Bayesian methods for analysing pesticide contamination with uncertain covariates

Al-Alwan, Ali A.

Bayesian Methods for Analysing Pesticide Contamination with Uncertain Covariates

Ali A. Al-Alwan

A Thesis presented for the degree of

Doctor of Philosophy

Statistics and Probability Group

Department of Mathematical Sciences

University of Durham

England

July 2008

Dedicated to

My parents

My wife

My children Murtada, Fatema, Muhammed, Maryam and Ohood

Bayesian Methods for Analysing Pesticide Contamination with Uncertain Covariates

Ali A. Al-Alwan

Submitted for the degree of Doctor of Philosophy

July 2008

Abstract

Two chemical properties of pesticides are thought to control their environmental fate. These are the adsorption coefficient k_{oc} and soil half-life $t^{soil}_{1/2}$. This study aims to demonstrate the use of Bayesian methods in exploring whether or not it is possible to discriminate between pesticides that leach from those that do not leach on the basis of their chemical properties, when the monitored values of these properties are uncertain, in the sense that there are a range of values reported for both k_{oc} and $t^{soil}_{1/2}$. The study was limited to 43 pesticides extracted from the UK Environment Agency (EA) where complete information was available regarding these pesticides. In addition, analysis of data from a separate study, known as "Gustafson's data", with a single value reported for k_{oc} and $t^{soil}_{1/2}$ was used as prior information for the EA data.

Bayesian methods to analyse the EA data are proposed in this thesis. These methods use logistic regression with random covariates and prior information derives from (i) available United States Department of Agriculture (USDA) data base values of k_{oc} and $t^{soil}_{1/2}$ for the covariates and (ii) Gustafson's data for the regression parameters. They are analysed by means of Markov Chain Monte Carlo (MCMC) simulation techniques via the freely available WinBUGS software and R package. These methods have succeeded in providing a complete or a good separation between leaching and non-leaching pesticides.

Declaration

The work in this thesis is based on research carried out at the University of Durham, the Department of Mathematical Sciences, the Statistics and Probability Group, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

First and foremost, I would like to express my thanks to my parents for their continual support and encouragement.

I would like to express my deep thanks, appreciation and gratitude to Dr. Allan Seheult, my supervisor, for his continual support and persistent encouragement during the period of my study over the past four and a half years.

I would like to thank King Faisal University, Saudi Arabia, for sponsoring my PhD.

I would like to thank the University of Durham, in particular, the Department of Mathematical Sciences for giving me the opportunity to complete my higher education.

Finally, there are many other people to thank, especially my family, my brother Saleh, and my friends in Saudi Arabia for their support and encouragement.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter describes the aims of the thesis and its objectives, data sources, research methodology, the relevant literature, and includes an outline of the thesis. The chapter is structured as follows. Section 1.1 describes the general aims of the thesis and outlines its objectives. Section 1.2 describes the data, its sources and obstacles. Section 1.3 provides a general description of the methodology used and its implementation. Section 1.4 documents the related literature. Section 1.5 outlines the structure of the thesis. Section 1.6 describes briefly which sections of this thesis are original and which are from literature.

1.1 The aims and objectives of the thesis

This study aims to demonstrate the use of Bayesian methods and modern statistical techniques for analysing contamination of groundwater as a consequence of using pesticides. More specifically, the aim is to explore whether or not it is possible to achieve pesticide discrimination on the basis of their available chemical properties.

This will benefit the national authorities when implementing the registration and regulation of uses of pesticides in order to maintain the quality of groundwater. Being able to predict that a manufactured pesticide will leach into the soil and contaminate the groundwater will help to end the use of this pesticide and protect the groundwater from contamination. In fact, such pesticides not only contaminate the groundwater but also threaten human health, contributing to the causes of several diseases, such as cancer and infertility; see [1].

1.2 Data description

The data used in this study comes from three sources. The main data set was collected by the UK Environment Agency and will be referred to throughout this thesis as the EA data. The second is from a United States Department of Agriculture (USDA) database consisting of a number of different values for certain chemical properties for more than 300 pesticides. The third data set, which we will refer to as Gustafson's data, was extracted from the California Department of Food and Agriculture (CDFA) database and analysed by Gustafson in [25]

1.2.1 The EA data

This data was described in [43] on which much of the following review is based. It consists of the levels of 112 different pesticides found in the UK groundwater. It was collected by sampling at a large number of sites across the UK between 1992 and 1995. Pesticides with levels in the groundwater exceeding a threshold of $0.1\mu g l^{-1}$ are considered as contamination pesticides which affect the quality of the groundwater and hence are classified as leachers. The EA data suffers from obstacles which restrict its usefulness in predicting the propensity of a pesticide to pollute. These obstacles as listed in [43] are:

1. The total number of samples taken differs for each pesticide. The compounds Atrazine and Chlorpyrifos are examples of this. In 1992, Atrazine was monitored as being above the threshold in 49 of 543 samples, while Chlorpyrifos was found not to be above the threshold in any of 20 samples.

2. There is no link or relationship of any kind between the levels of pesticides found in the groundwater and other environmental factors such as climate or rainfall patterns.

3. Not all of the pesticides were monitored in each year. For example, Chlorpyrifos was monitored in 1992 and 1993 but not in 1994.

4. Detection equipment may have caused errors in measuring the levels of pesticides in the groundwater.

5. Two reasons for a lack of evidence of a pesticide in a sample may be: (a) the pesticide has not been used in that area, or (b) it has been used but only recently, and leaching may become detectable at a later date.

Table 1.1 shows 43 compounds from the EA database, where complete information is available (as will be explained later), which are classified as leaching or non-leaching pesticides. These 43 are from a total of 112 different compounds which were monitored in several thousand groundwater samples taken in the period 1992-1995 [43]. The table also shows the total number of samples taken in 1993 and the number of samples in which the pesticide monitored as being above the threshold.

1.2.2 USDA chemical properties database

This database, published by the United States Department of Agriculture (USDA), contains information regarding chemical properties of each pesticide and other environmental factors such as soil types and climate patterns which have an effect on the tendency of a pesticide to leach and contaminate groundwater. Amongst the chemical properties, there are two believed to have the most influence on the leaching potential of a pesticide; see [25], [43] and [44]. The first, the adsorption coefficient (k_{oc}), is a measure of pesticide mobility through the soil. A pesticide with a high k_{oc} will be found in low levels in groundwater since it will be adsorbed into the soil as organic matter before it contaminates the groundwater. The second property is pesticide persistence in soil, measured by the estimated half-life of pesticide in the soil ($t_{1/2}^{soil}$), which is the time taken for the level of pesticide retained in the soil to decline by 50%. A high value of $t_{1/2}^{soil}$ tends to increase the leaching potential of pesticide into the groundwater. As stated in [43], measuring the k_{oc} and $t_{1/2}^{soil}$ at the sampling site where the levels of pesticides are determined is prohibitively expensive. As an alternative, published measurements (in particular, k_{oc} and $t_{1/2}^{soil}$) from the United States and Europe were aggregated into a database of chemical properties and are available from the USDA pesticide properties database. This database can be accessed via the Internet site `http:/www.ars.usda.gov`.

The USDA database lists more than 300 pesticides together with several physical and chemical properties. As stated above, the adsorption coefficient k_{oc} and the soil

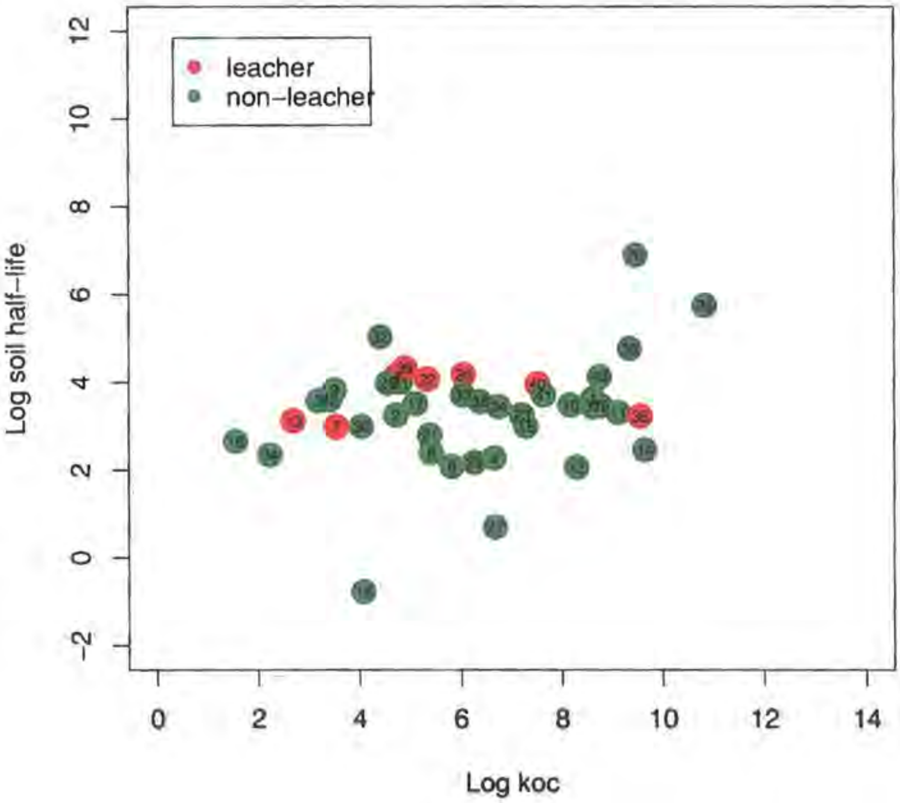

Figure 1.1: Means of the k_{oc} and $t^{soil}_{1/2}$ of the EA data; see Worrall et al. (1998).

half-life $t^{soil}_{1/2}$ are believed to be primarily responsible for the leaching potential of pesticides into groundwater. Therefore, the focus is on the published values of k_{oc} and $t^{soil}_{1/2}$ from this database. The following should be noted when scanning the USDA database.

1. Several chemical and physical properties are reported for each pesticide. Amongst these are the molecular formula; molecular weight; physical state (liquid, gas, solid); boiling, melting and decomposition points; vapour pressure; water solubility S_{H2O} (parts per million); organic solubility (parts per million); Henrys law (Pa m3/mol); Octanol/water partitioning; adsorption coefficient k_{oc}

NO	Pesticide	Leacher	samples.93	detected.93	koc.mean	koc.sd	shl.mean	shl.sd
1	2.4.DCPA	NO	1	0	8.6103	0.2739	3.6113	0.7596
2	2.4.5 T	NO	44	0	4.7194	0.6318	3.2441	0.6191
3	Aldicarb	NO	27	0	3.1848	0.6275	3.5701	0.7163
4	Atrazine	YES	603	66	4.7408	0.499	4.1429	0.6835
5	Azinphos.methyl	NO	233	0	6.6692	0.6117	2.2668	0.4358
6	Bendiocarb	NO	25	0	5.822	0.7406	2.0716	1.376
7	Bentazone	YES	34	5	3.5409	0.0205	2.9747	1.0485
8	Carbaryl	NO	27	0	5.414	0.7658	2.3844	0.6682
9	Carbofuran	NO	27	0	3.5056	0.7777	3.8208	0.5055
10	Chlorothalonil	NO	26	0	8.1611	0.8528	3.4708	0.8933
11	Chlorpyrifos	NO	39	0	9.1117	0.4655	3.3034	1.1482
12	Chlorpyrifos.methyl	NO	25	0	8.2866	0.3176	2.0557	1.2487
13	Clopyralid	YES	30	1	2.6876	1.267	3.1093	0.8988
14	Cyfluthrin	NO	3	0	9.6423	2.017	2.4537	1.1013
15	Diazinon	NO	336	0	7.2956	0.2645	2.9859	1.0461
16	Dicamba	NO	74	0	1.5486	1.1854	2.6458	0.5697
17	Dichlobenil	NO	96	0	5.1078	0.2708	3.5034	1.3279
18	Dichlorvos	NO	111	0	4.0812	0.8641	-0.7818	1.438
19	Endosulfan.a	NO	242	0	8.7621	1.6142	3.4527	1.3828
20	Endrin	NO	292	0	9.4626	0.8719	6.889	2.0893
21	EPTC	NO	25	0	5.3874	0.2169	2.7874	0.8416
22	Ethofumesate	YES	31	1	5.3421	0.7535	4.0599	0.7428
23	Fenthion	NO	225	0	7.2146	0.2829	3.2413	0.7711
24	Fonofos	NO	25	0	6.7618	1.6884	3.4363	0.5807
25	Heptachlor	NO	233	0	10.8235	1.7524	5.7443	1.0783
26	Linuron	YES	172	5	6.0577	0.6131	4.1642	0.7432
27	Malathion	NO	254	0	6.6883	1.2909	0.6931	2.4219
28	Metalaxyl	NO	25	0	4.5633	1.0256	3.9726	0.7931
29	Methiocarb	NO	27	0	6.2691	0.5364	2.1654	1.1018
30	Methomyl	NO	27	0	4.0421	1.1038	2.9894	0.9661
31	Monolinuron	NO	27	0	4.8115	0.8331	3.9985	0.1661
32	Monuron	NO	27	0	4.4153	0.6521	5.0278	0.3967
33	Napropamide	NO	25	0	6.0355	0.5091	3.7024	0.7609
34	Oxamyl	NO	27	0	2.2303	0.6159	2.3412	0.7089
35	Pendimethalin	NO	26	0	9.343	0.6251	4.7603	1.3649
36	Pentachlorophenol	YES	78	3	9.5542	2.3354	3.2151	0.9375
37	Phenmedipham	NO	12	0	8.6011	0.981	3.4369	0.3966
38	Propyzamide	NO	25	0	6.3874	0.7349	3.5496	0.9537
39	Simazine	YES	603	12	4.8989	0.2206	4.3083	0.6646
40	Terbutryn	YES	134	3	7.5237	0.9237	3.9472	1.4639
41	Triallate	NO	25	0	7.6399	0.3175	3.6996	0.96
42	Triclopyr	NO	29	0	3.4162	1.5004	3.5917	0.5807
43	Trifluralin	NO	241	0	8.7403	0.6133	4.1207	0.6786

Table 1.1: The 43 pesticides extracted from the EA database are classified as leachers or non-leacher together with the means, standard deviations of log k_{oc} and log $t_{1/2}^{soil}$ from the USDA database. “samples.93” indicates the total number of samples taken in 1993 and “detected.93” indicates the number of samples in which the pesticide monitored as being above the threshold

(cm^3/gm); hydrolysis half-life $t_{1/2}^{hyd}$ (days) and soil half-life $t_{1/2}^{soil}$ (days).

2. The published values of some physical properties, for example, k_{oc}, $t_{1/2}^{soil}$ and S_{H2O} are uncertain in the sense that for each pesticide there is a range of values published for each of them. These published values vary with soil type and climate, although this information is not always provided. However, the sources of the published values along with references are given; for example, whether the values came from a manufacturer, handbook, experiment or from specific calculations. All of the above explains why there is a range values and that the actual values are unknown. For instance, the pesticide Carbaryl has 20 reported values for k_{oc}, 9 values for $t_{1/2}^{soil}$, 5 values for vapour pressure, 4 values for S_{H2O}, 6 values for water partitioning and 1 value for Henrys law. Also, the soil type is provided for some of these values. For example, some of the k_{oc} values are tested for sand, loamy sand, silt loam and sandy clay loam.

   Table 1.1 shows the means and standard deviations of $\log k_{oc}$ and $\log t_{1/2}^{soil}$ for 43 pesticides published by the USDA database; see 3 below

3. Not all of the 112 different pesticides in the EA database have published values for both of k_{oc} and $t_{1/2}^{soil}$; for example, Chloridazon, has values for k_{oc} but not for $t_{1/2}^{soil}$. These pesticides were omitted from both the analysis in [43] and this study. In fact, only 43 pesticides from the EA database have published values for both k_{oc} and $t_{1/2}^{soil}$. The analysis in this study is restricted to these pesticides; see Table 1.1.

4. The latest USDA database update for two of the pesticides was in May 2001,

the rest having been updated in May 1999. The updated values for both the k_{oc} and $t_{1/2}^{soil}$ are the same as those used in [43].

Figure 1.1 which shows the means of the available values of k_{oc} and $t_{1/2}^{soil}$, in log-scale, for the 43 pesticides classified by the EA as leaching or non-leaching, demonstrates that discrimination based on these average values is poor.

1.2.3 Gustafson's data

This data was published by Gustafson in [25] and discussed in [44] from which most of the following is taken. The data was extracted from the the California Department of Food and Agriculture (CDFA) database, comprising of 44 pesticides, 22 of them which we will refer to as "Gustafson's data" and 7 as "transitional pesticides". Each pesticide from Gustafson's data was classified by CDFA as a leacher or non-leacher and single values for k_{oc}, $t_{1/2}^{soil}$, S_{H2O} and $t_{1/2}^{hyd}$ are given.

The transitional pesticides have a single value for both k_{oc} and $t_{1/2}^{soil}$, but unlike Gustafson's data the leaching potential for these pesticides was either inconclusive or conflicting.

The remaining 15 pesticides have some values reported for the above properties, but not for all, and so will be ignored in this study.

The CDFA classifies the pesticides as leachers and non-leachers by establishing specific numerical values for k_{oc}, $t_{1/2}^{soil}$, S_{H2O}, $t_{1/2}^{hyd}$ and other properties. CDFA has classified pesticides with the following values as contaminants; see [42]:

1. k_{oc} less than 512 cm^3/gm or S_{H2O} greater than 7 parts per million, and
2. $t_{1/2}^{hyd}$ greater than 13 days or $t_{1/2}^{soil}$ greater than 11 days.

NO	Pesticide	Leacher	adsorption.rate(koc)	soil half-life($t^{soil}_{1/2}$)
1	Aldicarb	Yes	2.8332	1.9459
2	Atrazine	Yes	4.6728	4.3041
3	Diuron	Yes	5.9636	5.2364
4	Metolachlor	Yes	4.5951	3.7842
5	Oxamyl	Yes	3.2581	2.0794
6	Picloram	Yes	3.2581	5.3279
7	Prometryn	Yes	6.42	4.5433
8	Simazine	Yes	4.9273	4.0254
9	Chlordane	No	9.8663	3.6109
10	Chlorothalonil	No	7.2298	4.2195
11	Chlorpyrifos	No	8.7136	3.989
12	2,4-D	No	3.9703	1.9459
13	DDT	No	12.2719	10.5506
14	Dicamba	No	6.2364	3.2189
15	Endosulfan	No	7.6207	4.7875
16	Endrin	No	9.3226	7.7142
17	Heptochlor	No	9.4978	4.6913
18	Lindane	No	7.4541	6.3439
19	Phorate	No	7.4146	3.6376
20	Propachlor	No	6.6771	1.3863
21	Toxaphene	No	11.4702	2.1972
22	Trifluralin	No	8.9809	4.4188
23	Alachlor	Transitional	5.081404	2.639057
24	Carbaryl	Transitional	6.047372	2.944439
25	Carbofuran	Transitional	4.007333	3.610918
26	Dieldrin	Transitional	9.400961	6.839476
27	Dinoseb	Transitional	8.682708	3.401197
28	Ethoprop	Transitional	3.258097	4.143135
29	Fonofos	Transitional	8.537976	3.218876

Table 1.2: 29 pesticides extracted from the CDFA are classified as leachers, non-leachers or transitional, together with their adsorption coefficients k_{oc} and soil half-life $t^{soil}_{1/2}$ in days, in log-scale.

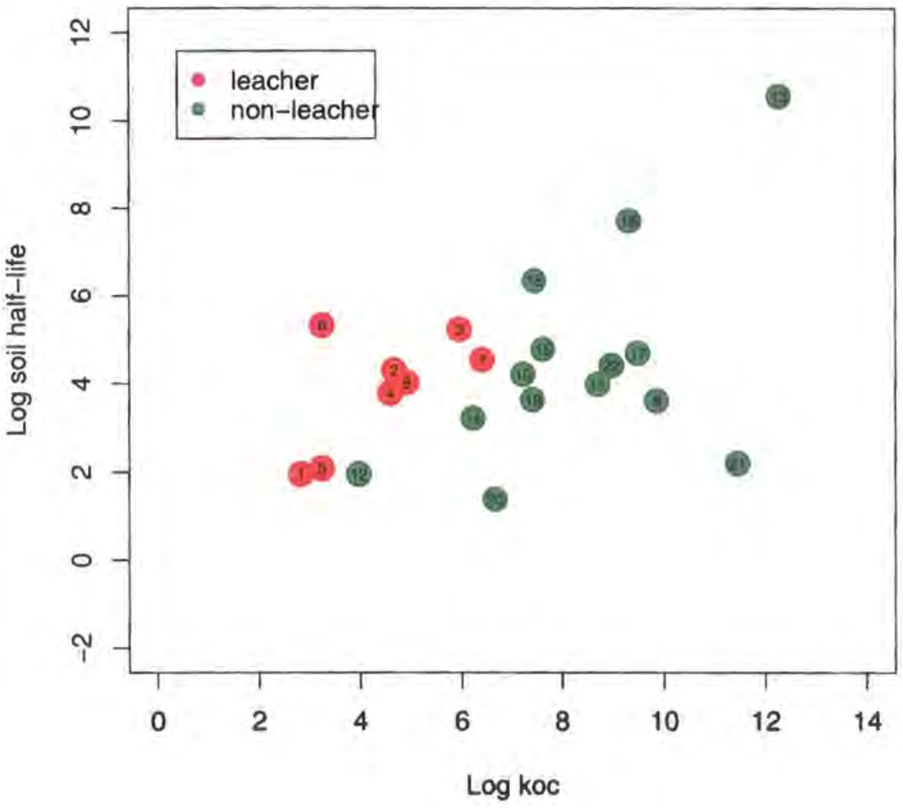

Figure 1.2: The 22 pesticides classified by the CDFA and known as Gustafson's data.

Gustafson's data and the transitional pesticides (a total of 29) are displayed in Table 1.2. Figure 1.2 plots the k_{oc} and $t_{1/2}^{soil}$ pairs for Gustafson's data, 22 pesticides, in a log-scale. It is apparent from this plot that these pesticides are separated into leacher and non-leacher groups according to their k_{oc} and $t_{1/2}^{soil}$ values.

1.2.4 Remarks and assumptions on the data

At this stage of the thesis, it is useful to summarise some remarks and considerations on the various types of the data.

1. The UK Environment Agency (EA) classifies the pesticides as leachers and non-leachers according to whether their levels in the groundwater exceed a threshold of $0.1\mu g l^{-1}$.

2. The California Department of Food and Agriculture (CDFA) classifies pesticides as leachers and non-leachers by establishing specific numerical values for specific physical properties.

3. From (1) and (2), each of EA and CDFA use a different classification basis. This means that we may need to account for uncertainty in the classification of leachers and non-leachers. However, in this thesis, we will use the CDFA as a source of prior information for analysing the EA data, assuming, as in [43], that the classification is secure and we will address the issue of accounting for any possible uncertainty in the classification in a future study; see Section 6.3.

4. As discussed in Section 1.2.1, a lack of evidence of a pesticide in a sample may be because it has not been used in that area or it has been used but not yet

reached the groundwater in a detectable amount. This kind of uncertainty needs to be accounted for. For example, if a given pesticide has not been used in a locality, then the probability of its leachability is 0 given any covariates values, i.e. P[leaches|any covariate] = 0. However, in this thesis, as in [43], we will not account for such uncertainty and we will address this in a future study.

1.3 The general methodology

This section describes briefly the methodology that will be used throughout the thesis. The core task of this study is to develop Bayesian methods to discriminate pesticides (classify them as leachers or non-leachers) on the basis of their chemical properties; in particular, the adsorption coefficient k_{oc} and the soil half-life $t_{1/2}^{soil}$. Therefore, the proposed models will be formulated using only the two covariates k_{oc} and $t_{1/2}^{soil}$; see [25] and Chapter 3. Throughout the analysis, the values of these covariates will be transformed to log-scale and will be denoted by z_1 and z_2 respectively, i.e. $z_1 = \log k_{oc}$ and $z_2 = \log t_{1/2}^{soil}$.

Some of this thesis is an extension of work found in the literature, particularly in [43] and [44]. In the main, this study concentrates on the analysis of the EA database where the available values for the covariates k_{oc} and $t_{1/2}^{soil}$ are uncertain. The first attempt to analyse the EA data was proposed in [43] using Bayes linear methods applied to a model linear in z_1 and z_2, and where a part of the prior information was derived from an analysis of Gustafson's data. This work is extended and investigated here using a z_2 term and an interaction term of z_1 and z_2 with the same source of prior information as in [43].

Part of the Bayesian methodology proposed in [44] was concerned with predicting the leaching probability of a given pesticide. It combined data from lysimeter experiments with a prior knowledge of the leaching probability, which was derived from the analysis of Gustafson's data where a logistic regression model was used. Again, this approach is extended and investigated using a model with a z_2 term and an interaction term of z_1 and z_2 with the same source of data and a similar method to derive the prior information as in [44].

A particular difficulty arises when trying to fit a logistic regression model with the interaction term. In this case, fitting a logistic regression model to Gustafson's data with non-overlapping groups of leachers and non-leachers, as appears in Figure 1.2, means that the maximum likelihood estimator (MLE) does not exist for a model with a z_2 term and an interaction term of z_1 and z_2, allowing for a curved discriminator that separates the leachers and non-leachers pesticides. This difficulty was tackled by firstly measuring the overlap in the logistic regression using the depth-based algorithm proposed in [7], which confirms that there is a complete separation in the covariate space of Gustafson's data as suggested by Figure 1.2. Secondly, alternative estimators such as the maximum estimated likelihood (MEL) and the weighted maximum likelihood estimator (WEMEL), which is robust against outliers, completely eliminate the overlap problem. These alternative estimators were proposed in [8].

Besides investigating the effect of introducing an interaction term on methods proposed in the literature, several Bayesian methods are developed to analyse the

EA data. These methods use logistic regression models and different types of prior information. Most of these methods were implemented using Markov Chain Monte Carlo (MCMC) simulation techniques using the WinBUGS software [40] and the R package [35]. These methods are compared using certain types of comparison tools. Related concepts such as the convergence of MCMC simulated values to a stationary distribution are discussed.

1.4 Literature review

This section documents studies that have contributed to the analysis of environmental fate, in particular the problem of groundwater contamination as a consequence of using pesticides. The review focuses on Bayesian methods which help in predicting the potential of pesticides to leach into soil and pollute the groundwater. The problem of groundwater pollution caused by pesticides has received much attention by both pesticide scientists and statisticians in recent years. These efforts have concentrated on determining environmental factors and chemical and physical properties which lie behind the tendency of pesticides to leach and contaminate groundwater. Several databases containing much information about pesticides have been published, such as those of the CDFA and the EA. Besides this, there have been attempts to develop methods to answer the question of whether it is possible to classify pesticides as leachers or non-leachers based on specified chemical properties.

Gustafson's attempt in [25] to classify pesticides in accordance with their chemical properties followed other attempts such as those developed by Cohen et al. [9] and Jury et al. [28]. However, Gustafson's attempt can be seen as an articulated

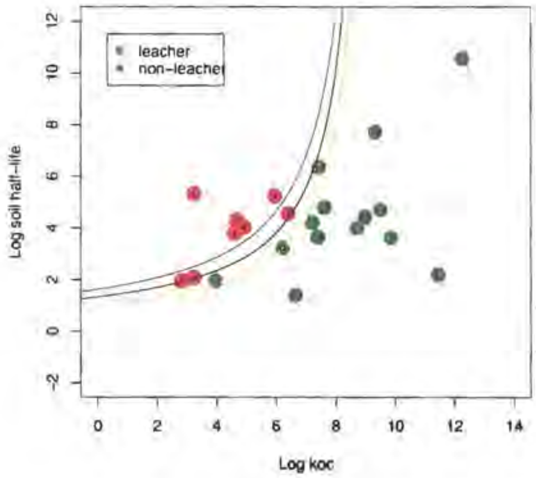

Figure 1.3: The 22 pesticides classified by CDFA together with three curves represent $GUS = 2.8$ (blue), $GUS = 1.8$ (yellow) and $GUS = 2.3$ (black).

phase in this area. Starting from Figure 1.2, Gustafson noticed (a) that the leachers occupy the left and upper portions; i.e. NW corner, corresponding to pesticides with low k_{oc} and high $t_{1/2}^{soil}$ and (b) the curved nature of the leachers corner suggests that a hyperbolic function should discriminate between the leaching and non-leaching pesticides. He devised a groundwater ubiquity score (GUS) to discriminate between leachers and non-leachers based on k_{oc} and $t_{1/2}^{soil}$. The score was derived using a functional combination of these two properties:

$$\text{GUS} = \log_{10}(t_{1/2}^{soil}) \times (P - \log_{10}(k_{oc})) \tag{1.1}$$

which can be written as

$$\text{GUS} = cz_2(P - cz_1) \tag{1.2}$$

where $z_1 = \log k_{oc}$, $z_2 = \log t_{1/2}^{soil}$, $c = \log_{10} e$, and P was set to 4 for the data he considered. As with regression models, the estimated value of P will depend on the data.

Gustafson developed a method for estimating the parameter P which can be described briefly as follows. For a given value of P, GUS values for the leachers and non-leachers can be calculated and a value Q which separates the two groups is defined as

$$Q = \frac{1}{2}\left(\text{Min GUS}_{L,i} + \text{Max GUS}_{N,j}\right) \tag{1.3}$$

where $\text{GUS}_{L,i}$ and $\text{GUS}_{N,j}$ are the GUS values for the ith leacher and jth non-leacher, respectively. A complete separation is achieved if all leachers have GUS values above Q and non-leachers have GUS values below Q. For example, given $P = 4$ in 1.1, then with $Q = 2.3$ we can achieve such separation. The penalty function f was defined as

$$f_{L,i} = \exp\left(5(Q - \text{GUS}_{L,i})/\hat{\sigma}_{L,N}\right) \tag{1.4}$$

for leachers, and

$$f_{N,j} = \exp\left(5(\text{GUS}_{N,j} - Q)/\hat{\sigma}_{L,N}\right) \tag{1.5}$$

for non-leachers, where $\hat{\sigma}_{L,N}$ is the estimate of the pooled within-class standard deviation. A combined penalty F, associated with a given estimate of P was defined as

$$F(P) = \frac{1}{n_L}\sum_i f_{L,i} + \frac{1}{n_N}\sum_j f_{N,j} \tag{1.6}$$

A simple iterative procedure was used to select the value of P that minimizes $F(P)$. Values in the range of 2 to 6 were examined and it was found that $P = 3.84$ minimizes $F(P)$, and perfect separation between the leachers and non-leachers was achieved for P anywhere from 3.6 to 4.1. "Use of $P = 4$ can be justified as the simplest numerically, although a slightly lower value may be somewhat more optimal" [25].

Using this score, Gustafson defined three zones in which transition occurs from leachers to non-leachers. These zones were defined according to two values of the GUS score, 2.8 and 1.8. Figure 1.3 shows the values of the k_{oc} and $t_{1/2}^{soil}$ for the 22 pesticides collected by the CDFA. It also shows three curves: the blue curve represents the function $GUS = 2.8$, the yellow curve represents the function $GUS = 1.8$ and the black curve represents the function $GUS = 2.3$, the average of 1.8 and 2.8. Gustafson argued that a pesticide with $GUS > 2.8$ can be considered as a leacher, $GUS < 1.8$ a non-leacher and $1.8 < GUS < 2.8$ as a transitional. He concluded in [25] that for the 22 pesticides classified by the CDFA, soil mobility and soil persistence are enough to predict the potential of a pesticide to leach. Other properties such as water solubility, the water partition coefficient and volatility do not appear to "provide any additional discriminating power in separating leachers from non-leachers" [25]. The form of Gustafson's curve in 1.1 and 1.2, which suggests a model with a z_2 term and an interaction term of z_1z_2, will be investigated in this thesis. However, Gustafson's method is confined to cases where the values of covariates are known. It does not address a situation where the values of these covariates are uncertain.

Worrall et al. in [44] proposed a Bayesian approach to discriminate pesticides as leachers or non-leachers and to predict the potential of pesticides to leach based on k_{oc} and $t_{1/2}^{soil}$. They provided a Bayesian approach to estimate the probability π that a pesticide with given chemical properties will leach and contaminate the ground-

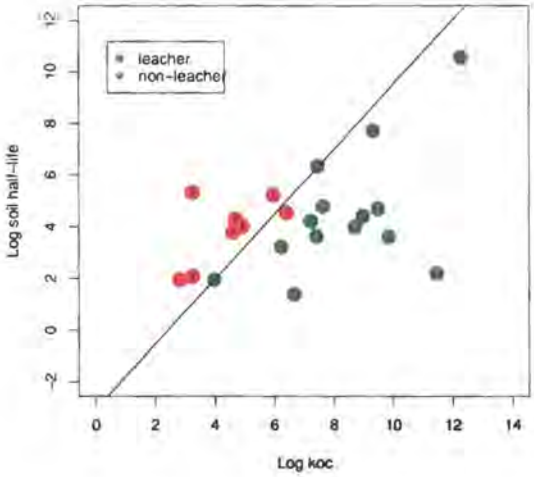

Figure 1.4: Classification for Gustafson's data using the logistic linear discriminant line proposed by Worrall et al. (1998).

water. As in any Bayesian method, the proposed model combines prior knowledge about π with available data (in the form of likelihood) to generate posterior knowledge about π. The data are from lysimeter experiments, see Section 3.3.1, which were used to discover whether or not a pesticide is observed to leach relative to a specified threshold. This data was represented in the form of likelihood using a binomial distribution with specified parameters. Worrall et al. [44] proposed the use of logistic regression to predict the binary outcome. Fitting this logistic regression to Gustafson's data, as in Figure 1.4, provided a prior distribution to predict the potential of a new pesticide to leach into groundwater given its values of k_{oc} and $t_{1/2}^{soil}$. This prior distribution was used in the Bayesian process proposed in this paper, in particular to generate the parameters of the assumed beta prior distribution for π. Combining the data from lysimeter experiments with the prior information leads to a beta posterior distribution of the probability that a pesticide under study will leach and contaminate the groundwater. They found that this method is not efficient if the values of the covariates are uncertain and suggested the use of an interaction term in logistic regression to improve the fit of logistic regression to the Gustafson's data. This suggestion, which will be expressed and formulated in Chapter 4, forms a major part of this thesis.

Wooff et al. in [43] developed a Bayes linear approach to discriminate pesticides as leachers and non-leachers based on k_{oc} and $t_{1/2}^{soil}$ where the monitored values for these pesticides are uncertain. The analysis was restricted to those 43 pesticides from the EA database where "complete data" is available. The complete data in this context means that it is known whether or not a pesticide has leached and values of k_{oc} and $t_{1/2}^{soil}$ can be extracted from the USDA database. They suggested the use of the available means and variances from the database to form the source of prior information for the uncertain values of transformed covariates, $z_1 = \log k_{oc}$ and $z_2 = \log t_{1/2}^{soil}$. They also suggested prior information for the parameter coefficients, $\boldsymbol{\beta}$, based on results of a linear model analysis of Gustafson's data, Figure 1.5. This approach can be seen as a first attempt to analyse the EA data where the covariate values are uncertain. However, an error was noted while reviewing this study. The plot shown in [43], Figure 19.3(b), which is supposed to depict the Bayes linear prediction taking into account uncertainty in the covariates, is incorrect. After investigation, it was discovered that the error was caused by using inappropriate variances. As the analysis shows, the values for the prior variances of β_0, β_1 and β_2 are $s_0^2 = 0.048, s_1^2 = 0.0010$ and $s_2^2 = 0.0017$, respectively. The incorrect plot shown here in Figure 1.7 was plotted using s_0^2 and s_1^2 instead of s_1^2 and s_2^2, as

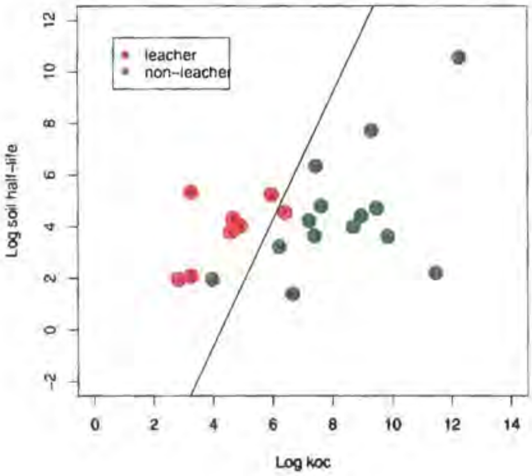

Figure 1.5: Linear discrimination based on Gustafson's data as analysed in [43] using least squares method.

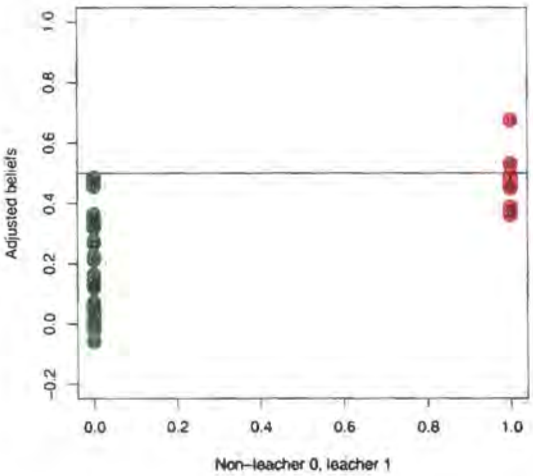

Figure 1.6: Predicted vs observed taking into account uncertainty in the covariates as analysed using Bayes linear estimate.

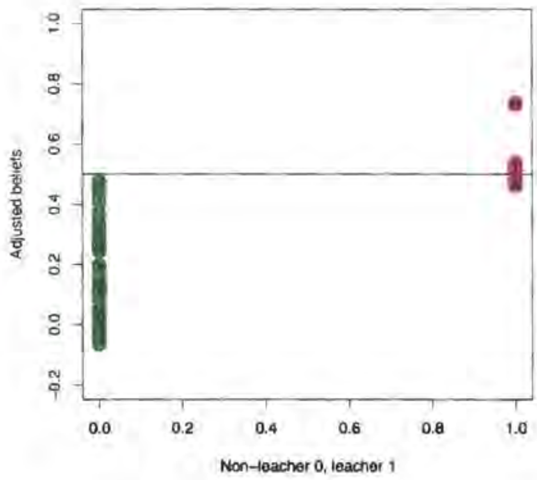

Figure 1.7: Predicted vs observed as plotted in [43], using Bayes linear estimate, where an error was encountered.

required to update the model via Bayes linear estimation. The standard deviations are approximately similar. They are range from 0.09 to 0.30. However, the corrected Bayes linear estimate still gives better discrimination than the means, as can be seen in the correct plot depicted in Figure 1.6.

In addition to the Bayes linear analysis, Worrall et al., in [45], proposed a prior contention that leaching pesticides (those with high estimated leaching probabilities) are pesticides with a low k_{oc} and high $t_{1/2}^{soil}$, and non-leaching pesticides (those with low estimated leaching probabilities) are those with a high k_{oc} and low $t_{1/2}^{soil}$, as suggested by Gustafson's data in Figure 1.2. According to this contention, leaching pesticides should appear in the NW corner and the non-leaching pesticides in the SE corner. They found that Gustafson's data is consistent with this contention, but not the means of EA data. They explained that the inconsistency is due to the limitations regarding the EA data; noted in [43] and listed on page 3 of this thesis. As an alternative, they suggest choosing a combination from the USDA database which would give the best possible separation. More specifically, the covariate pair in the NW corner is chosen for a leaching pesticide and the covariate pair in the SE corner is chosen for a non-leaching pesticide. This choice leads to complete separation of the leaching and non-leaching pesticides. They fit a logistic regression to the choice, but do not mention which method was used to derive the estimates. However, the use of maximum likelihood estimation is inappropriate because there is complete separation in the space of the covariates, rendering the MLE non-existent; and an alternative estimator should be used. Figure 1.8 shows the chosen combinations for the EA data together with the discriminant line derived from fitting the logistic regression using weighted maximum estimated likelihood (WEMEL), which will be detailed in Chapter 2.

The conclusion, in [45], was strengthened further by the multivariate runs test, described in Chapter 2, based on the total number of edges R that exist between the points in the leaching and non-leaching groups. R can be counted using a minimal spanning tree over all the points, as is shown in Figure 1.9. It is worth reporting that there is a missing edge (between cases 6 and 17) in the original figure in [45]. However, the analysis led to $R = 1$, indicating that the two groups are completely separated. The null hypothesis, whether the two groups are drawn from the same distribution, was tested using the Friedman and Rafsky (1979) statistic which is based on R. In particular, the expected value of R, $2mn/(m+n)$, and also the standard deviation of R, where m and n are the sample sizes of the two groups, were used to test the null hypothesis. In this example, R has expected value 13.02

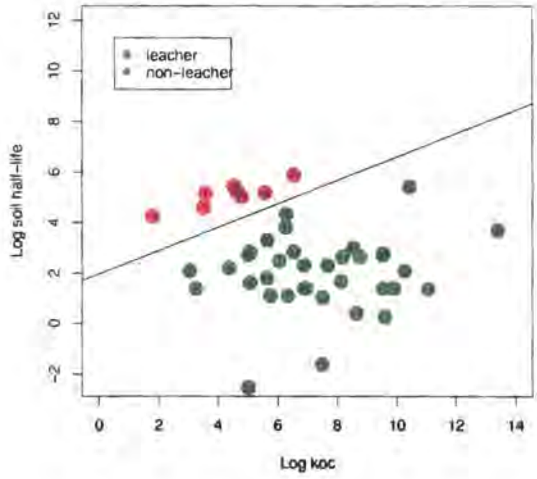

Figure 1.8: Specific combination of the EA data which support the prior contention that leachers correspond to (low k_{oc}, high $t_{1/2}^{soil}$) and non-leachers to (high k_{oc}, low $t_{1/2}^{soil}$) together with logistic discrimination line estimated by WEMEL.

and standard deviation of about 2.15, confirming that $R = 1$ is a small value, leading to a rejection of the null hypothesis.

They also tested whether it is possible to obtain a similar separation for any eight of the 43 pesticides. The test was carried out by simulation as follows.

1. Allocate at random 8 pesticides to group A and the remaining 35 to group B.

2. Apply the general methodology to separate the two groups as far as possible using the most supportive combination of the k_{oc} and $t_{1/2}^{soil}$.

3. Calculate the minimal spanning tree for each random allocation and count the number of edges, R, between the two groups.

The histogram of R for 5000 such random allocations is displayed in Figure 1.10. There are only 28 allocations with $R = 1$, indicating, as in [45], that the observed

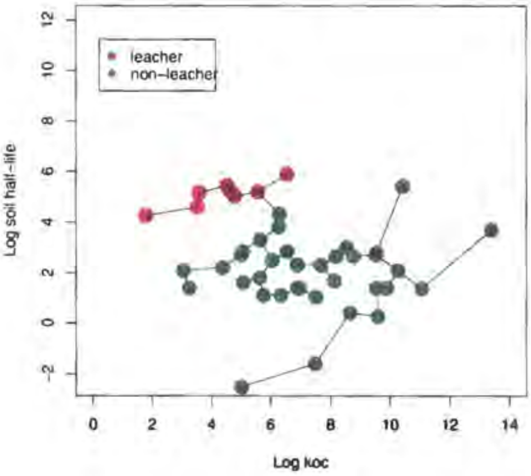

Figure 1.9: Minimal spanning tree of a specific combination of the EA data .

