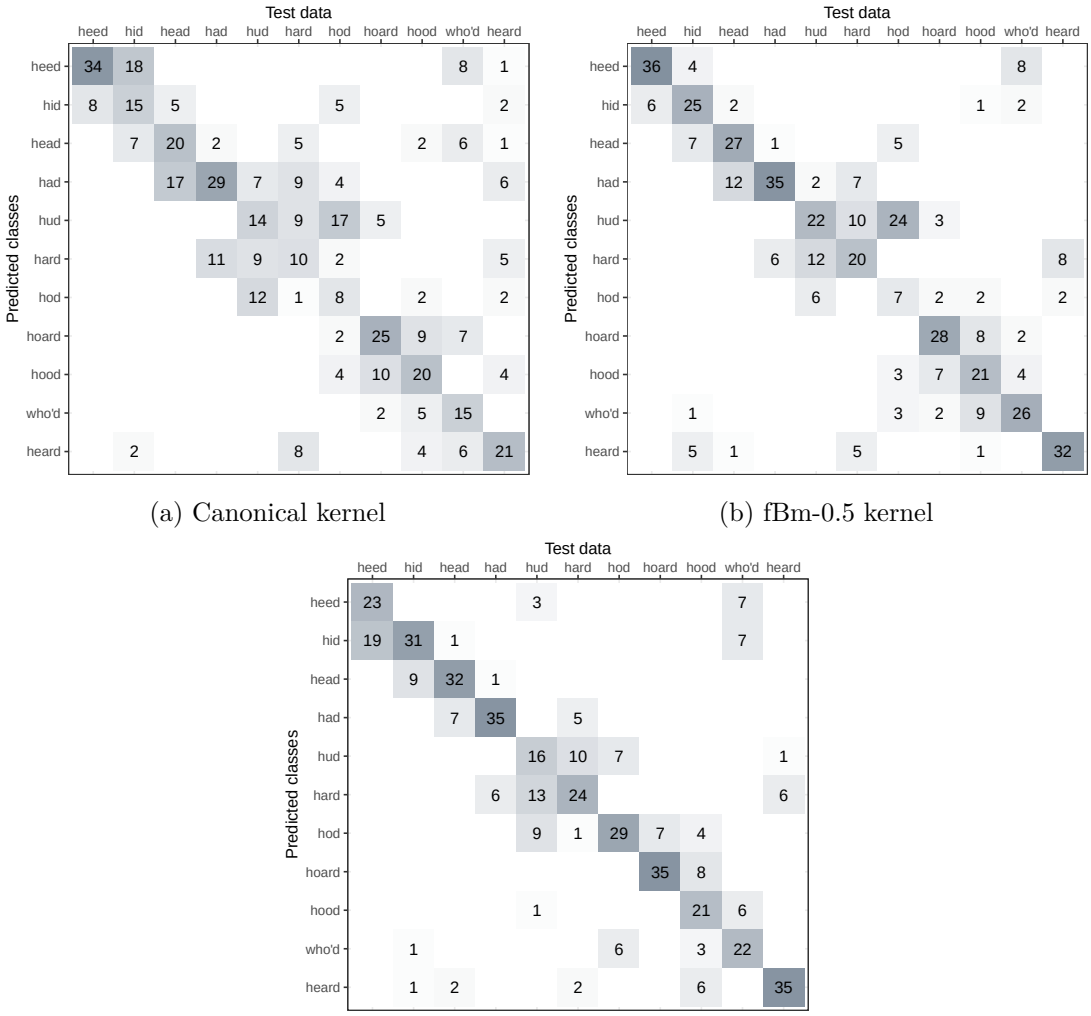
Class	Label	Vowel	Word	Class	Label	Vowel	Word
1	hid	iː	heed	7	hOd	ɒ	hod
2	hId	ɪ	hid	8	hod	ɔː	hoard
3	hEd	ɛ	head	9	hUd	ʊ	hood
4	hAd	a	had	10	hud	uː	who'd
5	hYd	ʌ	hud	11	hed	ɚː	heard
6	had	ɑː	hard				

