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 the R package **iprobit** developed by us. All of these examples had assumed a fixed error precision $\mathbf{\Psi} = \mathbf{I}_m$.

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).

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.

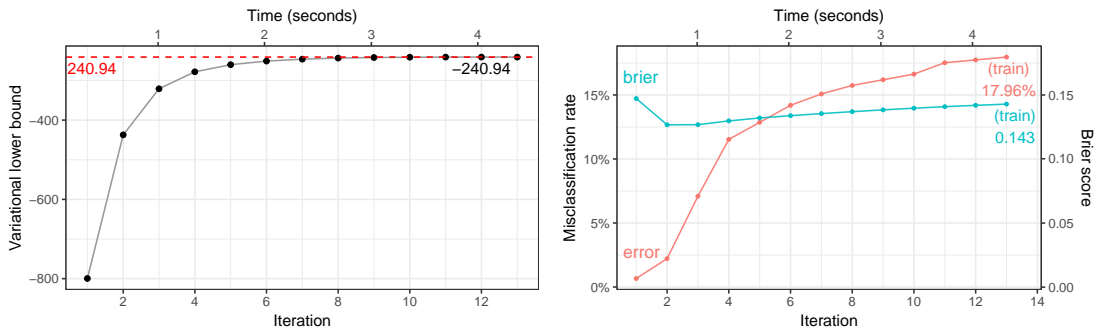[1]Data is made publicly available at https://archive.ics.uci.edu/ml/datasets/arrhythmia.

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

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 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.

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. 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.

2

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.

tab:cardiac

Method	Misclassification rate (%)		
	$s = 50$	$s = 100$	$s = 200$
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)

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

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 depending on the number of random decisions trees generated simultaneously, might be slow. The time complexity of a random forest algorithm is

$O(pqn \log(n))$, 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 to quit 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 5908 patients, and they are distributed roughly equally among the control and treatment groups (46.33% and 53.67% 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

$$\text{odds ratio} = \frac{\text{odds of quitting smoking in } treatment \text{ group}}{\text{odds of quitting smoking in } control \text{ group}}.$$
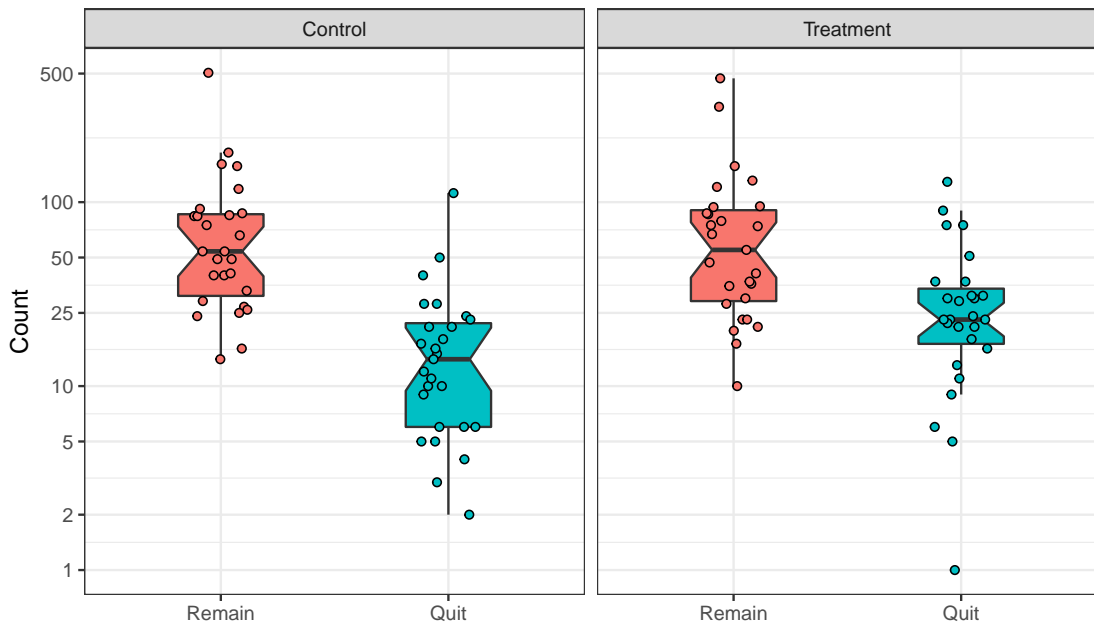
Figure 2: Comparative box-plots of the distribution of patients who successfully quit smoking and those who remained smokers, in the two treatment groups.