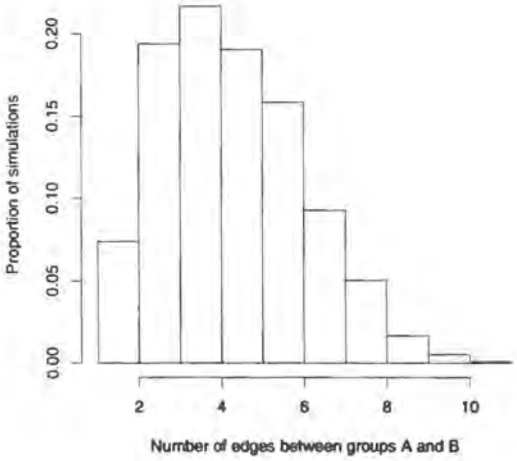

Figure 1.10: A simulated distribution of the number of edges between group A and B as conducted in [45] using 5000 simulations.

separation is real. The importance of this method is that it is applied to the EA pesticides where the values of the covariates are uncertain. However, it was limited to the use of specific values of covariates, those choices from the database values which maximises separation for each of the 5000 simulations.

Seheult [37], reviewed both classical and Bayesian discriminant rules. He used Gustafson's data to illustrate ideas. In particular, he used this data as an example in his discussion of Fisher's linear discriminant function and logistic discrimination. He used the logistic regression model proposed in [44] to fit Gustafson's data and also formulated for the first time a logistic model with an interaction term to fit Gustafson's data analogous to the GUS curve of Gustafson in [25]; see equation 1.2. This was done by formulating a model of the form:

$$\operatorname{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 z_2 + \beta_2 z_1 z_2 \tag{1.7}$$

where $z_1 = \log k_{oc}$ and $z_2 = \log t_{1/2}^{soil}$. He concluded that this model closely follows the model suggested by Gustafson in [25] and noted that it perfectly discriminates between leaching and non-leaching pesticides. However, there was no indication as to which method was used to derive the estimates of the regression parameters β_0, β_1 and β_2. However, as we have noted previously, the use of maximum likelihood estimator is inappropriate since there is complete separation of leachers and non-leachers in the space of the covariates z_1 and z_2, rendering the MLE non-existent. The model 1.7 will be fitted using the MEL and WEMEL schemes to be described in chapter 2.

1.5 Outline of the thesis

The chapters are organized as follows. In Chapter 2, the statistical concepts and computational techniques used in the thesis will be described. This will include a description of logistic regression and Bayes linear methods, including development of a general result for the Bayes linear estimate of any linear predictor with uncertain covariates. Furthermore, some aspects of Bayesian statistics, such as simulation techniques to draw samples from a posterior distribution, are reviewed.

In Chapter 3, the Bayesian method using logistic regression and lysimeter experiments proposed in [44] is both extended and modified.

In Chapter 4, the Bayes linear approach proposed in [43] is extended to include an interaction model, Bayes linear diagnostic and resulting prior variance modification, producing improved prediction.

In Chapter 5, alternative models are formulated to tackle the main research topic of the thesis, in particular, how to implement Bayesian analysis using MCMC simulation for a number of different prior specifications and number of models.

Finally, Chapter 6 concludes the thesis, including a summary of the research findings and suggestions for future work.

1.6 Originality of the thesis

This section describes briefly which sections of this thesis are original and which are from literature as follows. Most of Chapter 2 is a summary of relevant literature, except Sections 2.3.1 and 2.4.6 which are developed as parts of the thesis.

Chapter 3 extends the Bayesian analysis of Gustafson's data proposed in [44] by including an interaction term in the linear predictor. All the analyses in Section 3.2 are original. This includes the stepwise procedures to select the covariates and the most important model terms to be included in the linear predictor, interpretation of regression parameters estimates and checking the adequacy of the fitted model. Sections 3.3 (excluded 3.3.1), 3.4, 3.5, 3.6 and 3.7 are the same Bayesian components used in [44], but with appropriate modifications. The analysis in Section 3.8, which discusses an alternative Bayesian analysis, is original.

Chapter 4 extends the Bayes linear analysis of EA data proposed in [43] by including an interaction term in the linear predictor. Section 4.2 includes a discussion about regression analysis, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA): all are summarised from literature, and how these tools can be used to analyse the Gustafson data. Section 4.3 discusses whether the analyses of Gustafson's data provide good prediction for the EA pesticides. Section 4.4 implements the Bayes linear estimate to analyse the EA data including specifying the prior information, updating the model, using Bayes linear diagnostics to analyse the observed adjustments and re-structuring some of prior beliefs. Section 4.5 includes further analysis of the linear discriminant proposed in [43]. It also includes Bayes linear diagnostics to analyse the observed adjustments and re-structuring some of prior beliefs. All of above analyses form original parts of the thesis.

In Chapter 5, we develop alternative classical Bayesian models to analyse the EA data. All the sections in this chapter are original.

Finally, Chapter 6 suggests some original future research topics.

1.7 Conclusion

This chapter gives an overview of the aims and objectives of this thesis and the sources of the data used. It has also highlighted the obstacles presented by the data and the general methodology adopted throughout the work. In addition, it has provided a brief summary of some important studies concerned with pesticide discrimination; in particular, the study by Gustafson in [25], the Bayesian model proposed by Worrall et al. in [44], the Bayes linear model proposed by Wooff et al. in [43] and the Bayesian approaches proposed by Worrall et al. in [45].

some errors were noted in the documentation of some of these studies. An error was noted in the Bayes linear predictor in [43]. Furthermore, it is not clear how the estimates in [45] and [37] were achieved: the use of maximum likelihood estimator would be inappropriate since there is complete separation of leachers and non-leachers in the space of the covariates. Finally, an edge is missing from the plot of the minimal spanning tree for the most supportive combinations displayed in [45].

Chapter 2

Statistical concepts and computational techniques

2.1 Introduction

This chapter reviews the statistical concepts and computational techniques used in the thesis. The review concentrates on (a) logistic regression models, (b) Bayes linear methods and (c) classical Bayesian methods. The review of logistic regression includes a discussion regarding one of the deficiencies of maximum likelihood and how to tackle it. The statistical packages used to implement these methods will be described.

2.2 Logistic regression models

Logistic regression is widely used as a model for the analysis of binary data. Its importance stems from its straightforward implementation in studying and exploring many statistical concepts such as regression, classification and prediction. To set up this model, the following notation is needed.

Let $Y = (Y_1, \ldots, Y_n)$ denote an $(n \times 1)$ vector of binary responses or outcome variables where

$$Y_i = \begin{cases} 1 & \text{if the ith outcome is a success} \\ 0 & \text{if the ith outcome is a failure} \end{cases}$$

Associated with each Y_i $(i = 1, \ldots, n)$, there is a vector of model terms $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ each of them a known function of q explanatory variables $z_1, z_2, \ldots, z_q$, where the x_{i1} are fixed at 1 for $i = 1, \ldots, n$. Then, Y_i is modelled to have a Bernoulli distribution with probability of success $P(Y_i = 1|\boldsymbol{x}_i) = \pi_i$. In general, $\boldsymbol{x}_i$ is linked to the expectation of Y_i which is π_i by a link function $g(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$ such that $g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta})$ takes values in the interval (0,1). One possible choice of link function is the logit function, the logarithm of the odds $\pi_i/(1 - \pi_i)$ such that

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \eta_i \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of unknown parameters and η_i is called the linear predictor. This is equivalent to modelling the probability π_i as

$$\pi_i = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}} \tag{2.2}$$

where the logistic function form on the right-hand side is the inverse of the logit function. Thus, the probability of success π_i can be written as

$$p(Y_i = 1|\boldsymbol{x}_i, \boldsymbol{\beta}) = \pi_i = \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}} \tag{2.3}$$

The likelihood function of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \tag{2.4}$$

and the log-likelihood function is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\boldsymbol{x}_i^T \boldsymbol{\beta} y_i - \log\left(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})\right)\right] \tag{2.5}$$

To maximize $L(\boldsymbol{\beta})$, the derivative with respect to $\boldsymbol{\beta}$ is needed:

$$\frac{dL}{d\boldsymbol{\beta}} = S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \pi_i)\boldsymbol{x}_i^T = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\pi}) \tag{2.6}$$

where $\boldsymbol{x}_i^T$ is the i-th row of X and $S(\boldsymbol{\beta})$ is called the score function. To estimate the parameters $\boldsymbol{\beta}$, classically one uses the MLE, $\hat{\boldsymbol{\beta}}$, the solution to the equation $S(\boldsymbol{\beta}) = \mathbf{0}$, provided $S'(\hat{\boldsymbol{\beta}})$ is positive definite.

While logistic regression has several advantages, the MLE may not exist and it is influenced by extreme values in the design space; i.e., it is not robust to outliers in $\boldsymbol{x}$. The MLE does not exist for those data sets in which there is complete separation of successes and failures in the space of covariates. Santner and Duffy in [36] showed that the MLE does not exist if the data set is completely or quasicompletely separated and is unique if the data set has an overlap. We consider separation here because it arises when applying logistic regression to discriminating between leaching and non-leaching pesticides. The meanings of complete separation, quasicomplete separation and overlap and how to measure it are given below.

2.2.1 Measuring overlap

As mentioned above, the MLE does not exist for binary data, modelled by logistic regression, when there is a complete separation of successes and failures. As defined in [36], a data set of the form $Z_n = \left\{(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)\,; i = 1, \ldots, n\right\}$, where $x_{i1} = 1$ for all $(i = 1, \ldots, n)$, is said to have complete separation if there exists a vector

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ such that:

$$\boldsymbol{x}_i^T\boldsymbol{\beta} > 0 \qquad \text{if} \quad y_i = 1$$

$$\boldsymbol{x}_i^T\boldsymbol{\beta} < 0 \qquad \text{if} \quad y_i = 0$$

for $i = 1, \ldots, n$. Z_n, which does not have complete separation, is said to have a quasicomplete separation if there exists a vector $\boldsymbol{\beta} \in \mathbb{R}^p \setminus \{0\}$ such that:

$$\boldsymbol{x}_i^T\boldsymbol{\beta} \geq 0 \qquad \text{if} \quad y_i = 1$$

$$\boldsymbol{x}_i^T\boldsymbol{\beta} \leq 0 \qquad \text{if} \quad y_i = 0$$

for all i and if there exists some $j \in \{1, \ldots, n\}$ such that $\boldsymbol{x}_j^T\boldsymbol{\beta} = 0$. Z_n is said to have an overlap if there is no complete separation and no quasicomplete separation.

As shown in [2] and [36], the MLE of $\boldsymbol{\beta}$ exists if and only if the data set has an overlap. Consequently, in order to estimate $\boldsymbol{\beta}$, the amount of overlap needs to be measured. Christman et al. in [7], proposed an approach for measuring the amount of overlap using a depth-based algorithm. The proposed algorithm calculates the smallest number of observations whose removal destroys the overlap with the result that the MLE does not exist.

If the MLE does not exist, then alternative estimators similar to those proposed in [16], [12] and [8] can be adopted. In the last reference, two estimators were proposed using a hidden logistic regression model. The first estimator is the maximum estimated likelihood estimator (MEL) and the second is the weighted maximum estimated likelihood estimator (WEMEL). These two estimators have been used in the current study when complete separation arises. The MEL estimator helps to eliminate the overlap problem, but unlike the WEMEL estimator it is not robust against outliers. The following details the basis and derivation of the MEL and WEMEL estimators.

2.2.2 Maximum estimated likelihood (MEL) estimator

This estimator is constructed using a hidden logistic regression model. As depicted in Figure 2.1, this model has two responses described as follows. The first is the true response T which is assumed to be an unobservable variable having two outcomes: success (s) and failure (f). The second response, denoted by Y, takes values 0 and 1 and is assumed to be observable. The two responses are related as follows. If the true response is $T = s$, then the observed value will be $Y = 1$ with probability $\mathrm{P}\left[Y = 1|T = s\right] = \delta_1$ and hence $\mathrm{P}\left[Y = 0|T = s\right] = 1 - \delta_1$. Similarly, if the true response is $T = f$, then we observe $Y = 1$ with probability $\mathrm{P}\left[Y = 1|T = f\right] = \delta_0$ and hence $\mathrm{P}\left[Y = 0|T = f\right] = 1 - \delta_0$. The following restriction is assumed. $0 < \delta_0 < \frac{1}{2} < \delta_1 < 1$. It is shown in [8] that the maximum likelihood estimator of T, $\hat{T}_{ML}$, given $(Y = y)$ is:

$$\hat{T}_{ML}(Y = 0) = f$$

$$\hat{T}_{ML}(Y = 1) = s$$

and hence the conditional probability that $Y = 1$ given $\hat{T}_{ML}$ is:

$$\mathrm{P}\left[Y = 1|\hat{T}_{ML}\right] = \begin{cases} \delta_0 & \text{if y=0} \\ \delta_1 & \text{if y=1} \end{cases} \tag{2.7}$$

Denoting 2.7 by $\tilde{Y}$, then $\tilde{Y}$ can be written as:

$$\tilde{Y} = \delta_0 + (\delta_1 - \delta_0)Y = (1 - Y)\delta_0 + Y\delta_1$$

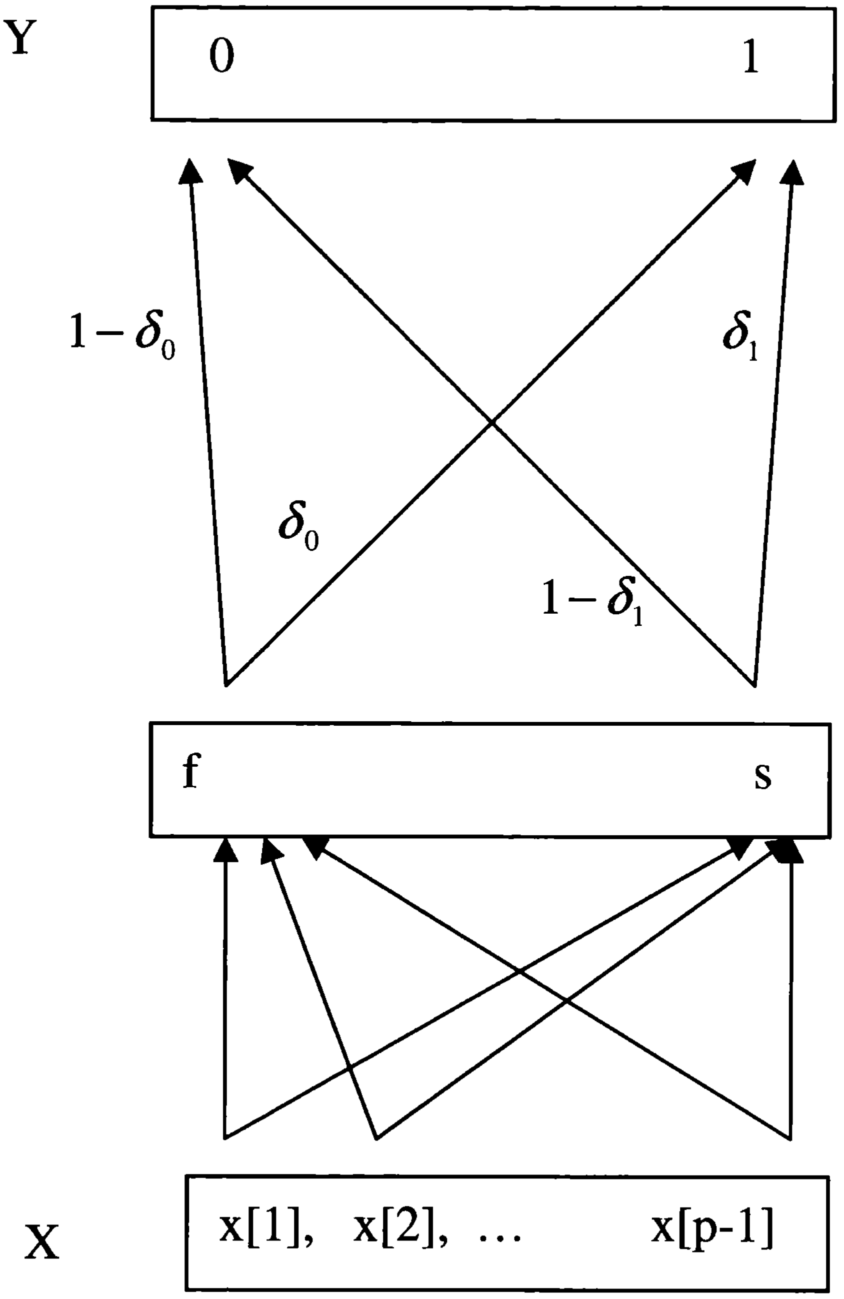

Figure 2.1: Hidden logistic regression model

and for the *ith* observation:

$$\tilde{y}_i = (1 - y_i)\delta_0 + y_i\delta_1. \tag{2.8}$$

So the pseudo-observation, $\tilde{y}_i$, is the result of a deterministic transformation of y_i. When $\delta_0 = 0$ and $\delta_1 = 1$, then $\tilde{y}_i = y_i$. To fit a logistic regression to $\tilde{y}_i$ using likelihood, $\tilde{y}_i$ was given a Bernoulli distribution. The estimated likelihood function of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ given $\tilde{y}_1, \ldots, \tilde{y}_n$ is

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{\tilde{y}_i} (1 - \pi_i)^{1-\tilde{y}_i} \tag{2.9}$$

where

$$\pi_i = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}$$

is the success probability. This likelihood is called the estimated likelihood because the true likelihood $\prod_{i=1}^{n} \pi_i^{t_i} (1 - \pi_i)^{1-t_i}$, which depends on the true observations $t_1, \ldots, t_n$ is unknown. It is known only when $\delta_0 = 0$ and $\delta_1 = 1$. The estimated log-likelihood is

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} [\tilde{y}_i \log \pi_i + (1 - \tilde{y}_i) \log(1 - \pi_i)] \tag{2.10}$$

and hence, the estimated score function is

$$S(\boldsymbol{\beta}|\tilde{y}_1, \ldots, \tilde{y}_n) = \sum_{i=1}^{n} (\tilde{y}_i - \pi_i)\, \boldsymbol{x}_i = X^T(\tilde{\boldsymbol{y}} - \boldsymbol{\pi}). \tag{2.11}$$

The value of $\boldsymbol{\beta}$ that maximizes equation 2.10 is called the maximum estimated likelihood (MEL) estimate; that is, estimate is obtained by solving equation $S(\boldsymbol{\beta}|\tilde{y}_1, \ldots, \tilde{y}_n) = 0$.

Choices of δ_0 and δ_1

In the absence of subject matter choices for δ_0 and δ_1, they are chosen in [8] to be

$$\delta_0 = \frac{\hat{\pi}\delta}{1+\delta} \quad \text{and} \quad \delta_1 = \frac{1+\hat{\pi}\delta}{1+\delta} \tag{2.12}$$

where

$$\hat{\pi} \; = \; \max(\delta, \min(1-\delta, \bar{\pi})) \quad \text{and} \quad \bar{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.13}$$

and δ is a small positive number with default choice $\delta = 0.01$.

2.2.3 Weighted maximum estimated likelihood (WEMEL) estimator

The MEL estimator helps to eliminate the overlap problem, but it is not robust against outliers. In [8] a robustification for the MEL estimator was proposed by down-weighting leverage points. This technique led to the weighted maximum estimated likelihood (WEMEL) estimator, which is defined as the solution $\hat{\boldsymbol{\beta}}$ to

$$\sum_{i=1}^{n} (\tilde{y}_i - \pi_i)\, w_i \boldsymbol{x}_i = \mathbf{0} \tag{2.14}$$

where the weight w_i depends on how far away $\boldsymbol{x}_i$ is from the bulk of the data. The following weights were proposed:

$$w_i = \frac{M}{\max\{RD^2(\boldsymbol{x}_i^*), M\}} \tag{2.15}$$

where $\boldsymbol{x}_i^* = (x_{i2}, \ldots, x_{ip})^T \in \mathbb{R}^{p-1}$, $RD(\boldsymbol{x}_i^*)$ is its robust distance and M is the 75^{th} percentile of all $RD^2(\boldsymbol{x}_j^*)$, $j = 1, \ldots, n$. In other words, this is equivalent to giving a weight less than 1 to all of the 25% most extreme design points. The WEMEL estimator can be computed in a straightforward manner using generalized linear model (GLM) algorithms (such as those in R and S-plus) with prior weights w_i.

The S-plus codes, for MEL and WEMEL can be downloaded from:

`http://win-www.uia.ac.be/u/statis/Robustn.htm`

`http://www.statistik.uni-dortmund.de/sfb475/berichte/rouschr2.zip`.

2.3 Bayes linear methods

Bayes linear methods are used as a simple approach to combining the prior knowledge of uncertainty with observational data using expectations. In this section, we follow aspects of the development detailed in [24] and summarised in [23]. Prior knowledge is collected and organised in the form of means, variances and covariances and is then updated via linear fitting. The Bayes linear approach is suitable for analysing the EA data, as it may be prepared in the form of means and variance structures, but full distributions, as required by classical Bayes methods, may be more difficult to specify.

Let $\boldsymbol{y}$ denote a vector of observed data and $\boldsymbol{x}$ denote an unobserved vector to be updated via $\boldsymbol{y}$. Then the **adjusted expectation** for $\boldsymbol{x}$ given $\boldsymbol{y}$, $\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x})$, is given by

$$\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{x}) + \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\mathrm{Var}(\boldsymbol{y})^{-1}[\boldsymbol{y} - \mathrm{E}(\boldsymbol{y})] \tag{2.16}$$

where $\mathrm{E}(\boldsymbol{x})$, $\mathrm{Var}(\boldsymbol{x})$, $\mathrm{Var}(\boldsymbol{y})$, $\mathrm{E}(\boldsymbol{y})$ and $\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})$ are specified a priori.

The **adjusted version** of $\boldsymbol{x}$ given $\boldsymbol{y}$, $\mathrm{A}_{\boldsymbol{y}}(\boldsymbol{x})$, is defined to be the residual vector

$$\mathrm{A}_{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{x} - \mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x}) \tag{2.17}$$

Therefore, the vector $\boldsymbol{x}$ can be partitioned as the sum of two uncorrelated vectors

$$\boldsymbol{x} = \mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x}) + \mathrm{A}_{\boldsymbol{y}}(\boldsymbol{x}) \tag{2.18}$$

Hence, the variance matrix of $\boldsymbol{x}$ is partitioned into two variance components

$$\mathrm{Var}(\boldsymbol{x}) = \mathrm{Var}(\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x})) + \mathrm{Var}(\mathrm{A}_{\boldsymbol{y}}(\boldsymbol{x})) \tag{2.19}$$

In 2.19, $\mathrm{Var}(\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x}))$ is called the **resolved variance matrix** for $\boldsymbol{x}$ by $\boldsymbol{y}$, written

$$\mathrm{RVar}_{\boldsymbol{y}}(\boldsymbol{x}) = \mathrm{Var}(\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{x})), \tag{2.20}$$

and $\mathrm{Var}(\mathrm{A}_{\boldsymbol{y}}(\boldsymbol{x}))$ is called the **adjusted variance matrix** for $\boldsymbol{x}$ by $\boldsymbol{y}$, written

$$\mathrm{Var}_{\boldsymbol{y}}(\boldsymbol{x}) = \mathrm{Var}(\mathrm{A}_{\boldsymbol{y}}(\boldsymbol{x})) \tag{2.21}$$

$\mathrm{Var}_{\boldsymbol{y}}(\boldsymbol{x})$ is calculated as

$$\mathrm{Var}_{\boldsymbol{y}}(\boldsymbol{x}) = \mathrm{Var}(\boldsymbol{x}) - \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\mathrm{Var}(\boldsymbol{y})^{-1}\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x}) \tag{2.22}$$

Thus,

$$\mathrm{RVar}_{\boldsymbol{y}}(\boldsymbol{x}) = \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\mathrm{Var}(\boldsymbol{y})^{-1}\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x}) \tag{2.23}$$

The **resolution transform matrix** is defined as

$$\begin{aligned} \mathrm{T}_{\boldsymbol{x}:\boldsymbol{y}} &= \mathrm{Var}(\boldsymbol{x})^{-1}\mathrm{RVar}_{\boldsymbol{y}}(\boldsymbol{x}) && (2.24) \\ &= \mathrm{Var}(\boldsymbol{x})^{-1}\mathrm{Cov}(\boldsymbol{x}, \boldsymbol{y})\mathrm{Var}(\boldsymbol{y})^{-1}\mathrm{Cov}(\boldsymbol{y}, \boldsymbol{x}) && (2.25) \end{aligned}$$

The **resolved uncertainty** for $\boldsymbol{x}$ given adjustment by $\boldsymbol{y}$ is defined to be

$$\mathrm{RU}_{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{i=1}^{r_x} \lambda_i = \mathrm{trace}\left\{\mathrm{T}_{\boldsymbol{x}:\boldsymbol{y}}\right\} \tag{2.26}$$

where $\lambda_1, \lambda_2, \dots, \lambda_{r_x}$ are the eigenvalues of $\mathrm{T}_{\boldsymbol{x}:y}$ and r_x is the rank of $\mathrm{Var}(\boldsymbol{x})$.

The **system resolution** for $\boldsymbol{x}$ is defined as

$$\mathrm{R}_{\boldsymbol{y}}(\boldsymbol{x}) = \frac{\mathrm{RU}_{\boldsymbol{y}}(\boldsymbol{x})}{r_x} = \frac{1}{r_x}\sum_{i=1}^{r_x} \lambda_i \tag{2.27}$$

$\mathrm{R}_{\boldsymbol{y}}(\boldsymbol{x})$ is used as a scale-free measure of the overall proportion of uncertainty explained by the model; see [43].

The **size of the adjustment** of $\boldsymbol{x}$ by $\boldsymbol{y} = y$ is defined as

$$\mathrm{Size}_y(\boldsymbol{x}) = [\mathrm{E}_y(\boldsymbol{x}) - \mathrm{E}(\boldsymbol{x})]^T \, \mathrm{Var}(\boldsymbol{x})^{-1} \, [\mathrm{E}_y(\boldsymbol{x}) - \mathrm{E}(\boldsymbol{x})] \tag{2.28}$$

$\mathrm{Size}_y(\boldsymbol{x})$ represents the maximal change in adjusted expectation relative to prior variation.

The **size ratio** for the adjustment of $\boldsymbol{x}$ by $\boldsymbol{y}$ is defined as

$$\begin{aligned} \mathrm{Sr}_y(\boldsymbol{x}) &= \frac{\mathrm{Size}_y(\boldsymbol{x})}{\mathrm{E}(\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{x}))} && (2.29) \\ &= \frac{[\mathrm{E}_y(\boldsymbol{x}) - \mathrm{E}(\boldsymbol{x})]^T \, \mathrm{Var}(\boldsymbol{x})^{-1} \, [\mathrm{E}_y(\boldsymbol{x}) - \mathrm{E}(\boldsymbol{x})]}{\sum_{i=1}^{r_T} \lambda_i} && (2.30) \end{aligned}$$

where $\mathrm{E}(\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{x})) = \sum_{i=1}^{r_T} \lambda_i = \mathrm{trace}\,\{T_{\boldsymbol{x}:y}\} = \mathrm{RU}_{\boldsymbol{y}}(\boldsymbol{x})$, and r_T is the rank of the resolution transform matrix. $\mathrm{Sr}_y(\boldsymbol{x})$ has an expectation of unity. A size ratio far away from unity may warn of possible conflicts between prior specification and adjusted beliefs. A simple rule to suggest warning levels for the size ratio is the following interval

$$P\left(1 - \frac{6\sqrt{\sum_{i=1}^{r_T}\lambda_i^2}}{\sum_{i=1}^{r_T}\lambda_i} \leq \mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{x}) \leq \frac{6\sqrt{\sum_{i=1}^{r_T}\lambda_i^2}}{\sum_{i=1}^{r_T}\lambda_i}\right) \leq 0.9444 \tag{2.31}$$

In the following section, a general formula for Bayes linear estimation applied to any linear model with uncertain covariates will be derived. A special version of the general result was derived in [43] and used as a linear discriminant.

2.3.1 Bayes linear estimation for linear models with uncertain covariates

The general linear model for a response y with q covariates $\boldsymbol{z} = (z_1, \ldots, z_q)^{\mathrm{T}}$ may be written

$$y = \sum_{j=1}^{p} f_j(\boldsymbol{z})\beta_j + \epsilon$$

where the f_j are specified functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ are parameters and ϵ is a random error.

The model can be written more compactly as

$$y = \sum_{j=1}^{p} x_j\beta_j + \epsilon = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} = \eta + \epsilon$$

where the $x_j = f_j(\boldsymbol{z})$ are model terms and $\eta = \sum_{j=1}^{q} x_j\beta_j$ is called the linear predictor.

When there are n cases, the response y_i for case i will be written

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} = \eta_i + \epsilon_i$$

where $x_{ij} = f_j(\boldsymbol{z}_i)$, corresponding to the covariate values $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iq})^{\mathrm{T}}$ for case i and $\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j$ is the linear predictor associated with the values $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ of the model terms for case i.

The model may be written in vector form as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} = \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

where $X = (x_{ij})$ is the $n \times p$ model matrix.

The statistical assumptions and specifications for the model are as follows:

1. $\boldsymbol{z}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are uncorrelated random vectors

2. $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are uncorrelated

3. $\mathrm{E}[\boldsymbol{x}_i] = \boldsymbol{m}_i$, $\mathrm{Var}[\boldsymbol{x}_i] = \Sigma_i$

4. $\mathrm{E}[\boldsymbol{\beta}] = \boldsymbol{b}$, $\mathrm{Var}[\beta] = \Sigma_\beta$

5. $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathrm{Var}[\epsilon] = \sigma^2 I$

Note that (a) the mean and variance structures for the $\boldsymbol{x}_i$ will be derived from those specified for the $\boldsymbol{z}_i$, in particular, the $\boldsymbol{x}$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are uncorrelated random vectors; and (b) the $\boldsymbol{m}_i$, Σ_i, $\boldsymbol{b}$, Σ_β, and σ^2 are specified from prior information.

Our aim is to derive the Bayes linear estimate $\hat{\boldsymbol{y}} = \mathrm{E}_{\boldsymbol{y}}[\boldsymbol{\eta}]$ of $\boldsymbol{\eta}$ and its adjusted variance $\mathrm{Var}_{\boldsymbol{y}}[\boldsymbol{\eta}]$.

Theorem The Bayes linear estimate of the linear predictor $\boldsymbol{\eta}$ is

$$\hat{\boldsymbol{y}} = (I + A)^{-1} M\boldsymbol{b} + A(I + A)^{-1}\boldsymbol{y}$$

with adjusted variance $\sigma^2 A(I + A)^{-1}$, $M = \mathrm{E}[X]$ and $\sigma^2 A = M\Sigma_\beta M^{\mathrm{T}} + D$ where D is a diagonal matrix with $D_{ii} = \boldsymbol{b}^{\mathrm{T}}\Sigma_i\,\boldsymbol{b} + \mathrm{trace}[\Sigma_i\Sigma_\beta]$.

Proof

It is straightforward to show that $\mathrm{E}[\boldsymbol{\eta}] = M\boldsymbol{b}$, $\mathrm{Cov}[\boldsymbol{\eta}, \boldsymbol{y}] = \mathrm{Var}[\boldsymbol{\eta}]$ and $\mathrm{Var}[\boldsymbol{y}] = \sigma^2 I + \mathrm{Var}[\boldsymbol{\eta}]$, so we only need to evaluate $\mathrm{Var}[\boldsymbol{\eta}]$ to evaluate the adjusted expectation and adjusted variance to complete the proof.

$$\mathrm{Cov}[\eta_i, \eta_j] = \mathrm{Cov}[\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta}] = \mathrm{E}[\mathrm{Cov}[\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta} \,|\, \boldsymbol{\beta}]] + \mathrm{Cov}[\mathrm{E}[\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \,|\, \boldsymbol{\beta}], \mathrm{E}[\boldsymbol{x}_j^{\mathrm{T}}\boldsymbol{\beta} \,|\, \boldsymbol{\beta}]]$$

As $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are uncorrelated, the first term on the right is zero, unless $i = j$, in which case it becomes

$$\mathrm{E}[\mathrm{Var}[\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \,|\, \boldsymbol{\beta}]] = \mathrm{E}[\boldsymbol{\beta}^{\mathrm{T}}\Sigma_i\boldsymbol{\beta}] = \boldsymbol{b}^{\mathrm{T}}\Sigma_i\boldsymbol{b} + \mathrm{trace}[\Sigma_i\Sigma_\beta]$$

The second term is

$$\mathrm{Cov}[\boldsymbol{m}_i^{\mathrm{T}}\boldsymbol{\beta},\, \boldsymbol{m}_j^{\mathrm{T}}\boldsymbol{\beta}] = \boldsymbol{m}_i^{\mathrm{T}}\Sigma_{\beta}\,\boldsymbol{m}_j$$

Thus

$$\mathrm{Var}[\eta_i] = \boldsymbol{b}^{\mathrm{T}}\Sigma_i\,\boldsymbol{b} + \mathrm{trace}[\Sigma_i\Sigma_{\beta}] + \boldsymbol{m}_i^{\mathrm{T}}\Sigma_{\beta}\,\boldsymbol{m}_i$$

Hence, $\mathrm{Var}[\boldsymbol{\eta}] = \sigma^2 A$ and we can now evaluate $\mathrm{E}_{\boldsymbol{y}}[\boldsymbol{\eta}]$ and $\mathrm{Var}_{\boldsymbol{y}}[\boldsymbol{\eta}]$. The adjusted expectation of $\boldsymbol{\eta}$ is

$$\mathrm{E}_{\boldsymbol{y}}[\boldsymbol{\eta}] = \mathrm{E}[\boldsymbol{\eta}] + \mathrm{Cov}[\boldsymbol{\eta},\, \boldsymbol{y}]\mathrm{Var}[\boldsymbol{y}]^{-1}\,[\boldsymbol{y} - \mathrm{E}[\boldsymbol{y}]] = M\boldsymbol{b} + \sigma^2 A(\sigma^2(I + A))^{-1}\,[\boldsymbol{y} - M\boldsymbol{b}]$$

which simplifies to give the required expression for $\hat{\boldsymbol{y}}$.

The adjusted variance of $\boldsymbol{\eta}$ is

$$\mathrm{Var}_{\boldsymbol{y}}[\boldsymbol{\eta}] = \mathrm{Var}[\boldsymbol{\eta}] - \mathrm{Cov}[\boldsymbol{\eta},\, \boldsymbol{y}]\,\mathrm{Var}[\boldsymbol{y}]^{-1}\mathrm{Cov}[\boldsymbol{y},\, \boldsymbol{\eta}] = \sigma^2 A - \sigma^2 A(\sigma^2(I + A))^{-1}\sigma^2 A$$

which simplifies to give $\sigma^2 A(I + A)^{-1}$. Notice that $\mathrm{Var}_{\boldsymbol{y}}[\boldsymbol{\eta}]$ does not depend on $\boldsymbol{y}$.

Examples

We evaluate $\hat{\boldsymbol{y}}$ for two examples of linear models that we use to discriminate between non-leaching and leaching pesticides, with $y = 0$ or $y = 1$, respectively. In both examples, there are $p = 3$ terms which are known functions of $q = 2$ covariates $z_1 = \log k_{oc}$ and $z_2 = \log t_{1/2}^{\mathrm{soil}}$, which we take to be uncorrelated.

Example 1 The model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ where $x_1 = z_1$ and $x_2 = z_2$.

Example 2 The model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ where $x_1 = z_2$ and $x_2 = z_1 z_2$.

For case i we have in both examples

$$\boldsymbol{m}_i^{\mathrm{T}} = (1, \mathrm{E}[x_{i1}], \mathrm{E}[x_{i2}])$$

$$\boldsymbol{b}^{\mathrm{T}}\Sigma_i\boldsymbol{b} = b_1^2\mathrm{Var}[x_{i1}] + 2b_1b_2\mathrm{Cov}[x_{i1},\, x_{i2}] + b_2^2\mathrm{Var}[x_{i2}]$$

$$\mathrm{trace}[\Sigma_i\Sigma_\beta] = \sigma_1^2\mathrm{Var}[x_{i1}] + 2\sigma_{12}\mathrm{Cov}[x_{i1},\, x_{i2}] + \sigma_2^2\mathrm{Var}[x_{i2}]$$

where $\sigma_1^2 = \mathrm{Var}[\beta_1]$, $\sigma_2^2 = \mathrm{Var}[\beta_2]$ and $\sigma_{12} = \mathrm{Cov}[\beta_1,\, \beta_2]$.

In example 1

$$\boldsymbol{m}_i^{\mathrm{T}} = (1, \mathrm{E}[z_{i1}], \mathrm{E}[z_{i2}])$$

$$\mathrm{Var}[x_{i1}] = \mathrm{Var}[z_{i1}], \quad \mathrm{Var}[x_{i2}] = \mathrm{Var}[z_{i2}], \quad \mathrm{Cov}[x_{i1},\, x_{i2}] = 0$$

In example 2

$$\boldsymbol{m}_i^{\mathrm{T}} = (1, \mathrm{E}[z_{i2}], \mathrm{E}[z_{i2}]\mathrm{E}[z_{i2}])$$

$$\mathrm{Var}[x_{i1}] = \mathrm{Var}[z_{i2}]$$

$$\mathrm{Var}[x_{i2}] = \mathrm{Var}[z_{i1}z_{i2}] = \mathrm{Var}[z_{i1}]\mathrm{Var}[z_{i2}] + \mathrm{E}[z_{i1}]^2\mathrm{Var}[z_{i2}] + \mathrm{E}[z_{i2}]^2\mathrm{Var}[z_{i1}]$$

$$\mathrm{Cov}[x_{i1},\, x_{i2}] = \mathrm{Cov}[z_{i2},\, z_{i1}z_{i2}] = \mathrm{E}[z_{i1}]\mathrm{Var}[z_{i2}]$$

These are all the results necessary to calculate $\hat{\boldsymbol{y}}$ and the adjusted variance in the two examples. To implement these results it is necessary to specify $\mathrm{E}[\boldsymbol{\beta}]$, Σ_β and $\mathrm{E}[z_{i1}]$, $\mathrm{Var}[z_{i1}]$, $\mathrm{E}[z_{i2}]$ and $\mathrm{Var}[z_{i2}]$ for $i = 1, \ldots, n$.

2.4 Bayesian inference

Bayesian statistical inference uses Bayes theorem to combine sample data (in the form of likelihood) with prior beliefs (in the form of a prior distribution) to arrive at posterior beliefs (in the form of a posterior distribution).

Let $\boldsymbol{\beta}$ denote an unobservable vector of parameters of interest with prior distribution $p(\boldsymbol{\beta})$ and let $\boldsymbol{y}$ denote an observable vector of sampled data with sampling distribution $p(\boldsymbol{y}|\boldsymbol{\beta})$. Then, from Bayes' theorem, the posterior probability distribution of $p(\boldsymbol{\beta}|\boldsymbol{y})$ can be expressed as:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})} \tag{2.32}$$

where $p(\boldsymbol{y}) = \int p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})d\boldsymbol{\beta}$ is called the prior predictive distribution of $\boldsymbol{y}$. Since the term $p(\boldsymbol{y})$ does not depend on $\boldsymbol{\beta}$, omitting it yields the unnormalized posterior distribution which can be expressed as

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta})p(\boldsymbol{\beta}) \tag{2.33}$$

The last equation shows that the data $\boldsymbol{y}$ affects the posterior distribution of $\boldsymbol{\beta}$ only through $p(\boldsymbol{y}|\boldsymbol{\beta})$, which, as a function of $\boldsymbol{\beta}$, is called the likelihood function. Hence, the last equation states that the posterior distribution of $\boldsymbol{\beta}$, $p(\boldsymbol{\beta}|\boldsymbol{y})$, is proportional to the product of the likelihood function $p(\boldsymbol{y}|\boldsymbol{\beta})$ and the prior distribution $p(\boldsymbol{\beta})$.

The posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$ can be used to make inferences about $\boldsymbol{\beta}$ or a future observations $\tilde{\boldsymbol{y}}$ conditional on the observed data $\boldsymbol{y}$. This inference is called "Bayesian predictive inference" and the distribution of $\tilde{\boldsymbol{y}}$ is called the "posterior predictive distribution" which can be evaluated as

$$\begin{aligned} p(\tilde{\boldsymbol{y}}|\boldsymbol{y}) &= \int p(\tilde{\boldsymbol{y}}, \boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta} \\ &= \int p(\tilde{\boldsymbol{y}}|\boldsymbol{\beta}, \boldsymbol{y})p(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta} \\ &= \int p(\tilde{\boldsymbol{y}}|\boldsymbol{\beta})p(\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta} \end{aligned} \tag{2.34}$$

provided $\boldsymbol{y}$ and $\tilde{\boldsymbol{y}}$ are conditionally independent, given $\boldsymbol{\beta}$ [18].

Another important feature of Bayesian inferences is the choice of prior information. In the absence of prior information, a "non-informative" prior or "vague" distribution can be used; see, for example, [18].

As we will see, in our Bayes and Bayes linear analyses of the EA data, part of our prior beliefs about some parameters of interest, namely regression parametes, derives from Bayesian analysis of Gustafson's data with a non-informative prior. In the case of linear regression models, using non-informative prior information and normal errors, results in least squares analysis; see [5], pages 146 and 154-155. Also, in the case of generalized linear models, using non-informative prior information, results in maximum likelihood analysis; see [29], page 104.

In practice, we may need to calculate the marginal posterior distributions of the parameters of interest. This computation may require a high dimensional integration which could be intractable analytically. In this case, simulation techniques such as the Markov Chain Monte Carlo method (MCMC) can be used to draw samples from the posterior distribution to approximate these marginal posterior distributions. As MCMC simulation is used extensively in this thesis, we give a brief description below.

A useful tool in Bayesian methods employs graphical models to represent the dependence structure among variables in a probability distribution, making the Bayesian inference straightforward without the need for algebraic manipulation of multivariate distributions. Graphical models and how to implement them using the WinBUGS software package will be discussed. We also discuss model selection, a statistical tool used to compare different models.

2.4.1 Markov Chain Monte Carlo

As mentioned above, Bayesian methods combine a prior distribution for unknowns with the study data represented in the form of likelihood. The result of this combination is a posterior distribution on which inferences about the unknowns are based. The posterior distribution is proportional to the product of the likelihood function and the prior distribution. In order to inquire into the parameters of interest and draw inferences, it is necessary to evaluate the marginal posterior distributions of these parameters. Computation of these marginal distributions often require high dimensional integration that is not always available in a closed form, making the performance of such calculations analytically impossible. However, these difficulties can be overcome by adopting approximation or simulation methods. This work concentrates on the use of simulation methods, in particular the MCMC method. MCMC is the simulation technique most widely used to handle such complex computations that can not be performed analytically. The main purpose of MCMC is to explore the posterior distributions of the parameters of interest by drawing or generating samples from marginal posterior distributions which can be used to describe or obtain specific information about these parameters. There are several methods or algorithms for MCMC. The most widely used method is the Metropolis-Hastings algorithm: most of the others are specific modifications of this method, such as the Gibbs sampler. The main idea behind MCMC is the construction of a stationary distribution with limiting distribution converging to the target posterior distribution.

As the Bayesian methods in this thesis are implemented via MCMC, a brief description of the most general algorithm, Metropolis-Hastings, is given below, followed by a brief description of the Gibbs sampler used in the WinBUGS software [40].

2.4.2 Metropolis-Hasting algorithm

The Metropolis-Hastings algorithm (M-H) is the most general approach used to draw samples from a posterior distribution. Let θ denote a parameter with posterior distribution $p(\theta|\boldsymbol{y})$ known up to a proportionality constant. The M-H algorithm can be implemented as follows; see, for example, [18] and [22].

1. initialize θ by starting at some value $\theta^{(0)}$.

2. For the current state $\theta^{(t)}$ at iteration t, where $t \geq 1$, generate a candidate value θ^{*} from a transition proposal distribution $q(\theta^{*}|\theta^{(t-1)})$.