We will fit the data using an I-probit model with the canonical linear kernel, fBm-0.5 kernel, and the SE kernel with lengthscale $l = 1$. Each model took roughly 13 seconds per iteration in fitting the training data set ($n = 528$). In particular, the canonical kernel model took a long time to converge, with each variational inference iteration improving the lower bound only slighly each time. In contrast, both the fBm-0.5 and SE model were quicker to converge. Multiple restarts from different random seeds were conducted, and we found that they all converged to a similar lower bound value. This alleviates any concerns that the model might have converged to different multiple local optima.

A good way to visualise the performance of model predictions is through a confusion matrix, as shown in Figure 4. The numbers in each row indicate the instances of a predicted class, while the numbers in the column indicate instances of the actual classes, while nil values are indicated by blank cells.

Comparisons to other methods that had been used to analyse this data set is given in Table 4. In particular, the I-probit model is compared against 1) linear regression; 2) logistic linear regression; 3) linear and quadratic discriminant analysis; 4) decision trees; 5) neural networks; 6) k-nearest neighbours; and 7) flexible discriminant analysis. All of these methods are described in further detail in Friedman et al. (2001, chs. 4 & 12, table 12.3). The I-probit model using both the fBm-0.5 and SE kernel offers one of the best out-of-sample classification error rates (34.4%) of all the methods compared. The linear I-probit model is seen to be comparable to logistic regression, linear and quadratic discrimant analysis, and also decision trees. It also provides significant improvement over multiple linear regression.

[2]Data is publicaly available from the UCI Machine Learning Repository, URL: https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Vowel+Recognition+-+Deterding+Data).



(a) Canonical kernel

(b) fBm-0.5 kernel

(c) SE kernel

Figure 4: Confusion matrices for the vowel classification problem in which predicted values were obtained from the I-probit models. The maximum value for any cell is 42 (seven speakers delivered six frames of speech per vowel). Blank cells indicate nil values.

Table 4: Results of various classification methods for the vowel data set.

	Error rate (%)	
Model	Train	Test
I-probit		
Linear	29	54
Smooth (fBm-0.5)	22	40
Smooth (SE-1.0)	7	34
Others		
Linear regression	48	67
Logistic regression	22	51
Linear discriminant analysis	32	56
Quadratic discriminant analysis	1	53
Decision trees	5	54
Neural networks		45
k-nearest neighbours		44
FDA/BRUTO	6	44
FDA/MARS	13	39

0.0.4 Spatio-temporal modelling of bovine tuberculosis in Cornwall

Data containing the number of breakdows of bovine tubercolosis (BTB) in Cornwall, the locations of the infected animals, and the year of occurence is analysed. The interest, as motivated by veterinary epidimiology, is to understand whether or not there is spatial segregation of the infection of the herds, and whether there is a time-element to the presence or absence of this spatial segregation. There has been previous work done to analyse this data set. Diggle et al. (2005) developed a non-parametric method to estimate spatial segregation using a multivariate point process. The occurrences are modelled as Poisson point processes, and spatial segregation is said to have occured if the model-estimated type-specific breakdown probabilities at any given location are not significantly different from the sample proportions. The authors estimated the probabilities via kernel regression, and the test statistic of interest had to be estimated via Monte Carlo methods. Other works include Diggle et al. (2013), who used a fully Bayesian approach for spatio-temporal multivariate log-Gaussian Cox processes, which is implemented in the R package **lgcp** (Taylor et al., 2013).