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 the 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 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 $i = 1, \ldots, n_j$ index the patients in study group $j \in \{1, \ldots, 27\}$. For patient i in study j, p_{ij} denotes the probability that the patient has successfully quit smoking. Additionally, x_{ij} is the centred dummy variable indicating patient i's treatment group in study j. 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 that $\sigma_{01} = 0$ so that, coupled with the contrast coding used for x_{ij}, the total variance $\mathrm{Var}(\beta_{0j} + \beta_{1j} x_{ij})$ would be constant in both treatment groups. The overall log odds ratio is represented by β_1, and this is estimated as $0.57 = \log 1.76$.

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

$$\Phi^{-1}(p_{ij}) = f(x_{ij}, j)$$
$$= f_1(x_{ij}) + f_2(j) + f_{12}(x_{ij}, j).$$

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_{ij} and j 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_{ij} 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 I-prior modelling, 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-prior model fit for three models.

Model	ELBO	Brier score	No. of parameters
f_1	-3210.76	0.179	1
$f_1 + f_2$	-3142.24	0.206	2
$f_1 + f_2 + f_{12}$	-3091.20	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 ($\mathrm{BF}(M_3, M_1)$ and $\mathrm{BF}(M_3, M_2)$ are bot 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.

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 j - call these $p_j(\text{treatment})$ and $p_j(\text{control})$. That is,

$$p_j(\text{treatment}) = \Phi\big(\tilde{\mu}(\text{treatment}, j)\big)$$
$$p_j(\text{control}) = \Phi\big(\tilde{\mu}(\text{control}, j)\big),$$

where $\tilde{\mu}$ represents the posterior mean estimate for the regression function given in Which section?. These 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 $0.55 = \log 1.73$, 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 in **??** for the log odds ratios under an I-prior are also noticeably narrower compared to the multilevel model fitted.
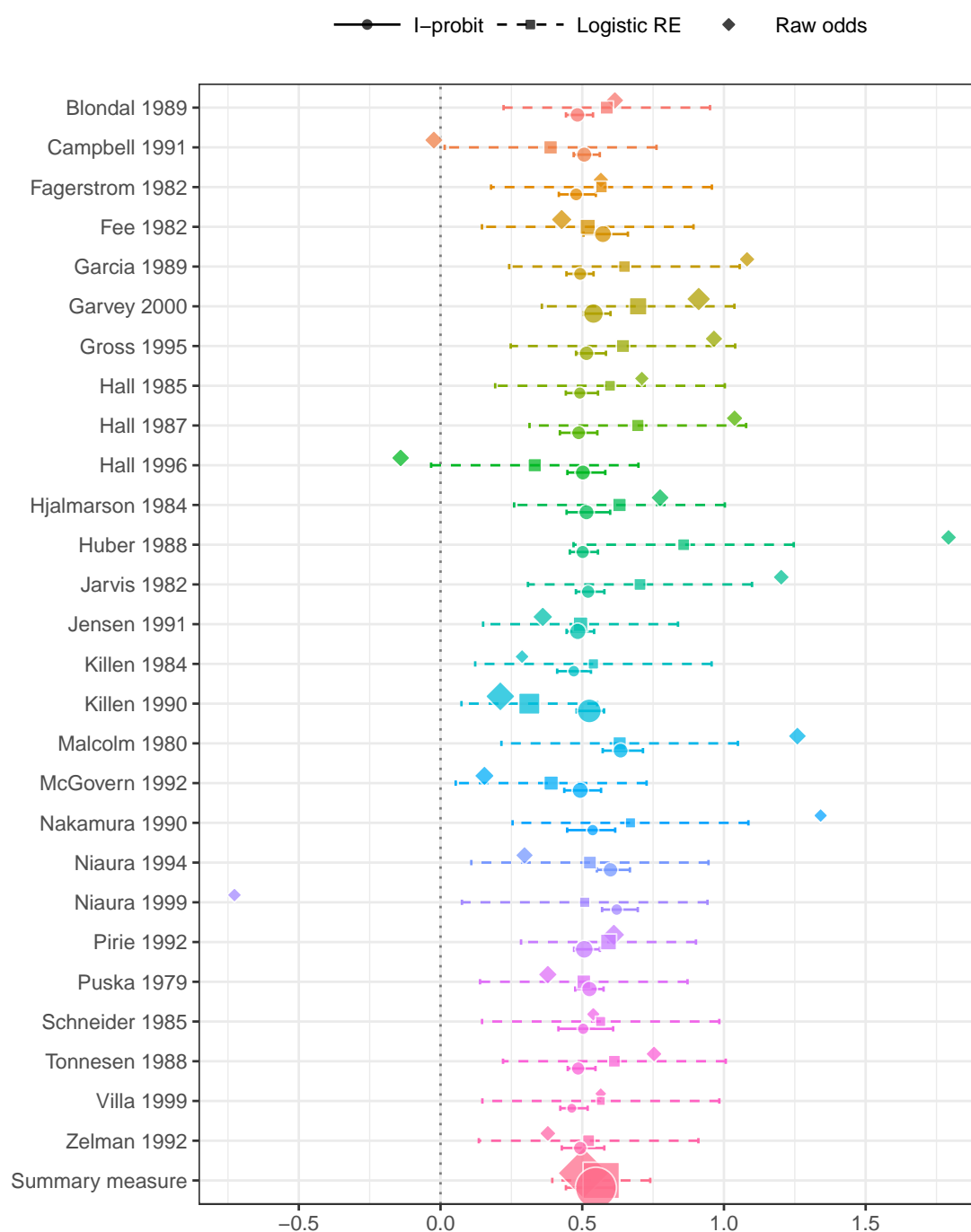
Figure 3: Forest plot of effect sizes (log odds ratios) in each group as well as the overall effect size together with their 95% confidence bands. The plot compares the raw log odds ratios, the logistic random-effect estimates, and the I-prior estimates. Sizes of the points indicate the relative sample sizes per study group.