3. Calculate the following ratio of densities:

$$r(\theta^{(t-1)}, \theta^{*}) = \frac{q(\theta^{(t-1)}|\theta^{*})p(\theta^{*}|\boldsymbol{y})}{q(\theta^{*}|\theta^{(t-1)})p(\theta^{(t-1)}|\boldsymbol{y})} \tag{2.35}$$

4. Calculate $\alpha = \min\left\{1, r(\theta^{(t-1)}, \theta^{*})\right\}$

5. Generate a uniform random quantity $U \in [0, 1]$.

6. Set

$$\theta^{(t)} = \begin{cases} \theta^{*} & \text{if } U < \alpha \\ \theta^{(t-1)} & \text{otherwise} \end{cases}$$

The choice of the transition proposal distribution $q(.)$ is arbitrary and can be chosen so that the convergence to the target distribution can be reached quickly.

A specific example of the M-H algorithm is the Metropolis algorithm, where the proposed distribution is chosen to be symmetric, i.e, $q(\theta^{(t-1)}|\theta^{*}) = q(\theta^{*}|\theta^{(t-1)})$. In this case, the ratio of densities, expressed in step (3) above, simplifies to

$$r(\theta^{(t-1)}, \theta^{*}) = \frac{p(\theta^{*}|y)}{p(\theta^{(t-1)}|y)} \qquad (2.36)$$

For example, the candidate value θ^{*} can be generated from a normal distribution with the mean $\theta^{(t-1)}$ and variance σ^2. In this case σ^2 acts as a tuning parameter.

Another example of the M-H algorithm is the Gibbs sampler algorithm which is the main algorithm used in the WinBUGS software. There follows a brief description of this algorithm.

2.4.3 The Gibbs sampler

Much of this is taken from [22] and [18]. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ denote a p length vector of parameters to be estimated, having a joint posterior density $p(\boldsymbol{\theta}|\boldsymbol{y})$. The Gibbs sampler is used to estimate posterior distributions by generating random samples from the full conditional distributions of each parameter given the rest of the parameters and the data. The general algorithm for this sampling method is as follows. Let $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_p^{(t)})$ denote the current state of the chain or the sampled value of $\boldsymbol{\theta}$ at iteration t. Now, the value of $\boldsymbol{\theta}$ at iteration $(t+1)$, $\boldsymbol{\theta}^{(t+1)}$, is obtained by drawing from the following conditional distributions:

$$\text{draw } \theta_1^{(t+1)} \text{ from } p_1(\theta_1|\theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)}, \boldsymbol{y})$$

$$\text{draw } \theta_2^{(t+1)} \text{ from } p_2(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_p^{(t)}, \boldsymbol{y})$$

$$\text{draw } \theta_i^{(t+1)} \text{ from } p_i(\theta_i|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{i-1}^{(t+1)}, \theta_{i+1}^{(t+1)}, \ldots, \theta_p^{(t)}, \boldsymbol{y})$$

$$\text{draw } \theta_p^{(t+1)} \text{ from } p_p(\theta_p|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{p-1}^{(t+1)}, \boldsymbol{y})$$

Thus, each parameter is updated, conditional on the latest values of the other parameters. Hence, the new sampled values or the new state at iteration t is $\boldsymbol{\theta}^{(t+1)} = (\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_p^{(t+1)})$. Establishing that the Gibbs sampler algorithm is a special case of the M-H algorithm is straightforward and appears in many references, such as [18].

There are several practical considerations and concepts regarding the implementation of MCMC methods. Amongst these issues are thinning the chain, the burn-in period, the starting points and convergence diagnostics. There follows a brief explanation of these concepts.

The burn-in period refers to the number of iterations before convergence of the chain is achieved: the burn-in values are discarded.

Thinning means running the chain normally but recording only every k^{th} value [40]. Thinning the chain is usually used to reduce the autocorrelation for a parameter.

Starting points or initial values, as referred to in WinBUGS, are associated with convergence diagnostics especially when assessing convergence by running multiple chains from different initial values.

Convergence tests refer to formal and informal tools used to assure that the chain has reached its stationary or target distribution. There are several convergence diagnostics but the focus here is on those practical or graphical tests built in to WinBUGS that are similar to the modified Gelman-Rubin diagnostic proposed in [19], autocorrelation, history or trace of the chain and kernel density estimates of posterior distributions. The following is a brief discussion of these.

Autocorrelation refers to the correlation between consecutive simulated realisations for a single parameter. High autocorrelation may cause a slow convergence to the target distribution and this can be reduced by thinning the chain. This test is used informally.

The history of the chain provides a visual assessment of convergence. It is simply obtained by plotting the sampled value, $\boldsymbol{\theta}^{(t)}$, of the parameter against the iteration number t. Convergence may be informally assessed by looking at simulation trends and in case of running more than one chain, history plots should appear to mix rapidly and overlap when they are depicted in one plot.

A diagnostic test to assess convergence efficiency, developed by Gelman and Rubin in [19] and modified by Brooks and Gelman in [6] is referred to in the WinBUGS manual [40] and here as the BGR diagnostic. This is a more formal test which has been reviewed and summarized in many books and papers, such as [22], on which much of the following discussion is based. The test uses within chain variance, pooled variance and their ratios. An understanding of how this test is constructed helps to define these concepts. BGR is set up by running several independent parallel chains using widely dispersed initial values following the steps presented in [22] as follows.

1. Run $m \geq 2$ chains of length $2n$ from widely dispersed initial values $\theta_{(1)}^{[0]}$, $\theta_{(2)}^{[0]}$, $\ldots, \theta_{(m)}^{[0]}$:

$$\begin{array}{c} \theta_{(1)}^{[0]}, \theta_{(1)}^{[1]}, \ldots, \theta_{(1)}^{[2n-1]}, \theta_{(1)}^{[2n]} \\ \theta_{(2)}^{[0]}, \theta_{(2)}^{[1]}, \ldots, \theta_{(2)}^{[2n-1]}, \theta_{(2)}^{[2n]} \\ \vdots \\ \theta_{(m)}^{[0]}, \theta_{(m)}^{[1]}, \ldots, \theta_{(m)}^{[2n-1]}, \theta_{(m)}^{[2n]} \end{array}$$

Now, discard the first n chain iterations for each of the m chains.

2. For each parameter of interest calculate the following:

- Within chain variance:

$$W = \frac{1}{m(n-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\theta_{(j)}^{[i]} - \bar{\theta}_{(j)} \right)^2$$

where $\bar{\theta}_{(j)}$ is the mean of the n values for the j^{th} chain.

- Between chain variance:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\bar{\theta}_{(j)} - \bar{\bar{\theta}} \right)^2$$

where $\bar{\bar{\theta}}$ is the grand mean (the mean of the means).

- Estimated variance (pooled variance):

$$\widehat{\mathrm{Var}}(\theta) = (1 - 1/n)\, W + (1/n)\, B$$

3. Evaluate the following ratio:

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\mathrm{Var}}(\theta)}{W}}$$

4. If the monitored values of $\sqrt{\hat{R}}$ are close to 1 so that W is approximately equal to $\widehat{\text{Var}}(\theta)$, then this indicates that convergence has been reached.

As explained in the WinBUGS manual [40], the BGR diagnostic test can be displayed in a plot of three quantities (a) the normalized width of the central 80% interval of the pooled runs, coloured in green, (b) the normalized average width of the 80% intervals within the individual runs, coloured in blue and (c) their ratio test R=(pooled/within), coloured in red. To judge MCMC convergence, R should converge to 1 and the pooled and within interval widths should stabilise.

The convergence diagnostic tests described above can be carried out using R packages such as CODA (Convergence Diagnostics Analysis) and BOA (Bayesian Output Analysis); see [35] for more details.

The following section discusses graphical models and their use in Bayesian statistics.

2.4.4 Graphical Models

Graphical models can be used to represent the dependence structure among variables in a probability distribution and via them Bayesian inference becomes straightforward without the need to carry out algebraic derivations.

The graphical models used here are directed acyclic graphs (DAG), which are directed graphs where there is no path from a node to itself. A DAG is usually displayed using plates, ellipses, rectangles and arrows. Plates are used to represent the levels of the model, ellipses to represent observed and unobserved variables and rectangles to represent constants. There are two types of arrows: solid and hollow.

The solid arrows represent stochastic links and the hollow ones are used for logical links.

The following definitions are important in Bayesian analysis and are needed here when a DAG is constructed to represent the joint posterior distribution for a logistic regression model with uncertain covariates to be described in the following section. These definitions are summarised from [13]

Definition 2.1: A graph consists of a set of nodes V and a set of edges E where an edge in E may be directed or undirected.

Definition 2.2: For a DAG, the parents (denoted by pa) of a node $v \in V$ are:

$$pa(v) = \{w \in V : (w, v) \in E\} .$$

Definition 2.3: In a DAG, the children (denoted by $ch(v)$) of v are:

$$ch(v) = \{w \in V : (v, w) \in E\}$$

and the non-descendants (denoted by $nd(v)$) of v are:

$$nd(v) = \{w \in V : \text{no path from } v \text{ to } w\} .$$

An important benefit of using a DAG is that the joint distribution of all variables V can be represented using the following factorization:

$$f(V) = \prod_{v \in V} f\left(v|pa(v)\right)$$

Another important advantage is that it is easy to derive the full conditional distribution (mainly required for the Gibbs sampler) of any variable $v \in V$ conditioning on the other variables $(V \backslash v)$ using the following equation:

$$f(v|V \backslash v) \propto f(v|pa(v)) \prod_{w \in ch(v)} f\left(w|pa(w)\right)$$

The following discussion focuses on the implementation of graphical models using the WinBUGS package.

2.4.5 Implementation of Graphical Models Using WinBUGS

The major steps needed to implement graphical models using WinBUGS are:

1. Constructing the graphical model.

2. Assigning a full probability distribution to all of the stochastic nodes. In fact, implementation of the Gibbs sampler as an MCMC simulation technique requires identifying the full conditional posterior distribution for each parameter of interest, which may not be easy, especially obtaining it in a closed form. Fortunately, the WinBUGS software performs this automatically without the need to derive the forms of the conditional posterior distribution. So, what is really needed is to assign a probability distribution to each of the stochastic nodes.

3. Specifying the number of chains needed to run the Gibbs sampler, which is important in assessing convergence.

4. Specifying different initial values for each chain.

5. Checking convergence to the target distribution.

6. Extracting the simulated values and looking at the desired statistics and estimates.

2.4.6 Joint posterior distribution representation for logistic regression with uncertain covariates using a DAG

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ be a data set of n independent and identically distributed binary observations. Assigning y_i a Bernoulli distribution with probability of success $p(y_i = 1|\boldsymbol{z}_i, \boldsymbol{\beta}) = \pi_i$ leads to the following likelihood for $\boldsymbol{\beta}$ and the $\boldsymbol{z}_i$

$$p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{z}_i, \boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

where

$$\pi_i = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}$$

where the structure of $\boldsymbol{Z}$, the $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ is given in section 2.3.1.

To derive a representation for the joint posterior distribution of $\boldsymbol{Z}$ and $\boldsymbol{\beta}$, we assume, as in our application in this thesis, that $\boldsymbol{Z}$ and $\boldsymbol{\beta}$ are uncorrelated and also that the n rows of $\boldsymbol{Z}$ are independent; i.e.

$$p(\boldsymbol{Z}) = p(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n) = p(\boldsymbol{z}_1)p(\boldsymbol{z}_2)\ldots p(\boldsymbol{z}_n)$$

We will also assume,

$$p(\boldsymbol{z}_i) = p(z_{i1}, z_{i2}, \ldots, z_{iq}) = p(z_{i1})p(z_{i2})\ldots p(z_{iq})$$

for all $i = 1, 2, \ldots, n$. Thus all nq components of $\boldsymbol{Z}$ are assumed to be a priori independent.

As shown in the DAG display in Figure 2.2, the joint posterior distribution of $\boldsymbol{Z}$

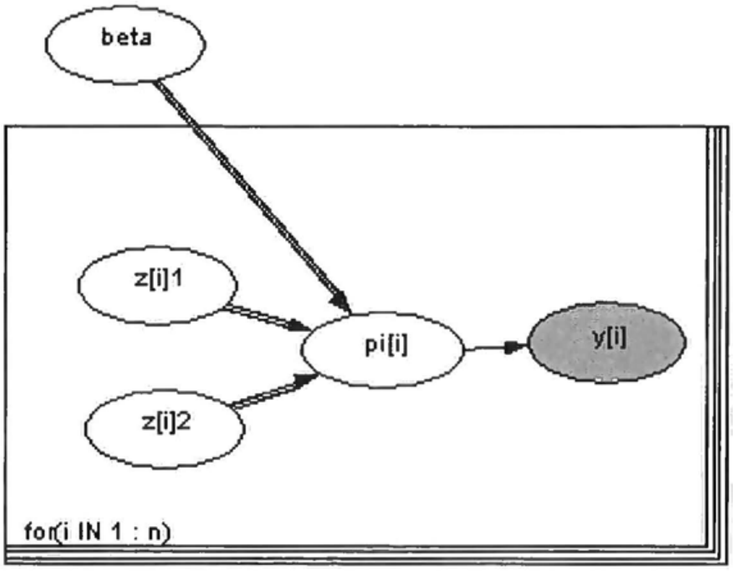

Figure 2.2: Directed acyclic graphs (DAG) for logistic regression with uncertain covariates

and $\boldsymbol{\beta}$ given the data $\boldsymbol{y}$ can be expressed as

$$\begin{aligned} p(\boldsymbol{Z},\boldsymbol{\beta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{Z},\boldsymbol{\beta})p(\boldsymbol{Z},\boldsymbol{\beta}) \\ &\propto p(\boldsymbol{y}|\boldsymbol{Z},\boldsymbol{\beta})p(\boldsymbol{Z})p(\boldsymbol{\beta}) \\ &\propto p(\boldsymbol{y}|\boldsymbol{Z},\boldsymbol{\beta})p(\boldsymbol{z}_1)p(\boldsymbol{z}_2)\dots p(\boldsymbol{z}_n)p(\boldsymbol{\beta}) \\ &\propto \prod_{i=1}^{n}\left[p(y_i|z_{i1},z_{i2},\dots,z_{iq},\boldsymbol{\beta})p(z_{i1})p(z_{i2})\dots p(z_{iq})\right]p(\boldsymbol{\beta}) \end{aligned}$$

The next section, which discusses the multivariate runs test, uses some of the above assumptions.

2.5 Multivariate runs test

The aim of the multivariate runs test, proposed by Friedman and Rafsky in [17], is to test the null hypothesis of whether two groups are drawn from the same distribution. The multivariate runs test is used in this thesis as a tool (a) to strengthen the derived results by testing the degree of separation between leaching and non-leaching pesticides and (b) to compare different models.

The following definitions are needed in order to describe this test. These definitions can be found in many references, such as [26] and [17], from which many of the following are taken.

Definition 2.4: A tree is a connected graph with no cycles.

Definition 2.5: Let G be a graph. A spanning tree T is a connected subgraph of G with no cycles and contains all the vertices of G. The length of T, $R(T)$, is the total number of edges in T.

Definition 2.6: A minimal spanning tree (MST) is a spanning tree T with $R(T) <$

$R(T')$ for all spanning trees T'.

To describe the test, let $X_m = (x_1, x_2, \ldots, x_m)$ and $Y_n = (y_1, y_2, \ldots, y_n)$ be two independent samples from distributions F_x and F_y, respectively. Hence, a graph $G(X_m, Y_n)$ can be formed using all of the vertices $x_1, x_2, \ldots,\ x_m,\ y_1, y_2, \ldots, y_n$ and edges between these vertices. A spanning tree $T(X_m, Y_n)$ is any connected subgraph of $G(X_m, Y_n)$, containing all of the above vertices, and it does not contain any cycle. Let $R(X_m, Y_n)$ denote the number of edges of $T(X_m, Y_n)$ which connect a point of X_m to a point of Y_n.

Now, the null hypothesis to be tested is

$$H_0 : F_x = F_y$$

against the general alternative hypothesis

$$H_1 : F_x \neq F_y$$

The Friedman and Rafsky test statistic $R_{m,n}$ is given by

$$R_{m,n} = R(X_m, Y_n)$$

They conjecture that 'small' values of $R_{m,n}$ lead to a rejection of the null hypothesis. Thus, models with small values of $R_{m,n}$ may reflect good separation between leaching and non-leaching pesticides than models with large values of $R_{m,n}$.

A further important step is the assessment of model complexity and comparing different models, which is the focus of the next section.

2.6 Model selection

There are several tools for comparing and assessing the fit of the different models. Among these are the Deviance Information Criterion (DIC), stepwise selection and the likelihood ratio statistic. What follows are brief descriptions of these tools.

2.6.1 Deviance Information Criterion

This tool as proposed in [39] and implemented in WinBUGS can be described briefly using the following definitions which are summarized from the WinBUGS manual [40]. DIC is defined as:

$$DIC = \bar{D} + PD = \hat{D} + 2PD$$

where $\bar{D}$ is the posterior mean of the deviance $-2 *$ log(likelihood), $\hat{D}$ is the point estimate of the deviance obtained by substituting in the posterior means of parameters: thus $\hat{D} = -2 * \log p(y|\text{posterior means of parameters})$, and PD is the effective number of parameters calculated as $PD = \bar{D} - \hat{D}$. The values of DIC are important when comparing different models for the same problem, which is the case in this thesis. Models with small values of DIC indicate better fits. The program for calculating DIC in R is given in Appendix A.1.1.

2.6.2 Stepwise selection

The objective is to select the most relevant model terms which give best prediction. It can be implemented in three options: forward, backward and in both directions. In forward selection, we start by selecting a single model term which provides the best fit to the data according to a predictive criterion; see below. Then, each model term is examined to see if adding it will significantly improve the overall fit. In backward selection, we start with the full model which includes all of the model terms. Then, each model term is removed and tested to see whether its removal will significantly improve the overall fit. The option "both directions", allows inclusion of model terms using either forward or backward selection at each stage according to a significant improvement of including or deleting the current model term.

The Akaike information criterion (AIC) is used judge the adequacy of the selected model in the stepwise procedure. The AIC is defined, see [34], to be

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + 2p \tag{2.37}$$

where $l(\hat{\boldsymbol{\beta}})$ is the log-likelihood maximum, $\hat{\boldsymbol{\beta}}$ is MLE and p is the dimension of $\boldsymbol{\beta}$. The smaller the value of AIC, the better the model fits the data. The stepwise procedure is stopped if adding or eliminating a model term from the current sub-model will increase the value of AIC.

2.6.3 Likelihood ratio statistic

Much of following is summarised from [14] and [34]. The log likelihood ratio statistic or deviance is used to compare different models and to select the most adequate from available models. The model with the maximum number of parameters that can be estimated is called the saturated model. As in [14], let m be the number of such parameters and $\boldsymbol{\beta}_{\text{max}}$ be the parameter vector for the saturated model with $\boldsymbol{b}_{\text{max}}$ as the maximum likelihood estimator. Let $\boldsymbol{\beta}$ be the parameter vector for the model of interest with n parameters and $\boldsymbol{b}$ as the maximum likelihood estimator. Then, the likelihood ratio (LR) is

$$LR = \frac{l\left(\boldsymbol{b}_{\max}, \boldsymbol{y}\right)}{l(\boldsymbol{b}, \boldsymbol{y})} \tag{2.38}$$

where $l()$ denotes the likelihood function. The logarithm of LR is

$$\log LR = L\left(\boldsymbol{b}_{\max}, \boldsymbol{y}\right) - L(\boldsymbol{b}, \boldsymbol{y}) \tag{2.39}$$

where $L()$ is the log-likelihood function. Large values of $\log LR$, indicate poor fit of the model of interest to the data.

The deviance (D) is $2 \log LR$, i.e.

$$D = 2\left[L\left(\boldsymbol{b}_{\max}, \boldsymbol{y}\right) - L(\boldsymbol{b}, \boldsymbol{y})\right] \tag{2.40}$$

It can be shown, see, for example, [14], that D has an approximate chi-squared distribution with $m - n$ degrees of freedom, i.e.

$$D \sim \chi^2(m - n) \tag{2.41}$$

This result can be used to calibrate model adequacy.

2.7 Statistical packages

Two statistical packages have been used to analyse the data and implement the models developed in this thesis, R and BUGS. R, which provides an environment in which to perform the statistical analysis, is freely available at `http://www.r-project.org`. BUGS (Bayesian inference Using Gibbs Sampling) is used to analyse complex Bayesian models using MCMC simulation. WinBUGS, the Windows version of BUGS, used in this thesis is freely available at `http://www.mrc-bsu.cam.ac.uk/bugs`. In addition, packages such as CODA (Convergence Diagnostics Analysis), BOA(Bayesian

Output Analysis) and R2WinBUGS, which are available in R, were also used. The R2WinBUGS package was used to run WinBUGS from R; see [20].

2.8 Conclusion

This chapter has provided discussions of the statistical concepts and techniques used throughout the thesis. It reviewed logistic regression models and the MLE of the parameters. In this regard, a discussion about one of the logistic regression deficiencies which arose during this research and how to tackle it is provided. This deficiency appears when fitting data with complete separation between successes and failures in the covariates space, where the MLE does not exist. Alternatives such as the MEL and WEMEL estimates can be used.

Another important concept reviewed in this chapter is Bayes linear estimation. A general formula for computing the Bayes linear estimate for a general linear model with uncertain covariates is derived.

Full Bayesian inference is discussed. The discussion includes MCMC simulation including its most general implementation via the Metropolis-Hastings algorithm. It also includes a brief description of graphical models and how they can be implemented using the WinBUGS software package. A general formula to represent the joint posterior distribution for logistic regression with uncertain covariates using a DAG is given. there is a brief discussion about the diagnostic tests used to assess convergence of MCMC algorithms.

The Deviance Information Criterion (DIC), proposed in [39] and implemented in WinBUGS, stepwise selection procedures and the likelihood ratio statistic, which will be used as comparison tools in this thesis, have been described. Finally, the statistical packages used in this study are referenced.

Chapter 3

Discrimination using a model with an interaction term

3.1 Introduction

This chapter investigates the Bayesian method proposed in reference [44] applied to a model with an interaction term. It was suggested in [44] that the use of logistic regression with a linear predictor non-linear in the covariates could serve to discriminate between leaching and non-leaching pesticides. This idea is formulated here using an interaction term in the linear predictor. The investigation is straightforward except for one difficulty which has arisen in trying to fit logistic regression using maximum likelihood to the interaction model. Fitting the logistic regression model to Gustafson's data using an interaction model leads to non-overlapping groups of leachers and non-leachers so that the MLE does not exist. This problem is tackled here by firstly measuring the overlap using the depth-based algorithm proposed in [7], which showed that there is a complete separation in the covariate space of

Gustafson's data. Secondly, the MEL estimator, which completely eliminates the overlap problem, and the WEMEL estimator, which is also robust against outliers, were used as alternative estimators. These alternative estimators were proposed in [8].

In addition, an alternative Bayesian analysis will be proposed here. The analysis benefits from 'direct' use of lysimeter experiments and logistic regression methods to predict the potential of a given pesticide to leach.

3.2 Formulation of the discriminant model

The idea of improving discrimination by using a non-linear predictor in logistic regression, such as a curve drawn to discriminate between leaching and non-leaching pesticides, was suggested in [44]. The model uses the same sources of data from lysimeter experiments, and hence the same likelihood, the same prior information and a similar method to generate the parameters of the prior distribution. The only difference is the logistic regression used to predict pesticide leachability.

There are several specifications and considerations that should be taken into account in formulating a statistical model. Among these are two questions that should be answered (a) which covariates are needed and (b) which model terms involving these covariates should be included. These questions should not be answered without taking into account other consideration, in particular the use to which the model will be put. In our application, the model will be used to discriminate between leaching and non-leaching pesticides, i.e. it is constructed for predictive purposes.

As discussed in Chapter 1, certain chemical properties are believed to control the potential for a given pesticide to leach into the soil and pollute the groundwater. Among these properties are adsorption coefficient (k_{oc}), soil half-life ($t_{1/2}^{soil}$) and water solubility (S_{H2O}). Now, we discuss the choice of a statistical model to be used for discrimination of pesticides as leachers or non-leachers on the basis of the above chemical properties..

The Gustafson data is used and a logistic regression model, as described in Chapter 2, is adopted with

$$y_i = \begin{cases} 1 & \text{if ith pesticide is monitored as a leacher} \\ 0 & \text{if ith pesticide is monitored as a non-leacher.} \end{cases}$$

and the covariates are $z_1 = \log k_{oc}$, $z_2 = \log t_{1/2}^{soil}$ and $z_3 = \log S_{H2O}$

The stepwise procedures, described in Chapter 2, are used to select the covariates and the most important model terms to be included in the linear predictor. We begin with a full model containing z_1, z_2, z_3 and their interactions, i.e. the linear predictor η is

$$\eta = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_1 z_2 + \beta_5 z_1 z_3 + \beta_6 z_2 z_3 + \beta_7 z_1 z_2 z_3 \tag{3.1}$$

First of all, to fit this model, the amount of leacher/non-leacher overlap, as described in Section 2.2.1, in covariates space needs to be measured. The use of the depth-based algorithm in [7] shows that there is complete separation between leachers and non-leachers in the space of the covariates z_1, z_2, z_3, so that the MLE of the parameters in 3.1 does not exist. Therefore, we adopt an alternative estimator. The alternatives we consider are MEL, which resolves the separation problem, and its robust version WEMEL, which is robust against outliers.

The objective now is to find the best model to predict the potential of a given

Model	Df	Deviance	Change	P-value
Null	21	21.847		
z_1	20	10.969	10.878	0.001
z_2	19	5.990	4.979	0.026
z_3	18	5.953	0.037	0.848
$z_1 : z_2$	17	1.826	4.127	0.042
$z_1 : z_3$	16	1.344	0.482	0.487
$z_2 : z_3$	15	1.036	0.308	0.579
$z_1 : z_2 : z_3$	14	0.973	0.063	0.801

Table 3.1: Analysis of deviance where terms are added sequentially (first to last).

pesticide to leach using the fewest number of terms. The stepwise selection procedure in "both directions", using, for example, the function step in R with AIC, leads to a final model with an intercept, z_1, z_2 and the interaction term z_1z_2. This model has estimated logit

$$\text{estimated logit}(\pi|z_1, z_2) = -22.762 + 0.664z_1 + 17.610z_2 - 2.045z_1z_2 \tag{3.2}$$

The estimated model and the analysis of deviance summarised in Table 3.1 suggest that the covariate z_3 (logs of solubility) and its interaction terms with the other covariates are not useful discriminants, as including the model terms z_3, z_1z_3, z_2z_3 and $z_1z_2z_3$ does not cause significant change in deviance. This conclusion is consistent with Gustafson's suggestion in [25] that water solubility is not relevant in discriminating leachers and non-leachers pesticides.

The signs of the estimates of the coefficients of z_1 and z_2 are consistent with Gustafson's contention that leaching pesticides are those with low k_{oc} and high $t_{1/2}^{soil}$

Effect	Estimate	Standard error	z-value	P-value
Intercept	-22.762	21.706	-1.049	0.294
logkoc	0.664	3.127	0.212	0.832
logt	17.610	12.882	1.367	0.172
logkoc:logt	-2.045	1.608	-1.272	0.203

Table 3.2: Regression coefficients estimates for the model as selected via the stepwise procedure.

values and non-leaching pesticides are those with low $t_{1/2}^{soil}$ and high k_{oc} values. This can be explained using either η or the estimated leacher/non-leacher *odds*

$$odds = \exp(-22.762)\exp(0.664z_1)\exp(17.610z_2)\exp(-2.045z_1z_2) \qquad (3.3)$$

For example, fixing the covariate z_2 at a small value and letting the covariate z_1 vary over its range decreases the odds and linear predictor η, and hence decreases the leaching probability as depicted in Figures 3.1 (a) and (b). This is consistent with Gustafson's contention that non-leaching pesticides are those with low $t_{1/2}^{soil}$ and high k_{oc} values. Also, fixing the covariate z_1 at a small value and letting the covariate z_2 vary over its range increases the odds and linear predictor η, and hence increases the leaching probability as depicted in Figures 3.1 (c) and (d). These observations are consistent with Gustafson's contention that leaching pesticides are those with low k_{oc} and high $t_{1/2}^{soil}$ values.

Table 3.2, which displays some statistical summaries for the model in 3.2, shows that none of the model terms are significant. Moreover, the z_1 main effect differs from zero by only 0.212 standard errors. Taking this into account and the form, equation 1.2, proposed by Gustafson in [25] for discrimination, we consider the

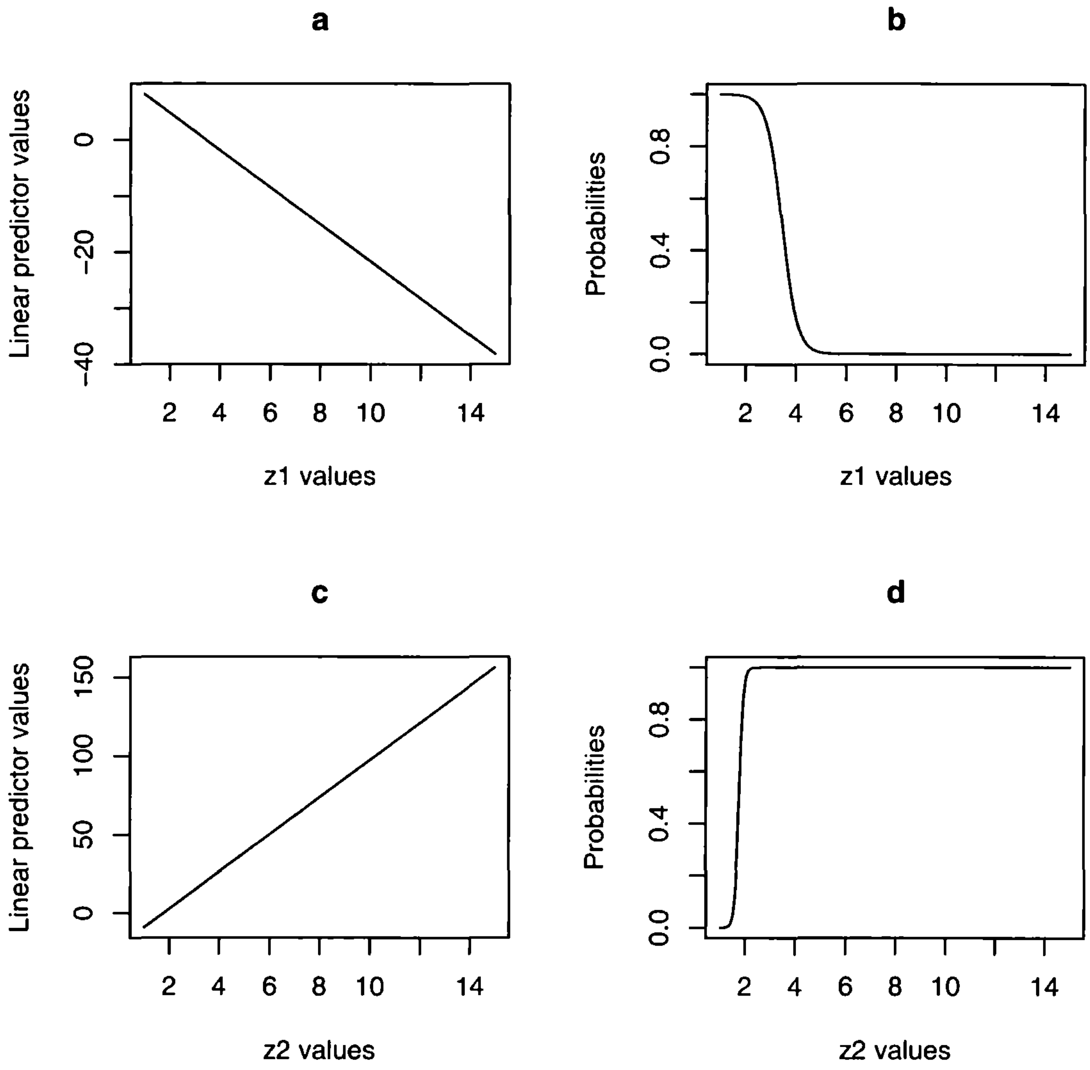

Figure 3.1: Plots of linear predictor η and leaching probability. In (a) and (b) the covariate z_2 is fixed at a certain value and the covariate z_1 varies over a range of values. In (c) and (d) the covariate z_1 is fixed at a certain value and the covariate z_2 varies over a range of values.

Effect	Estimate	Standard error	z-value	P-value
Intercept	-20.384	14.381	-1.417	0.156
logt	17.380	11.593	1.499	0.134
logkoc:logt	-1.947	1.275	-1.527	0.127

Table 3.3: Regression coefficients estimates after eliminating the term z_1.

further step of removing the term z_1 from 3.2. This leads to

$$\text{estimated logit}(\pi|z_1, z_2) = -20.384 + 17.380z_2 - 1.947z_1z_2 \tag{3.4}$$

Table 3.3 displays statistical summaries for the reduced model in 3.4. We will use this model rather than the model in 3.2, although it is not in general recommended to include the higher degree term z_1z_2 without including both z_1 and z_2; see [15]. However, we justify our choice as follows.

1. The objective of constructing the model is to use it to discriminate or predict the leachability of a given pesticide, not to select an optimal model to fit the data.

2. It is apparent from both Tables 3.2 and 3.3 that none of the model terms are significant.

3. It is apparent from Table 3.2 that the magnitude of coefficient estimate for z_1 is small relative to the other estimates and it is much smaller relative to its standard error.

4. Omitting the term z_1 from the model in 3.2 decreases the variances of regression coefficient estimates of the other terms; see Table 3.3.

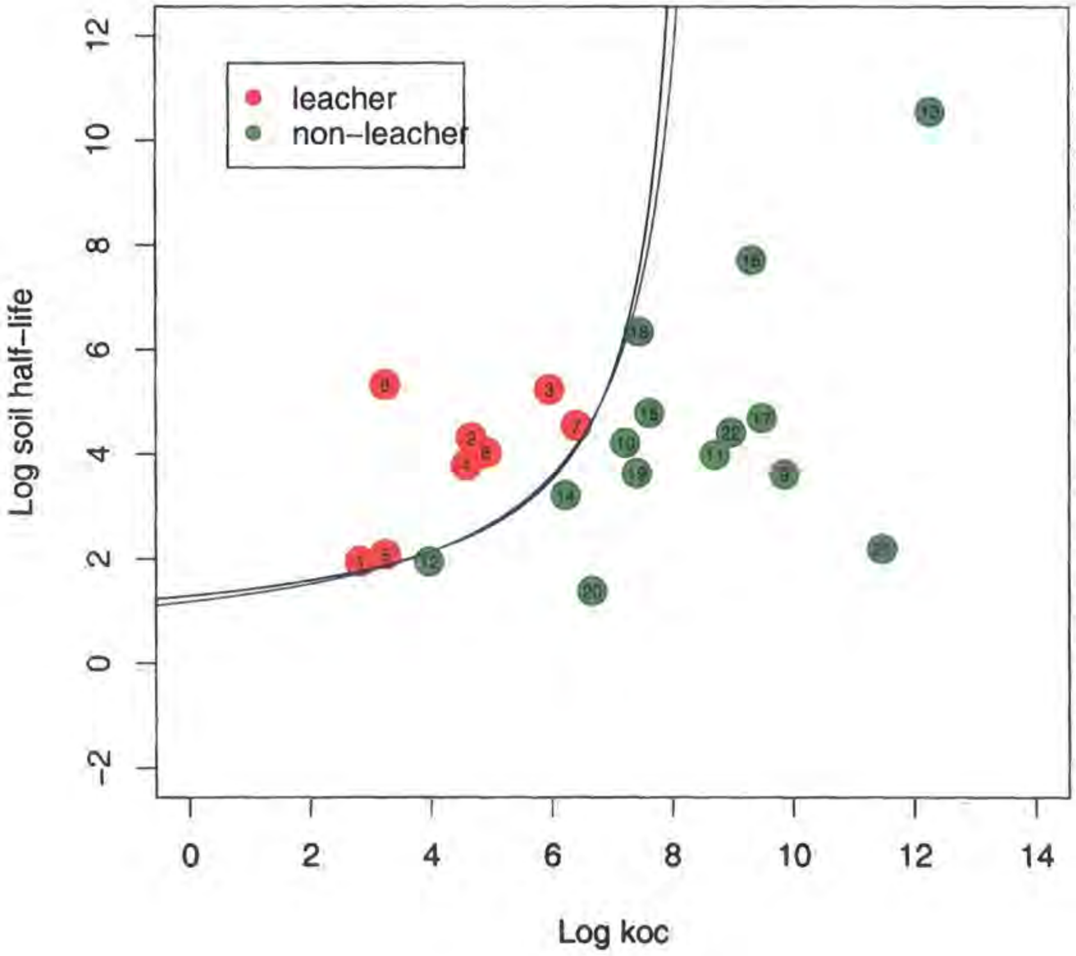

Figure 3.2: Scatter plot of $z_2 = \log t_{1/2}^{soil}$ against $z_1 = \log k_{oc}$ together with discriminant curves using the logistic regression model. The black curve is the discriminant curve for Model 3.2 as selected via the stepwise procedure. The blue curve is discriminant curve for Model 3.4 after eliminating the term z_1. The two curves give similar predictive results.

5. The model 3.4 is consistent with the form of Gustafson's GUS curve in 1.2.

6. The discriminant curves

$$-22.762 + 0.664z_1 + 17.610z_2 - 2.045z_1z_2 = 0 \tag{3.5}$$

and

$$-20.384 + 17.380z_2 - 1.947z_1z_2 = 0 \tag{3.6}$$

depicted in Figure 3.2 are almost similar. Furthermore, both models give almost equal estimated predictive probabilities.

3.2.1 Model checking

After selecting the model, there are still essential processes that should be performed. These are, checking the adequacy of the fitted model and studying robustness of the results. There are several techniques to check the fitted models such as residual patterns and to study robustness such as influence of observations.

The residuals r can be plotted either against fitted values or covariates to detect any deficiency in the fitted model. The null pattern (random scatter) indicates that the fit is good and there is no relation between residuals and fitted values or covariates.

Robustness can be assessed using measures of leverage and influence. A measure of leverage of observation i is given by the diagonal entry h_i of the hat matrix

$$H = W^{-1/2}X(X^TWX)^{-1}X^TW^{-1/2} \tag{3.7}$$

where X is the model matrix and $W$is the prior weight matrix. An observation i with h_i greater than two or three times p/n, where p is the length of the vector $\boldsymbol{\beta}$, might be a potentially influential point, see [14] page 89. The standardized residual sr_i is

$$sr_i = \frac{r_i}{\sqrt{\hat{\phi}(1 - h_i)}} \tag{3.8}$$

where $\hat{\phi}$ is an estimate of the dispersion. An observation i with sr_i "greater in magnitude than 2 or 3 will suggest a possible outlier" [38].

Cook's distance d_i can be used to measure the influence of observation i where

$$d_i = \frac{1}{p} \cdot sr_i \cdot \frac{h_i}{1 - h_i} \tag{3.9}$$

where p is the length of the vector $\boldsymbol{\beta}$. Figure 3.3 plots the standardized residuals against the covariates, the fitted values and the quantiles of the standard normal distribution of Model 3.4. These figures indicate that

1. The plot of the standardized residuals against the covariates $z_1 = \log k_{oc}$, $z_2 = \log t_{1/2}^{soil}$ and the fitted values have the null pattern indicating a good model fit.

2. The q-q plot shows that the standardized residuals have, approximately, a standard normal distribution.

However, such plots are "generally uninformative for binary data because all the points lie on one of two curves according as $y = 0$ or $y = 1$" [33].

Figure 3.4 plots leverage values, standardized residuals and Cook's distance. These figures show that

1. The leverage plot suggests that observations 7, 12 and 18 are possibly potentially influential points.

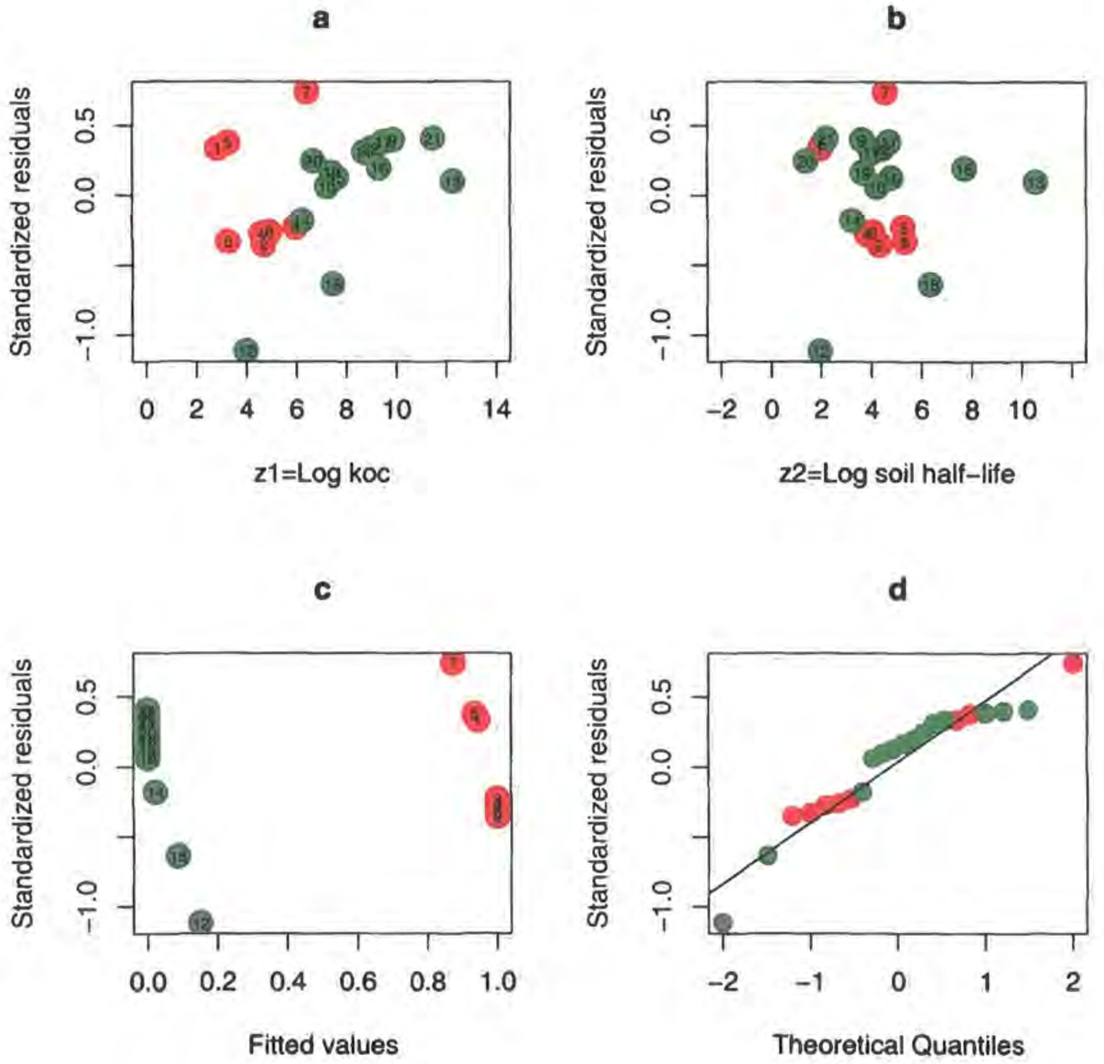

Figure 3.3: For model 3.4, plots of (a) standardized residuals against $z_1 = \log k_{oc}$, (b) standardized residuals against $z_2 = \log t^{soil}_{1/2}$, (c) standardized residuals against fitted values, and (d) standardized residuals against the quantiles of the standard normal distribution.