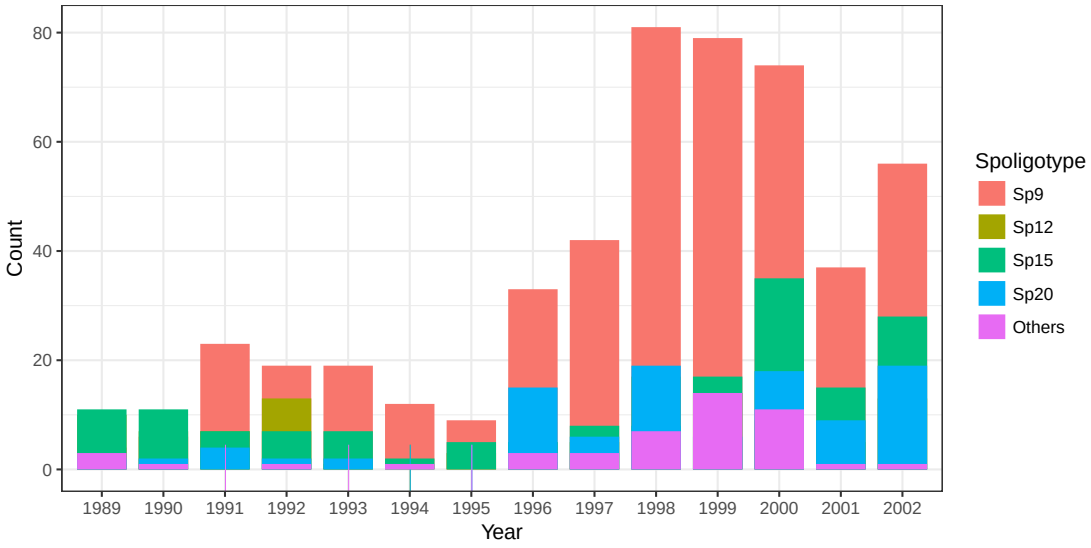


Figure 5: Distribution of the different types (Spoligotypes) of bovine tubercolosis affecting herds in Cornwall over the period 1989 to 2002.

The data set contains $n = 919$ recorded cases over a span of 14 years. For each of the cases, spatial data pertaining to the location of the farm (Northings and Eastings, measured in kilometres) are available. Originally, 11 unique spoligotypes were recorded in the data, with the four most common spoligotypes being Sp9 ($m = 1$), Sp12 ($m = 2$), Sp15 ($m = 3$) and Sp20 ($m = 4$), as shown by the histogram in Figure 5. We had grouped the remaining seven spoligotypes into an 'Others' category ($m = 5$), so that the problem becomes a multinomial regression with five distinct outcomes.