fig:smoke.f orest.plot

tab:vowel

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	əː	heard
6	had	ɑː	hard				

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 528, while 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, 1989). Machine learning methods such as neural networks and nearest neighbour methods were analysed by Robinson (1989).

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 slightly 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.

---

[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

Test data

Predicted classes	heed	hid	head	had	hud	hard	hod	hoard	hood	who'd	heard
heed	34	18								8	1
hid	8	15	5				5				2
head		7	20	2		5		2	6	1	
had			17	29	7	9	4				6
hud				14	9	17	5				
hard				11	9	10	2				5
hod				12	1	8		2			2
hoard						2	25	9	7		
hood						4	10	20			4
who'd							2	5	15		
heard		2				8		4	6	21	

(b) fBm-0.5 kernel

Test data

Predicted classes	heed	hid	head	had	hud	hard	hod	hoard	hood	who'd	heard
heed	36	4								8	
hid	6	25	2						1		2
head		7	27	1			5				
had			12	35	2	7					
hud					22	10	24	3			
hard					6	12	20				8
hod					6		7	2	2		2
hoard								28	8	2	
hood							3	7	21	4	
who'd	1						3	2	9	26	
heard	5	1				5				1	32

(c) SE kernel

Test data

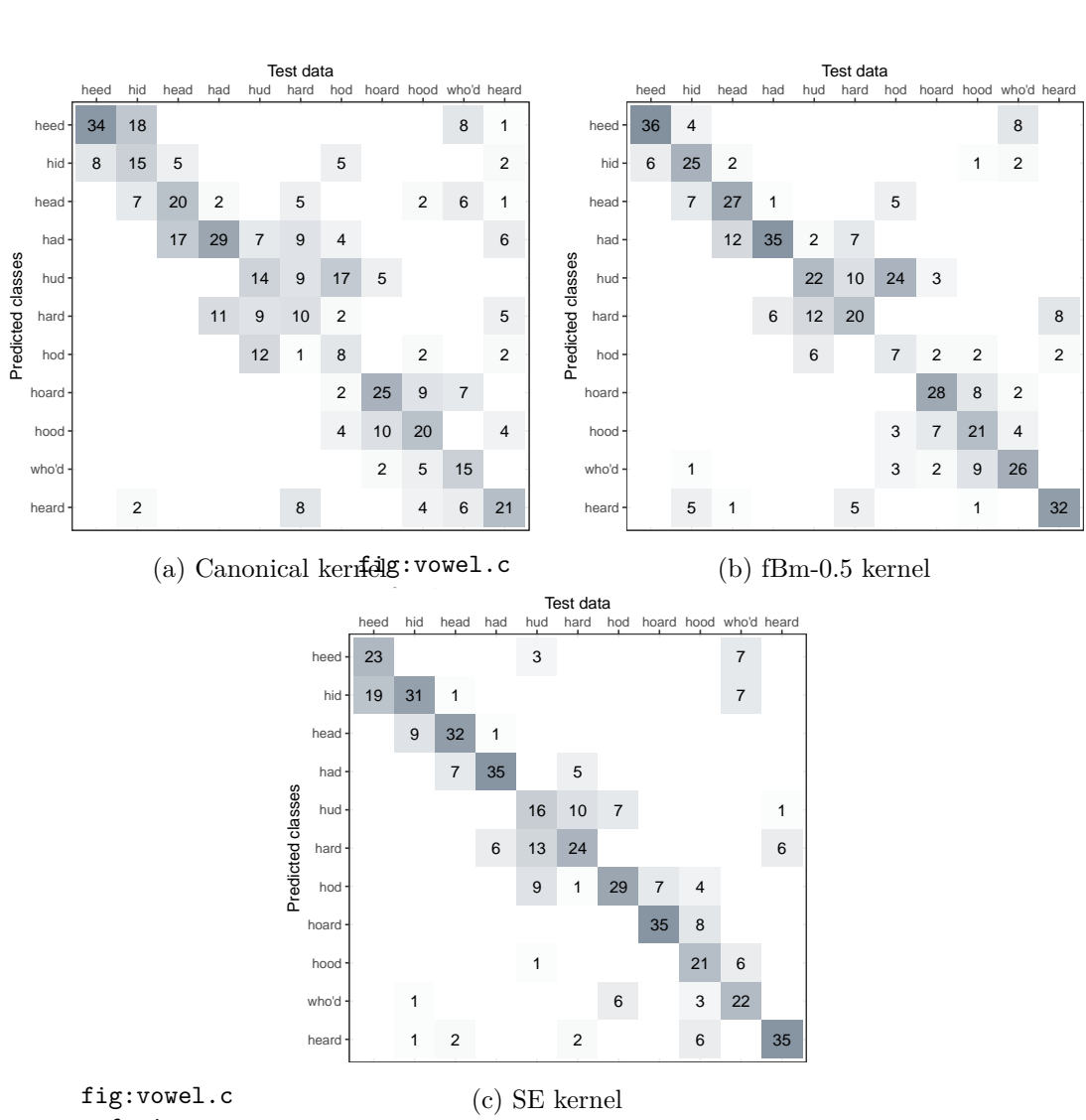| Predicted classes	heed	hid	head	had	hud	hard	hod	hoard	hood	who'd	heard