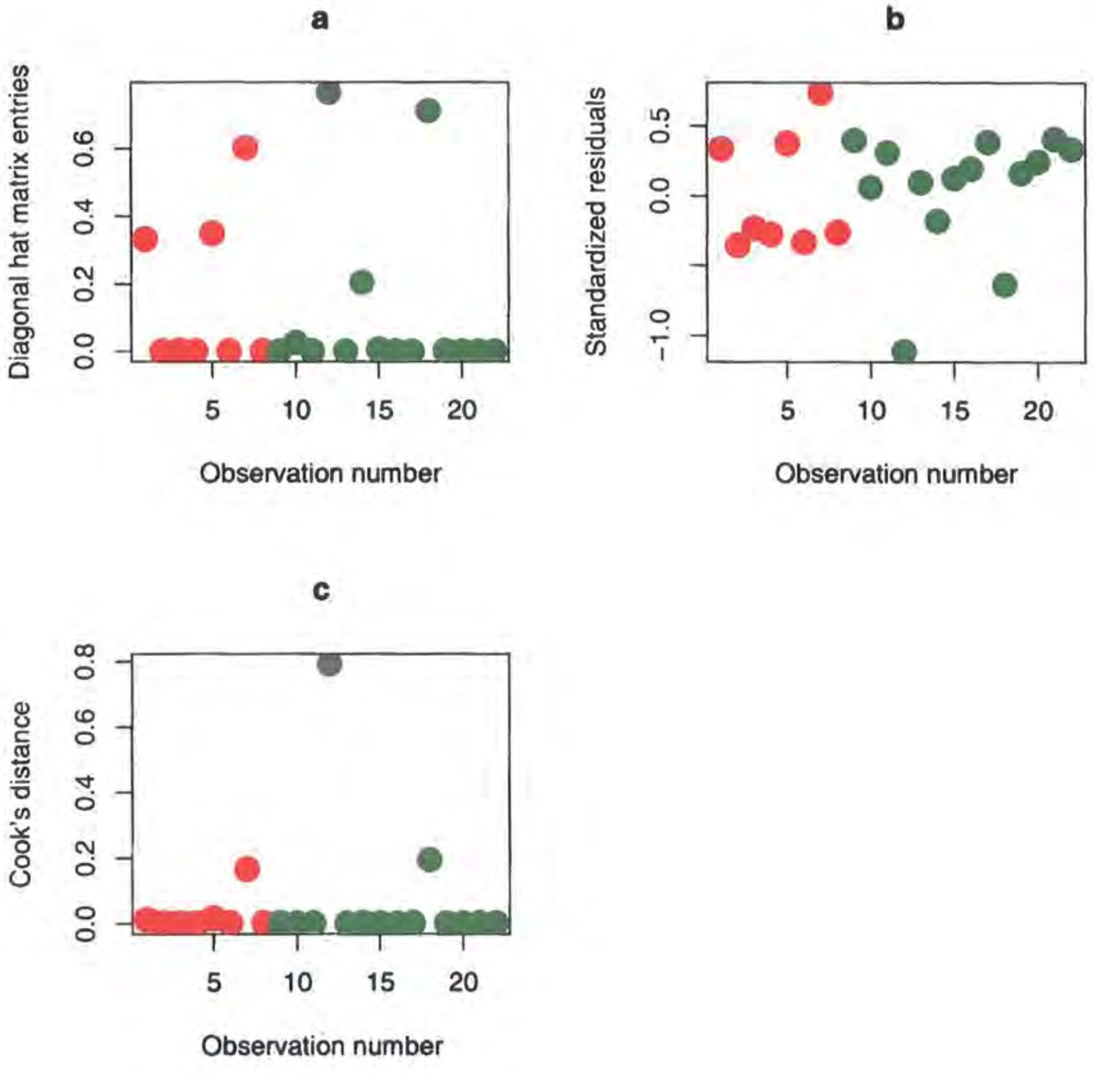

Figure 3.4: For model 3.4, plots of (a) leverage values h_i, (b) the standardized residuals, and (c) Cook's distance.

2. All standardized residuals are less in magnitude than 1, except observation 12, which is slightly bigger, indicating that none of observations is an outlier.

3. The Cook's distance plot suggests that observation 12 with $d_{12} = 0.791$ is possibly influential while observations 7 and 18 have small Cook's distance, $d_7 = 0.165$ and $d_{18} = 0.194$.

In addition to the above measures, the scatter plot of $z_2 = \log t_{1/2}^{soil}$ against $z_1 = \log k_{oc}$, Figure 3.5, shows that observation 13 is far away from the bulk of data, so it might be considered as an outlier. However, the WEMEL estimator is robust against outliers. Instead of deleting outliers, WEMEL gives less weight to these points depending on how far away they are from the bulk of data. For example, in fitting Model 3.4, observation 13 was given weight of 0.043 while observation 19, for example, was given weight of 1. On the other hand, observations 12 and 18 are inconsistent with other observations from their group (non-leaching pesticides). The typical non-leachers are those with high k_{oc} and small $t_{1/2}^{soil}$ values. But, observation 12 has small k_{oc} value and observation 18 has high $t_{1/2}^{soil}$. This is true also for observation 7 which has a relatively high k_{oc} value which makes it inconsistent with the leacher group which has small k_{oc} and high $t_{1/2}^{soil}$ values. All three of these observations, 7, 12 and 18, are close to the boundary of the discrimination curve, explaining why some of the above measures show that these observations are influential.

In conclusion, the above measures show that observations 7, 12 and 18 are possibly influential points. These points are close to the boundary of the discrimination curve. Because of separation, their removal will change the curve substantially, see Figure 3.5 when we remove all three points, for example. However, for discrimi-

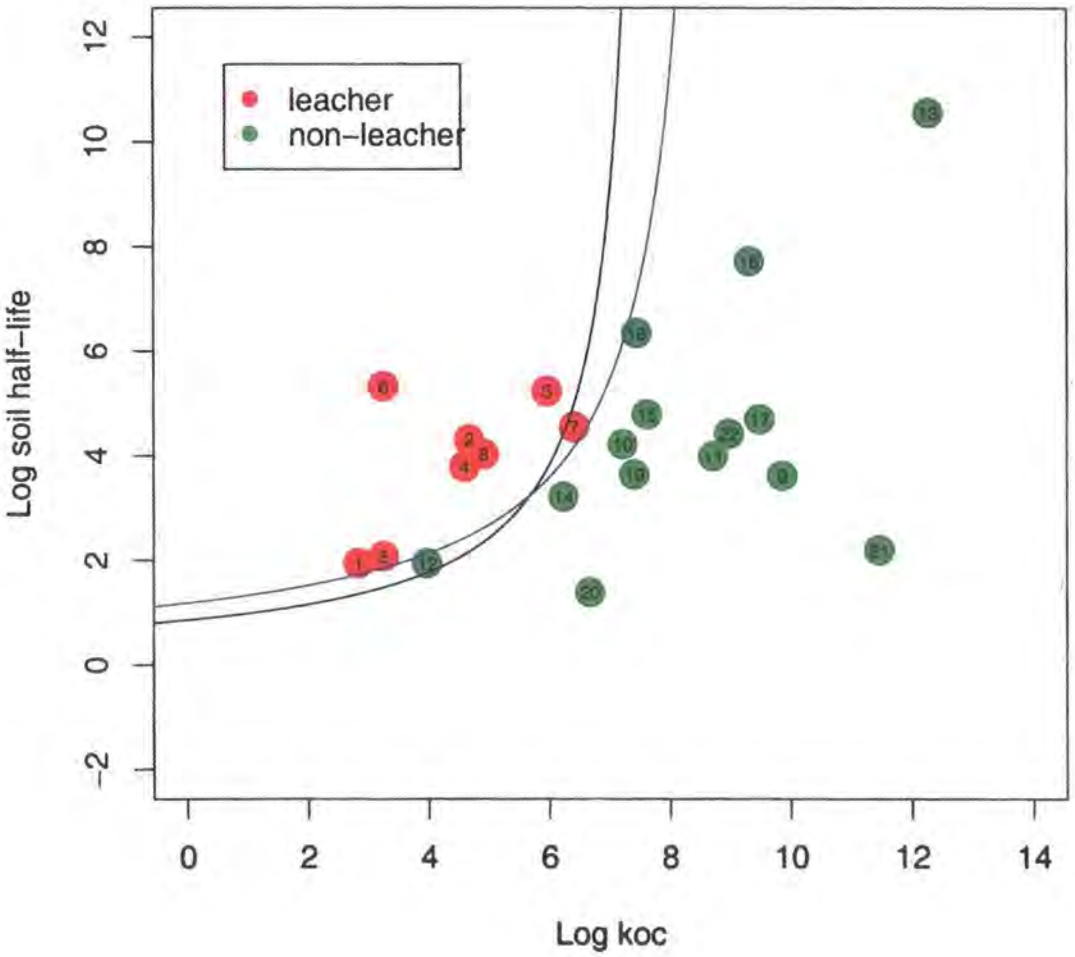

Figure 3.5: Scatter plot of $z_2 = \log t_{1/2}^{soil}$ against $z_1 = \log k_{oc}$. The blue curve is WEMEL fit for Model 3.4 where all observations are included. The black curve is WEMEL fit where observations 7, 12 and 18 are excluded.

nation purposes, we would not want to remove these points, because they are not outliers. However, while this residuals analysis can be understood, it is important to be careful when using residuals from binary data analysis; see, for example, [14] page 128 and [33] page 399.

The following sections describe the main components of the model, which are similar to those in [44]. The examples given in [44] will be reanalysed here for comparison.

3.3 Likelihood

As in [44], the data are from lysimeter experiments (see below) which are used to discover whether or not a pesticide is observed to leach relative to a specified threshold. The data can be represented in the form of a likelihood function as follows. Let m denote the number of lysimeter experiments and r denote the number out of m where a given pesticide is observed to leach, so that $s = m - r$ is the number of experiments where the pesticide is observed not to leach. In addition, we regard r as an observation on a Binomial random variable Y with distribution

$$p(Y = r|m, \pi) = \binom{m}{r} \pi^r \, (1 - \pi)^s \qquad r = 0, \dots, m \tag{3.10}$$

where π is the leaching probability. For given r, 3.10 is the likelihood function of π.

3.3.1 What is a lysimeter?

A lysimeter is "a device for measuring the percolation of water through soils and for determining the soluble constituents removed in the drainage" [1]. It has a use

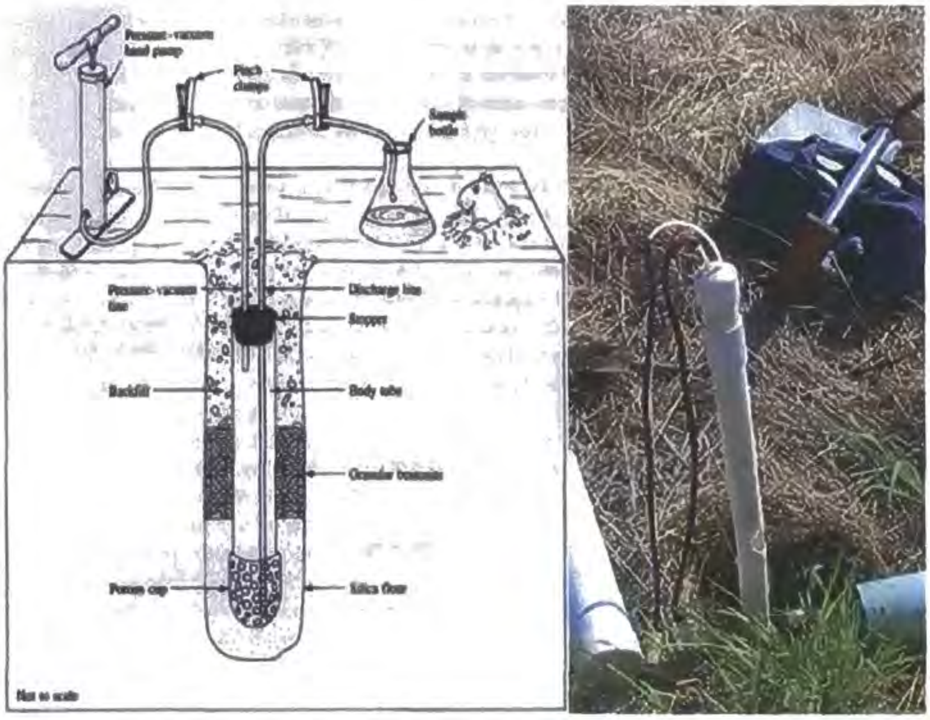

Figure 3.6: A lysimeter diagram as appears in [1]

in environmental site assessments to study contamination of groundwater. Simply, it can be constructed using a cylindrical device, a sample bottle and a vacuum. As described in [1], it works as follows; see Figure 3.6,

1. Using the pressure-vacuum pump, first apply vacuum to suck the moisture in.

2. Then apply pressure to pump it up to the sample bottle.

3.4 Prior knowledge

As suggested in [44], a conjugate prior distribution in the form of a beta distribution is chosen to represent prior knowledge for π, i.e.

$$p(\pi|a,b) = \frac{1}{B(a,b)}\pi^{a-1}(1-\pi)^{b-1} \quad 0 \leq \pi \leq 1 \tag{3.11}$$

where a and b are specified positive constants reflecting the current knowledge about π prior to obtaining the relevant data m and r. The prior mean and standard deviation of π are $\frac{a}{a+b}$ and $\sqrt{\frac{ab}{(a+b)^2\,(a+b+1)}}$, respectively.

3.5 Updating the model

In general, Bayesian analysis combines data $\boldsymbol{y}$ in the form of likelihood $p(\boldsymbol{y}|\pi)$ with the prior distribution $p(\pi)$ to generate the posterior distribution $p(\pi|\boldsymbol{y})$ for π as follows.

$$\begin{aligned} p(\pi|\boldsymbol{y}) &= p(\boldsymbol{y}|\pi)\cdot p(\pi)/p(\boldsymbol{y}) \\ &\propto \pi^r\,(1-\pi)^s\ \cdot \pi^{a-1}\,(1-\pi)^{b-1} \\ &= \pi^{a+r-1}\,(1-\pi)^{b+s-1}. \end{aligned}$$

This shows that the posterior distribution is also a beta distribution but with new parameters $(a+r, b+s)$. The posterior mean and standard deviation of π are therefore

$$\frac{a+r}{a+b+m} \quad \text{and} \quad \sqrt{\frac{(ab)\,(b+s)}{(a+b+m)^2\,(a+b+m+1)}} \tag{3.12}$$

respectively.

3.6 Specifying parameters of the prior distribution

Worrall et al. in [44] argued that the parameter a can be seen to play a role similar to that of r and b plays a role similar to that of s, hence $a+b$ plays a role similar to that of m where r, s and m are as described in Section 3.3.

As in [44], we describe how the above methodology can be implemented to predict the potential of a given pesticide to leach and contaminate groundwater.

3.7 Pesticide discrimination

In this section, the Bayesian analysis will be applied to predict the potential of a given pesticide to leach into groundwater on the basis of two of its chemical properties, k_{oc} and $t_{1/2}^{soil}$. In this regard, two steps are needed: (a) predicting pesticide leachability and (b) using this as prior knowledge to generate parameters for the beta prior distribution.

3.7.1 Predicting pesticide leachability

As in [44], logistic regression will be used to predict the potential of a given pesticide to leach and contaminate the groundwater. This stage of the process is the main difference between the original method in [44] and this study. As discussed in Section 3.2, the logistic regression uses the explanatory variable $z_2 = \log t_{1/2}^{soil}$ and an interaction term of $z_1 = \log k_{oc}$ and $z_2 = \log t_{1/2}^{soil}$ to fit Gustafson's data to improve the discrimination power. The analysis in Section 3.2 led to the estimated logit of the probability π, given the values of z_1 and z_2, using the MEL approach,

$$\text{estimated logit}(\pi_i|z_1, z_2) = -17.145 + 14.728z_2 - 1.658z_1z_2 \tag{3.13}$$

so that the leaching probability π is estimated as

$$\hat{\pi}_{MEL} = \frac{e^{-17.145+14.728z_2-1.658z_1z_2}}{1+e^{-17.145+14.728z_2-1.658z_1z_2}} \tag{3.14}$$

In the case of using the WEMEL approach:

$$\hat{\pi}_{WEMEL} = \frac{e^{-20.384+17.380z_2-1.947z_1z_2}}{1+e^{-20.384+17.380z_2-1.947z_1z_2}}. \tag{3.15}$$

Equations 3.14 and 3.15 can be used to predict the potential of a new pesticide to leach given its values of z_1 and z_2. In addition, these equations will play an important role in the process of generating parameters a and b for the beta prior distribution.

Figure 3.7 shows $\log k_{oc}$ and $\log t_{1/2}^{soil}$ with discriminant curves

$$-17.145 + 14.728z_2 - 1.658z_1z_2 = 0 \tag{3.16}$$

and

$$-20.384 + 17.380z_2 - 1.947z_1z_2 = 0 \tag{3.17}$$

where the black curve is the discriminant curve corresponding to MEL and the blue curve corresponds to WEMEL. The two curves are indistinguishable, but we prefer to use WEMEL since it is robust against outliers.

Table 3.4 shows three estimated leaching probabilities for Gustafson's data. $\hat{\pi}_{Worrall}$ is the original estimate evaluated by MLE in [44]. $\hat{\pi}_{MEL}$ and $\hat{\pi}_{WEMEL}$ are the predicted leaching probabilities calculated from the logistic regression model with an interaction term. A pesticide would be classified as a leacher if its estimated leaching probability exceeded some threshold, such as $\hat{\pi} > 0.5$. According to this rule, all Gustafson's pesticides were classified accurately using the WEMEL analysis of the logistic regression model with an interaction term. Three pesticides were misclassified when using the model in [44]. The known leacher Prometryn was misclassified with estimated leaching probability $\hat{\pi}_{Worrall} = 0.2005$, but it has

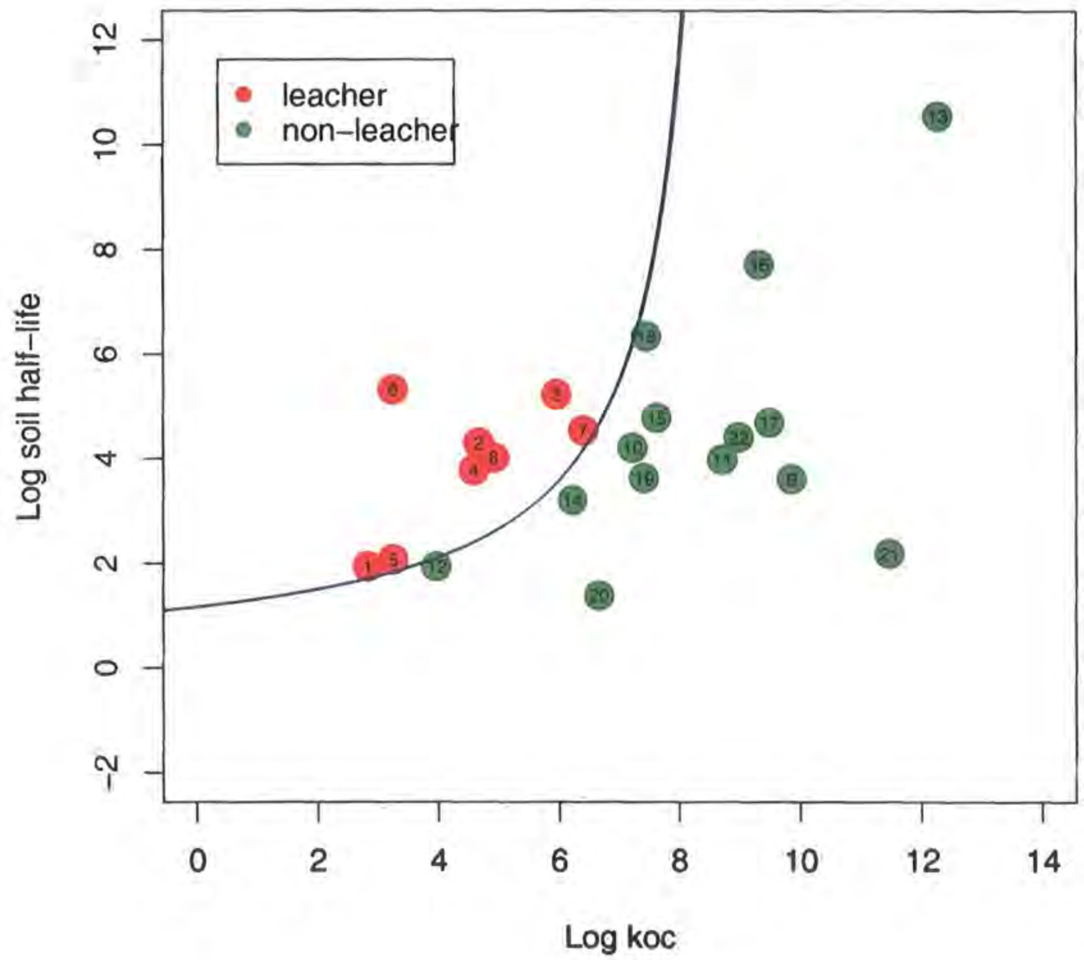

Figure 3.7: Discriminant curves using the logistic regression model with an interaction term. The black curve is a discriminant curve using the MEL estimator and the blue curve uses the WEMEL estimator. The two curves are almost indistinguishable.

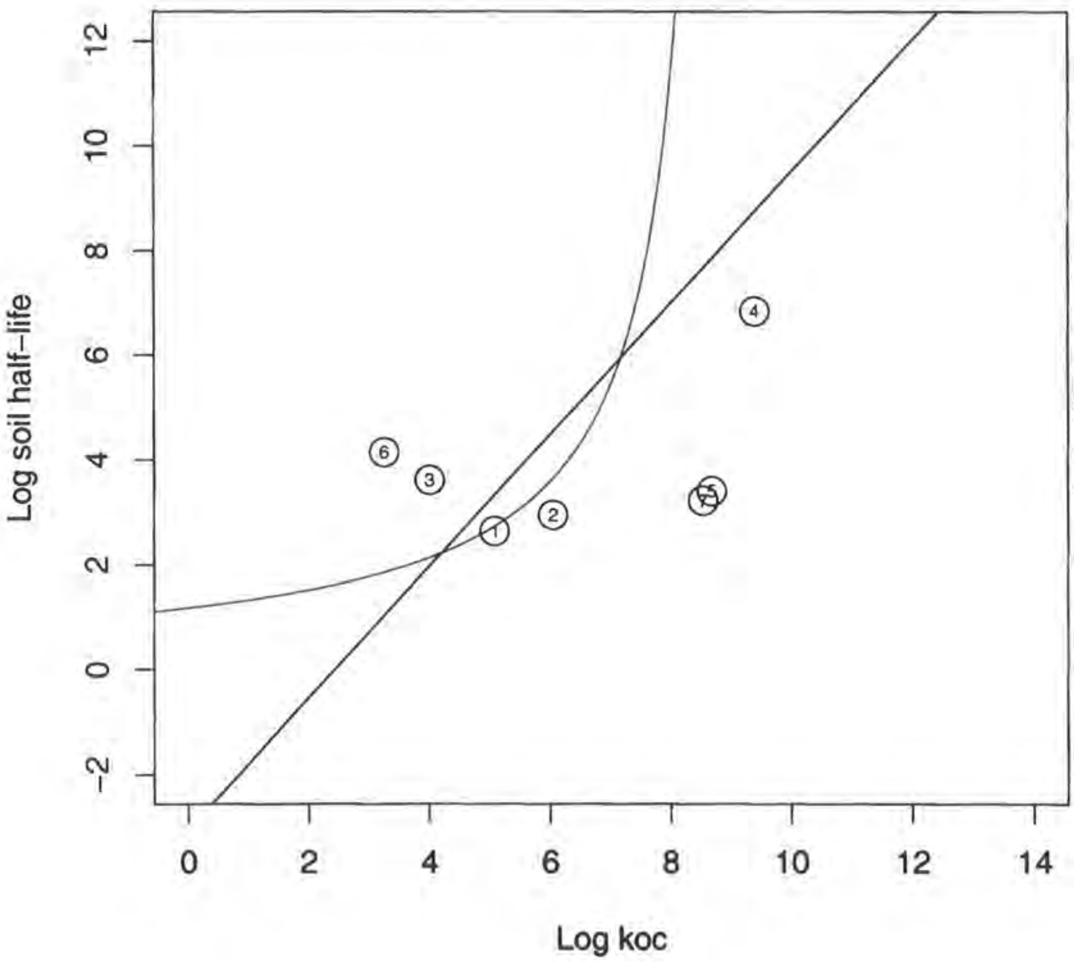

Figure 3.8: $\log k_{oc}$ versus $\log t_{1/2}^{soil}$ for transitional pesticides with the non-linear discriminant curve, blue, derived from analysis of WEMEL and linear discriminant line, black, derived from analysis of MLE as in [44].

Pesticide	Leacher	Ads.rat(koc)	soil half-life($t^{soil}_{1/2}$)	$\hat{\pi}_{Worrall}$	$\hat{\pi}_{MEL}$	$\hat{\pi}_{WEMEL}$
Aldicarb	Yes	17	7	0.9835	0.9148	0.9371
Atrazine	Yes	107	74	0.9854	0.99999	0.99999
Diuron	Yes	389	188	0.9035	0.9997	0.9999
Metolachlor	Yes	99	44	0.9529	0.9999	0.99999
Oxamyl	Yes	26	8	0.9500	0.9045	0.9286
Picloram	Yes	26	206	0.9999	1	1
Prometryn	Yes	614	94	0.2005	0.8035	0.8567
Simazine	Yes	138	56	0.9243	0.9999	0.99998
Chlordane	No	19269	37	7.32e-08	9.94e-11	1.90e-12
Chlorothalonil	No	1380	68	0.0053	0.0038	0.0016
Chlorpyrifos	No	6085	54	1.36e-05	1.09e-07	7.34e-09
2,4-D	No	53	7	0.5009	0.2150	0.1671
DDT	No	213600	38200	0.0051	6.44e-34	2.01e-39
Dicamba	No	511	25	0.0109	0.0465	0.0286
Endosulfan	No	2040	120	0.0066	0.0008	0.0003
Endrin	No	11188	2240	0.0595	1.29e-10	3.66e-12
Heptochlor	No	13330	109	6.02e-06	3.00e-10	7.61e-12
Lindane	No	1727	569	0.5105	0.1075	0.09998
Phorate	No	1660	38	5.29e-04	0.0003	6.28e-05
Propachlor	No	794	4	1.21e-05	5.70e-06	6.08e-07
Toxaphene	No	95816	9	4.05e-12	2.89e-12	2.64e-14
Trifluralin	No	7950	83	1.77e-05	1.75e-08	8.79e-10
Alachlor	Transitional	161	14	0.1181	0.3750	0.3483
Carbaryl	Transitional	423	19	0.0098	0.0355	0.0203
Carbofuran	Transitional	55	37	0.9903	0.99999	0.99999
Dieldrin	Transitional	12100	934	0.0039	1.01e-10	2.53e-12
Dinoseb	Transitional	5900	30	2.82e-06	1.11e-07	7.06e-09
Ethoprop	Transitional	26	63	0.9998	1	1
Fonofos	Transitional	5105	25	2.82e-06	2.26e-07	1.60e-08

Table 3.4: The CDFA data together with adsorption coefficients k_{oc} and soil half-life $t^{soil}_{1/2}$ in days and three predicted leaching probabilities estimated using logistic regression without ($\hat{\pi}_{Worrall}$) and with ($\hat{\pi}_{MEL}$ and $\hat{\pi}_{WEMEL}$) an interaction term.

$\hat{\pi}_{MEL} = 0.8035$ and $\hat{\pi}_{WEMEL} = 0.8567$, which is evidence that Prometryn should be classified as a leacher. Furthermore, the non-leaching pesticides 2,4-D and Lindane have $\hat{\pi}_{Worrall} = 0.5009$ and $\hat{\pi}_{Worrall} = 0.5105$ respectively, and were also misclassified. By contrast, the predicted values $\hat{\pi}_{MEL}$ and $\hat{\pi}_{WEMEL}$ for each support their classification in the CDFA database as non-leachers; see Table 3.4.

Table 3.4 also includes predicted leaching probabilities for transitional pesticides. Carbaryl, Dieldrin, Dinoseb and Fonofos all have estimated leaching probabilities which are close to zero providing strong evidence that they should be classified as non-leachers. Compounds like Carbofuran and Ethoprop have leaching probabilities close to one, which suggests they should be classified as leachers. The remaining transitional pesticide, Alachlor, has $\hat{\pi}_{Worrall} = 0.1181$, $\hat{\pi}_{MEL} = 0.3750$ and $\hat{\pi}_{WEMEL} = 0.3483$. These three leaching probabilities, especially the latter two, are relatively high and this adds some doubts about classifying Alachlor as a non-leacher. A pictorial view of these results are displayed in Figure 3.8 which shows the covariate values z_1 and z_2 of the transitional pesticides with discriminant curve expressed by equation 3.17 and the linear discriminant line derived from the MLE analysis in [44].

3.7.2 Generating parameters for a beta prior distribution

For a given pesticide, equations 3.14 and 3.15 can be used to estimate the leaching probability $\hat{\pi}$ for a particular pesticide given the values of k_{oc} and $t_{1/2}^{soil}$. The estimated leaching probability $\hat{\pi}$ with information from lysimeter experiments can be used to generate the parameters a and b of the beta prior distribution. It was suggested in [44] that the value of the prior evidence should not exceed three lysimeter experiments. This led to the suggestion that the prior evidence corresponds to approximately $m = 2$ lysimeter experiments for a new compound. As in [44], the parameters a and b of the beta prior distribution can be derived from the relationships

$$a + b = m \tag{3.18}$$

$$\mathrm{E}(\pi) = \frac{a}{a+b} = \hat{\pi} \tag{3.19}$$

This leads to $a = m\hat{\pi}$ and $b = m(1 - \hat{\pi})$. Therefore, the prior mean and standard deviation of π are $\mathrm{E}(\pi) = \hat{\pi}$ and $\mathrm{sd}(\pi) = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{m+1}}$. Taking $m = 2$, as suggested previously, leads to

$$a = 2\hat{\pi} \tag{3.20}$$

$$b = 2(1 - \hat{\pi}) \tag{3.21}$$

The next section describes how these results may be applied.

Example 1

This example can been seen as an extension of example 1 given in [44]. It illustrates how the above can be used to assess the environmental fate of the pesticide Triclopyr. As in [44], the values of k_{oc} and $t_{1/2}^{soil}$ are 41 and 6, respectively. Also, the source of the data were from two lysimeter experiments, where it was found that the annual leachate concentrations were below $0.1\mu g$ litre^{-1} in both experiments. Thus, there were $m = 2$ observations, $r = 0$ leachers and $s = 2$ non-leachers. Using equation 3.15, the predicted leaching probability is $\hat{\pi}_{WEMEL} = 0.100$, which suggests that Triclopyr does not leach. Furthermore, equations 3.20 and 3.21 lead to $a = 0.20$ and $b = 1.80$ for the parameters of the beta prior distribution. As in [44], statistical tables [32] can be used to find a highest density prior probability interval for the true leaching probability π. This leads to (0.027-0.602) as a 90% interval for π for Triclopyr. Combining the data with the prior distribution leads to a posterior beta distribution with parameters $a + r = 0.20$ and $b + s = 3.80$. This leads to a revised leaching probability of 0.050 with (0.008-0.416) as a 90% Bayesian confidence interval, which supports the classification of Triclopyr as a non-leacher. [The same analysis could be followed using $\hat{\pi}_{MEL}$ to generate the parameters of the prior beta distribution.] Table 3.5 displays results for the analysis of Triclopyr. It shows the prior and posterior leaching probabilities calculated for the original model in [44], where MLE was used, and in the case of the interaction model where WEMEL is used. The table shows that the interaction model gives better results in the following sense. The prior evidence using MLE suggests that Triclopyr leaches with predicted probability 0.619, contradicting the posterior estimate, 0.310. On the other hand, the evidence in the case of the interaction model shows that both the prior and posterior estimates suggest that Triclopyr should be classified as a non-leacher.

Example 2

This example is an extension of example 2 given in [44]. It concerns the pesticide Bentazone. As in [44], this pesticide has a range of values reported for both k_{oc} and $t_{1/2}^{soil}$. These values (log scale) are listed in Table 3.6 and also displayed in Figure 3.9 together with various prior discriminant curves and lines derived from

					prior	Prior	posterior	posterior
				GUS	leaching	probability	leaching	probability
Method	k_{oc}	$t_{1/2}^{soil}$	GUS	classification	probability	90% interval	probability	90% interval
MLE	41	6	1.86	transitional	0.619	0.214-0.882	0.310	0.043-0.504
WEMEL	41	6	1.86	transitional	0.100	0.027-0.602	0.050	0.008-0.416

Table 3.5: Prior screening and posterior probabilities for Triclopyr in case of using MLE, as in [44], or WEMEL to generate the parameters for the beta prior distribution.

different analyses of Gustafson's data. The predicted leaching probabilities $\hat{\pi}_{WEMEL}$, calculated for each pair of k_{oc} and $t_{1/2}^{soil}$ values, are listed in Table 3.6. The data came from four lysimeter experiments, resulting in $r = 0$ leachers and $s = 4$ non-leachers. The results derived from analyses similar to those used in Example 1 are summarized in Table 3.6. For example, the analysis using the values of $k_{oc} = 13.3$ and $t_{1/2}^{soil} = 3$ leads to a prior leaching probability of 0.001 which leads to $a = 0.002$ and $b = 1.998$ from which we obtain (0.000-0.537) as a 90% probability interval. Updating the model leads to $a + r = 0.002$, $b + s = 5.998$ and a posterior leaching probability of 0.000 with (0.000-0.281) as a 90% Bayesian confidence interval, suggesting that Bentazone is not a leacher.

It seems that there is a conflict between the data and prior leaching probability in three of the five pesticides. While the prior evidence strongly indicates that Bentazone leaches into groundwater, the evidence from the lysimeter experiments ($r = 0$ leachers from $m = 4$ experiments) is to the contrary. On the other hand, the posterior evidence in each of the five cases seems to support Bentazone as not being a leacher, more strongly reflecting the data rather than the prior evidence. Hence, the prior evidence is no longer helpful here. This may be because there is a range of

				Prior	Prior	Posterior	Posterior	Posterior
			GUS	leaching	probability	leaching	probability	predictive
k_{oc}	$t_{1/2}^{soil}$	GUS	classification	probability	90% interval	probability	90% interval	probability
13.3	3	1.37	non-leacher	0.001	0.000-0.537	0.000	0.000-0.281	0.279
13.3	21	3.80	leacher	1.000	0.464-1	0.333	0.108-0.632	0.646
34.0	20	3.21	leacher	1.000	0.464-1.000	0.333	0.108-0.632	0.629
175.6	3	0.84	non-leacher	0.000	0.000-0.536	0.0000	0.000-0.280	0.006
175.6	21	2.32	transitional	0.869	0.378-0.965	0.290	0.086-0.588	0.085

Table 3.6: Prior screening and posterior probabilities for Bentazone where WEMEL is used to generate the parameters for the beta prior distribution. The last column is the posterior predictive probability calculated from the alternative Bayesian analysis proposed in section 3.8.

				prior	Prior	posterior	posterior
			GUS	leaching	probability	leaching	probability
k_{oc}	$t_{1/2}^{soil}$	GUS	classification	probability	90% interval	probability	90% interval
13.3	3	1.37	non-leacher	0.927	0.417-0.978	0.309	0.095-0.607
13.3	21	3.80	leacher	0.999	0.464-1.000	0.333	0.108-0.632
34.0	20	3.21	leacher	0.990	0.458-0.997	0.330	0.106-0.628
175.6	3	0.84	non-leacher	0.0012	0.000-0.537	0.0004	0.000-0.281
175.6	21	2.32	transitional	0.238	0.064-0.683	0.079	0.011-0.368

Table 3.7: Prior screening and posterior probabilities for Bentazone as analysed in [44] where MLE is used to generate the parameters for the beta prior distribution.

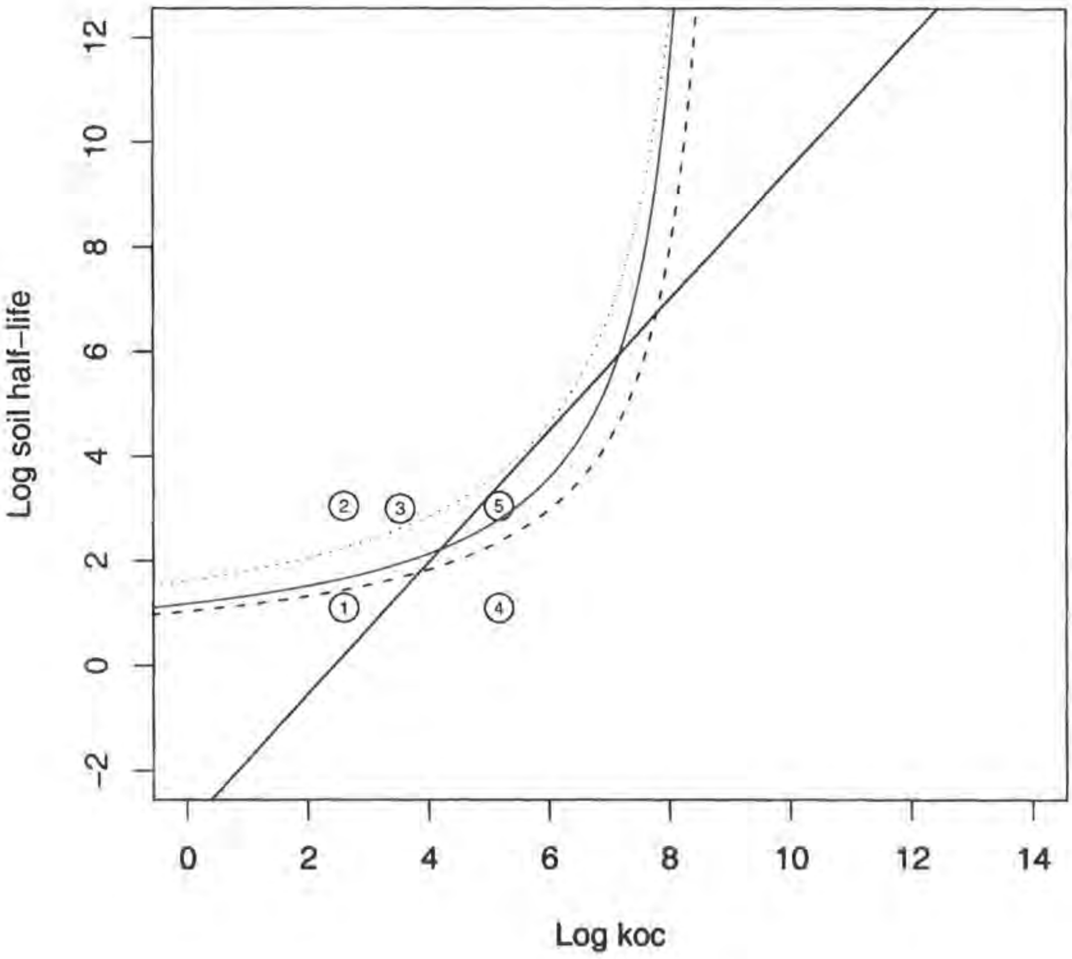

Figure 3.9: The available values of k_{oc} and $t_{1/2}^{soil}$ in a log-scale for Bentazone together with various discriminant curves and lines derived from analyses of Gustafson's data. The dotted and dashed curves represent the discriminant curves GUS=2.8 and GUS=1.8 as derived in [25]. The blue curve represents a logistic discriminant curve estimated by WEMEL. The black line is a logistic discriminant line estimated by MLE as in [44].

values rather than a single value reported for each covariate. Pesticide discrimination in the case of uncertain covariates (the focus of this thesis) was addressed and tackled for the first time in [43] using Bayes linear methods.

For comparison, Table 3.7 shows prior screening and posterior probabilities for Bentazone as analysed in [44] where MLE was used.

3.8 An alternative Bayesian analysis

As discussed above, the focus in [44] was on Bayesian inferences about the probability π that a given pesticide will leach. The likelihood function of π derives from modelling the results of lysimeter experiments as a Binomial variate. Then, a conjugate prior distribution, in the form of a beta distribution with a specified parameters, for π was chosen, so that the resulting posterior distribution is also a beta distribution.

In this section, an alternative Bayesian analysis is proposed. It uses Gustafson's data, results from lysimeter experiments and logistic regression in a 'direct' approach as follows.

The idea is to combine the 22 cases from Gustafson's data with the results from lysimeter experiments for a given pesticide in one data set, which we denote by $\boldsymbol{y}$. This means that for a given pesticide with m lysimeter experiments, $\boldsymbol{y}$ is a $(22+m) \times 1$ vector with values 0 or 1, depends on whether the case is classified or monitored as a non-leacher or as a leacher, respectively.

The method is implemented using "Bayesian predictive inference", as described in Chapter 2. This inference uses the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$ of the regression parameters $\boldsymbol{\beta}$ in a logistic model for the leaching probabilities to make inferences about the leaching status y_{new} of a new pesticide conditional on the data $\boldsymbol{y}$, which can be calculated (see Section 2.4) as:

$$p(y_{new}|\boldsymbol{y}) \;=\; \int p(y_{new},\boldsymbol{\beta}|\boldsymbol{y})d\boldsymbol{\beta} \tag{3.22}$$

provided $\boldsymbol{y}$ and y_{new} are conditionally independent, given $\boldsymbol{\beta}$ [18], which we assume here. This posterior predictive probability, $p(y_{new}|\boldsymbol{y})$, requires a three dimensional integral for each value of y_{new}. Alternatively, simulation techniques can be used to draw samples from $p(y_{new}|\boldsymbol{y})$. The simulation technique used here comprises the following steps:

1. Simulate N values of $\boldsymbol{\beta}$ from the posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y})$. This can be done using MCMC simulation using, for example, the function `MCMClogit` in the R package. A non-informative prior distribution may be chosen for $\boldsymbol{\beta}$.

2. For each simulated value $\boldsymbol{\beta}^*$ we calculate the leaching probability

$$\pi_{new} = \frac{\exp(\boldsymbol{x}_{new}^T\boldsymbol{\beta}^*)}{1+\exp(\boldsymbol{x}_{new}^T\boldsymbol{\beta}^*)} \tag{3.23}$$

3. Draw U from a uniform [0, 1] distribution.

4. Set

$$y_{new} = \begin{cases} 1 & \text{if} \quad U < \pi_{new} \\ 0 & \quad \text{otherwise.} \end{cases}$$

5. $p(y_{new}|\boldsymbol{y})$ is estimated as the proportion of 1's in the MCMC sample.

Example 3

This example illustrates how the above can be used to analyse the pesticide Triclopyr. As mentioned in Example 1, this pesticide has $z_{new,1} = 41$, $z_{new,2} = 6$ and two lysimeter experiments, $m = 2$, were undertaken resulting in $r = 0$ leachers and $s = 2$ non-leachers. Hence, two cases will be added to the Gustafson data in which each case is classified as a non-leacher. This leads to new data, $\boldsymbol{y}$, with 24 cases in which 8 are classified as leachers and 16 as non-leachers. Now, we wish to make inference about a future observation y_{new} conditional on the data $\boldsymbol{y}$. A MCMC sample of 10000 leads to the estimate $p(y_{new} = 1|\boldsymbol{y}) = 0.2512$ which supports the classification of Triclopyr as a non-leacher, a more convincing conclusion than that for the model in [44] which gives 0.310 as a posterior estimate of the leaching probability.

Example 4

This example concerns the pesticide Bentazone which has a range of values reported for both k_{oc} and $t_{1/2}^{soil}$. There are also $m = 4$ lysimeter experiments resulting in $r = 0$ leachers and $s = 4$ non-leachers. Thus, four non-leaching cases will be added to the Gustafson data. The last column in Table 3.6 shows the estimated posterior predictive probability $p(y_{new} = 1|\boldsymbol{y})$ of a future observation y_{new} for each available pair of k_{oc} and $t_{1/2}^{soil}$. In cases 1, 4 and 5, the posterior predictive probability suggests that the Bentazone does not leach. In cases 2 and 3, the posterior predictive probability suggests that the Bentazone is a leacher. However, these two cases are located in the NW corner, as in Figure 3.9, which means that they are pairs with low k_{oc} and high $t_{1/2}^{soil}$, which is consistent with the conjecture proposed in [43] and [45] that leaching pesticides are those with a low k_{oc} and high $t_{1/2}^{soil}$, and the non-leacher pesticides are those with high k_{oc} and low $t_{1/2}^{soil}$.