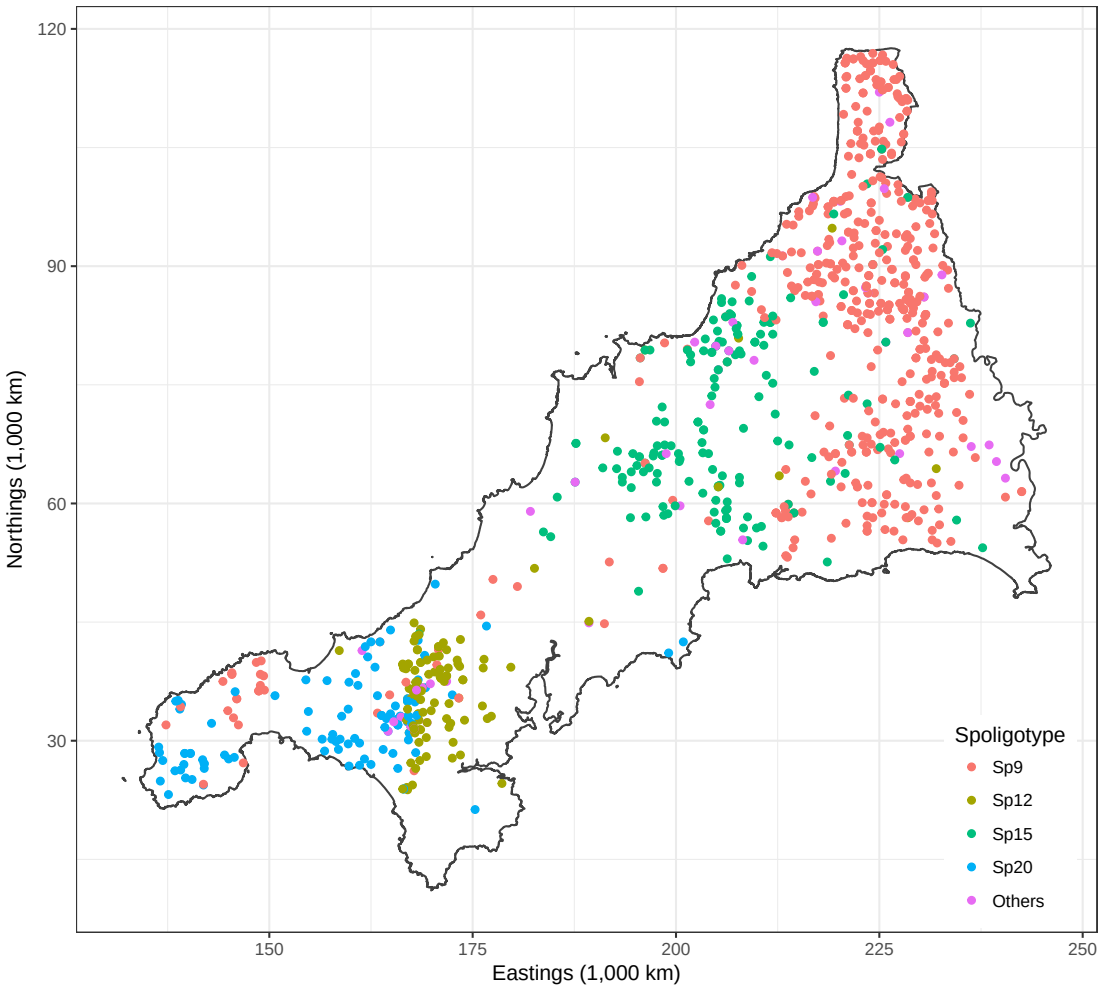


Figure 6: Spatial distribution of all cases over the 14 years.

We are able to investigate any spatio-temporal patterns of infection using I-priors rather simply. Let p_{ij} denote the probability that a particular farm i is infected with a BTB disease with spoligotype $j \in \{1, \dots, 5\}$. We model the transformed probabilities $g_j(p_{ij})$ as following a function which takes two covariates, i.e. the spatial data $x_1 \in \mathbb{R}^2$, and the temporal data x_2 (year of infection):

$$\begin{aligned} p_{ij} &= g_j^{-1}\big(f_k(x_1, x_2)\big)_{k=1}^m \\ &= g_j^{-1}\big(f_{1k}(x_1) + f_{2k}(x_2) + f_{12k}(x_1, x_2)\big)_{k=1}^m, \end{aligned}$$

where the function $g_j^{-1} : \mathbb{R}^m \to [0, 1]$ is the same squashing function used in equation (5.10). We assume a smooth effect of space and time on the probabilities, and appropriate RKHSs for the functions $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$ are the fBm-0.5 RKHS. Alternatively, as per Diggle et al. (2005), divide the data into four distinct time periods: 1) 1996 and earlier; 2) 1997 to 1998; 3) 1999 to 2000; and finally 4) 2001 to 2002. In this case, x_2 would indicate which period the infection took place in, and thus would have a nominal

effect on the probabilities. An appropriate RKHS for f_2 in such a case would be the Pearson RKHS. In either case, the function $f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$ would be the "interaction effect", meaning that with such an effect present, the spatial distribution of the diseases are assumed to vary across the years.

We fitted four different models:

- $\boldsymbol{M_0}$**: Intercept only**.

$$p_{ij} = g_j^{-1}(\alpha_k)_{k=1}^m$$

- $\boldsymbol{M_1}$**: Spatial segregation**.

$$p_{ij} = g_j^{-1}(\alpha_k + f_{1k}(x_i))_{k=1}^m$$

  $f_{1k} \in \mathcal{F}_1$ Pearson RKHS.