heed	23				3					7	
hid	19	31	1							7	
head		9	32	1							
had			7	35		5					
hud					16	10	7				1
hard				6	13	24					6
hod					9	1	29	7	4		
hoard								35	8		
hood					1				21	6	
who'd		1					6		3	22	
heard		1	2			2			6		35

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 one cell is 42. Blank cells indicate nil values.

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)

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

tab:vowel.tab

| | 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

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, Ch.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.

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 between the herds, and whether there is a time-element to presence or absence of this spatial segregation. There have been previous work done to analyse this data set: P. 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 <mark>I think this is what they did - recheck</mark>. The authors estimated the probabilities via kernel regression, and the resulting test statistic had to be estimated via Monte Carlo methods. Other work includes P. J. Diggle et al. (2013), who used a fully Bayes scheme for spatio-temporal multivariate log-Gaussian Cox processes, and implemented in the R package ??? (Taylor et al., 2013).
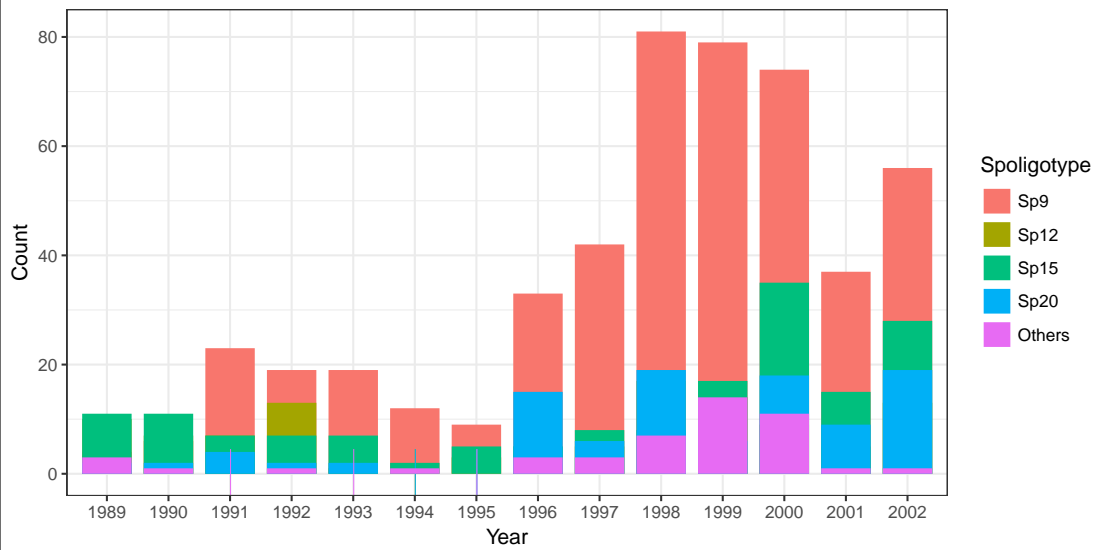