3.9 Conclusion

This chapter extends the Bayesian methods proposed in [44]. One extension incorporates an interaction term in the linear predictor of the logistic regression model as follows for pesticide i

$$\eta_i = \beta_0 + \beta_1 z_{i2} + \beta_2 z_{i1} z_{i2} \qquad i = 1, 2, ..., n$$

where the leaching status Y_i is regarded as a binary response having a Bernoulli distribution with probability of success $p(Y_i = 1) = \pi_i = \exp(\eta_i)/(1 + \exp(\eta_i))$.

Here, a deficiency arises in fitting the logistic regression by maximum likelihood. The deficiency is that the MLE does not exist when there is a complete separation in the space of the explanatory variables relative to the model. This difficulty was resolved by first measuring the overlap in the logistic regression model via the depth-based algorithm proposed in [7], and then the alternative estimators MEL and WEMEL proposed in [8] were used.

Stepwise procedures are used to select covariates and the most important model terms to be included in the linear predictor. These procedures show that k_{oc} and $t_{1/2}^{soil}$, but not S_{H2O}, are important in discriminating between the leaching and non-leaching pesticides. In addition, some model diagnostics, such as residual patterns and influence measures are used for checking model adequacy.

The interaction model fitted Gustafson's data better than the main effects model proposed in [44], in the sense that none of the pesticides were misclassified, whereas three pesticides were misclassified using the main effects model where there is an overlap. As in [44], the estimated leaching probabilities were used as a prior screening to assess the parameters for the beta prior distribution, and the same lysimeter data was used.

An alternative Bayesian analysis was proposed which combines the results from lysimeter experiments and Gustafson's data. Then, Bayesian predictive inference is used to draw inferences about a future observation for any particular pesticide. These methods were illustrated with examples discussed in [44] and showed improved classification.

A difficulty that arises in both the original study and here concerns analysis of a given pesticide with a range of covariate values for both k_{oc} and $t_{1/2}^{soil}$, that is, a pesticide with uncertain covariates. The first attempt to tackle discrimination with uncertain covariates was given in [43], where Bayes linear methods were used. These ideas will be extended in the next chapter to any linear model, including one with an interaction term.

Chapter 4

Bayes linear discrimination with uncertain covariates

4.1 Introduction

The use of Bayes linear estimation for the simple linear model proposed in [43] was the first attempt to discriminate pesticides as leachers or non-leachers based on two of their chemical properties where the published values of these properties are uncertain. This chapter applies the general Bayes linear estimate derived in Chapter 2 to a model with an interaction term.

The chapter is organised as follows. In Section 4.2, we discuss discriminant models. This includes a discussion about regression analysis, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). In Section 4.3, we investigate whether the prior discriminant curve derived from analysis of Gustafson's data provides good prediction for the EA pesticides. In Section 4.4, we implement the Bayes linear approach to analyse the EA data. This includes specifying prior beliefs structures for the unknown quantities, adjusting these beliefs, analysing these adjusted beliefs and investigating whether re-structuring the prior beliefs will result in better prediction for the EA data. In Section 4.5, the linear model proposed in [43] will be subjected to further diagnostic analysis to improve it. Section 4.6 concludes the chapter.

4.2 Formulation of the discriminant models

Let $\boldsymbol{y}$ denote a vector of binary responses where $y_i = 1$ if the i^{th} pesticide is monitored as a leacher and $y_i = 0$ if it is a non-leacher. We consider a linear model with main effects terms $z_1 = \log k_{oc}$ and $z_2 = \log t_{1/2}^{soil}$ and their interaction $z_1 z_2$,

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_1 z_2 + \epsilon \tag{4.1}$$

We now discuss the relationship between fitting this model by least squares and discriminant analysis.

Much of the following is taken from [27]. Linear Discriminant analysis (LDA) arises (in our case) from considering two classes, non-leachers (Π_0) and leachers (Π_1). Let $f_0(\boldsymbol{z})$ and $f_1(\boldsymbol{z})$ be the class-conditional density of $\boldsymbol{z} = (z_1, z_2)$ in populations Π_0 and Π_1, respectively, and let π_0 and π_1 be the prior probabilities of populations Π_0 and Π_1, respectively, with $\pi_0 + \pi_1 = 1$. Now, we assume that each class density, $f_k(\boldsymbol{z})$, is modelled as multivariate Gaussian with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\sum_k$

$$f_k(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_k)} \tag{4.2}$$

for $k = 0, 1$. Linear discriminant analysis (LDA) arises when we assume that the classes have a common covariance matrix $\sum_k = \sum$ for $k = 0$ and 1. From 4.2, the discriminant functions $\delta_k(\boldsymbol{z})$ derive from the log-ratio of the posterior probabilities of the two classes and are given by

$$\delta_k(\boldsymbol{z}) = \boldsymbol{z}^T \Sigma^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \Sigma^{-1} \boldsymbol{\mu}_k + \log \pi_k \tag{4.3}$$

Each $\delta_k(\boldsymbol{z})$ is linear in z_1 and z_2, where we see that the linear coefficients are functions of the parameters $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1$ and Σ of the Gaussian distributions and π_0 and π_1. In practice, these parameters are estimated as the corresponding moment estimates from the sets of sample values of (z_1, z_2) from the two populations; see [27], for example.

With two classes, there is a simple correspondence between linear discriminant analysis and classification by linear least squares; see, [27] and [43]. The LDA rule classifies to class 1 if

$$\boldsymbol{z}^T \hat{\Sigma}^{-1} (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0) > \frac{1}{2} \hat{\boldsymbol{\mu}}_1^T \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_1 - \frac{1}{2} \hat{\boldsymbol{\mu}}_0^T \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_0 + \log (N_0/N) - \log (N_1/N) \tag{4.4}$$

and class 0 otherwise, where N_k is the number of sample values from population k, and $\hat{\mu}_k$ and $\hat{\Sigma}$ are the moment estimates. [In particular, $\hat{\Sigma}$ is a pooled estimate of $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$, the individual variance matrix estimates from the two samples.] As shown in [27], the coefficients vector from least squares is proportional to the LDA direction given in 4.4, but unless $N_0 = N_1$, the intercepts are different. However, in practice, we can choose the intercept or cut-point that empirically minimizes error rates for a given dataset, see, [27], page 88, and Section 5.9 of this thesis. "Optimal scoring" can also be used to establish a correspondence between regression and LDA; see [27], page 88.

It turns out that the linear discriminant between non-leachers ($y = 0$) and leachers ($y = 1$) can be derived by ordinary least squares fitting of the linear model

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \epsilon \tag{4.5}$$

Quadratic Discriminant Analysis (QDA) arises under the same assumptions for LDA, except that the Σ_k in 4.2 are not assumed to be equal and 4.3 becomes

$$\delta_k(\boldsymbol{z}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_k)^T \Sigma_k^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_k) + \log \pi_k \tag{4.6}$$

which is quadratic in z_1 and z_2.

Hastie et al. in [27] demonstrate that QDA is well approximated by least squares analysis of the appropriate quadratic model, in our case model 4.1, which is equivalent to LDA for three discriminants $x_1 = z_1$, $x_2 = z_2$ and $x_3 = z_1 z_2$. This approximation is appropriate here, as our main purpose is to use the quadratic fit to Gustafson's data and estimated variance matrix as a source of prior information about $\boldsymbol{\beta}$ in the Bayes linear analysis of the 43 pesticides from the EA data base.

By analogy with the logistic model in Chapter 3 and Gustafson's form in 1.2, model 4.1 will be reduced to

$$y = \beta_0 + \beta_2 z_2 + \beta_3 z_1 z_2 + \epsilon \tag{4.7}$$

This simplified form can be further justified using similar procedures to those used in Chapter 3 for the logistic model.

Approximating QDA by least squares analysis of the quadratic model 4.7 leads to the fit

$$\hat{y} = -0.03 + 0.31 z_2 - 0.03 z_1 z_2 \tag{4.8}$$

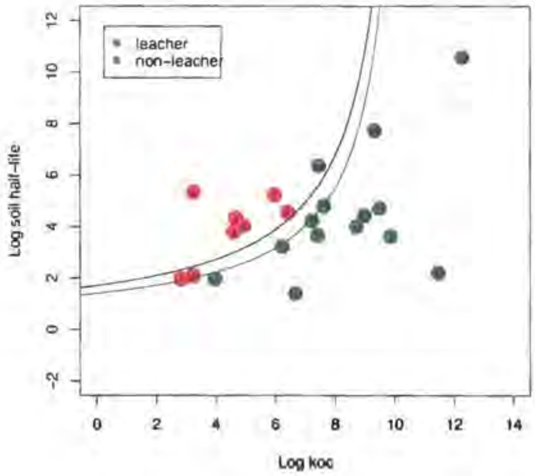

Figure 4.1: The Gustafson data together with the discriminant curve 4.10, black. The blue curve is the discriminant curve 4.10, but with cut point of 0.4 instead of 0.5. The blue curve results in better discrimination, where only one pesticide is misclassified.

and estimated variance-covariance matrix

$$s^2(\mathrm{G}^T\mathrm{G})^{-1} = \begin{bmatrix} 0.0562 & -0.0195 & 0.001 \\ -0.0195 & 0.0096 & -7E-4 \\ 0.001 & -7E-4 & 1E-4 \end{bmatrix} \tag{4.9}$$

where $s^2 = 0.15$ is the estimated error variance and G is the model matrix.

Figure 4.1 shows the values of $z_1 = \log k_{oc}$ and $z_2 = \log t^{soil}_{1/2}$ of Gustafson's data and the discriminant curve (black)

$$-0.03 + 0.31z_2 - 0.03z_1z_2 = 0.5 \tag{4.10}$$

The curve shows good discrimination between leaching and non-leaching pesticides with misclassification of three pesticides. However, as mentioned above, this fit can be improved by choosing a different cut-point to minimise error rate. For example, the choice 0.4 instead of 0.5 in 4.10 leads to the blue curve in Figure 4.1, where only one pesticide is misclassified.

In the following section, we investigate whether the non-linear discrimination based on Gustafson's data provides good prediction for the EA data.

4.3 Prior discrimination for the EA data

As discussed in Chapter 1, 43 pesticides were aggregated from the EA database. These 43 pesticides have complete information in the sense that for each pesticide the leachability status (whether or not it was monitored as a leacher in a specified year) and a range of reported values for both of the explanatory variables k_{oc} and $t_{1/2}^{soil}$ are known. This information is displayed in Table 1.1. The covariate means are plotted in Figure 4.2 with the discriminator curve in 4.10. There is poor discrimination apparent in both Figure 4.2 and Figure 4.3 which shows the plot of $\hat{\boldsymbol{y}} = \boldsymbol{Mb}$ vs. $\boldsymbol{y}$ where $\boldsymbol{M} = (\mathbf{1}, \boldsymbol{m}_2, \boldsymbol{m}_1\boldsymbol{m}_2)$, $\boldsymbol{b}^T = (-0.03, 0.31, -0.03)$ and $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are the vectors of the two covariate means, and $\boldsymbol{m}_1\boldsymbol{m}_2$ stands for a component-wise product.

As in [43], the poor prediction for the EA data using the prior non-linear discriminant can be improved by a Bayes linear estimate which accounts for uncertainty in the covariates. The solution begins by investigating whether the interaction model in 4.7 relates the observed binary vector $\boldsymbol{y}$ for the 43 pesticides to their corresponding unobserved vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ with suitable prior information on $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. The 43 pesticides will be considered as not inconsistent with the Gustafson's model if $\hat{\boldsymbol{y}} = \mathrm{E}_{\boldsymbol{y}}(\boldsymbol{\eta})$ is highly correlated with $\boldsymbol{y}$, where $\boldsymbol{\eta} = X\boldsymbol{\beta}$. This means that a large

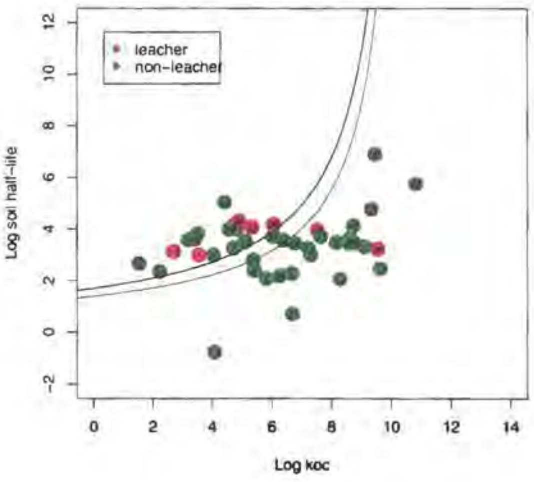

Figure 4.2: Mean values of log k_{oc} and log $t_{1/2}^{soil}$ for the EA data with the discriminant curve 4.10, black, obtained from Gustafson's data. The blue curve is the discriminant curve 4.10, but with cut point of 0.4. Both prior discriminant curves result in poor discrimination.

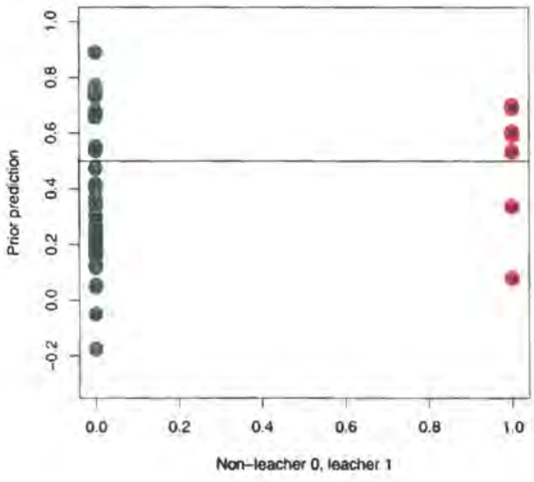

Figure 4.3: Predicted vs. observed state for the EA data based on the non-linear discriminant obtained from Gustafson's data.

percentage of the leachers (non-leachers) should have $\hat{\boldsymbol{y}}$-values greater (less) than some cut-off point such as 0.5.

The next section discusses the steps needed to compute the Bayes linear estimate.

4.4 Implementing the Bayes linear model

Two steps are needed to implement a Bayes linear model: (a) choosing prior information and organising them in form of means, variances and covariances, and (b) updating the prior information by combining them with observed data using adjusted expectation, as given in 2.16.

4.4.1 Prior information for z_1, z_2, $\boldsymbol{\beta}$ and ϵ

As mentioned earlier, the values of the covariates z_1 for $\log k_{oc}$ and z_2 for $\log t_{1/2}^{soil}$ are uncertain in the sense that there is a range of values reported for each of them for each pesticide. The Bayes linear approach is based on using prior information on z_1, z_2, $\boldsymbol{\beta}$ and ϵ. First of all, the prior information on z_1 and z_2 are chosen in the forms of their USDA database means and variances. Let $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ denote the vector of means of the components of z_1 and z_2, respectively. Also, let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ denote the diagonal matrices of the variances of the components of z_1 and z_2, respectively. We choose to specify the prior information for $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ as:

$$\mathrm{E}(\boldsymbol{z}_1) = \boldsymbol{m}_1, \quad \mathrm{Var}(\boldsymbol{z}_1) = \boldsymbol{v}_1$$

$$\mathrm{E}(\boldsymbol{z}_2) = \boldsymbol{m}_2, \quad \mathrm{Var}(\boldsymbol{z}_2) = \boldsymbol{v}_2 \quad \text{and } \mathrm{Cov}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \boldsymbol{0}$$

where the last equation represents the assumption that the covariates $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are uncorrelated.

The prior information on $\boldsymbol{\beta}$ and ϵ are derived from the least-squares analysis of the interaction model in equation 4.7 applied to Gustafson's data. In particular, the following prior information is assumed.

$$\mathrm{E}(\boldsymbol{\beta}) = \boldsymbol{b}, \quad \mathrm{Var}(\boldsymbol{\beta}) = \Sigma_\beta = s^2(\mathrm{G}^T\mathrm{G})^{-1}$$
$$\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}, \quad \mathrm{Var}(\boldsymbol{\epsilon}) = s^2\boldsymbol{I}$$

where $\boldsymbol{b} = (-0.03, 0.31, -0.03)$, $s^2 = 0.15$ (error variance), $\boldsymbol{I}$ is the identity matrix, G is the model matrix and Σ_β is given in 4.9.

Finally, $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are assumed to be uncorrelated.

4.4.2 Updating the model

The Bayes linear estimate $\hat{\boldsymbol{y}} = \mathrm{E}_{\boldsymbol{y}}(\boldsymbol{\eta})$, where $\boldsymbol{\eta} = X\boldsymbol{\beta}$, is used to update the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{4.11}$$

where $x_1 = z_2$ and $x_2 = z_1 z_2$. Using Theorem 1 and Example 2 in Section 2.3.1, the Bayes linear estimate can be computed as:

$$\hat{\boldsymbol{y}} = (I + A)^{-1}M\boldsymbol{b} + A(I + A)^{-1}\boldsymbol{y} \tag{4.12}$$

where $M = \mathrm{E}[X]$, $\sigma^2 A = M\Sigma_\beta M^{\mathrm{T}} + D$ and D is a diagonal matrix with $D_{ii} = \boldsymbol{b}^{\mathrm{T}}\Sigma_i\,\boldsymbol{b} + \mathrm{trace}[\Sigma_i\Sigma_\beta]$, where $\Sigma_i = \mathrm{Var}[\boldsymbol{x}_i]$, the variance matrix of the i-th row of X, namely $\boldsymbol{x}_i^T = (1, z_{i2}, z_{i1}z_{i2})$; see Example 2 in Section 2.3.1 for further detail.

Note that $\hat{\boldsymbol{y}}$ in 4.12 is of the form $\hat{\boldsymbol{y}} = \hat{\boldsymbol{y}}_{Gustafson} + \boldsymbol{a}$, where $\boldsymbol{a}$ are the adjustments to the prior values $\hat{\boldsymbol{y}}_{Gustafson} = \boldsymbol{Mb}$.

Figure 4.4, panel (a), where $\hat{\boldsymbol{y}}$ is plotted against $\boldsymbol{y}$, shows good posterior discrimination between leachers and non-leachers in comparison with the poor discrimination for the unadjusted values $\hat{\boldsymbol{y}}_{Gustafson}$ depicted in Figure 4.3.

As discussed in [27], the cut-off point can be chosen in a way which minimizes the error rate. For example, a cut-off point such as

$$0.5 * (\max\{\text{adjusted values of non-leachers}\} + \min\{\text{adjusted values of leachers}\})$$

is preferable to 0.5, when the adjusted values are separated.

4.4.3 Diagnostic analysis for belief adjustment

After adjusting our prior beliefs, we use Bayes linear diagnostics to analyse the observed adjustments. This will help us to examine any conflict between data and prior specification. In this regard, useful diagnostic tools are, for example, the system resolution and the size ratio diagnostics, described in Section 2.3.

As in 2.24, the resolution transform matrix is

$$\begin{aligned} \mathrm{T}_{\boldsymbol{\eta}:\boldsymbol{y}} &= \mathrm{Var}(\boldsymbol{\eta})^{-1}\mathrm{RVar}_{\boldsymbol{y}}(\boldsymbol{\eta}) \\ &= \mathrm{Var}(\boldsymbol{\eta})^{-1}\mathrm{Cov}(\boldsymbol{\eta},\boldsymbol{y})\mathrm{Var}(\boldsymbol{y})^{-1}\mathrm{Cov}(\boldsymbol{y},\boldsymbol{\eta}) \end{aligned} \tag{4.13}$$

which can be computed as $A(I + A)^{-1}$, where A is given below 4.12.

As in 2.27, the system resolution for $\boldsymbol{\eta}$, $\mathrm{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$, or the overall proportion for uncertainty explained by the model is the trace of $\mathrm{T}_{\boldsymbol{\eta}:\boldsymbol{y}}$ divided by 43, the rank of the matrix $\mathrm{Var}(\boldsymbol{\eta})$. Here, the value of $\mathrm{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$ is 0.19, the proportion of uncertainty explained by the model, reflects the large degree of uncertainty about the values of k_{oc} and $t_{1/2}^{soil}$; see also [43].

As in 2.29, the size ratio for the adjustment of $\boldsymbol{\eta}$ by the data $\boldsymbol{y}$, $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$, is defined to be

$$\begin{aligned} \mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta}) &= \frac{\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{\eta})}{\mathrm{E}(\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{\eta}))} \\ &= \frac{[\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{\eta})-\mathrm{E}(\boldsymbol{\eta})]^T \mathrm{Var}(\boldsymbol{\eta})^{-1}[\mathrm{E}_{\boldsymbol{y}}(\boldsymbol{\eta})-\mathrm{E}(\boldsymbol{\eta})]}{\sum_{i=1}^{rT} \lambda_i} \end{aligned} \tag{4.14}$$

where $\mathrm{E}(\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{\eta})) = \sum_{i=1}^{rT} \lambda_i = \mathrm{trace}\{T_{\boldsymbol{\eta}:\boldsymbol{y}}\} = \mathrm{RU}_{\boldsymbol{y}}(\boldsymbol{\eta})$. Here, $\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{\eta}) = 9.961$, $\mathrm{E}(\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{\eta})) = 8.180$, and hence $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta}) = 1.218$ which is bigger than its expectation of unity. The lower and upper thresholds for $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$, as calculated by 2.31, are 0 and 2.158, so that the value of $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ is within this interval. [Note that the lower bound replaces the negative computed value.]

Note also that the posterior standard deviations of the components of $\boldsymbol{\eta}$, calculated from the posterior variance $\sigma^2 A(I+A)^{-1}$ (see Theorem 1 in Chapter 2), range from 0.09 to 0.26, compared to their prior standard deviations which range from 0.12 to 0.41.

However, the combination of (a) a value of $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{x\beta})$ greater than one, (b) the imperfect separation of leachers and non-leachers and (c) a concern that the prior variance matrix for $\boldsymbol{\beta}$, from the least squares analysis of Gustafson's data is overly precise, suggested investigating the effect of increasing the prior uncertainty for $\boldsymbol{\beta}$ in various ways.

4.4.4 Modifying prior beliefs about $\boldsymbol{\beta}$

The applicability of using the Gustafson's data as a source of prior information for $\boldsymbol{\beta}$ to analyse the EA data may be limited, but still usable; see Section 5.3 for more details. We investigate whether down-weighting prior information will improve the Bayes linear prediction for the EA data. Down-weighting may be achieved by modifying the prior variance-covariance matrix Σ_β, the estimated variance-covariance matrix 4.9 from the least squares analysis of 4.7. For example, Σ_β can be modified to be of the form $a\Sigma_\beta + bD_{\Sigma_\beta}$, where D_{Σ_β} is the diagonal of Σ_β and a and b are factors which can be chosen in different ways to reflect our uncertainty about prior information; see Section 5.3 for more details and a list of possible types of choice of a and b. The Bayes linear model is updated using the modified variance-covariance form $a\Sigma_\beta + bD_{\Sigma_\beta}$ for Σ_β. Minimum values of a and b resulting in perfect discrimination between leaching and non-leaching pesticides were determined experimentally as follows.

1. If $b = 0$, then any real value of $a \geq 7.7$ results in complete separation between leaching and non-leaching pesticides. For example, Figure 4.4, panel (b), shows the results for $a = 8$. We believe that the smallest value of a that gives complete separation is the best choice so as to avoid any exaggeration of our prior uncertainty using larger values; see the analysis of the choice of $a = 100$, below, where there is complete separation but the size ratio is further from the ideal value of unity than for the choice $a = 8$.

2. If $a = 1$, then any real value of $b \geq 3.2$ results in complete separation between leaching and non-leaching pesticides. For example, Figure 4.4, panel (c), shows the results for $b = 4$.

3. If $a = 0$, then any real value of $b \geq 4$ results in complete separation between leaching and non-leaching pesticides. For example, Figure 4.4, panel (d), shows the results for $b = 4$.

Var($\boldsymbol{\beta}$)	$\text{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$	$\text{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$	Size ratio interval
$\Sigma_{\boldsymbol{\beta}}$	0.190	1.218	(0, 2.158)
$8\Sigma_{\boldsymbol{\beta}}$	0.305	0.698	(0, 2.093)
$\Sigma_{\boldsymbol{\beta}} + 4D_{\Sigma_{\boldsymbol{\beta}}}$	0.366	0.604	(0, 2.044)
$4D_{\Sigma_{\boldsymbol{\beta}}}$	0.354	0.625	(0, 2.049)

Table 4.1: Various structure types of variance-covariance matrix of $\boldsymbol{\beta}$ together with summaries of diagnostic analysis for belief adjustment for the quadratic model, namely the system resolution $\text{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$, the size ratio $\text{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ and the size ratio interval. $D_{\Sigma_{\boldsymbol{\beta}}}$ denotes the diagonal of $\Sigma_{\boldsymbol{\beta}}$.

Furthermore, each of these choices for a and b resulted in an acceptable size ratio; see Table 4.1.

Note also from Figure 4.4 that the results for (1), (2) and (3) (above) are almost identical; see also Table 4.1.

Jointly choosing a and b to minimize the separation, or some form of Bayesian choice for a and b, is possible, but beyond the scope of this thesis.

Table 4.1 displays the system resolution $\text{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$, the size ratio $\text{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ and the size ratio interval, where negative lower bounds are replaced by zero. We notice that the models with modified variance-covariance matrices have (a) larger system resolution values and hence larger proportions of uncertainty are explained by the models, (b) acceptable size ratios which fall within approximate 95% credible limits, and (c) complete separation between the leaching and non-leaching pesticides.

It is worth to noting here that very large values for a and b were considered and resulted in complete separation between non-leachers and leachers, but some of the

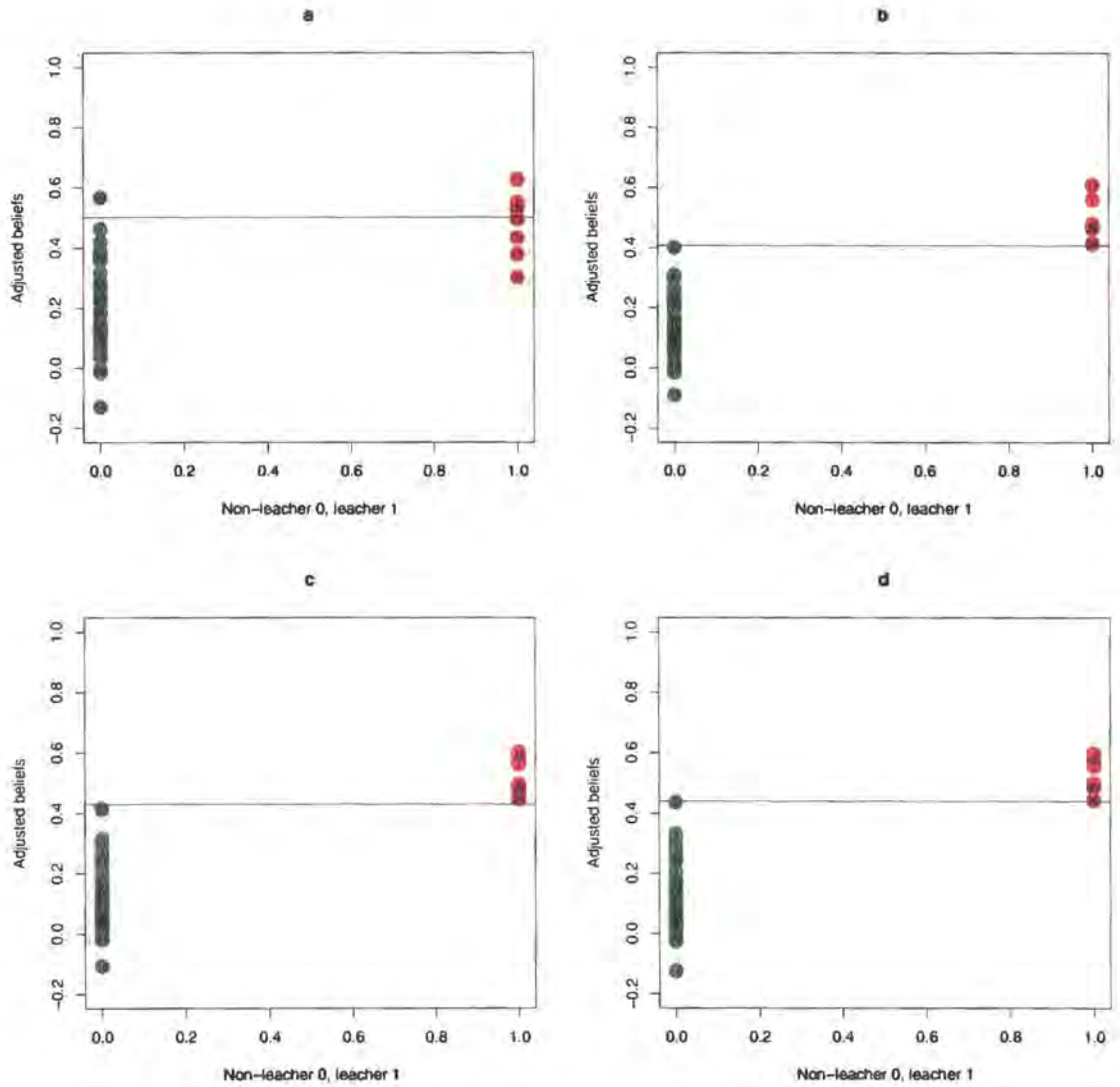

Figure 4.4: Bayes linear predictions for the EA data using the interaction model taking into account uncertainty in the covariates, where in panel (a) the original prior variance-covariance matrix $\Sigma_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is used, in panels (b), (c) and (d) the modified prior variance-covariance matrices $8\Sigma_{\boldsymbol{\beta}}$, $\Sigma_{\boldsymbol{\beta}}+4D_{\Sigma_{\boldsymbol{\beta}}}$ and $4D_{\Sigma_{\boldsymbol{\beta}}}$, respectively, are used, where $D_{\Sigma_{\boldsymbol{\beta}}}$ is the diagonal of $\Sigma_{\boldsymbol{\beta}}$. The cut-off points in panels (a), (b), (c) and (d) are 0.5, 0.41, 0.43 and 0.44, respectively.

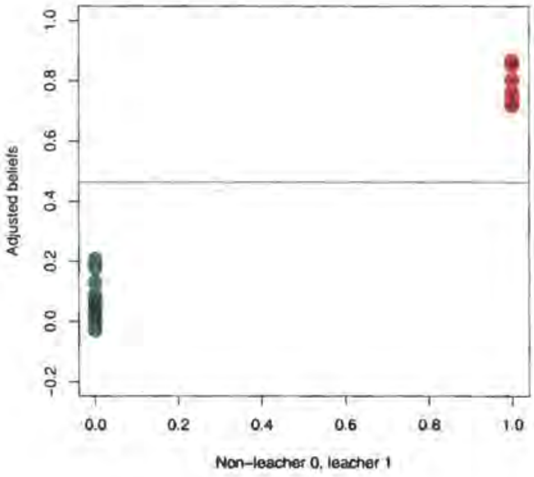

Figure 4.5: Bayes linear predictions for the EA data using the interaction model taking into account uncertainty in the covariates, where the prior variance-covariance matrix $\Sigma_{\boldsymbol{\beta}}$ is modified to $100\Sigma_{\boldsymbol{\beta}}$. The cut-off point is 0.46.

analyses of beliefs adjustment warn us of using such large values. As an example, the form $100\Sigma_{\boldsymbol{\beta}}$ results in complete separation, as depicted in Figure 4.5, with system resolution of 0.708. However, the size of the adjustment is 7.911 with expected value of 30.435, so the size ratio is 0.260 which is distant from its expectation of unity, but it is within its size ratio interval (0.066, 1.934), which "suggests that we have exaggerated our prior uncertainty"; see [24].

4.5 Further analysis of the linear discriminant

4.5.1 Background

The linear model, in 4.5, proposed in [43], will be subjected here to further diagnostic analysis. As in [43] and explained above, the linear regression analysis led to the

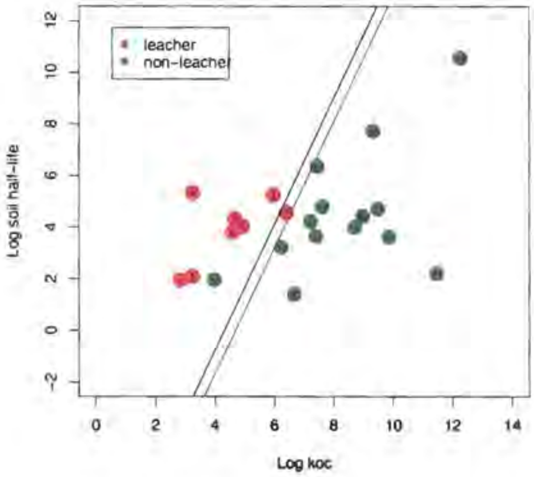

Figure 4.6: Linear discriminations based on Gustafson's data, see [43]. The black discriminant line is plotted using cut-off point of 0.5. The blue discriminant line is plotted using the cut-off point 0.44, which results in better discrimination, where only one pesticide is misclassified.

least squares fit

$$\hat{y} = 1.17 - 0.157z_1 + 0.064z_2 \tag{4.15}$$

and estimated variance-covariance matrix

$$s^2(\mathrm{G}^T\mathrm{G})^{-1} = \begin{bmatrix} 0.0479 & -0.0044 & -0.0029 \\ -0.0044 & 0.0010 & -0.0006 \\ -0.0029 & -0.0006 & 0.0017 \end{bmatrix} \tag{4.16}$$

where $s^2 = 0.116$ is the estimated error variance and G is the model matrix.

Figure 4.6 shows the values of $z_1 = \log k_{oc}$ and $z_2 = \log t_{1/2}^{soil}$ of Gustafson's data and the discriminant line (black)

$$1.17 - 0.157z_1 + 0.064z_2 = 0.5 \tag{4.17}$$

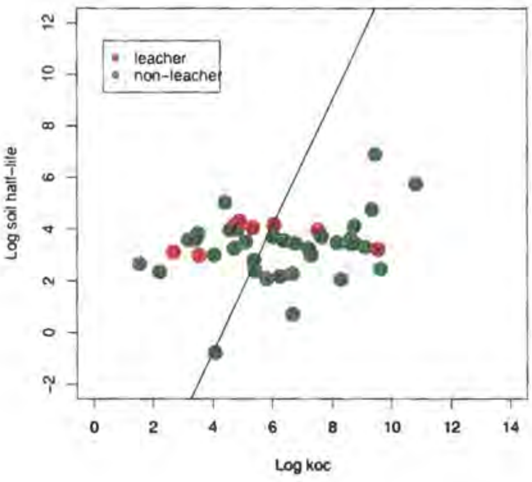

Figure 4.7: Linear discrimination for the means of $\log k_{oc}$ and $z_2 = \log t_{1/2}^{soil}$ where the discriminant line is derived from the analysis of Gustafson's data; see [43].

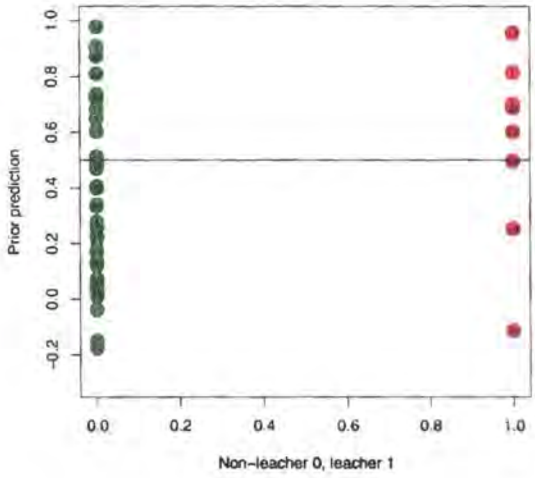

Figure 4.8: Predicted vs. observed leaching state for the EA data based on prior analysis of Gustafson's data; see [43].

The plot shows good discrimination between leaching and non-leaching pesticides with misclassification of two pesticides. However, as suggested in [27], this fit can be improved by choosing a different cut-point that minimizes error rate. For example, the choice 0.44 instead of 0.5 in 4.17 leads to a better discrimination as indicated by the blue line in Figure 4.6, where only one pesticide is misclassified.

Figure 4.7 shows the scatter plot of the means of k_{oc} and $t_{1/2}^{soil}$ for the 43 pesticides with the discriminator line in 4.17. There is poor discrimination apparent in Figure 4.7 and Figure 4.8 which shows the plot of $\hat{\boldsymbol{y}} = \boldsymbol{Mb}$ vs $\boldsymbol{y}$ where $\boldsymbol{M} = (1, \boldsymbol{m}_1, \boldsymbol{m}_2)$, $\boldsymbol{b}^T = (1.17, -0.157, 0.064)$ and $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ are the vectors of covariate means.

The poor discrimination was then tackled by means of Bayes linear estimation. The same prior specifications as in Section 4.4.1 were used with $\boldsymbol{b}^T = (1.17, -0.157, 0.064)$ and Σ_β in 4.16.

Then, the Bayes linear estimate $\hat{\boldsymbol{y}} = \mathrm{E}_{\boldsymbol{y}}(\boldsymbol{\eta})$ was used to update the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{4.18}$$

where $x_1 = z_1$ and $x_2 = z_2$. The Bayes linear estimate was computed as in 4.12 with $\boldsymbol{x}_i^T = (1, z_{i1}, z_{i2})$; see Example 1 in Section 2.3.1 for more calculation details.

The adjustment of $\boldsymbol{\eta}$ by $\boldsymbol{y}$ results in better discrimination, as depicted in Figure 4.9, panel (a), in comparison with the poor discrimination for the unadjusted values depicted in Figure 4.8.

In the following, diagnostic analyses for belief adjustment will be carried out. These analyses indicate conflict between the data and prior specifications. Where conflict does occur, we may choose to re-consider our prior specifications.

4.5.2 Diagnostic analyses

The same analyses used in Section 4.4.3 will be applied here, namely the system resolution and the size ratio diagnostics.

The system resolution $\mathrm{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$ is 0.21, reflecting the large degree of uncertainty about the values of k_{oc} and $t_{1/2}^{soil}$. The system resolution was calculated in [43], but a different value was obtained due to an error there; see Section 1.4 for more detail.

The size of the adjustment $\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{\eta})$ is 17.114 with an expected size $\mathrm{E}(\mathrm{Size}_{\boldsymbol{y}}(\boldsymbol{\eta}))$ of 8.869, giving a size ratio $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ of 1.930 which is bigger than its expectation of unity. The lower and upper thresholds for $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ are 0 and 2.187, so that the value of $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ is within its 95% limits.

As in the quadratic model analysis above, a value of $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ greater than one, the modest Bayes prediction shown in Figure 4.9, panel (a), and the limitation of Gustafson's data as a source of prior information, all support the view of possible conflict between the prior beliefs and the data. This leads us to re-structure our prior beliefs about $\boldsymbol{\beta}$ to improve the Bayes linear prediction.

As in Section 4.4.4, the form $a\Sigma_{\boldsymbol{\beta}} + bD_{\Sigma_{\boldsymbol{\beta}}}$ can be used to re-structure our prior beliefs about $\Sigma_{\boldsymbol{\beta}}$. Table 4.2, displays the best forms of prior beliefs together with results from the diagnostic analyses for belief adjustment. As seen from the table, the modifications of $\Sigma_{\boldsymbol{\beta}}$ help to increase slightly the system resolutions which means that the proportions of uncertainty explained by the models are slightly increased. Furthermore, the size ratios are close to their expectations of unity and are all within their lower and upper limits. Finally, the modifications of $\Sigma_{\boldsymbol{\beta}}$ result in better Bayes linear predictions as shown in Figure 4.9, panels (b), (c) and (d), where complete

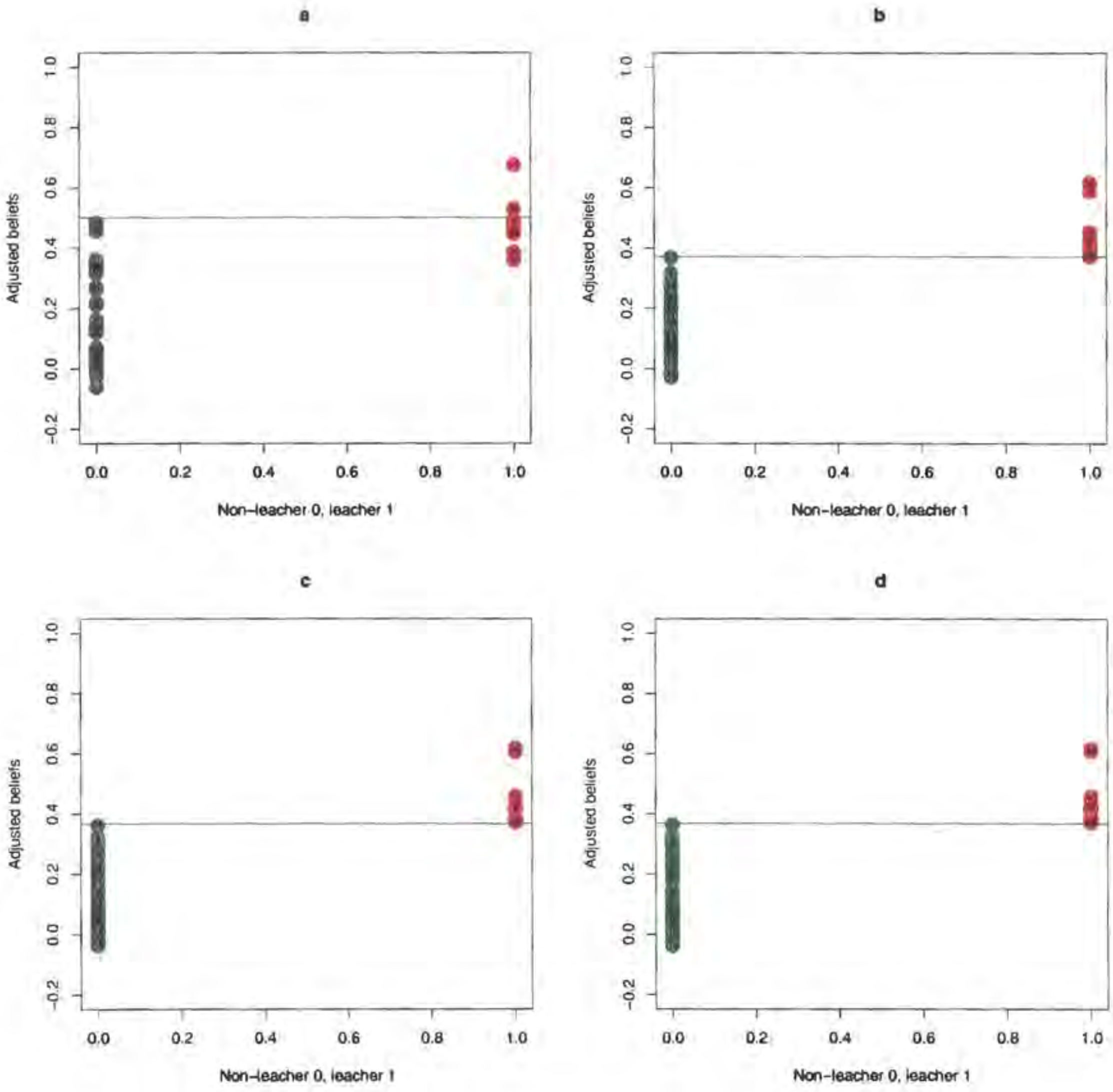

Figure 4.9: Bayes linear predictions for the EA data for the linear model proposed in [43] taking into account uncertainty in the covariates, where in panel (a) the original prior variance-covariance matrix $\Sigma_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is used, in panels (b), (c) and (d) the modified prior variance-covariance matrices $7\Sigma_{\boldsymbol{\beta}}$, $\Sigma_{\boldsymbol{\beta}} + 7D_{\Sigma_{\boldsymbol{\beta}}}$ and $8D_{\Sigma_{\boldsymbol{\beta}}}$, respectively, are used, where $D_{\Sigma_{\boldsymbol{\beta}}}$ is the diagonal of $\Sigma_{\boldsymbol{\beta}}$. The cut-off points in panels (a), (b), (c) and (d) are 0.5, 0.37, 0.37 and 0.37, respectively.