- $\boldsymbol{M_2}$**: Spatio-temporal**.

$$p_{ij} = g_j^{-1}(\alpha_k + f_{1k}(x_i) + f_{2k}(t_i) + f_{12k}(x_i, t_i))_{k=1}^m$$

  $f_{1k} \in \mathcal{F}_1$ Pearson RKHS, $f_{2k} \in \mathcal{F}_2$ fBm-0.5 RKHS, and $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

- $\boldsymbol{M_3}$**: Spatio-period**.

$$p_{ij} = g_j^{-1}(\alpha_k + f_{1k}(x_i) + f_{2k}(t_i) + f_{12k}(x_i, t_i))_{k=1}^m$$

  $f_{1k} \in \mathcal{F}_1$ Pearson RKHS, $f_{2k} \in \mathcal{F}_2$ Pearson RKHS, and $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

Model M_0 corresponds to a model which ignores any spatial or temporal effects (the baseline intercept only model). Model M_1 takes into account only spatial effects. Both models M_2 and M_3 account for spatio-temporal effects, but M_2 assumes a smooth effect of time, while M_3 segregates the points into four distinct time periods for analysis. Model comparison is easily done, and Table 5 indicates that model M_2 has the highest ELBO of the four models, making it the preferable model.

For a more visual approach, we can look at the plots of the surface probabilities. To obtain these probabilities, we first determined the spatial points (Northings and Eastings) which fall inside the polygon which makes up Cornwall. We then obtained predicted probabilities for each class of disease at each location. Figure 7 was obtained using the model with spatial covariates only, thus ignoring any temporal effects. In the case of the spatio-temporal model, we used the model which had the period formulation for time (model M_3). This way, we can display the surface probabilities of the time periods in four plots only, which is more economical to exhibit within the margins of this thesis. Note that there is no issue with using the continuous time model—we have

Table 5: Results of the fitted I-probit models. Estimates of the class intercepts and scale parameters, together with their respective bootstrap standard errors, are presented. For model comparison, we can look at ELBOs, error misclassification rates, and Brier scores.

	M_0: Intercepts only		M_1: Spatial only		M_2: Spatio-temporal		M_3: Spatio-period	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept (Sp9)	0.948	0.000	1.364	0.015	1.401	0.079	1.395	0.103
Intercept (Sp12)	-0.173	0.000	-0.435	0.013	-0.506	0.017	-0.463	0.045
Intercept (Sp15)	0.103	0.000	-0.020	0.011	-0.008	0.059	-0.010	0.094
Intercept (Sp20)	-0.202	0.000	-0.775	0.051	-0.795	0.223	-0.783	0.343
Intercept (Others)	-0.676	0.000	-0.134	0.016	-0.091	0.077	-0.139	0.104
Scale (spatial)			0.194	0.008	-0.176	0.178	0.172	0.169
Scale (temporal)					-0.006	0.003	-0.004	0.006
ELBO	-1187.47		-564.33		-537.23		-543.94	
Error rate (%)	46.25		19.26		18.06		18.50	
Brier score	0.249		0.143		0.136		0.138	

produced an animated gif image at http://phd.haziqj.ml/examples/, showing the yearly evolution of the surface probabilities between 1989 and 2002.

As the plots suggests, there is indeed spatial segregation for the four most common spoligotypes, and this is also very prominently seen from Figure 7. In comparing the distribution of the spoligotypes over the years, we may refer to Figure 8, a series of predicted probability surface plots over the four time periods obtained from model M_3. For each time period, we also superimposed the actual observations onto the predicted surface probabilities. In addition, coloured dotted lines are displayed to indicate the "decision boundaries" for each of the four spoligotypes. The most evident change is seen to the spatial distribution of spoligotype 12, with the decision boundary giving it a large area in years 1996 and earlier, but this steadily shrunk over the years. Spoligotype 9, which is most commonly seen in the east of Cornwall, seems to have made its way down to the south-west over the years. The other two spoligotypes seem to be rather constant over the years.