Figure 5: Distribution of the different types (Spoligotypes) of bovine tubercolosis affecting herds in Cornwall over the period 1989 to 2002.

fig:plot.cow

<mark>Explain data set. $n = 919$ cases in total. Originally there are 11 spoligotypes, but of these, four are most common. Therefore, the rest are combined into a separate class of 'Others'. Total 14 years of data, so total number of classes is $m = 5$.</mark>

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 animal i is infected with the disease with spoligotype $j \in \{1, \ldots, m\}$. We model the transformed probabilities $g(p_{ij})$ (as described in the categorical response chapter) as following a smooth function f which takes two covariates: the spatial data x_1 (Northings and Eastings, measured in
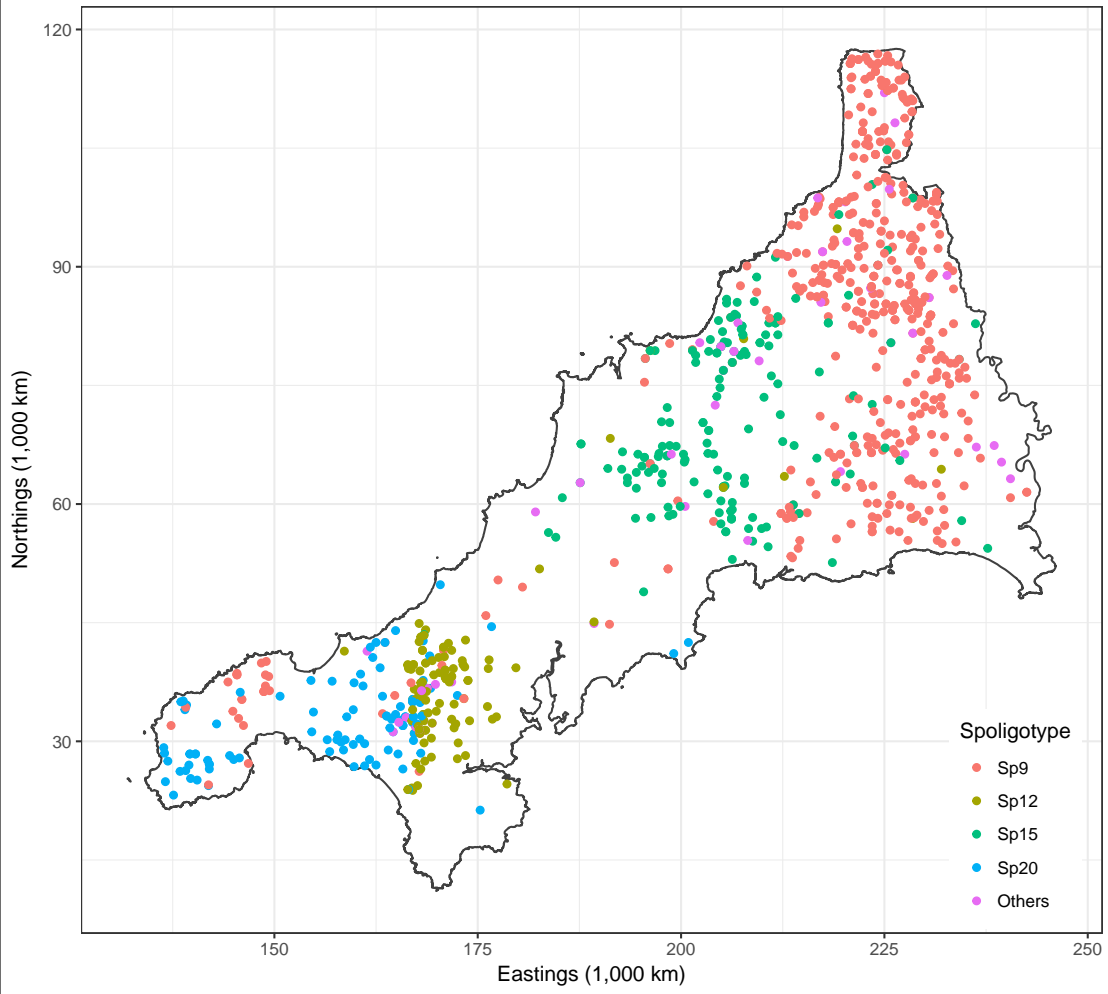
Figure 6: Spatial distribution of all cases over the 14 years.