$\mathrm{Var}(\boldsymbol{\beta})$	$\mathrm{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$	$\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$	Size ratio interval
$\Sigma_{\boldsymbol{\beta}}$	0.206	1.930	(0, 2.187)
$7\Sigma_{\boldsymbol{\beta}}$	0.291	0.925	(0, 2.134)
$\Sigma_{\boldsymbol{\beta}} + 7D_{\Sigma_{\boldsymbol{\beta}}}$	0.303	0.932	(0, 2.126)
$8D_{\Sigma_{\boldsymbol{\beta}}}$	0.302	0.946	(0, 2.125)

Table 4.2: Various structure types of variance-covariance matrix of $\boldsymbol{\beta}$ together with summaries of diagnostic analysis for belief adjustment for the linear model, namely the system resolution $\mathrm{R}_{\boldsymbol{y}}(\boldsymbol{\eta})$, the size ratio $\mathrm{Sr}_{\boldsymbol{y}}(\boldsymbol{\eta})$ and the size ratio interval, where the negative lower bound is replaced by zero. $D_{\Sigma_{\boldsymbol{\beta}}}$ denotes the diagonal of $\Sigma_{\boldsymbol{\beta}}$.

separation between the two groups of pesticides is achieved.

4.6 Conclusion

The primary aim of this chapter has been to investigate the Bayes linear approach suggested in [43] with an interaction term in the linear predictor. The Bayes linear approach combines prior knowledge of uncertainty with observational data using expectation. The general results developed in Chapter 2 were used to compute the Bayes linear estimate of the linear predictor. Based on misclassification statistics, the interaction model is slightly better than the original linear model proposed in [43]. A brief discussion on regression analysis, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) was provided.

The beliefs adjustments were analysed using certain Bayes linear diagnostic tools, such as system resolution and size ratio. This analysis suggests that the prior beliefs can be reformulated to improve the belief adjustments. Furthermore, such a reformulation resulted in complete separation between the leaching and non-leaching pesticides.

Finally, additional analyses were carried out for the linear model proposed in [43]. These analyses led also to complete separation between the two groups of the EA data.

Chapter 5

Full Bayes methods for analysing pesticide contamination

5.1 Introduction

The aim of this chapter is to demonstrate the use of classical Bayesian methods (as distinct from Bayes linear methods) in exploring whether or not it is possible to achieve discriminate between leachers and non-leachers on the basis of two of their chemical properties, the adsorption coefficient k_{oc} and soil half-life $t_{1/2}^{soil}$, when the monitored values of these properties are uncertain, in the sense that we only have a range of values reported for both k_{oc} and $t_{1/2}^{soil}$ for each pesticide.

The Bayesian methods use logistic regression and prior information from (i) analysis of Gustafson's data (where a single value is reported for both k_{oc} and $t_{1/2}^{soil}$) for the model parameters and (ii) the ranges of the chemical properties for these covariates from USDA database. The proposed models are analysed by means of Markov Chain Monte Carlo (MCMC) simulation techniques using the freely available

WinBUGS software and R package.

The chapter is organised as follows. Section 5.2 describes the methodology used and Section 5.3 provides a general description of the proposed models. Section 5.4 discusses discrimination for Gustafson's data and whether or not this discrimination provides a good prediction for the EA data. Section 5.5 describes the Bayesian methods developed to analyse the EA data. Sections 5.6 and 5.7 explain how to implement the proposed models in the WinBUGS software and the R package. Section 5.8 assesses convergence of the MCMC algorithm and Section 5.9 discusses the results obtained from the MCMC analysis and how to strengthen these results using the multivariate runs test described in Chapter 2.

5.2 Methodology

The main purpose of this chapter is to develop Bayesian methods for logistic regression with uncertain covariates using MCMC simulation techniques. Classical Bayesian analysis (unlike Bayes linear methods) requires specifying a full probability model for the data to be used as a likelihood function and a prior distribution for all of the unknown parameters. The core of Bayesian statistics is the combination of the likelihood together with the prior distribution to produce a posterior distribution for the parameters of interest from which inferences can be drawn. However, calculating marginal posterior distributions requires high dimensional integrations which may not be available in a closed form. One suggestion for making this kind of calculation more tractable is to use MCMC simulation. MCMC is implemented here using the WinBUGS software and the R package. Graphical models, as described in sections 2.4.4 and 2.4.6, are used to provide a non-algebraic description of a proposed model.

Logistic models with and without an interaction term will be investigated. An awkward additional feature of Bayesian analysis for the interaction model is the complete separation of leachers and non-leachers in Gustafson's data, a situation where MLE fails; see Chapter 2 for details.

The study starts with discrimination for the Gustafson data. This includes a brief discussion of the method proposed by Worrall et al. in [44] using logistic regression with an interaction term to analyse this data. Then the study moves on to investigate whether the discriminant line proposed in [44] or the discriminant curves derived from Gustafson's data provide good prediction for the EA data. This leads to different Bayesian methods to analyse the EA database pesticides. The analysis starts by considering three main effects models (Models 1, 2 and 3) and their corresponding extensions to include an interaction term (Models 1*, 2* and 3*). Each model uses the same sources of prior information. The models are compared using the Deviance Information Criterion (DIC) proposed in [39]. Finally, as in [45], the conclusions of the proposed models will be strengthened using the multivariate runs test proposed in [17] to test here the degree of separation between leaching and non-leaching pesticides.

5.3 General descriptions of the proposed models

The models are implemented in two stages as follows.

Stage 1: Likelihood

In this stage the following logistic regression model is adopted. Let $z_1 = \log k_{oc}$, $z_2 = \log t_{1/2}^{soil}$ and

$$y_i = \begin{cases} 1 & \text{if ith pesticide is monitored as a leacher} \\ 0 & \text{if ith pesticide is monitored as a non-leacher.} \end{cases}$$

For the unstarred models $y_i \sim \text{Bernoulli}(\pi_i)$, where $\boldsymbol{z}_i = (z_{i1}, z_{i2})$ is linked to $\text{E}(y_i) = \pi_i$ by the logit function,

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = z_i^T\beta = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} = \eta_i \tag{5.1}$$

where π_i is the probability that the ith pesticide will leach into the groundwater, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ is a vector of unknown parameters and η_i is the linear predictor.

The starred models use a linear predictor with an interaction term, leading to a discrimination curve $\pi = 0.5$ in the (z_1, z_2) plane. One suggestion for drawing this curve is to model the logit link function as an additive function of z_2 and an interaction term of z_1 and z_2, by analogy with the GUS curve of Gustafson in [25]. In this case, the linear predictor is

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{z}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 z_{i2} + \beta_2 z_{i1} z_{i2} = \eta_i \tag{5.2}$$

A Bernoulli distribution for each component of $\boldsymbol{y}$ is common for all proposed models, except for models 3 and 3*, where each component is assigned a Binomial distribution. Moreover, for all models, the components of $\boldsymbol{y}$ are assumed to be conditionally independent given $\boldsymbol{\beta}$ and the values of k_{oc} and $t_{1/2}^{soil}$, so that likelihood for $\boldsymbol{\beta}$, k_{oc} and $t_{1/2}^{soil}$ can be written

$$p(\boldsymbol{y}|\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i|z_{i1}, z_{i2}, \boldsymbol{\beta}). \tag{5.3}$$

Stage 2: Prior distributions

The prior information for all of the unknowns is chosen at this stage, namely for $\boldsymbol{\beta}$ and the $\boldsymbol{z}_i$. The form of prior information will differ for each model.

The prior information for the vector $\boldsymbol{\beta}$ is derived from Gustafson's data. However, it is believed that there are differences between the local environments of California, where the Gustafson data was collected, and the United Kingdom, where the EA data was collected. For example, "California climate and soil types are relatively homogeneous compared with the general range of soils and types encountered across Europe and the United States" [43]. Hence, the applicability of using the Gustafson data as prior information to analyse the EA data may be limited, but still usable. The perfect separation of leachers and non-leachers for Gustafson's data, based on k_{oc} and $t_{1/2}^{soil}$, see Figure 1.2, suggests what might be expected from our analysis of the EA data. Also, we investigate whether down weighting the (limited) information derived from Gustafson data will lead to improved discrimination for the EA data. Down weighting information from previous studies because of limitations of applying that information to a current study is discussed in [11] and [5], for example.

Down weighting prior information may be achieved in diffrent ways. For example, in the case of several unknowns, like here, a prior variance-covariance matrix, derived from a previous study, may be modified to down weight prior information. For example, in the present application, the prior variance-covariance matrix $\Sigma_{\boldsymbol{\beta}}$ can be modified to the form $a\Sigma_{\boldsymbol{\beta}} + bD_{\Sigma_{\boldsymbol{\beta}}}$, where $D_{\Sigma_{\boldsymbol{\beta}}}$ is the diagonal of $\Sigma_{\boldsymbol{\beta}}$ and a and b are factors which can be chosen in different ways to reflect our uncertainty about prior information. In this thesis, we consider three types of choice of a and b as follows.

1. $b = 0$ with different choices of a, for example $a = 4, 25, 64$ and 100. This choice, discussed in [11] and [5] for a scalar parameter, inflates the variances but the correlation structure is the same as for Σ_{β}.

2. $a = 1$ with different choices of b, for example $b = 0, 3, 24, 63$ and 99. In this choice, only variances are inflated, but the magnitudes of correlations between the different β parameters will be reduced. Notice that $b = 0$ corresponds to no down-weighting.

3. $a = 0$ with different choices of b, for example $b = 4, 25, 64$ and 100. In this choice, the variances are inflated and covariances are omitted.

We will investigate, experimentally, the above forms of variance-covariance modification to determine a suitable choice of prior information. Choice will be based on issues such as discrimination power and MCMC convergence.

As mentioned earlier, we start with Gustafson's data to see whether or not the discriminant there provides good prediction for the EA data.

5.4 Predicting the EA data using Gustafson's data

Figure 5.1 plots the values of k_{oc} and $t_{1/2}^{soil}$ for the 22 pesticides from Gustafson's data together with various discriminant lines and curves for discriminating between leaching and non-leaching pesticides. The dotted curve is from [25] representing $GUS = 2.3$, where GUS is the groundwater ubiquity score described by equation 1.1. It gives good separation with two cases, case 1 with $GUS = 2.334$ and case 5 with

$GUS = 2.327$, located close to the boundary. The black line $\hat{\eta}_{Worrall} = 0$ is from [44], where

$$\hat{\eta}_{Worrall} = 8.72 - 3.59z_1 + 2.86z_2$$

It misclassifies 3 cases: the leacher Prometryn (case 7), the non-leacher 2, 4-D (case 12) and the non-leacher Lindane (case 18). This discriminant line is improved in this thesis by fitting a logistic regression model with an interaction term, as described in Chapter 3. The coefficients were estimated using the WEMEL scheme [8]. The fitted linear predictor is

$$\hat{\eta}_{WEMEL} = -20.384 + 17.380z_{i2} - 1.947z_{i1}z_{i2}$$

and the blue curve corresponds to $\hat{\eta}_{WEMEL} = 0$. It gives complete separation between leaching and non-leaching pesticides. The red discriminant line $1.17 - 0.157z_1 + 0.0642z_2 = 0.5$ is from [43], which was estimated by least squares. This model misclassifies two cases: the leacher Prometryn (case 7) and the non-leacher 2, 4-D (case 12). The least squares fit to the interaction model gives the yellow discriminant curve $-0.028 + 0.310z_2 - 0.029z_1z_2 = 0.5$. It misclassifies 3 cases: the leacher Aldicarb (case 1), the leacher Oxamyl (case 5) and the non-leacher Lindane (case 18). We see that the WEMEL fit to the interaction model is the best of these five discriminants with zero misclassification rate.

Now, we move to explore the ability of the above discriminant lines and curves to discriminate the EA pesticides. As mentioned previously, each pesticide from the EA data has a range of values reported by the USDA for k_{oc} and $t_{1/2}^{soil}$. The plot of the k_{oc} and $t_{1/2}^{soil}$ means for the 43 EA pesticides in Figure 5.2 shows very poor

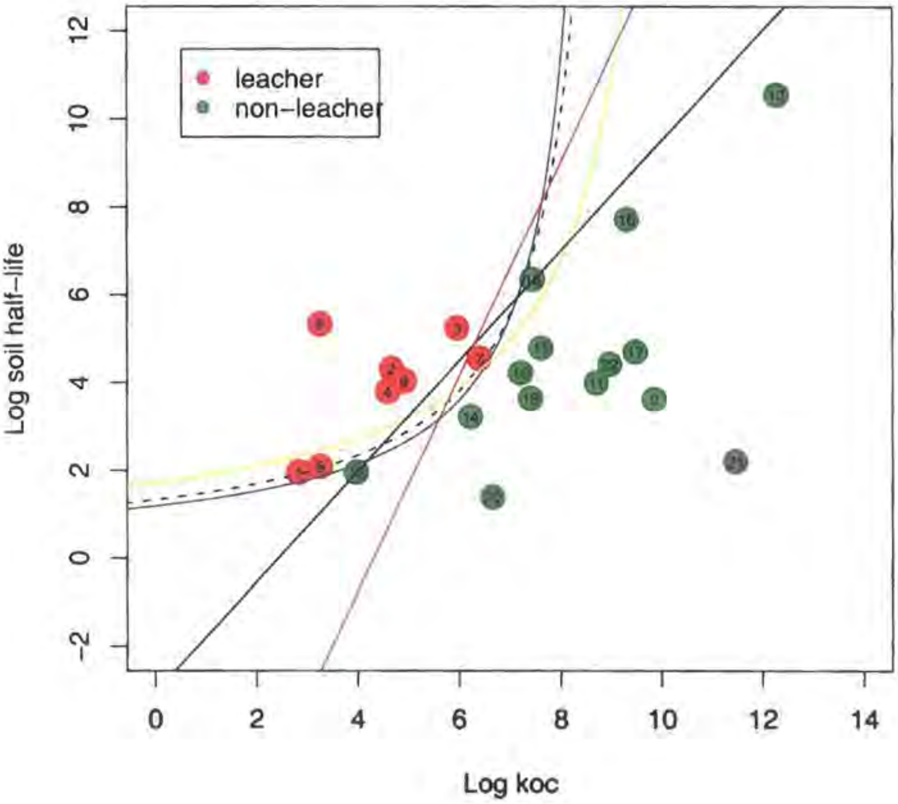

Figure 5.1: The various lines and curves used to discriminate Gustafson's data. The dotted curve is from [25] representing $GUS = 2.3$, the black line $\hat{\eta}_{Worrall} = 0$ is from [44], the blue curve corresponds to $\hat{\eta}_{WEMEL} = 0$, the red line $1.17 - 0.157z_1 + 0.0642z_2 = 0.5$ is from [43], which was estimated by least squares and the yellow curve represents model 4.10; see Chapter 4.

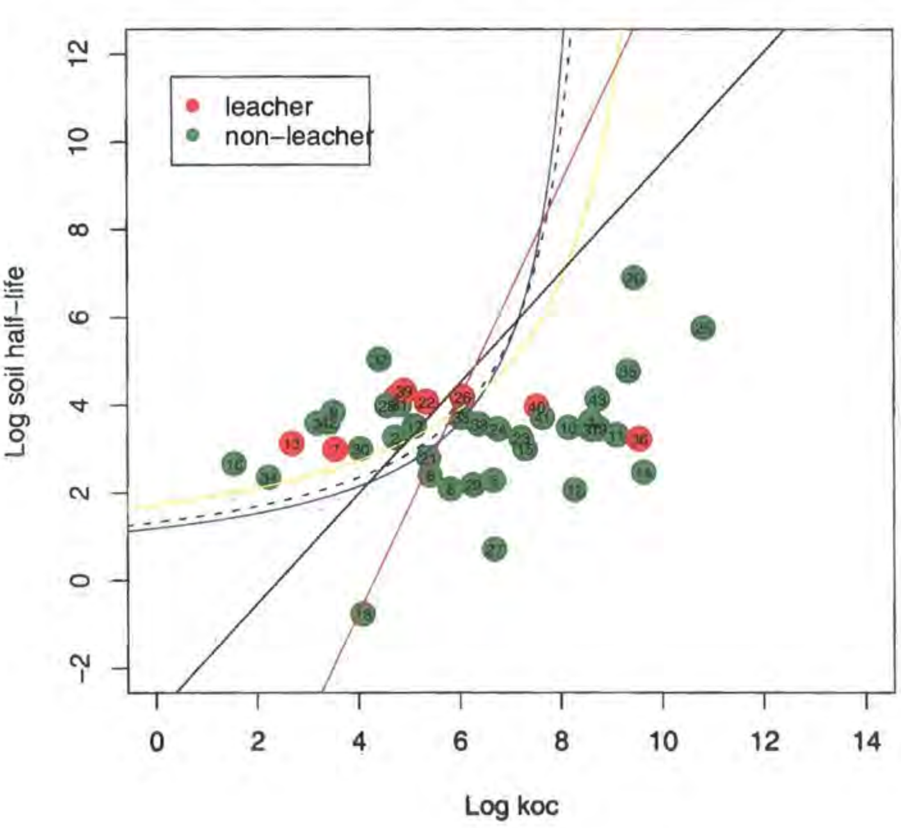

Figure 5.2: The EA data together with the various discriminant lines and curves derived from the analysis of Gustafson's data, depicted in Figure 5.1.

discrimination for each of the discriminant lines or curves derived from Gustafson's data and hence poor prediction for the means of the EA data. This figure also demonstrates that the prior contention, that leachers appear in the NW corner and non-leachers in the SE corner, proposed in [43] and [45], is not the case for the EA data means. This inconsistency, which was referred to in [43] and [45], was discussed in Chapter 1.

5.5 Bayesian methods for analysing the EA data

The poor discrimination for the EA data has been addressed in two references. In [43], a Bayes linear method was developed where the available means and variances for k_{oc} and $t_{1/2}^{soil}$ were chosen as prior information for these covariates. Also, a linear regression model was fitted to Gustafson's data (the red line in Figure 5.1). The parameter estimates and their standard errors derived from this linear regression were chosen to provide the prior information regarding the coefficient parameters. The posterior prediction, Figure 1.6, shows better discrimination for the EA data.

An alternative simple data-analytic attempt to discriminate the pesticides of the EA data was proposed in [45]. A data combination was chosen to make the EA data most consistent with the prior contention that the leaching pesticides are found in the NW corner and the non-leaching pesticides are found in the SE corner of the k_{oc} and $t_{1/2}^{soil}$ space. This contention was implemented by choosing for each leaching pesticide the combination with the lowest k_{oc} and the highest $t_{1/2}^{soil}$ and for non-leaching pesticide the highest k_{oc} with lowest $t_{1/2}^{soil}$. The method leads to a complete separation of the leaching and non-leaching pesticides; see Figure 1.8.

The Bayesian methods developed here tackle the problem of uncertain covariates, and are implemented by means of MCMC simulation. A logistic regression model with a logit link is modelled either as an additive function of the explanatory variables k_{oc} and $t_{1/2}^{soil}$ as in equation 5.1 or by equation 5.2. In both cases, there are two sources of uncertainty, namely (a) uncertainty in the model parameters $\boldsymbol{\beta}$ and (b) uncertainty in the values of z_{i1} and z_{i2} for each pesticide.

5.5.1 Model 1

The crucial idea behind this model is to regard the linear predictor η described in equation 5.1 as an unknown random quantity. This leads to a choice of prior information for η. A multivariate normal distribution is chosen to be the prior for $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{43})$, where the means and covariances will be derived from those for z_1, z_2 and $\boldsymbol{\beta}$. The two stages as follows.

Stage 1:

As described in Section 5.3, y_i is modeled as a Bernoulli random quantity with the leaching probability π_i, such that

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i. \tag{5.4}$$

where the vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)$ $(n = 43)$ is an unknown random quantity. Furthermore, η will be modeled as the additive function of z_1 and z_2 as follows:

$$\eta = \boldsymbol{z}^T\boldsymbol{\beta} = \beta_0 + \beta_1 z_1 + \beta_2 z_2. \tag{5.5}$$

Stage 2:

As mentioned in Stage 1 above, η is an unknown random quantity, so that prior information for η needs to be assessed. A multivariate normal distribution with a mean vector $\mathrm{E}(\boldsymbol{\eta})$ and variance matrix $\mathrm{Var}(\boldsymbol{\eta})$ is chosen as the prior distribution for $\boldsymbol{\eta}$. Treating η as a bi-linear function of z_1, z_2 and $\boldsymbol{\beta}$ in equation 5.5 requires a choice of prior information for $\boldsymbol{\beta}$, z_1 and z_2 in order to derive the parameters of multivariate normal prior distribution for $\boldsymbol{\eta}$. In the absence of full distributional assumptions for z_1 and z_2, the same prior information suggested in [43] will be chosen here. In particular, prior information for z_1 and z_2 are chosen in the form of means and variances, i.e.

$$\mathrm{E}(\boldsymbol{z}_1) = \boldsymbol{m}_1, \quad \mathrm{Var}(\boldsymbol{z}_1) = \boldsymbol{v}_1$$

$$\mathrm{E}(\boldsymbol{z}_2) = \boldsymbol{m}_2, \quad \mathrm{Var}(\boldsymbol{z}_2) = \boldsymbol{v}_2 \quad \text{and } \mathrm{Cov}(\boldsymbol{z}_1, \boldsymbol{z}_2) = \mathbf{0}$$

where $\boldsymbol{m}_i$ is the mean vector of $\boldsymbol{z}_i$ and $\boldsymbol{v}_i$ is the diagonal matrix of the variances of $\boldsymbol{z}_i$ for $i = 1, 2$.

Prior information for $\boldsymbol{\beta}$ is derived from the logistic regression model analysis of Gustafson's data, suggested in [44]. Specifically,

$$\mathrm{E}(\boldsymbol{\beta}) = \boldsymbol{b} \quad \text{and} \quad \mathrm{Var}(\boldsymbol{\beta}) = \left[\frac{1}{2}Dev''(\boldsymbol{\beta})\right]^{-1}_{\boldsymbol{\beta}=\boldsymbol{b}} = \Sigma_\beta$$

where $\boldsymbol{b}$ is the MLE estimate $\hat{\beta}$ of $\boldsymbol{\beta}$ and Dev is the deviance function. In particular,

$$\boldsymbol{b}^T = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (8.72, -3.59, 2.86) \tag{5.6}$$

and

$$\Sigma_\beta = \begin{bmatrix} 21.48 & -7.75 & 5.64 \\ -7.75 & 3.90 & -3.43 \\ 5.64 & -3.43 & 3.28 \end{bmatrix}. \tag{5.7}$$

By experiment it was found that the 'best', in the sense of best discrimination result, prior variance-covariance type of $\boldsymbol{\beta}$ is type (2), discussed in Section 5.3, with $a = 1$ and $b = 0$. This means that Σ_{β} in 5.7 will be chosen.

Hence, using Theorem 1 and Example 1 in Section 2.3.1 with $x_1 = z_1$ and $x_2 = z_2$, the mean vector $\mathrm{E}(\boldsymbol{\eta})$ and the variance matrix $\mathrm{Var}(\boldsymbol{\eta})$ of the prior multivariate normal distribution for $\boldsymbol{\eta}$ can be derived using all the prior information as follows:

$$\mathrm{E}(\boldsymbol{\eta}) = \mathrm{E}(X\boldsymbol{\beta}) = \boldsymbol{M}\boldsymbol{b} \qquad (5.8)$$

$$\mathrm{Var}(\boldsymbol{\eta}) = \mathrm{Var}(X\boldsymbol{\beta}) = \boldsymbol{M}\Sigma_{\beta}\boldsymbol{M}^T + D \qquad (5.9)$$

where $\boldsymbol{M} = \mathrm{E}[X]$ and D is a diagonal matrix with $D_{ii} = \boldsymbol{b}^{\mathrm{T}}\Sigma_i\,\boldsymbol{b} + \mathrm{trace}[\Sigma_i\Sigma_{\beta}]$, where $\Sigma_i = \mathrm{Var}[\boldsymbol{x}_i]$, the variance matrix of $\boldsymbol{x}_i^T = (1, z_{i1}, z_{i2})$, the i-th row of X; see Example 1 in Section 2.3.1 for more calculation details.

5.5.2 Model 1*

This model is a modification of Model 1 with an interaction term. As with Model 1, there are two stages

Stage 1:

This stage is as in Model 1 with the exception that η is treated as an additive function of z_2 and an interaction term of z_1 and z_2 as follows:

$$\eta = \beta_0 + \beta_1 z_2 + \beta_2 z_1 z_2$$

Stage 2:

This model uses the same sources of prior information used in Model 1 for the covariates z_1 and z_2 and $\boldsymbol{\beta}$. However, the difference is that the MLE estimator for $\boldsymbol{\beta}$ for the Gustafson data does not exist due to complete separation of leachers and non-leachers in the space of the covariates z_1 and z_2. As suggested in [7], this problem can be resolved by using MEL or the robust version, WEMEL. Adopting the latter estimator gives:

$$\boldsymbol{b}^T = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (-20.384, 17.380, -1.947) \tag{5.10}$$

and

$$\Sigma_\beta = \begin{bmatrix} 206.81 & -164.51 & 17.88 \\ -164.51 & 134.39 & -14.75 \\ 17.88 & -14.75 & 1.63 \end{bmatrix}. \tag{5.11}$$

As in Model 1, by experiment, it was found that Σ_β in 5.11 is the best choice to express our prior beliefs about our uncertainty for $\boldsymbol{\beta}$.

Using Theorem 1 and Example 2 in Section 2.3.1 with $x_1 = z_2$ and $x_2 = z_1 z_2$, the vector mean and covariance matrix for the prior multivariate normal distribution of $\boldsymbol{\eta}$ are

$$\begin{aligned} \mathrm{E}(\boldsymbol{\eta}) &= \mathrm{E}(X\boldsymbol{\beta}) = \boldsymbol{M}\boldsymbol{b} \\ \mathrm{Var}(\boldsymbol{\eta}) &= \mathrm{Var}(X\boldsymbol{\beta}) = \boldsymbol{M}\Sigma_\beta \boldsymbol{M}^T + D \end{aligned}$$

where $\boldsymbol{M} = \mathrm{E}[X]$ and D is a diagonal matrix with $D_{ii} = \boldsymbol{b}^{\mathrm{T}}\Sigma_i \boldsymbol{b} + \mathrm{trace}[\Sigma_i \Sigma_\beta]$, where $\Sigma_i = \mathrm{Var}[\boldsymbol{x}_i]$, the variance matrix of $\boldsymbol{x}_i^T = (1, z_{i2}, z_{i1} z_{i2})$, the i-th row of X; see Example 2 in Section 2.3.1 for more calculation details.

5.5.3 Model 2

This model differs from Model 1 in the sense that each of the unknown random quantities z_1, z_2 and $\boldsymbol{\beta}$ will be assigned a prior probability distribution.

The two stages are

Stage 1:

As before, y_i is modeled as a Bernoulli variate with the leaching probability π_i and linear predictor η given by

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{z}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} = \eta_i \tag{5.12}$$

where the z_{i1}, z_{i2} and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2)$ are all unknown random quantities.

Stage 2:

Since the z_{i1}, z_{i2} and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2)$ are all unknown, prior information should be chosen for each of them. In this regard, independent normal distributions are chosen to represent prior information for $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ and each of their components. The means and variances of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ from the USDA data base are used as the means and variances for the prior normal distributions, i.e. z_{i1} will be assigned a normal distribution Normal(m_{i1}, v_{i1}) with mean m_{i1} and variance v_{i1} and z_{i2} will be assigned a normal distribution Normal(m_{i2}, v_{i2}) with mean m_{i2} and variance v_{i2}.

A multivariate normal distribution with a mean vector $\boldsymbol{b}$ and a variance-covariance matrix $\boldsymbol{V}$ of the form $a\Sigma_{\boldsymbol{\beta}} + bD_{\Sigma_{\boldsymbol{\beta}}}$ is chosen to represent prior information for $\boldsymbol{\beta}$, where $\Sigma_{\boldsymbol{\beta}}, a, b$ and $D_{\Sigma_{\boldsymbol{\beta}}}$ are explained in Section 5.3. Both of $\boldsymbol{b}$ and $\Sigma_{\boldsymbol{\beta}}$ are derived from the logistic regression analysis of Gustafson's data, as suggested in [44]. The derived values are given in equations 5.6 and 5.7 for Model 1.

This model was investigated experimentally using the various types of prior variance-covariance matrices for $\boldsymbol{\beta}$, i.e. using the same choices of factors a and b as in Section 5.3. The 'best' choices of a and b are:

1. If $b = 0$, then the best choice of a is 100, i.e. $\boldsymbol{V} = 100\Sigma_{\boldsymbol{\beta}}$

2. If $a = 1$, then the best choice of b is 24, i.e. $\boldsymbol{V} = \Sigma_{\boldsymbol{\beta}} + 24D_{\Sigma_{\boldsymbol{\beta}}}$

3. If $a = 0$, then the best choice of b is 4, i.e. $\boldsymbol{V} = 4D_{\Sigma_{\boldsymbol{\beta}}}$

As mentioned in Section 5.3, these choices give the best discrimination between leaching and non-leaching pesticides and have the additional benefit of making the MCMC algorithm converge quickly to the target posterior distribution.

All of the above best choices give very similar discriminant results. Hence, the second choice, $a = 1$ and $b = 24$, is chosen to illustrate this current model in detail. From now onward, the prior variance-covariance matrix of $\boldsymbol{\beta}$ in Model 2 is chosen to be:

$$\boldsymbol{V} = \Sigma_{\boldsymbol{\beta}} + 24D_{\Sigma_{\boldsymbol{\beta}}} = \begin{bmatrix} 537 & -8 & 6 \\ -8 & 98 & -3 \\ 6 & -3 & 82 \end{bmatrix} \tag{5.13}$$

The above prior distributions are summarised as follows:

$$z_{i1} \sim Normal\left(m_{i1}, v_{i1}\right) \tag{5.14}$$

$$z_{i2} \sim Normal\left(m_{i2}, v_{i2}\right) \tag{5.15}$$

$$\boldsymbol{\beta} \sim MVN\left(\boldsymbol{b}, \boldsymbol{V}\right) \tag{5.16}$$

Furthermore, we assume that the z_{i1}, z_{i2} and $\boldsymbol{\beta}$ are a priori independent. Then, as in section 2.4.6, the posterior distribution of $\boldsymbol{z}_1, \boldsymbol{z}_2$ and $\boldsymbol{\beta}$ given the data $\boldsymbol{y}$ can be written as:

$$p\left(\boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{\beta} | \boldsymbol{y}\right) \propto p\left(\boldsymbol{y} | \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{\beta}\right) \cdot p(\boldsymbol{z}_1) \cdot p(\boldsymbol{z}_2) \cdot p(\boldsymbol{\beta}) \tag{5.17}$$

where $p\left(\boldsymbol{y} | \boldsymbol{z}_1, \boldsymbol{z}_2, \boldsymbol{\beta}\right)$ is the likelihood function and $p(\boldsymbol{z}_1), p(\boldsymbol{z}_2)$ and $p(\boldsymbol{\beta})$ are the prior distributions for $\boldsymbol{z}_1, \boldsymbol{z}_2$ and $\boldsymbol{\beta}$, respectively. This posterior distribution can be expressed as:

$$\begin{aligned} p(z_1, z_2, \beta|y) \propto & \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} \cdot \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\left(\tfrac{z_{i1}-m_{i1}}{v_{i1}}\right)^2\right\} \\ & \cdot \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\left(\tfrac{z_{i2}-m_{i2}}{v_{i2}}\right)^2\right\} \\ & \cdot \exp\left\{-\tfrac{1}{2}(\beta-b)^T V^{-1}(\beta-b)\right\} \end{aligned} \qquad (5.18)$$

where $n = 43$ and

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2})}{1 + \exp(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2})}.$$

5.5.4 Model 2*

This is a modification of Model 2 with an interaction term. It uses the same prior distributions for both z_1 and z_2 as in Model 2. As in Model 2, a multivariate normal distribution with a mean vector $\boldsymbol{b}$ and variance matrix $\boldsymbol{V}$ is chosen as the prior information for $\boldsymbol{\beta}$, where $\boldsymbol{V}$ has the same form with the same choices of the factors a and b. However, the mean vector $\boldsymbol{b}$ and the variance-covariance matrix $\Sigma_{\boldsymbol{\beta}}$ are from the analysis of Gustafson's data using the WEMEL scheme.

The two stages are

Stage 1:

This stage is as in stage 1 in Model 2, but with linear predictor given by

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = z_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 z_{i2} + \beta_2 z_{i1} z_{i2} = \eta_i. \qquad (5.19)$$

Stage 2:

The covariates z_1 and z_2 are assigned the same prior distributions as in Model 2 and $\boldsymbol{\beta}$ is assigned a multivariate normal prior distribution with a mean vector $\boldsymbol{b}$ and variance-covariance matrix $\boldsymbol{V}$ of the form $a\Sigma_{\boldsymbol{\beta}} + bD_{\Sigma_{\boldsymbol{\beta}}}$, where $\boldsymbol{b}$ and $\Sigma_{\boldsymbol{\beta}}$, derived from the WEMEL logistic regression analysis of Gustafson's data, are given in 5.10 and 5.11.

As in Model 2, this model was investigated experimentally using various choices of the factors a and b and the best were as in Model 2. Furthermore, as in Model 2, all of these best choices give very similar discriminant results and the choice with $a = 1$ and $b = 3$ is chosen to illustrate the model in detail. Thus, the prior variance-covariance matrix of $\boldsymbol{\beta}$ is chosen to be:

$$\boldsymbol{V} = \Sigma_{\beta} + 3D_{\Sigma_{\beta}} = \begin{bmatrix} 827 & -165 & 18 \\ -165 & 538 & -15 \\ 18 & -15 & 7 \end{bmatrix} \tag{5.20}$$

The posterior distribution of z_1, z_2 and $\boldsymbol{\beta}$ has the same form as the posterior distribution of Model 2, represented by equation 5.18, but with π_i given as

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 z_{i2} + \beta_2 z_{i1} z_{i2})}{1 + \exp(\beta_0 + \beta_1 z_{i2} + \beta_2 z_{i1} z_{i2})}. \tag{5.21}$$

5.5.5 Model 3

In all of the previous models, the data y_i was treated as a dichotomous response and hence was assigned a Bernoulli distribution with the leaching probability π_i modelled via a logit link function. This was done by assigning y_i a value of 1 if the *ith* pesticide was observed in the groundwater with levels exceeding a threshold of $0.1\mu g l^{-1}$ in at least one sample, and a value of 0 if the *ith* was not detected in any sample with levels exceeding the previous threshold. This means that all leaching and non-leaching pesticides are given the same weight without taking into account the number of samples that were tested or the number of times they were observed in the groundwater with levels exceeding the threshold. For example, in the previous models, there was no distinction between the leaching pesticide Atrazine, which was detected with levels exceeding the threshold in 66 of 603 samples, and the leaching pesticide Terbutryn, which was detected with levels exceeding the same threshold in 3 of 134 samples. In this current model, the number of times y_i where the *ith* pesticide was observed with levels exceeding the threshold in n_i samples will be taken into account. In this case, the appropriate model is to assign y_i a Binomial distribution with parameters n_i (sample size) and π_i (leaching probability).

As with the previous models, this model will be implemented in two stages model as follows.

Stage 1:

As mentioned above, $y_i \sim \text{Binomial}(\pi_i, n_i)$, where π_i is given in 5.1.

Stage 2:

The prior information regarding z_1, z_2 and $\boldsymbol{\beta}$ are chosen as in Model 2. Also, the best choices for the factors a and b in the prior variance-covariance matrix form, $a\Sigma_{\boldsymbol{\beta}} + bD_{\Sigma_{\boldsymbol{\beta}}}$, are the same as Model 2 and the variance-covariance matrix $\boldsymbol{V}$ in 5.13, is chosen to illustrate the model in detail.

Combining the data with the prior information leads to the following joint posterior distribution regarding z_1, z_2 and $\boldsymbol{\beta}$:

$$\begin{aligned} p(z_1, z_2, \boldsymbol{\beta}|\boldsymbol{y}) \;\propto\; & \prod_{i=1}^{n} \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} \\ & \cdot \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^{n} \left(\tfrac{z_{i1} - m_{i1}}{\sigma_{z_{i1}}} \right)^2 \right\} \\ & \cdot \exp\left\{ -\tfrac{1}{2} \sum_{i=1}^{n} \left(\tfrac{z_{i2} - m_{i2}}{\sigma_{z_{i2}}} \right)^2 \right\} \\ & \cdot \exp\left\{ -\tfrac{1}{2} (\boldsymbol{\beta} - \boldsymbol{b})^T \boldsymbol{V}^{-1} (\boldsymbol{\beta} - \boldsymbol{b}) \right\} \end{aligned} \qquad (5.22)$$

5.5.6 Model 3*

This model investigates Model 3 with an interaction term. The two stages are

Stage 1:

This stage is as stage 1 in Model 3, but with the logit link function in 5.2

Stage 2:

The prior information regarding z_1, z_2 and $\boldsymbol{\beta}$ is chosen as in Model 2*. The prior variance-covariance matrix V in 5.20 will be used to illustrate the model since all the best choices of prior variance-covariance matrices give similar results.

5.5.7 Further models

Two further models were investigated. These models suffer from some obstacles which need further analysis. These obstacles will be addressed as a future work. What follows is a description of these models, but their analysis will not be given.

Predictive model

This model differs from the previous models (especially Models 2 and 3) in the sense that the monitored values of z_1 and z_2 are considered as i.i.d observations from normal distributions with unknown means and variances. The crucial idea is to choose the posterior predictive distributions of z_1 and z_2 as prior information for these covariates. Analytically, as in [18] and [3], if $\boldsymbol{z}$ is a random sample of size n from a normal distribution with an unknown mean and variance (with the usual non-informative independent prior distributions), then the posterior predictive distribution of a future observation $\tilde{z}$, denoted by $p(\tilde{z}|\boldsymbol{z})$, is a student-t distribution with location $\bar{z}$, scale $\left(1+\frac{1}{n}\right)^{1/2} s$ and $n-1$ degree of freedom, where $\bar{z}, s$ and n are the mean, standard deviation and the sample size of $\boldsymbol{z}$ respectively. The two stages required to implement this model are as follows:

Stage 1:

This stage is as stage 1 of model 2.

Stage 2:

As mentioned above, the posterior predictive distributions, in the form of student-t distributions, of covariates z_1 and z_2 are chosen as prior information regarding these covariates. This prior information can be summarised as follows:

$$z_{i1} \sim \text{st}\left(\bar{z}_{i1}, \left(1+\frac{1}{n_{i1}}\right)^{1/2} s_{i1}, n_{i1}-1\right) \tag{5.23}$$

$$z_{i2} \sim \text{st}\left(\bar{z}_{i2}, \left(1+\frac{1}{n_{i2}}\right)^{1/2} s_{i2}, n_{i2}-1\right) \tag{5.24}$$

The model can be investigated using prior information for $\boldsymbol{\beta}$ of the form $a\Sigma_{\boldsymbol{\beta}}+bD_{\Sigma_{\boldsymbol{\beta}}}$, for example, the factor a and b can be chosen as in Model 3. Combining the data with this prior information leads to the following joint posterior distribution of z_1, z_2 and $\boldsymbol{\beta}$:

$$\begin{aligned} p\left(z_1, z_2, \boldsymbol{\beta} | \boldsymbol{y}\right) \propto & \prod_{i=1}^{n} \pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i} \\ & \cdot \prod_{i=1}^{n}\left[1+\frac{1}{n_{i1}}\left(\frac{z_{i1}-\bar{z}_{i1}}{s_{i1}}\right)^2\right]^{-n_{i1}/2} \\ & \cdot \prod_{i=1}^{n}\left[1+\frac{1}{n_{i2}}\left(\frac{z_{i2}-\bar{z}_{i2}}{s_{i2}}\right)^2\right]^{-n_{i2}/2} \\ & \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{b})^T V{-}1(\boldsymbol{\beta}-\boldsymbol{b})\right\} \end{aligned} \tag{5.25}$$

where $n = 43$. However, if a random variable θ has a student-t distribution with location μ, scale $\sigma > 0$ and degree of freedom $\nu > 0$, i.e. $\theta \sim$ Student-t(μ, σ, ν), then θ has the following mean and variance:

$$\begin{aligned} \mathrm{E}(\theta) &= \mu, \quad \text{for } \nu > 0 && (5.26) \\ \mathrm{Var}(\theta) &= \frac{\nu}{\nu - 2}\sigma^2, \quad \text{for } \nu > 2. && (5.27) \end{aligned}$$

So, for θ to have finite variance, the degrees of freedom ν should exceed 2. Consequently, WinBUGS provides a student-t distribution but with degrees of freedom greater than 2. Unfortunately, there are some pesticides that only have two monitored values for one of the covariates, so the covariates for these pesticides will be assigned a student-t distribution with just one degree-of-freedom (a Cauchy distribution) which can not be implemented in the WinBUGS software. To try to avoid this difficulty, the model was implemented directly by MCMC simulation in R. However, there were convergence difficulties. These difficulties will be addressed in a future study.

Discrete model

This model differs from the previous models in the sense that the covariates z_1 and z_2 are assigned discrete prior distributions over their data-base values. Furthermore, this model is implemented using the Metropolis-Hastings algorithm with a "weighted uniform distribution" transition proposal distribution. It can be implemented in the R package in two stages as follows.

Stage 1:

This stage is as stage 1 for Model 2.

Stage 2:

Discrete uniform distributions are chosen to represent prior information for z_1 and

z_2. These discrete uniform distributions are chosen over the available values of the covariates; i.e. if the ith pesticide has n_{i1} and n_{i2} possible values for the covariates z_1 and z_2, respectively, then the prior distributions for z_1 and z_2 are

$$\begin{aligned} p(z_{i1}) &= \frac{1}{n_{i1}} && (5.28) \\ p(z_{i2}) &= \frac{1}{n_{i2}} && (5.29) \end{aligned}$$

In addition, the model uses the same prior information for $\boldsymbol{\beta}$ as in Model 3.