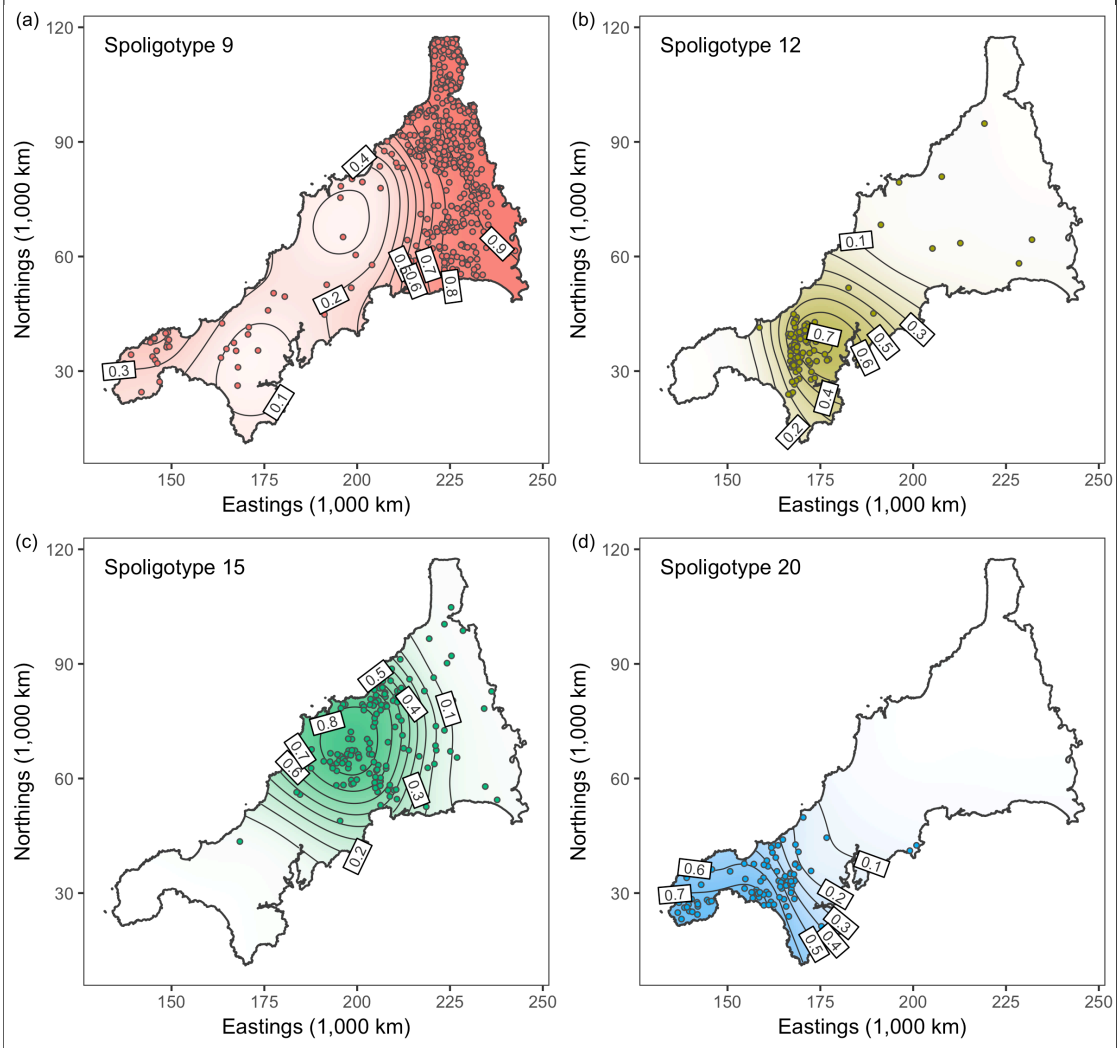


Figure 7: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium *Mycobacterium bovis* over the entire time period using model M_1.