fig:plot.co
rnwall

kilometres), and the temporal data x_2 (year of infection):

$$g(p_{ij}) = f_j(x_1, x_2)$$
$$= f_{1j}(x_1) + f_{2j}(x_2) + f_{12j}(x_1, x_2)$$

We assume a smooth effect of space and time on the probabilities, and an appropriate RKHS for the functions f_1 and f_2 are the fBm-0.5 RKHS. Alternatively, as per Diggle et al., 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

13

appropriate RKHS for f_2 in such a case would be the Pearson RKHS. In either case, the function f_{12} 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.

Let h_k, $k \in \{1, 2\}$ denote the reproducing kernel of the spatial and temporal RKHSs respectively. Then, an I-prior on f_j takes the form

$$f_j(x_1, x_2) = \lambda_{1j} \sum_{i=1}^{n} h_1(x_1, x_{i1}) w_{ij} + \lambda_{2j} \sum_{i=1}^{n} h_2(x_2, x_{i2}) w_{ij} + \lambda_{1j}\lambda_{2j} \sum_{i=1}^{n} h_1(x_1, x_{i1}) h_2(x_2, x_{i2}) w_{ij}$$

where $\mathbf{w}_j = (w_{1j}, \ldots, w_{nj})^\top \sim \mathrm{N}(0, \mathbf{I}_n)$ and each of the $\mathbf{w}_j$ are also independent of each other. The parameters λ_{1j} and λ_{2j} are the RKHS scale parameters for the spatial and temporal covariates respectively. Notice that the functions are indexed by the classes j, such that there would be $2m$ scale parameters to estimate. This is the more general case, in which we assume *separate scale* parameters in each class. However, we may also restrict the scale parameters to be equivalent in each class, so that this so-called *shared scale* model has only two parameters to estimate, which is simpler to do inference. Note that there are also intercept parameters to estimate (one in each class), but these will not be reported as they are irrelevant to the discussion at hand.

Spatio-temporal effects of the BTB breakdowns can be easily inferred through the RKHS scale parameters. The hypothesis of temporal significance is the same as testing the significance of the λ_2 parameter, while the test of both spatial and temporal effects are conducted on λ_1 and λ_2 simultaneously (equivalent to modelling f with a constant). For these tests, it is simpler to infer from the shared scale model, for which we can read the results directly of off Table 5. The said table displays the posterior mean estimate of the scale parameters, and together with its posterior standard deviation. From Chapter X, we know that these scale parameters follow a normal posterior distribution, so we can calculate the Z-scores by dividing the mean by its corresponding s.d.. Absolute values greater than three would satisfy a Bayesian hypothesis test of significance at the 0.01 level, for which we see all parameters satisfy in the shared scale model.

Table 5: Results of the fitted I-probit models.

	Intercepts only		Spatial only		Spatio-temporal		Spatio-period			
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Intercept (Sp9)	0.948	0.033	1.364	0.033		0.033	1.395	0.033	1.401	0.033
Intercept (Sp12)	-0.173	0.033	-0.435	0.033		0.033	-0.463	0.033	-0.506	0.033
Intercept (Sp15)	0.103	0.033	-0.020	0.033		0.033	-0.010	0.033	-0.008	0.033
Intercept (Sp20)	-0.202	0.033	-0.775	0.033		0.033	-0.783	0.033	-0.795	0.033
Intercept (Others)	-0.676	0.033	-0.134	0.033		0.033	-0.139	0.033	-0.091	0.033
Scale (spatial)			0.194	0.003			0.172	0.003	-0.176	0.003
Scale (temporal)							-0.004	0.000	-0.006	0.000
Log-likelihood	-1187.47		-564.33				-543.94		-537.23	
Error rate (Brier score)	0.25		0.14				0.14		0.14	

tab:table.b
tb

A similar conclusion is reached when inferring from the separate scale model. Instead of individual tests of significance, we now need to test

$$H_0 : \lambda_1 = \cdots = \lambda_m = 0.$$

We know that by the mean-field approximation used, the λ_js are independent of each other, and therefore a χ^2 test statistic can be built via

$$\chi^2 = \sum_{j=1}^{m} Z_j^2$$

which is then compared against extreme values of the χ_m^2-distribution. As is often the case, separate scale models tend to fit the data better as it gives more generality due to having different scale parameters in each class. This is also the case for the BTB data, where we see from the footnotes of Table 5 that the Brier scores for the separate scale models are better than the Brier scores in the shared scale models. For all following plots, we made use of the separate scale model for predicting the surface probabilities. Another comment regarding the models is that the conclusion remains the same if we had used the periodic formulation for x_2.