The posterior distribution of $\boldsymbol{z}_1 = (z_{11}, z_{21}, \ldots, z_{n_1 1})$, $\boldsymbol{z}_2 = (z_{12}, z_{22}, \ldots, z_{n_2 2})$ and $\boldsymbol{\beta}$ can be expressed as

$$\begin{aligned} p\left(z_1, z_2, \boldsymbol{\beta}|\boldsymbol{y}\right) \propto\ & \textstyle\prod_{i=1}^{n} \pi_i^{y_i} \left(1-\pi_i\right)^{1-y_i} \cdot \prod_{i=1}^{n} \frac{1}{n_{i1}} \cdot \prod_{i=1}^{n} \frac{1}{n_{i2}} \\ & \cdot \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta}-\boldsymbol{b}\right)^T \mathrm{Var}^{-1}(\boldsymbol{\beta})\left(\boldsymbol{\beta}-\boldsymbol{b}\right)\right\} \end{aligned} \qquad (5.30)$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2})}{1 + \exp(\beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2})}.$$

Possible choice of the proposal distribution

To describe our choice of proposal distribution, consider the ith pesticide and let $(z_{i1,1}, z_{i1,2}, \ldots, z_{i1,n_{i1}})$ and $(z_{i2,1}, z_{i2,2}, \ldots, z_{i2,n_{i2}})$ be the possible values of the covariates z_{i1} and z_{i2}, respectively, sorted from the smallest to the largest values. We chose a discrete transition proposal distribution which gives more weight to combinations of the values of z_{i1} and z_{i2} located in the NW corner if this ith pesticide is a leacher and more weight to those combinations in the SE corner if the ith pesticide is a non-leacher. Different ways can be used to assign the weights; for example, if the ith pesticide is a leacher (non-leacher), then the first value of z_{i1} (z_{i2}) can be chosen as a candidate value with prior weight p, the second with a reduced prior weight

$\frac{p}{2}$, the third with $\frac{p}{4}$ and the last value will be assigned prior weight $\frac{p}{2^{n_{i1}-1}}(\frac{p}{2^{n_{i2}-1}})$. Then, p is chosen such that the sum of all prior weights is 1, i.e.

$$p+\frac{p}{2}+\frac{p}{4}+\ldots+\frac{p}{2^{n_{i1}-1}}=1$$

from which we obtain

$$p=\frac{2^{n_{i1}-1}}{2^{n_{i1}}-1} \tag{5.31}$$

Also, if the ith pesticide is a leacher (non-leacher), then the first value of z_{i2} (z_{i1}) can be chosen with prior weight $\frac{p}{2^{n_{i2}-1}}(\frac{p}{2^{n_{i1}-1}})$ and the last value with prior weight p, where again p is calculated using equation 5.31.

However, some convergence difficulties in implementing this model arose. In particular, the simulated chains of z_1 and z_2 do not converge to their target distributions. This may due to our choice of proposal distributions. The weighted uniform proposal distributions make the sampler 'stick' to the most likely value, the value with the largest weight, as the chains run, and do not move to choose from the other possible values. This convergence difficulty will be addressed as a future study.

After describing the different models above, further steps must be taken to complete the Bayesian analysis, such as implementation using MCMC simulation with WinBUGS or R software, checking convergence, monitoring and summarising the simulated values and drawing inferential conclusions. A description on how to implement the proposed models in both WinBUGS and R will be given below. Results of diagnostic tests (to assess convergence) and comparison tools (to compare models) will be discussed.

5.6 Implementing Bayesian analysis using WinBUGS

The focus of this section is on the implementation of Bayesian analysis of the models described in the previous section using WinBUGS. A number of steps are needed to implement any Bayesian analysis using WinBUGS. Among these steps are assigning a full probability distribution to all of the stochastic nodes of the associated DAG, model diagnostics, assessing model complexity and comparing different models. As regards assigning a full probability distribution, implementation of the Gibbs sampler as an MCMC simulation technique requires identifying the full conditional posterior distribution for each parameter of interest which may not be easy, especially obtaining it in a closed form. Fortunately, the WinBUGS software performs this automatically without the need to derive the forms of the conditional posterior distribution. So, what is really needed is to assign a probability distribution to each of the stochastic nodes.

Model diagnostics involve specifying the number of chains to run using the Gibbs sampler, specifying different initial values for each chain, and assessing convergence to the target distribution. The focus of this study is on those diagnostic tools built in WinBUGS, such as the tracing or history of the chains, autocorrelation plots and the modified Gelman-Rubin diagnostic (BGR). A further important step is the assessment of model complexity and the comparison of different models. There are several tools for comparing different models. The one used in this study is the Deviance Information Criterion (DIC), proposed in [39] and implemented in WinBUGS, and discussed in Chapter 3.

The code used to implement model 1*, as an example, is given in Appendix A.2. It (as with other WinBUGS codes) consists of three parts. The first expresses likelihood, the second is for specifying the prior distribution and the third is for including data. Also initial values must be specified, especially when running multiple chains. The first code illustrates these specifications. In this example, y_i is assigned a Bernoulli distribution with the leaching probability π_i using the command `dbern(`π_i`)`, the symbol $\sim$ indicating that the node y_i is defined as a stochastic node. The command `logit(`π_i`) <- eta[`i`]` represents the link function (logit) and the symbol (<-) is used to indicate that π_i is treated as a logical node. The command `eta[`i`]`$\sim$ `dmnorm(mean[ ], precision[ , ])` represents the chosen prior information for η which is a multivariate normal with a mean vector `mean[ ]` and precision matrix `precision[ , ]`. (The mean vector and precision matrix were calculated using the R package.) Running the code requires specification of initial values for each chain, encoded under `inits`. Two different sets of initial values are used since two chains will be run. After the specifications have been encoded the model can be run and convergence assessed.

5.7 Implementing the proposed models using R

This section aims to show how to implement Bayesian analysis of the proposed models using MCMC simulation in R. Because of its generality and simplicity, the Metropolis-Hastings algorithm (M-H) is chosen to draw samples from the joint posterior distribution. Another reason for using the M-H algorithm is that there is no requirement to derive the full conditional distributions for all the parameters of interest as with the Gibbs sampler algorithm. Moreover, deriving full conditional distributions in closed form is intractable for the models considered. The following discussion illustrates how the M-H algorithm is used to implement the proposed models. The R code for the discrete model (as an example) is given in Appendix A.1. The aim is to draw samples from the joint posterior distributions of z_1, z_2 and $\boldsymbol{\beta}$ given the data $\boldsymbol{y}$ described in 5.18. These posterior distributions are given up to proportionality without the normalizing constant, as the M-H algorithm can be implemented in the absence of the normalizing constant. What follows are the general steps to perform the M-H algorithm:

Step 1:

Assign starting values for the Markov chain to each parameter: $\boldsymbol{z}_1^{(0)}, \boldsymbol{z}_2^{(0)}$ and $\boldsymbol{\beta}^{(0)}$.

Step 2:

Update each parameter in turn as follows.

At time t:

(a) Update $\boldsymbol{z}_1$:

(1) Sample a candidate value $\boldsymbol{z}_1^*$ for the vector $\boldsymbol{z}_1$ from a normal proposal distribution with mean $\boldsymbol{z}_1^{*(t-1)}$ and variance $\sigma^2_{\boldsymbol{z}_1}$.

(2) Compute the Metropolis-Hastings ratio:

$$r(\boldsymbol{z}_1^*) = \frac{p\left(\boldsymbol{y}|\boldsymbol{z}_1^*, \boldsymbol{z}_2^{(t-1)}, \boldsymbol{\beta}^{(t-1)}\right) \cdot p\left(\boldsymbol{z}_1^*\right) \cdot p\left(\boldsymbol{z}_2^{(t-1)}\right) \cdot p\left(\boldsymbol{\beta}^{(t-1)}\right)}{p\left(\boldsymbol{y}|\boldsymbol{z}_1^{(t-1)}, \boldsymbol{z}_2^{(t-1)}, \boldsymbol{\beta}^{(t-1)}\right) \cdot p\left(\boldsymbol{z}_1^{(t-1)}\right) \cdot p\left(\boldsymbol{z}_2^{(t-1)}\right) \cdot p\left(\boldsymbol{\beta}^{(t-1)}\right)} \tag{5.32}$$

(3) Calculate $\alpha(\boldsymbol{z}_1^*) = min\left\{1, r(\boldsymbol{z}_1^*)\right\}$

(4) Draw the components of $\boldsymbol{U}$ independently from uniform $[0, 1]$ distribution.

(5) Set

$$\boldsymbol{z}_1^{(t)} = \begin{cases} \boldsymbol{z}_1^* & \text{if } \boldsymbol{U} < \alpha(\boldsymbol{z}_1^*) \\ \boldsymbol{z}_1^{(t-1)} & \text{otherwise.} \end{cases}$$

(b) Update $\boldsymbol{z}_2$:

(1) Sample a candidate value $\boldsymbol{z}_2^*$ for the vector $\boldsymbol{z}_2$ from a normal proposal distribution with mean $\boldsymbol{z}_2^{*(t-1)}$ and variance $\sigma^2_{\boldsymbol{z}_2}$.

(2) Compute the Metropolis-Hastings ratio:

$$r(\boldsymbol{z}_2^*) = \frac{p\left(\boldsymbol{y}|\boldsymbol{z}_1^{(t)}, \boldsymbol{z}_2^*, \boldsymbol{\beta}^{(t-1)}\right) \cdot p\left(\boldsymbol{z}_1^{(t)}\right) \cdot p\left(\boldsymbol{z}_2^*\right) \cdot p\left(\boldsymbol{\beta}^{(t-1)}\right)}{p\left(\boldsymbol{y}|\boldsymbol{z}_1^{(t)}, \boldsymbol{z}_2^{(t-1)}, \boldsymbol{\beta}^{(t-1)}\right) \cdot p\left(\boldsymbol{z}_1^{(t)}\right) \cdot p\left(\boldsymbol{z}_2^{(t-1)}\right) \cdot p\left(\boldsymbol{\beta}^{(t-1)}\right)} \tag{5.33}$$

(3) Calculate $\alpha(\boldsymbol{z}_2^*) = min\left\{1, r(\boldsymbol{z}_2^*)\right\}$.

(4) Draw the components of $\boldsymbol{U}$ independently from uniform $[0,1]$ distribution.

(5) Set

$$\boldsymbol{z}_2^{(t)} = \begin{cases} \boldsymbol{z}_2^* & \text{if } \boldsymbol{U} < \alpha(\boldsymbol{z}_2^*) \\ \boldsymbol{z}_2^{(t-1)} & \text{otherwise.} \end{cases}$$

(c) Update $\boldsymbol{\beta}$:

(1) Sample a candidate vector $\boldsymbol{\beta}^*$ for $\boldsymbol{\beta}$ from a multivariate normal proposal distribution with mean vector $\boldsymbol{\beta}^{*(t-1)}$ and variance matrix $\Sigma_{\boldsymbol{\beta}}$.

(2) Compute the Metropolis-Hastings ratio:

$$r(\boldsymbol{\beta}^*) = \frac{p\left(\boldsymbol{y}|\boldsymbol{z}_1^{(t)}, \boldsymbol{z}_2^{(t)}, \boldsymbol{\beta}^*\right) \cdot p\left(\boldsymbol{z}_1^{(t)}\right) \cdot p\left(\boldsymbol{z}_2^{(t)}\right) \cdot p\left(\boldsymbol{\beta}^*\right)}{p\left(\boldsymbol{y}|\boldsymbol{z}_1^{(t)}, \boldsymbol{z}_2^{(t)}, \boldsymbol{\beta}^{(t-1)}\right) \cdot p\left(\boldsymbol{z}_1^{(t)}\right) \cdot p\left(\boldsymbol{z}_2^{(t)}\right) \cdot p\left(\boldsymbol{\beta}^{(t-1)}\right)} \tag{5.34}$$

(3) Calculate $\alpha(\boldsymbol{\beta}^*) = min\left\{1, r(\boldsymbol{\beta}^*)\right\}$.

(4) Draw the components of $\boldsymbol{U}$ independently from uniform $[0,1]$ distribution.

(5) Set

$$\boldsymbol{\beta}^{(t)} = \begin{cases} \boldsymbol{\beta}^* & \text{if } \boldsymbol{U} < \alpha(\boldsymbol{\beta}^*) \\ \boldsymbol{\beta}^{(t-1)} & \text{otherwise.} \end{cases}$$

Note that the above Metropolis-Hastings ratios include distributions that appear in both the denominator and numerator and hence these distributions can be cancelled to simplify the ratios.

The above algorithm can be used to draw samples from the posterior distributions for the discrete model taking into account the following considerations and modifications.

1. In updating z_1, the candidate values are sampled from a weighted uniform proposal distribution as described in Subsection 5.5.7.

2. The above transition proposal distribution is not symmetric, hence the Metropolis-Hastings ratio is modified to be

$$r(z_1^*) = \frac{q\left(z_1^{(t-1)}|z_1^*\right) \quad p\left(y|z_1^*, z_2^{(t-1)}, \beta^{(t-1)}\right) \cdot p\left(z_1^*\right) \cdot p\left(z_2^{(t-1)}\right) \cdot p\left(\beta^{(t-1)}\right)}{q\left(z_1^*|z_1^{(t-1)}\right) p\left(y|z_1^{(t-1)}, z_2^{(t-1)}, \beta^{(t-1)}\right) \cdot p\left(z_1^{(t-1)}\right) \cdot p\left(z_2^{(t-1)}\right) \cdot p\left(\beta^{(t-1)}\right)}$$

By cancelling the distributions that appear in both the denominator and numerator and noting that $p(z_1^*) = p(z_1^{(t-1)})$, which follows from assigning a uniform prior distribution for z_1, the last ratio becomes

$$r(z_1^*) = \frac{q\left(z_1^{(t-1)}|z_1^*\right) \quad p\left(y|z_1^*, z_2^{(t-1)}, \beta^{(t-1)}\right)}{q\left(z_1^*|z_1^{(t-1)}\right) p\left(y|z_1^{(t-1)}, z_2^{(t-1)}, \beta^{(t-1)}\right)}$$

and the same modification is applied when updating z_2.

5.8 Assessing convergence and model selection

Assessing convergence is a crucial part of simulation using MCMC methods. The focus will be on the use of those diagnostic tools built into WinBUGS, such as tracing the history of the chains, autocorrelation plots and the modified Gelman-Rubin diagnostic (BGR).

There are 89 stochastic variables in the models that we consider (43 for each of z_1 and z_2 and the 3 regression coefficients $\boldsymbol{\beta}$). On the other hand, there are only 43 data points all of them zeros or ones. In addition, the regression coefficients $\beta_0, \beta_1, \beta_2$ are a priori highly correlated. All of this leads to a need to process long MCMC runs with consequent increasing storage (memory) requirements. A possible resolution, to reduce the impact on the memory requirements of processing long runs, is to thin the chains. Two types of thinning, available in WinBUGS, are used. As in [40], the implementation of these steps in our simulation are as follows.

1. In the first step of thinning, 1^{st}-thin, the samples from every k_1^{th} iteration are stored and the other samples are permanently discarding as the MCMC simulation runs, helping to reduce memory requirements.

2. In the second step of thinning, 2^{nd}-thin, the samples from every k_2^{th} iteration are selected from the already generated (and stored) posterior samples from the first thinning. Inferences will be based on these second thinned samples. The other samples from the 2^{nd}-thin may be temporarily discarded, as we may wish to base our inferences on these discarded samples if we decide to change the chosen value of k_2. The best choice of k_2 is the value that makes successive samples approximately independent, see for example [21].

Two parallel chains were run from dispersed starting values for each model. Each model used (a) different initial iterations which were discarded (burn-in) after reaching convergence status, (b) different values of k for each step of thinning, (c) different size of posterior stored sample (after applying 1^{st}-thin), and (d) different size of the retained posterior samples (after discarding the initial iterations and applying the

Model	s-sample	1^{st}-thin	2^{nd}-thin	d-sample	r-sample	DIC	Threshold (η)	CC-NL	CC-L	CR	R
1	12000	3	1	2000	20002	29.91	$\eta = 0$	100	100	100	-
1*	11000	3	1	2000	18002	17.01	$\eta = 0$	100	100	100	-
2	81000	6	30	35000	3066	21.63	$\eta = 0$	100	100	100	5
2*	101000	6	35	40000	3484	17.91	$\eta = 0$	100	100	100	2
3	121000	6	35	42000	4514	57.30	$\eta = -5$	97.14	100	97.67	5
3*	241000	6	50	40000	8040	48.52	$\eta = -5$	97.14	100	97.67	7

Table 5.1: Statistical summaries: s-sample (stored posterior sample size), 1^{st}-thin (the first thinning step), 2^{nd}-thin (the second thinning step), d-sample (the discarded sample size), r-sample (the retained posterior sample size), DIC (Deviance Information Criterion), CC-NL (correctly classified non-leaching pesticides), CC-L (correctly classified leaching pesticides), CR (classification rate) and R, the number of edges of minimal spanning tree which connect points from different groups.

2^{nd}-thin). The inferences are based on these retained posterior samples.

The size of posterior stored sample (s-sample), amount of thinning for the first (1^{st}-thin) and second (2^{nd}-thin) thinning, size of the discarded sample (d-sample) and size of the retained posterior sample (r-sample) for each model are displayed in Table 5.1. For example, for Model 2, the size of the stored posterior sample is 81000 (for each of the two chains) after applaying a first thinning with $k_1 = 6$. The size of the retained posterior sample is 3066 (each chain has a size of 1533) after burn-in the first 35000 iterations (from each chain) and the second thinning with $k_2 = 30$. The summary statistics are based on this retained posterior sample from the two chains (size of 3066). Models 1 and 1* have 43 stochastic variables and each of the remaining models has 89 stochastic variables. Consequently, diagnostic tests will be illustrated for only some of these stochastic variables.

A figure will be provided for each of Models 1 and 1* showing diagnostic tests

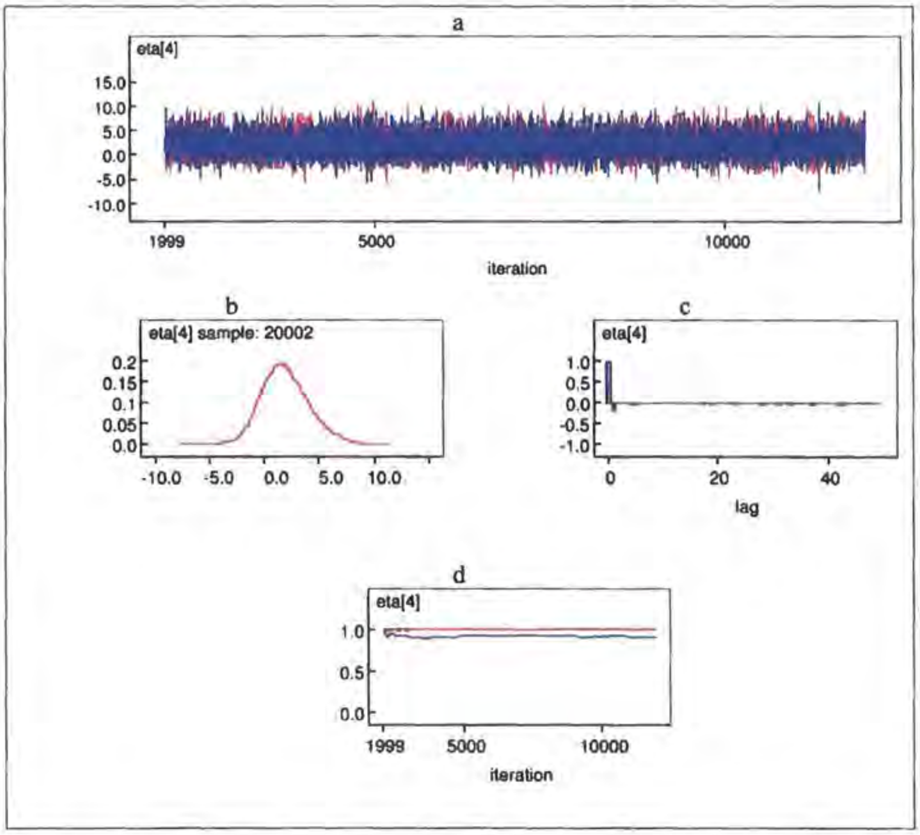

Figure 5.3: Diagnostic tests for η_4 from MCMC analysis of Model 1: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

for a selected stochastic variable. Three figures will be provided for each of the remaining models. The first two figures will illustrate the diagnostic tests for z_1 and z_2 for a selected pesticide. The third figure will show the diagnostic tests for one of regression parameters β_0, β_1 and β_2.

Figures 5.3 and 5.4 show history plots, posterior densities, autocorrelation functions and the formal BGR tests for Models 1 and 1* for η_4 and η_{17}, respectively.

Figures 5.5 - 5.16 show history plots, posterior densities, autocorrelation functions and the formal BGR tests for Models 2, 2*, 3 and 3* for the variables z_1, z_2,

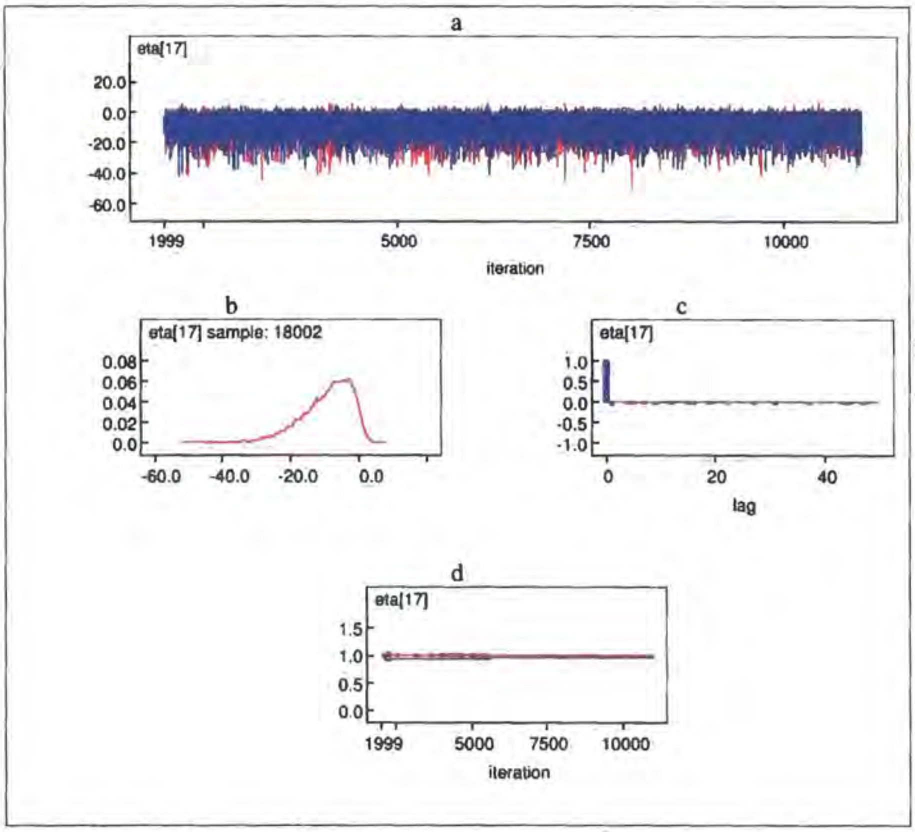

Figure 5.4: Diagnostic tests for η_{17} from MCMC analysis of Model 1*: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

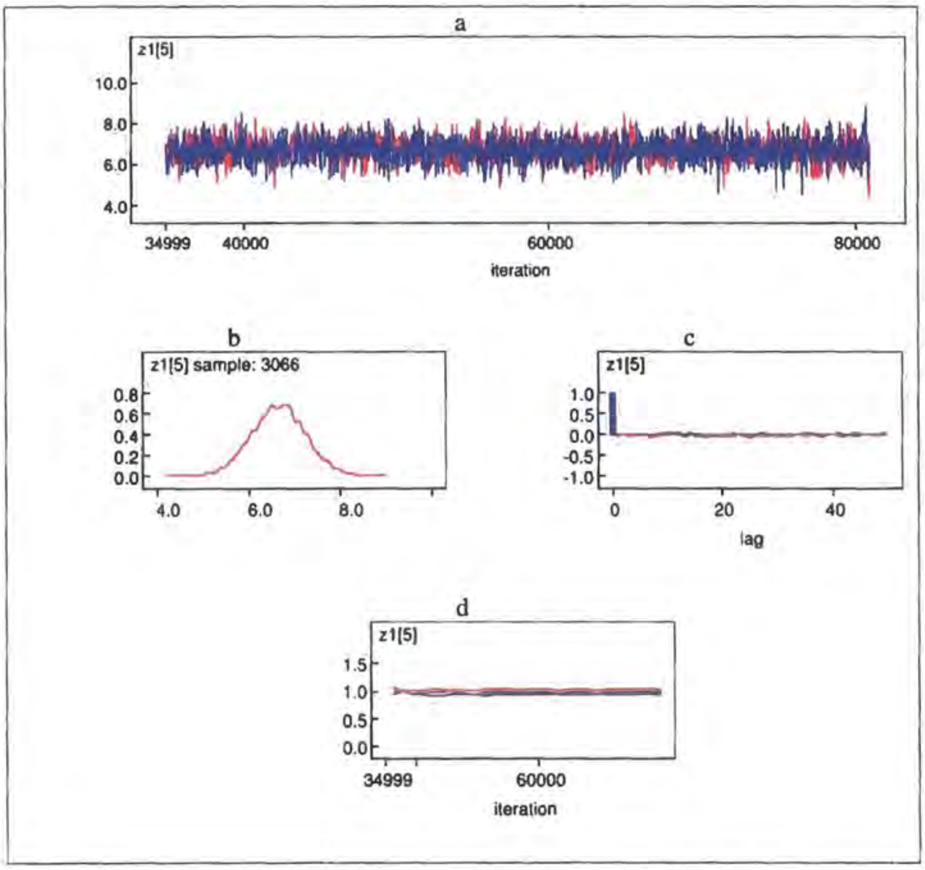

Figure 5.5: Diagnostic tests for $z_{5,1}$ from MCMC analysis of Model 2: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

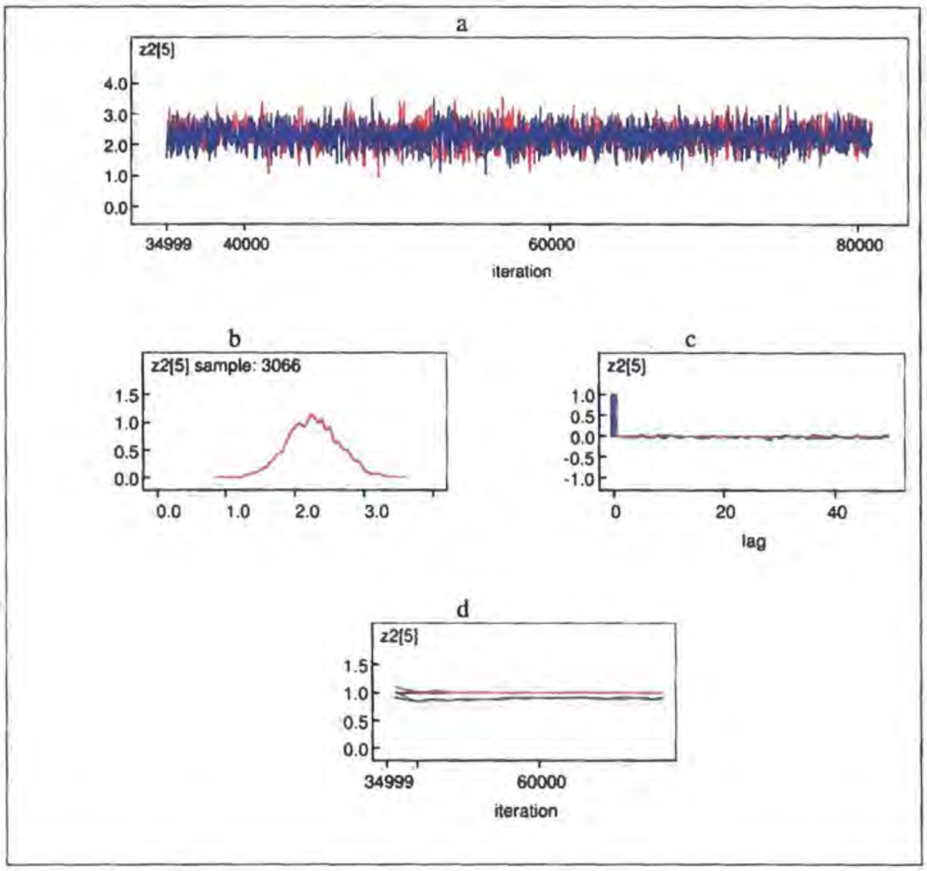

Figure 5.6: Diagnostic tests for $z_{5,2}$ from MCMC analysis of Model 2: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

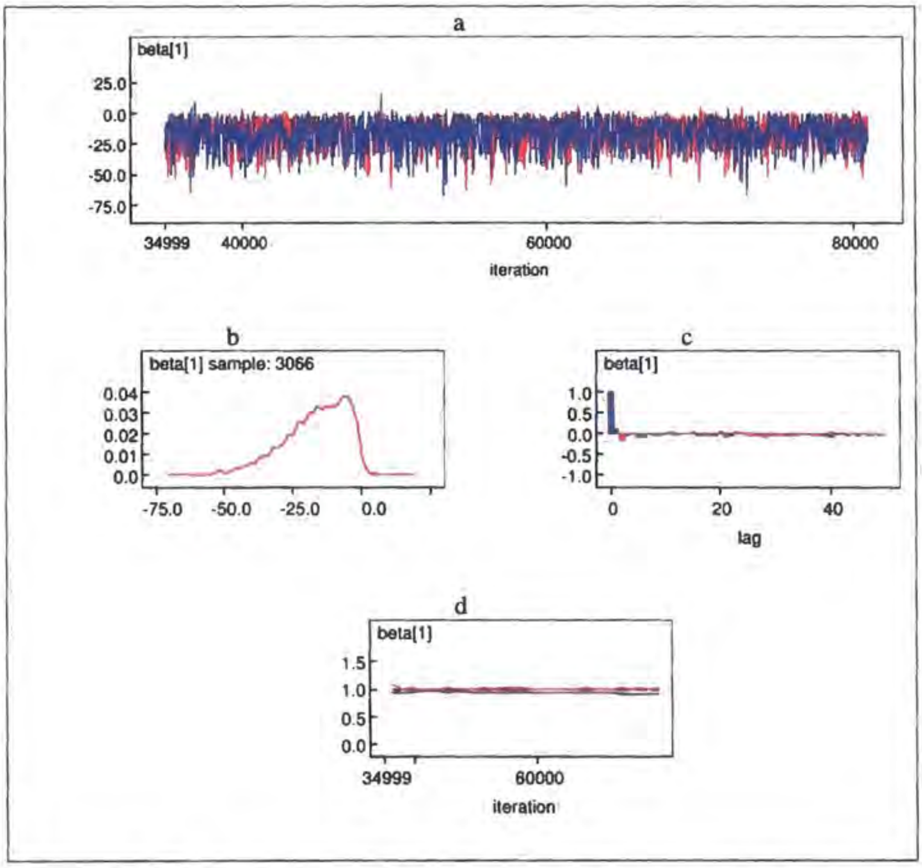

Figure 5.7: Diagnostic tests for β_0 from MCMC analysis of Model 2: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

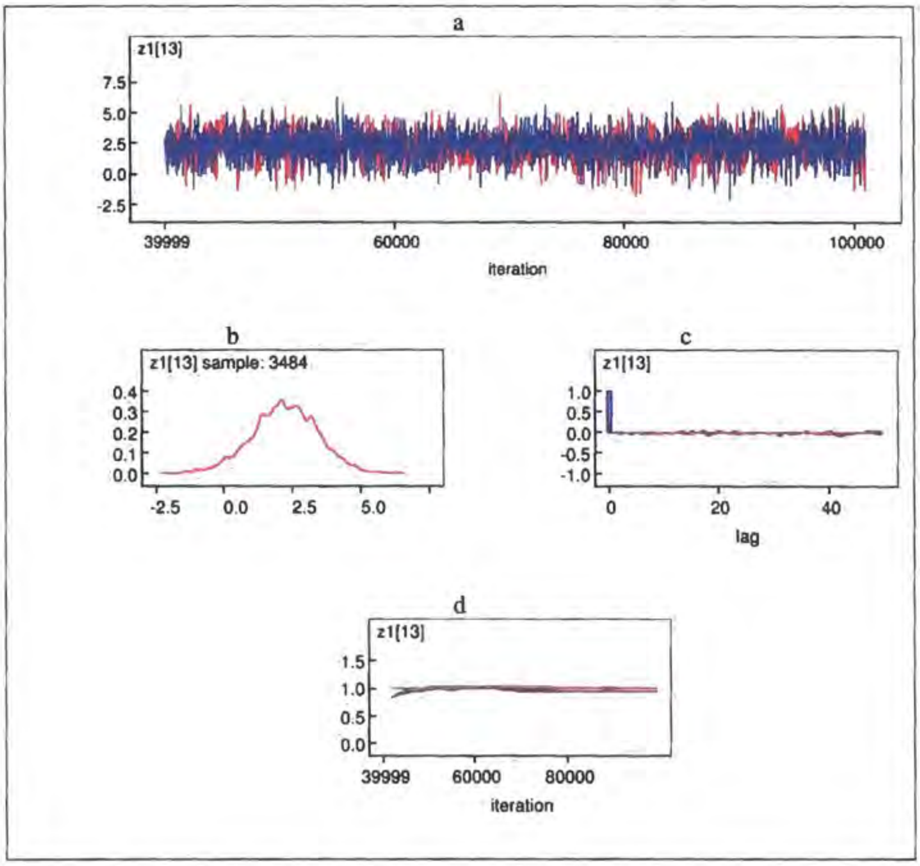

Figure 5.8: Diagnostic tests for $z_{13,1}$ from MCMC analysis of Model 2*: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

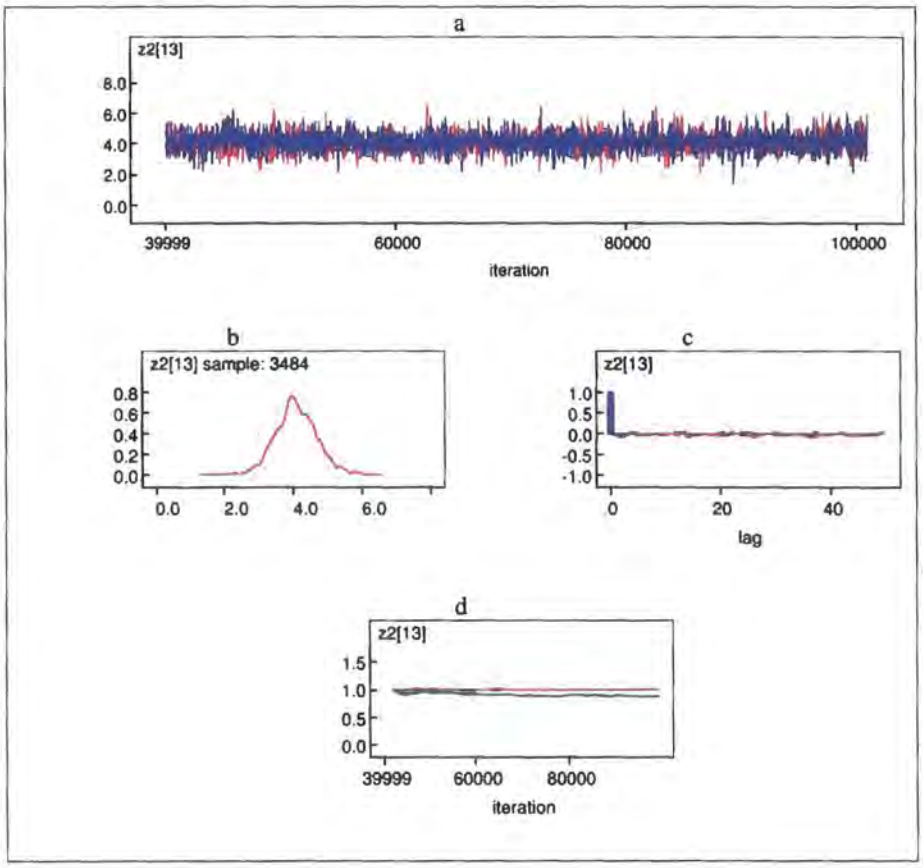

Figure 5.9: Diagnostic tests for $z_{13,2}$ from MCMC analysis of Model 2*: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

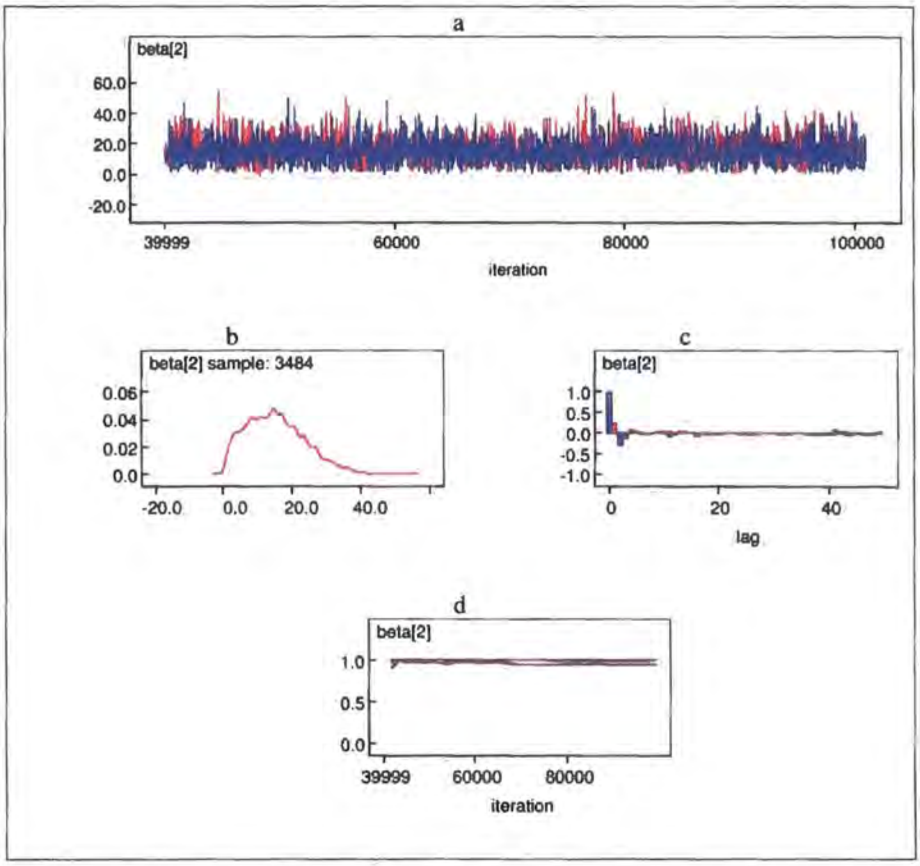

Figure 5.10: Diagnostic tests for β_1 from MCMC analysis of Model 2*: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

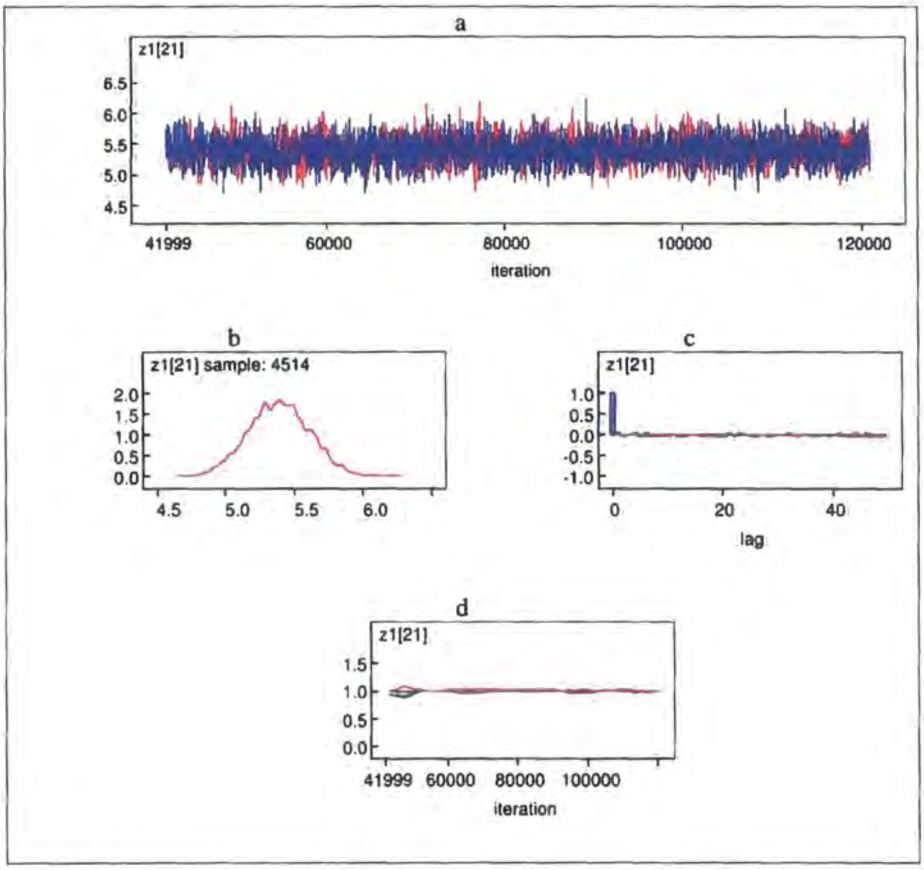

Figure 5.11: Diagnostic tests for $z_{21,1}$ from MCMC analysis of Model 3: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

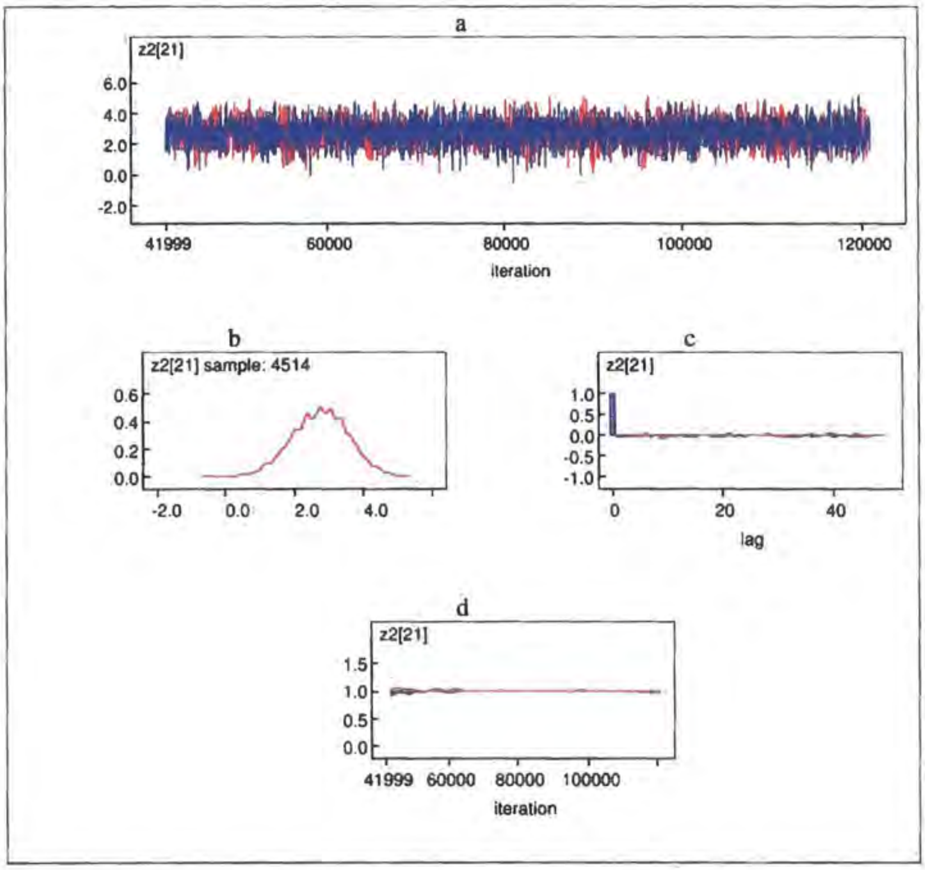

Figure 5.12: Diagnostic tests for $z_{21,2}$ from MCMC analysis of Model 3: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