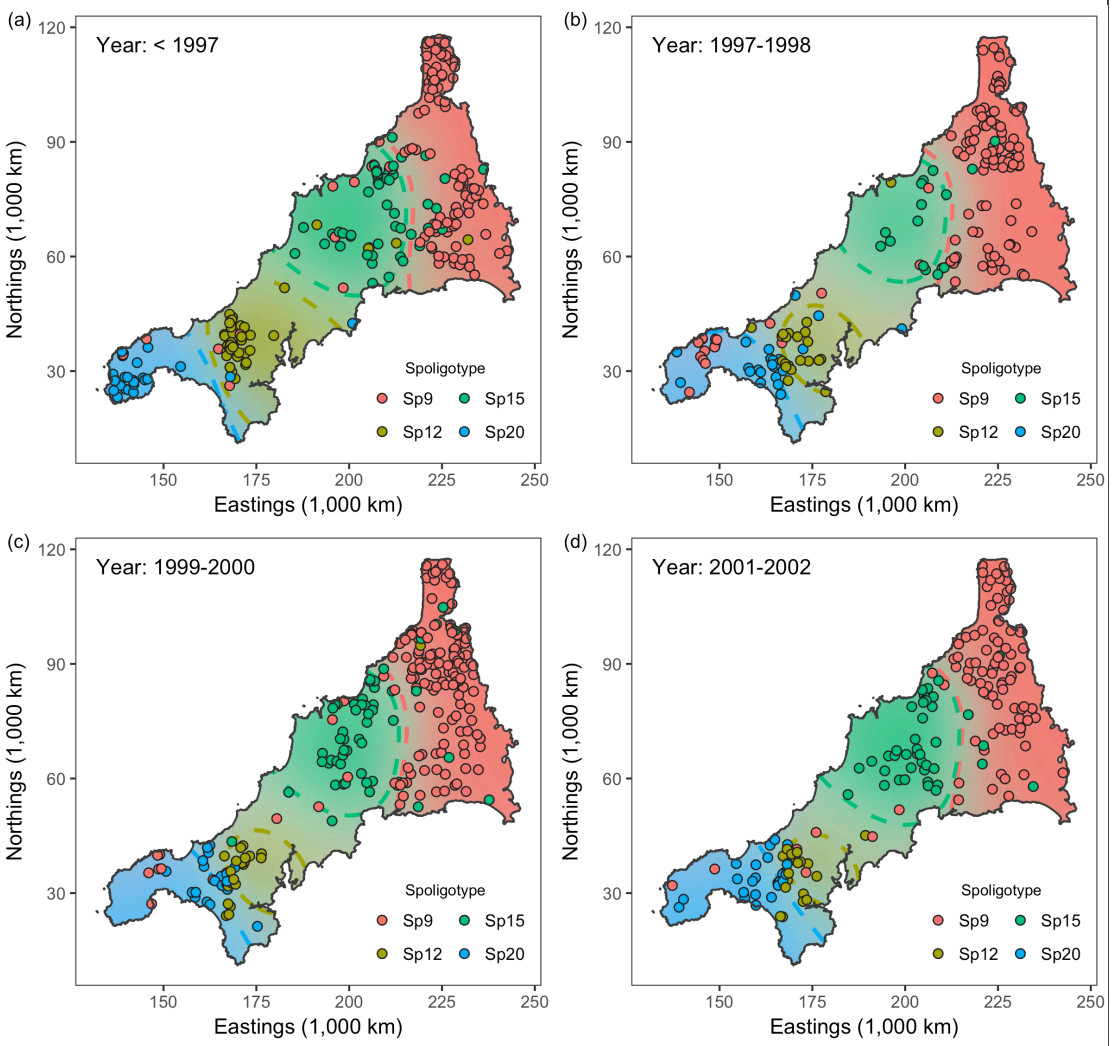


Figure 8: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium *Mycobacterium bovis* over four different time periods using model M_3.