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. This way, we can obtain the surface probabilities in only four time periods, although there is no issue with using the continuous time model. It is more economical to display four plots rather than the 14 yearly plots within the margins of this thesis.

As the model 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. For each time period, we superimpose 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. This is supported also by the spatio-period model results in Table 5, where the test of nullity for the scale parameters of these two spoligotypes are not rejected.

```
## Error in `[.data.frame`(contour.df, , 2 + j): undefined columns selected
```

```
## Error in b[[year]]: subscript out of bounds
```

```
## Error in b[[year]]: subscript out of bounds
```

```
## Error in b[[year]]: subscript out of bounds
```
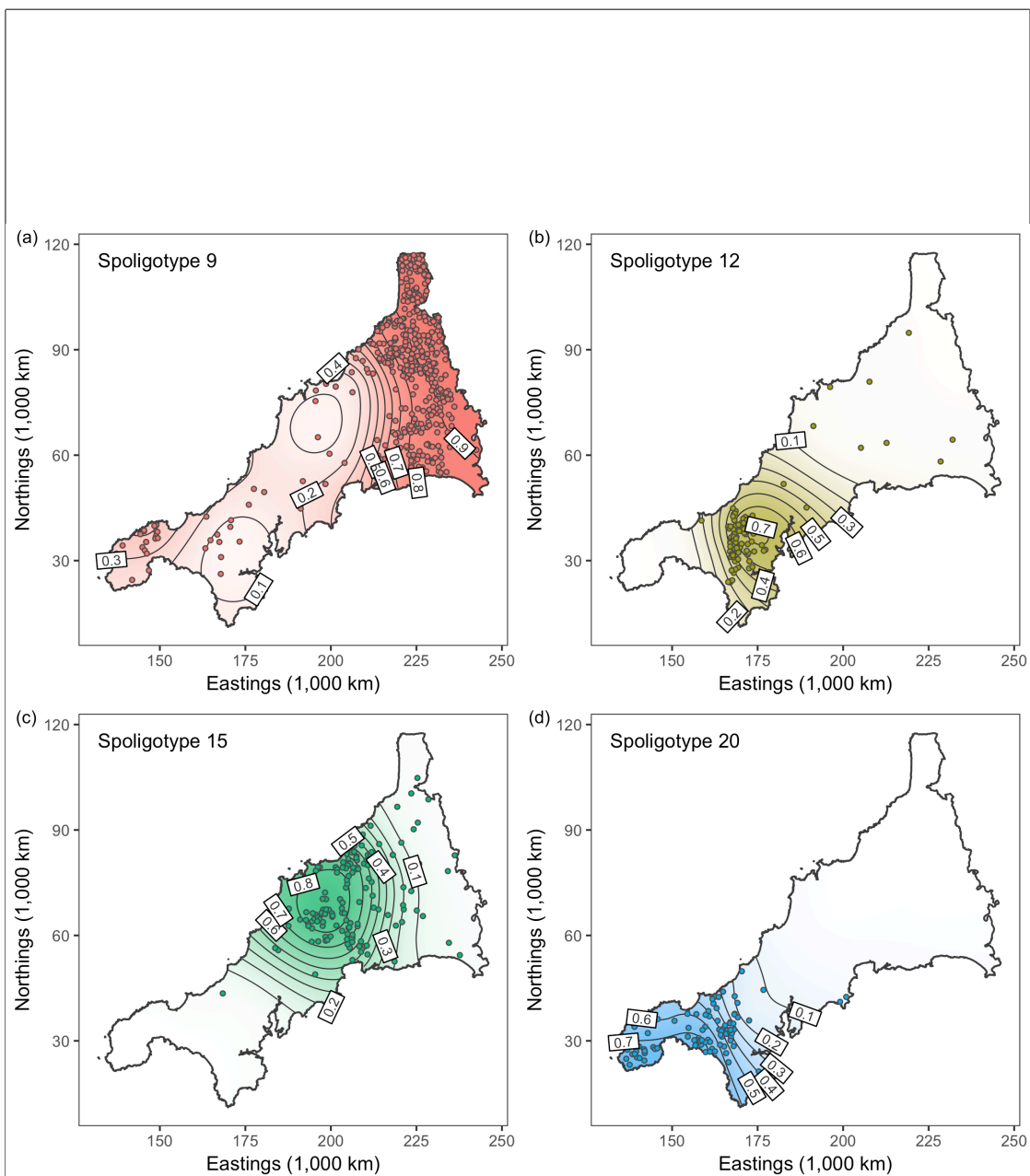
```
## Error:  `quo` must be a quosure
```
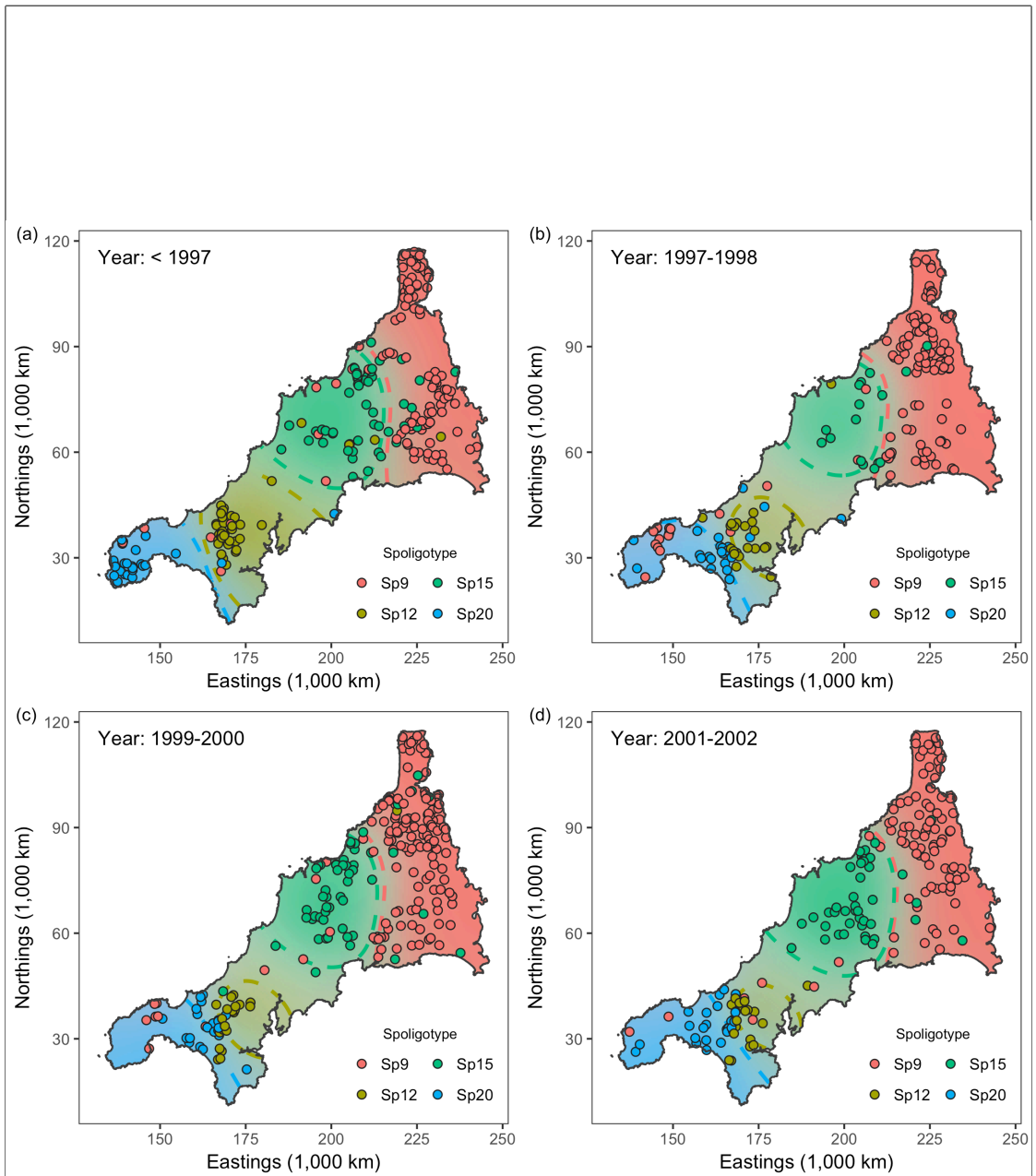
Figure 7: Caption.

fig:plot.btb

fig:plot.te
mporal.btb

Figure 8: Caption.