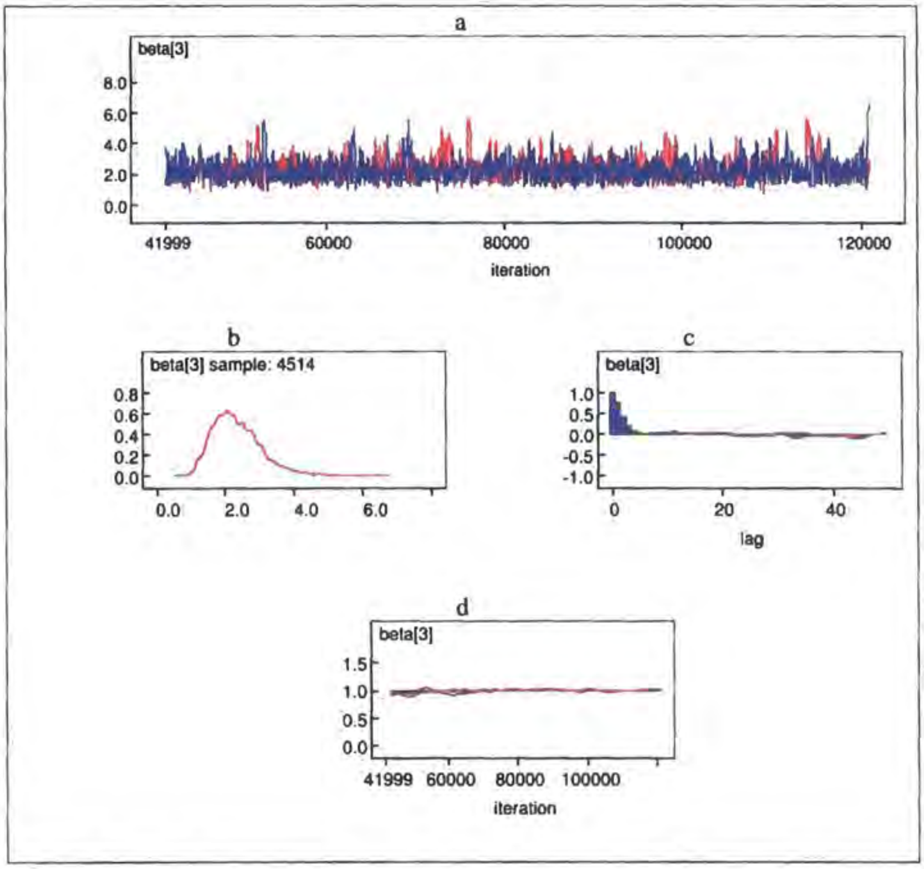

Figure 5.13: Diagnostic tests for β_2 from MCMC analysis of Model 3: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

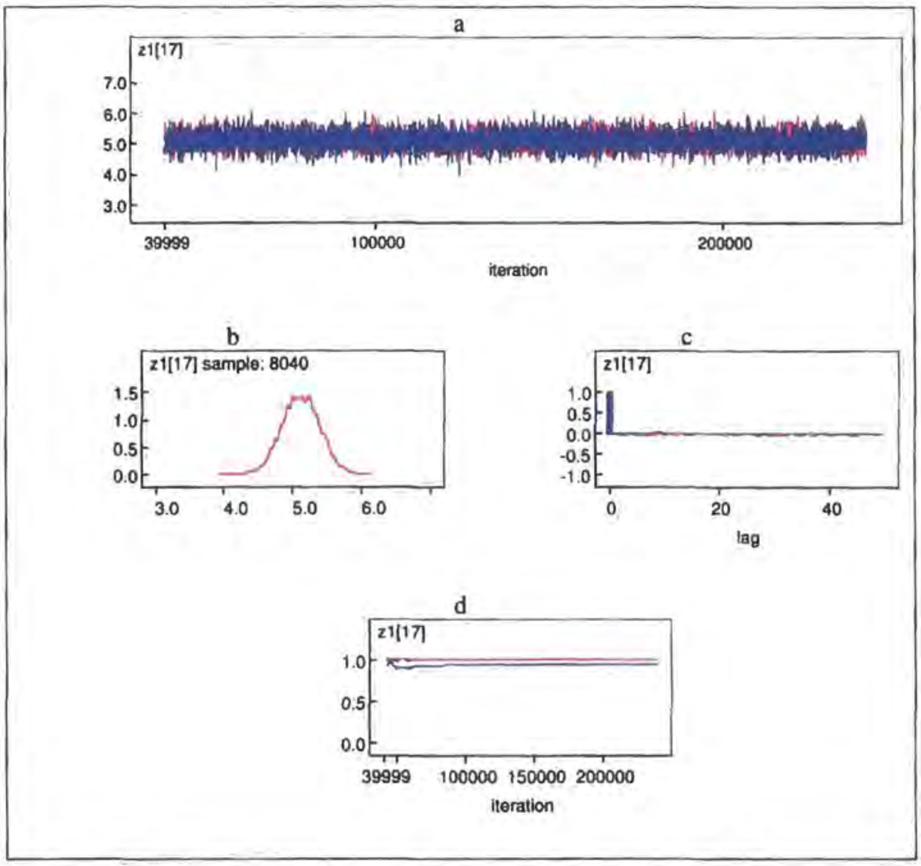

Figure 5.14: Diagnostic tests for $z_{17,1}$ from MCMC analysis of Model 3*: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

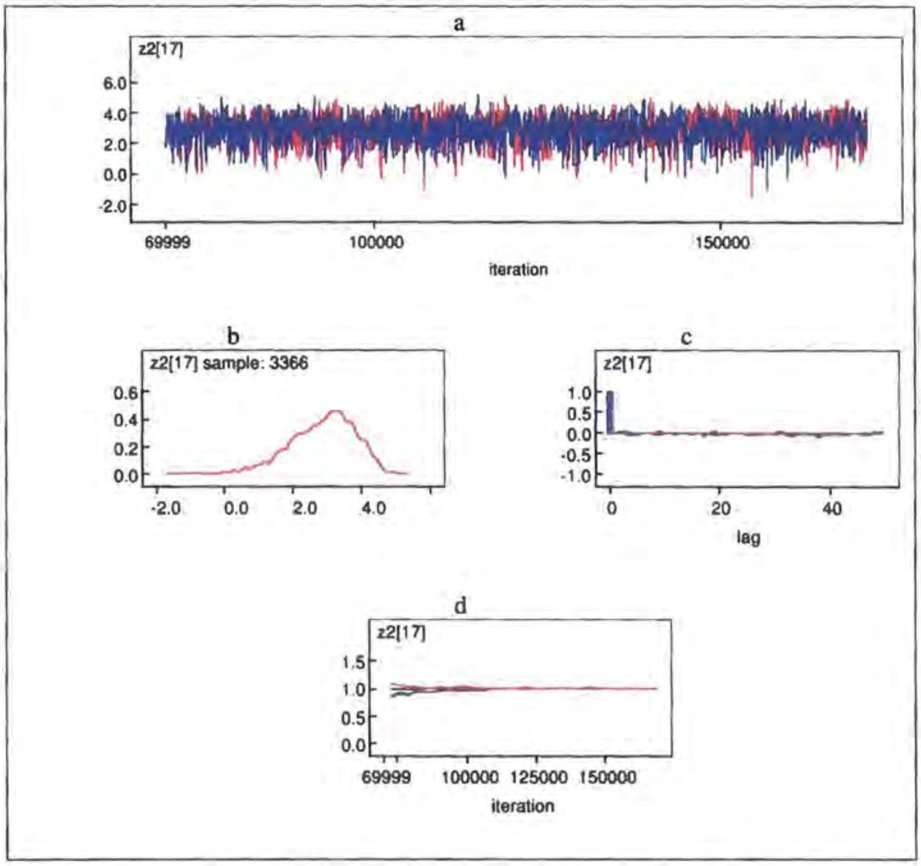

Figure 5.15: Diagnostic tests for $z_{17,2}$ from MCMC analysis of Model 3*: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

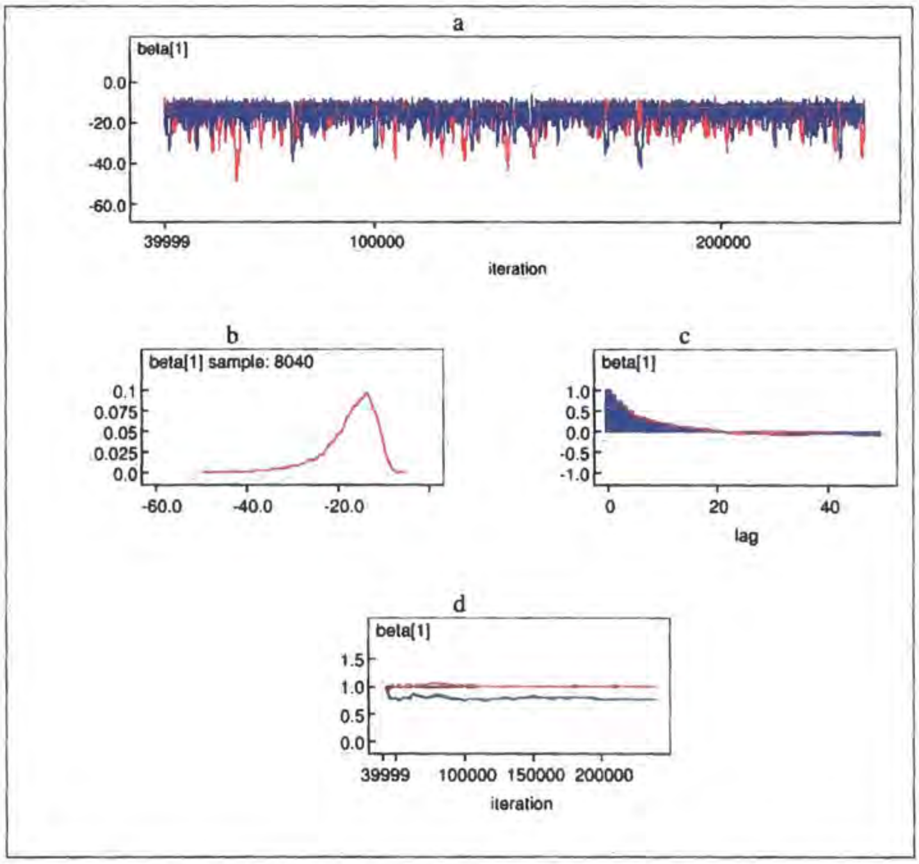

Figure 5.16: Diagnostic tests for β_0 from MCMC analysis of Model 3*: (a) history plot of two superimposed chains, (b) smoothed posterior density, (c) autocorrelation function, and (d) Gelman-Rubin test (BGR).

β_0, β_1 and β_2. All of the history plots, see panels a, with two superimposed chains, show that the chains mixed and converged to the same estimated posterior distributions, informally indicating that convergence had been reached. The autocorrelation plots, see panels c, clearly indicate that the within-chain correlations are negligible, so that sampled values are approximately independent suggesting that the sampler moves quickly around the posterior distributions. The formal BGR tests, see panel d, show that the monitored values of the ratio (coloured in red) converge to 1 and the values of both pooled and within interval widths (coloured in green and blue, respectively) converge, indicating that convergence had been reached. Convergence was similarly assessed and achieved for the other random quantities.

Two comparison tools are used to compare models: DIC and misclassification statistics. Table 5.1 displays the DIC values and misclassification rates for all models. As can be seen from this table, Models 1* and 2* have the smallest values of DIC, 17.01 and 17.91, respectively, indicating that they provide the best description. Also, we can notice, in general, that the starred models have smaller DIC values in comparison with unstarred models. This might be expected since modeling Gustafson's data using the interaction term yields complete separation between leachers and non-leachers.

5.9 Results

After achieving convergence, the monitored values can be regarded as random draws from the desired posterior distributions. Using WinBUGS, these posterior distributions, their means, standard deviations, percentiles, 95% posterior credible interval

(CI) and other related statistics can be displayed.

Since all of the proposed models use logistic regression, the linear predictor η and the leaching probability π are important for predicting pesticide leachability. In particular, the posterior means of η and π for any of the 43 pesticides should be used as a basis for discrimination. The posterior distribution of $\boldsymbol{\eta} = \boldsymbol{X\beta}$, where $\boldsymbol{X}$ is the model matrix for either starred or unstarred models, is estimated as follows. As discussed in Section 5.7, at iteration t, the vectors $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and $\boldsymbol{\beta}$ are updated to $\boldsymbol{z}_1^{(t)}$, $\boldsymbol{z}_2^{(t)}$ and $\boldsymbol{\beta}^{(t)}$, respectively, and $\boldsymbol{\eta}$ is updated to $\boldsymbol{\eta}^{(t)} = \boldsymbol{X}^{(t)}\boldsymbol{\beta}^{(t)}$, using $\boldsymbol{z}_1^{(t)}$, $\boldsymbol{z}_2^{(t)}$ and $\boldsymbol{\beta}^{(t)}$. Similarly, the leaching probability $\boldsymbol{\pi}$ is updated to $\boldsymbol{\pi}^{(t)}$, using $\boldsymbol{\eta}^{(t)}$. For N iterations, the posterior mean, $\hat{\boldsymbol{\eta}}$, of $\boldsymbol{\eta}$ is estimated by the average of the N values of $\boldsymbol{\eta}^{(t)}$

$$\hat{\boldsymbol{\eta}} = \frac{\sum \boldsymbol{\eta}^{(t)}}{N} = \frac{\sum \boldsymbol{X}^{(t)}\boldsymbol{\beta}^{(t)}}{N} \tag{5.35}$$

Also, the posterior mean, $\hat{\boldsymbol{\pi}}$, of $\boldsymbol{\pi}$, in which discrimination is based on, is estimated by the average of the N values of $\boldsymbol{\pi}^{(t)}$

$$\hat{\boldsymbol{\pi}} = \frac{\sum e^{\boldsymbol{\eta}^{(t)}}/(1+e^{\boldsymbol{\eta}^{(t)}})}{N} = \frac{\sum e^{\boldsymbol{X}^{(t)}\boldsymbol{\beta}^{(t)}}/(1+e^{\boldsymbol{X}^{(t)}\boldsymbol{\beta}^{(t)}})}{N} \tag{5.36}$$

To predict pesticides leachabilities we use the posterior means of $\boldsymbol{\pi}$, as calculated in 5.36. Of course, all of this applies to any of the models 1, 1*, 2, 2*, 3 and 3*. Other posterior summaries, such as standard deviations, are calculated similarly.

Beside the posterior means of $\boldsymbol{\pi}$, we can also use the posterior probabilities of $\mathrm{P}[\pi_i > cp]$, where cp is an identified threshold, which might be more informative than the posterior means of $\boldsymbol{\pi}$.

Results are summarised using (a) scatter plots of the posterior means of z_1 and z_2, (b) plots of posterior means of $\boldsymbol{\eta}$ vs $\boldsymbol{y}$, (c) ranked boxplots for $\boldsymbol{\eta}$, and (d) ranked boxplots for $\boldsymbol{\pi}$. The boxplots summarise the posterior distributions of components of $\boldsymbol{\eta}$ and $\boldsymbol{\pi}$. As described in [40], the boxes represent inter-quartile ranges and the solid black line in each box is the posterior mean. The arms of each box cover the central 95% per cent of the distributions. The posterior distributions are ranked by the ranks of the posterior means.

In addition to the figures, the results about π will be summarised in tables. Each table will include the posterior means of π as calculated using 5.36, posterior standard deviations, 95% posterior credible interval (CI) and posterior probabilities of $\mathrm{P}[\pi_i > cp]$.

For each model, a posterior mean threshold of leaching probability π is identified. For example, we can identify a threshold for π of 0.5 such that any pesticide with a mean posterior leaching probability greater than 0.5, will be classified as a leacher, otherwise it will be classified as a non-leacher.

Figures 5.17 and 5.19 show the posterior means of the predictor η for Models 1 and 1*. It is apparent that $\eta = 0$ splits the posterior means into two non-overlapping groups. Scanning posterior means of π for both models, see Tables 5.4 and 5.5, according to a threshold of 0.5, shows that all pesticides are correctly classified. The posterior distributions for π and η for both models are summarised using boxplots as in Figures 5.18 and 5.20. From the above tables and boxplot figures, we can observed that there is large uncertainty regarding the posterior distributions of π and η. For example, the non-leacher Napropamide, number 33, as analysed using Model 1, has posterior mean of 0.1106 with posterior standard deviation of 0.175, 95% posterior credible interval (CI) of (2e-04, 0.6719) and posterior probability

$P[\pi_i > 0.5] = 0.0551$; see Table 5.4.

For Models 2, 2*, 3 and 3*, we have to be careful using the posterior means of z_1, z_2 and β to predict leaching probability since this should be calculated using 5.36. Instead, we use the posterior means of π to assess leachability of a given pesticide.

Figure 5.21 plots the posterior means of z_2 against posterior means of z_1 together with the discriminant line $\eta = 0$ as analysed using Model 2. The discriminant line $\eta = 0$ was plotted using the posterior means of z_1, z_2 and β; i.e. $\hat{\beta}_0 + \hat{\beta}_1 \hat{z}_1 + \hat{\beta}_2 \hat{z}_2 = 0$, while the posterior means, $\hat{\eta}$, of η and $\hat{\pi}$, of π should be calculated using 5.35 and 5.36. For this reason, the leacher pesticide number 36, Pentachlorophenol, seems to be misclassified. However, this pesticide has a posterior mean of leaching probability of 0.6191 suggesting that it is correctly classified; see Table 5.6 and Figure 5.23. It has 95% posterior credible interval (CI) for the posterior mean of (0.0191, 1) reflecting large uncertainty. This result is also confirmed using posterior probability which turned out to be $P[\pi_i > 0.5] = 0.6132$. The same thing happens for the non-leacher pesticide number 32, Monuron. It seems to be misclassified, but it has posterior mean of leaching probability of 0.4766 with CI of (7e-04, 0.9745), reflecting large uncertainty, and posterior probability $P[\pi_i > 0.5] = 0.4866$, confirming that it is correctly classified. The final conclusion is that this model results in complete separation between the leaching and non-leaching pesticides on the basis of posterior means of $\boldsymbol{\pi}$; see Table 5.6.

For Model 2*, Figure 5.24 plots the posterior means of z_2 against posterior means of z_1 together with the discriminant curve $\eta = 0$. This model gives a complete separation between leaching and non-leaching pesticides based on posterior means

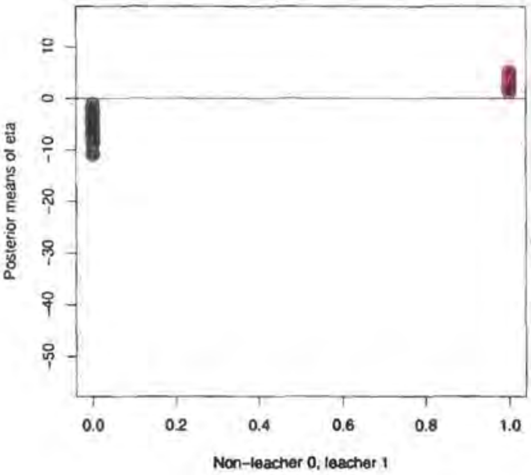

Figure 5.17: The posterior means of η with the discriminant line $\eta = 0$ using model 1.

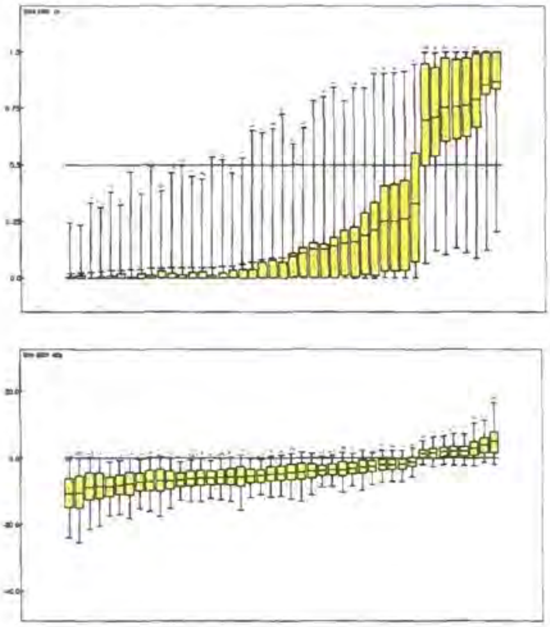

Figure 5.18: Above, the ranked boxplot of π with the discriminant line $\pi = 0.5$; below, the ranked boxplot of η with the discriminant line $\eta = 0$, using model 1.

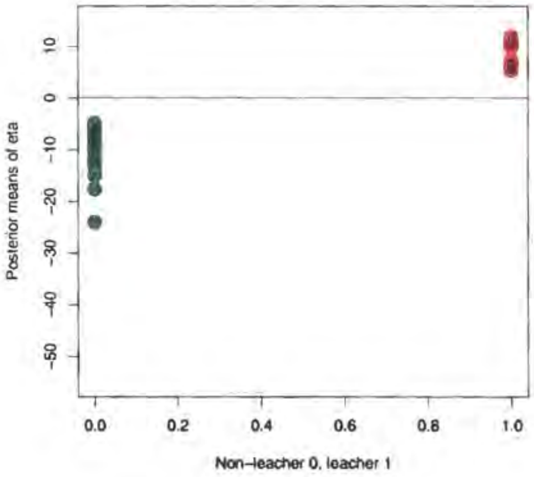

Figure 5.19: The posterior means of η with the discriminant line $\eta = 0$ using model 1*.

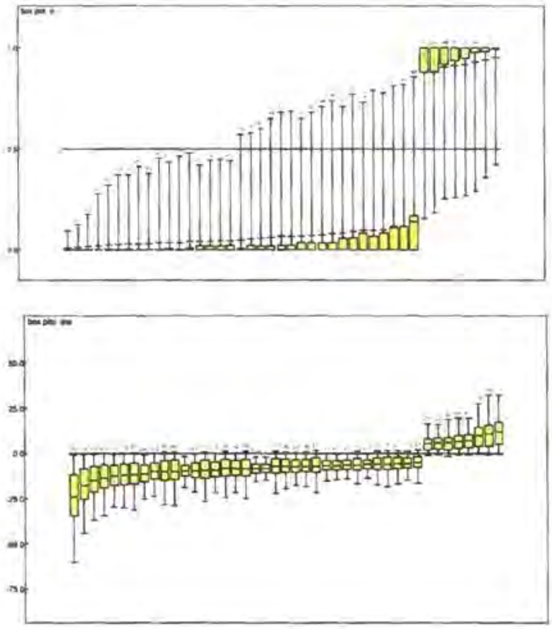

Figure 5.20: Above, the ranked boxplot of π with the discriminant line $\pi = 0.5$; below, the ranked boxplot of η with the discriminant line $\eta = 0$, using model 1*.

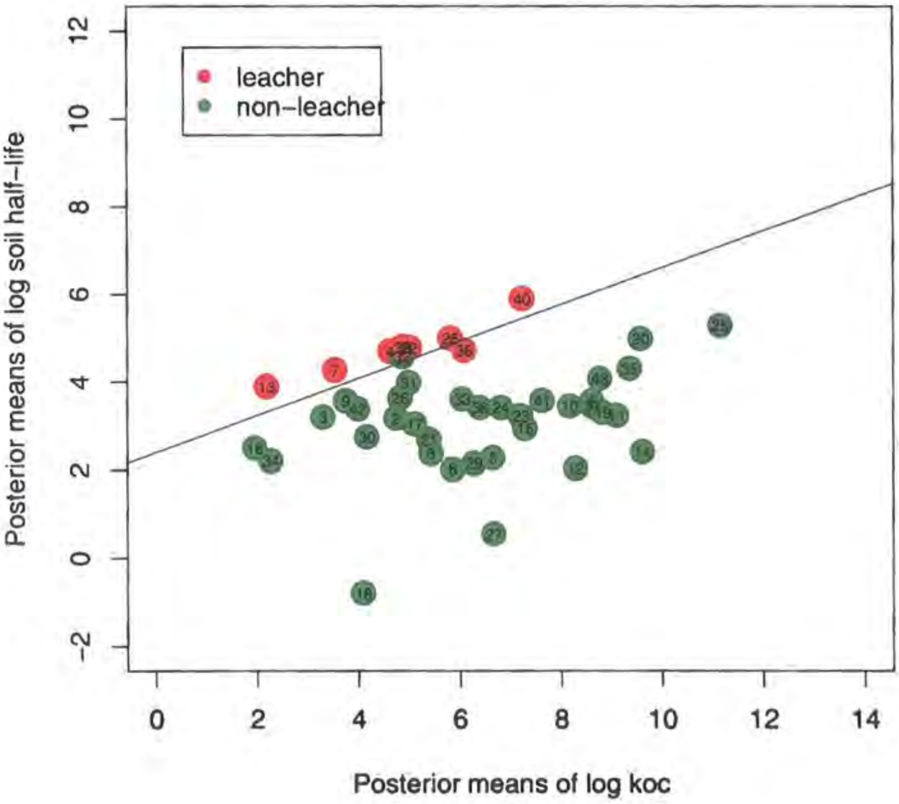

Figure 5.21: The posterior means of logs of k_{oc} and $t_{1/2}^{soil}$ together with the discriminant line $\eta = 0$ using model 2.

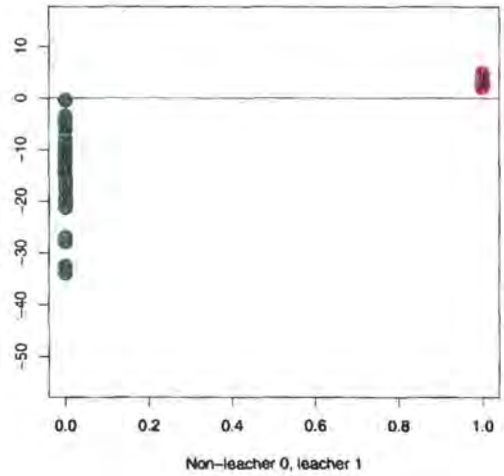

Figure 5.22: The posterior means η with the discriminant line $\eta = 0$ using model 2.

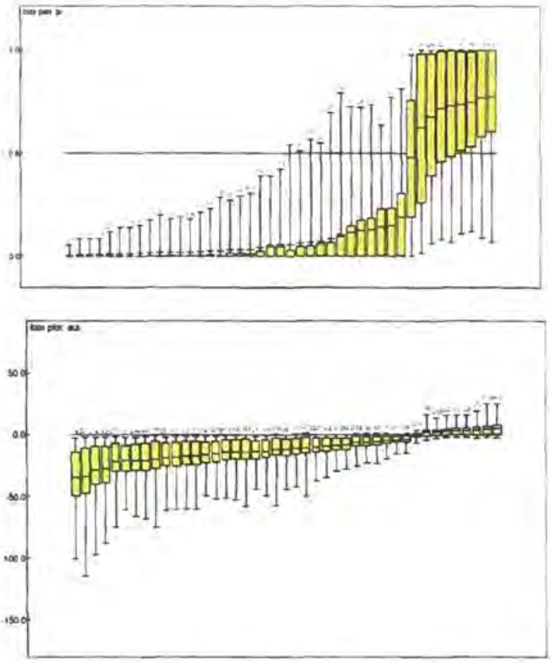

Figure 5.23: Above, the ranked boxplot of π with the discriminant line $\pi = 0.5$; below, the ranked boxplot of η with the discriminant line $\eta = 0$, using model 2.

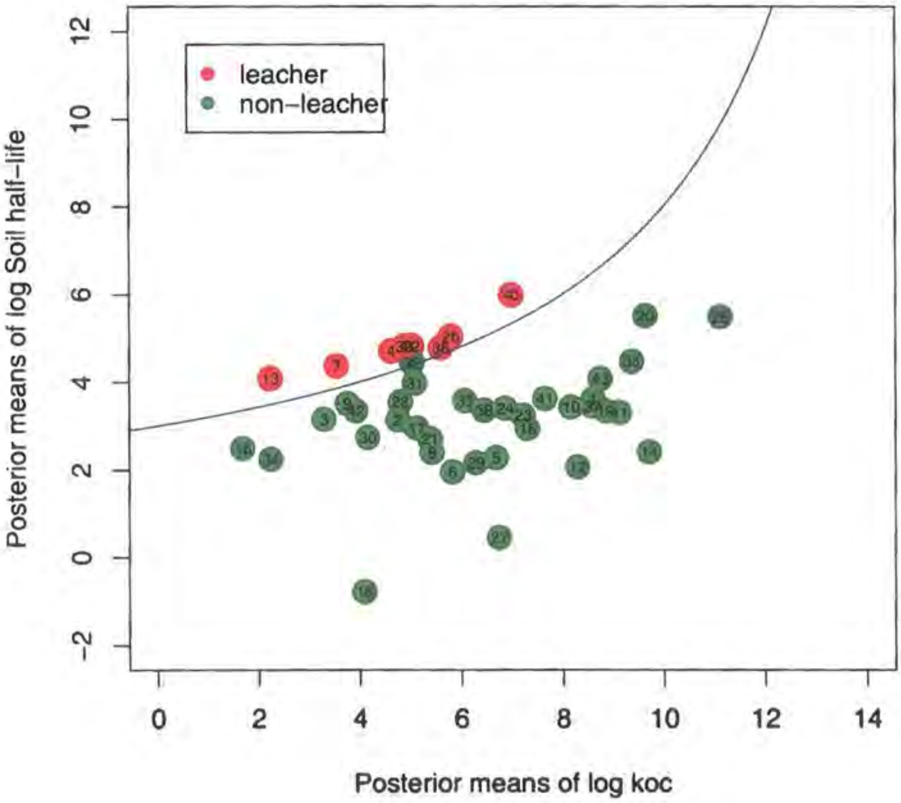

Figure 5.24: The posterior means of logs of k_{oc} and $t_{1/2}^{soil}$ together with the discriminant curve $\eta = 0$ using model 2*.

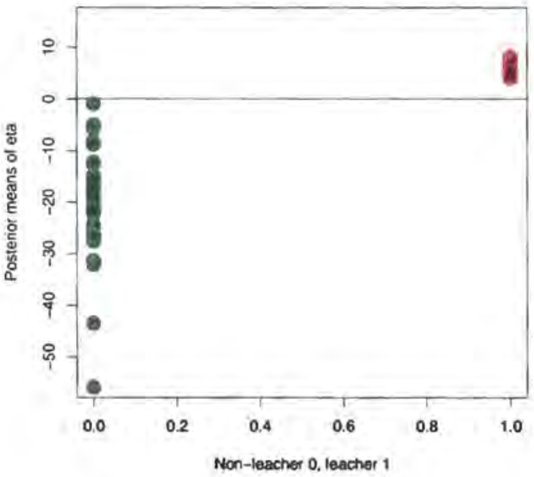

Figure 5.25: The posterior means η together with the discriminant line $\eta = 0$ using model 2*.

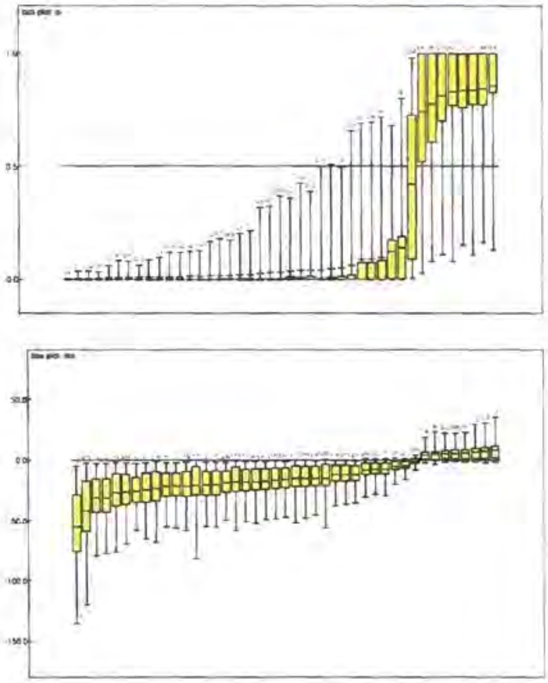

Figure 5.26: Above, the ranked boxplot of π with the discriminant line $\pi = 0.5$; below, the ranked boxplot of η with the discriminant line $\eta = 0$, using model 2*.

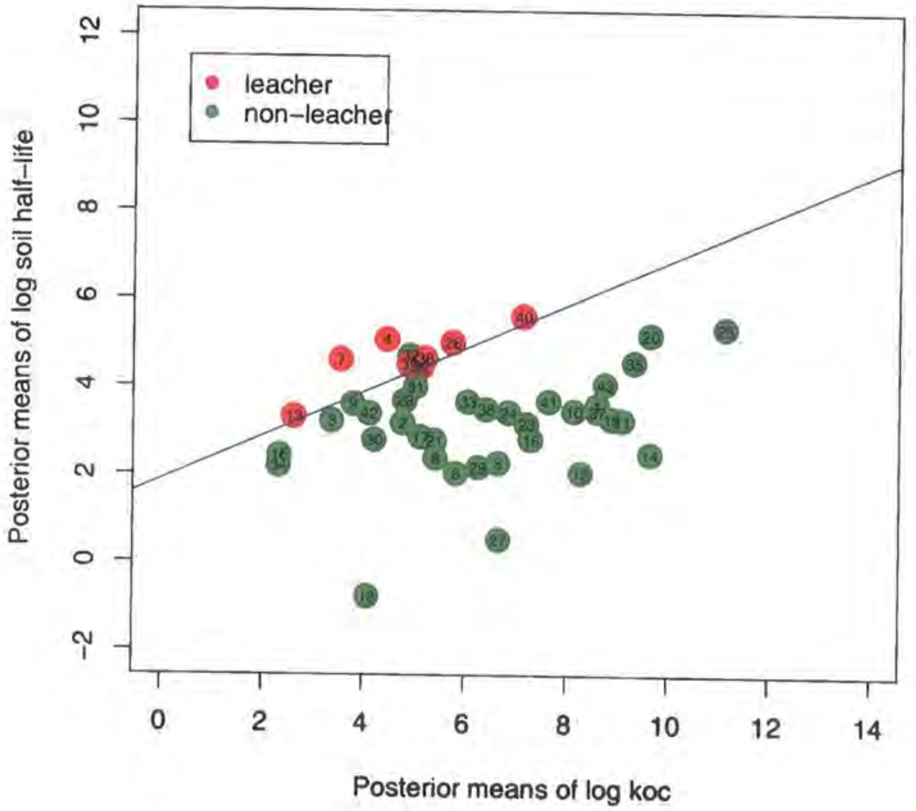

Figure 5.27: The posterior means of logs of k_{oc} and $t_{1/2}^{soil}$ together with the discriminant line $\eta = -5$ using model 3.

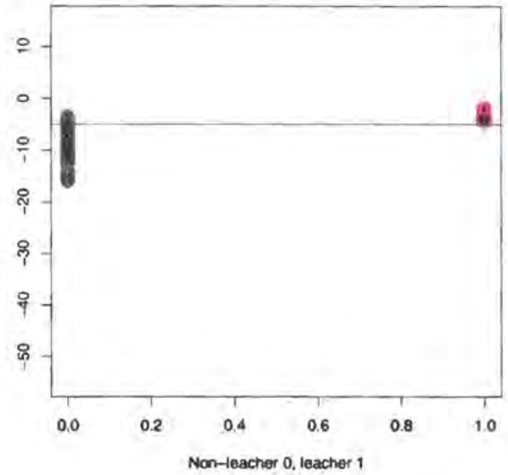

Figure 5.28: The posterior means of η together with the discriminant line $\eta = -5$ using model 3.

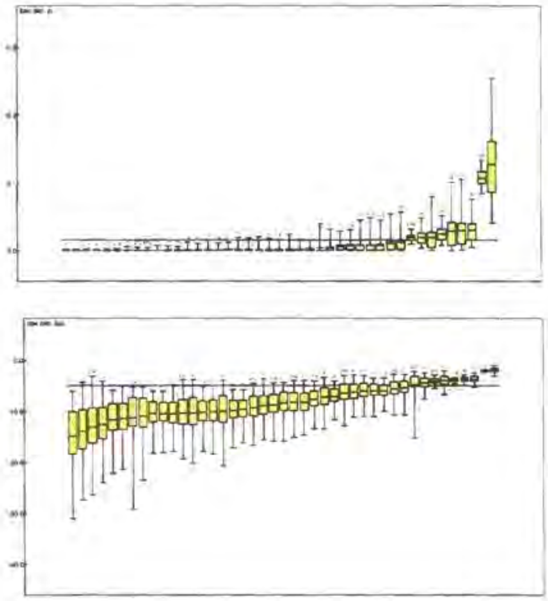

Figure 5.29: Above, the ranked boxplot of π with the discriminant line $\pi = 0.016$; below, the ranked boxplot of η with the discriminant line $\eta = -5$, using model 3.

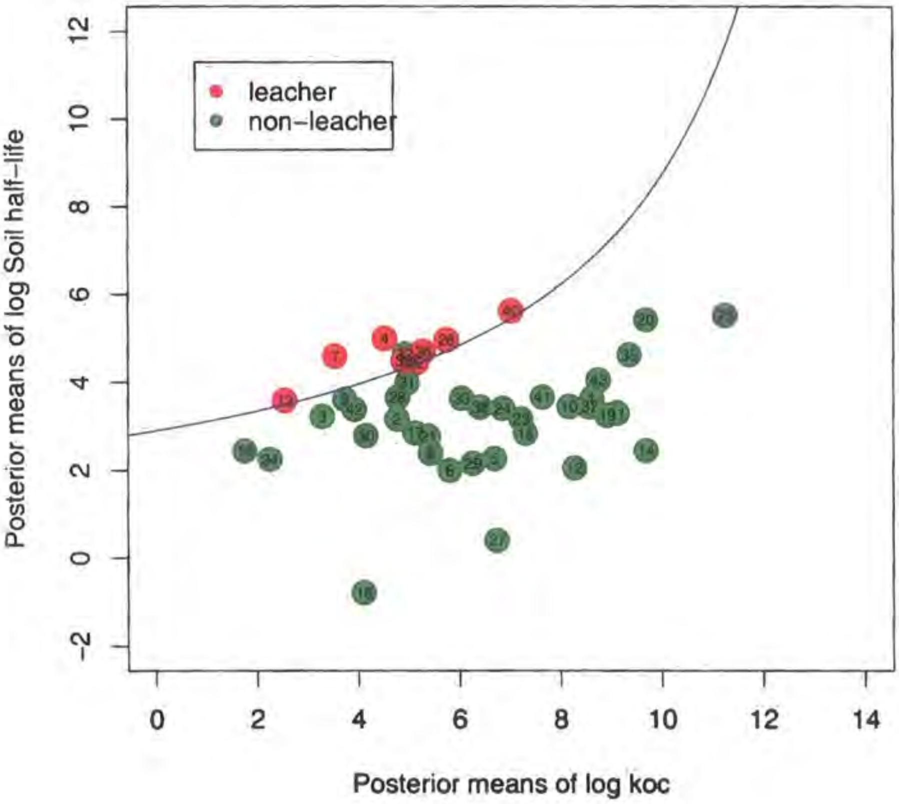

Figure 5.30: The posterior means of logs of k_{oc} and $t_{1/2}^{soil}$ together with the discriminant curve $\eta = -5$ using model 3*.

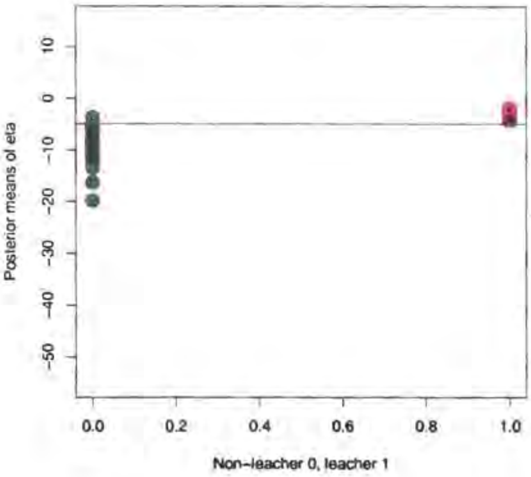

Figure 5.31: The posterior means of η together with the discriminant line $\eta = -5$ using model 3*.

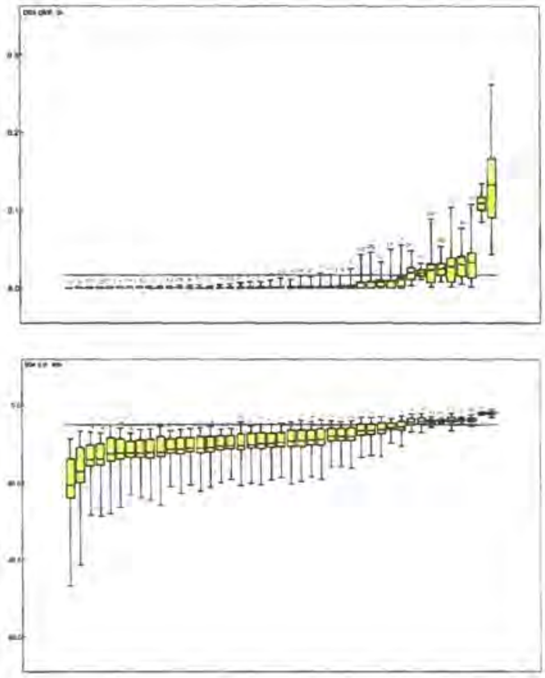

Figure 5.32: Above, the ranked boxplot of π with the discriminant line $\pi = 0.016$; below, the ranked boxplot of η with the discriminant line $\eta = -5$, using model 3*.

of leaching probability π; see Figure 5.26 and Table 5.7. However, notice that the non-leacher pesticide number 32 seems to be misclassified when it is displayed in the plane of posterior means of z_1 and z_2. However, this pesticide has a posterior mean of leaching probability of 0.4198, suggesting that it is correctly classified as shown in Figure 5.26. The 95% CI for the posterior mean is (3e-04, 0.9797), reflecting again large uncertainty. The posterior probability $\mathrm{P}[\pi_i > 0.5] = 0.4202$, confirming that it is correctly classified as a non-leacher. The same explanation for this conflict as for Model 2 applies.

Figure 5.27 plots the posterior means of z_2 against those of z_1 as simulated using Model 3. However, in this case, identifying $\pi = 0.5$ as a discriminant threshold leads to a poor discrimination. However, as discussed in [27], we can improve the discrimination by choosing a different cut-point that minimizes error rate. It was observed from scanning the posterior means of η and π that identifying a threshold of -5 for η (for example) and 0.016 for π, which is not the corresponding value 0.007 of $\eta = -5$, each of these two cut-off values leads to good discrimination between leaching and non-leaching pesticides, as shown in Figures 5.27 and 5.28, although we have to be careful using the discriminant line $\hat{\beta}_0 + \hat{\beta}_1\hat{z}_1 + \hat{\beta}_2\hat{z}_2 = -5$ for the same reasons discussed in Model 2. According to the above specified threshold, only the non-leacher pesticide number 32, Monuron, is misclassified. It has posterior leaching probability mean of 0.0346 with 95% CI of (0.0028, 0.1049) and posterior probability $\mathrm{P}[\pi_i > 0.016] = 0.7337$; see Table 5.8, confirming that it is misclassified as a non-leacher. Figure 5.29, summarise the posterior distributions for π and η, where it is apparent that the model correctly classifies each pesticide except number

32, on the basis of the above cut-points.

The same threshold identified for Model 3 applies for Model 3* as shown in Figures 5.30 and 5.31. As in Model 3, only the non-leacher pesticide number 32 is misclassified with posterior leaching probability of 0.0331 with 95% CI of (0.0019, 0.1088) and posterior probability $P[\pi_i > 0.016] = 0.6769$; see Table 5.9, confirming that it is misclassified as a non-leacher. Figure 5.32, summarise the posterior distributions for π and η, where it is apparent that the model correctly classifies each pesticide except number 32, where the cut-points or the baselines are as in Model 3.

We notice that Models 3 and 3* misclassify the non-leacher pesticide number 32, Monuron. One possible explanation for this misclassification is its prior mean is located in the extreme NW corner, the corner of the leachers; see Figure 5.2.

The fact that that all of the leaching pesticides have posterior probability means of less than 0.5 in Models 3 and 3* should be expected from a probabilistic view for the following reasons. It is apparent from the EA data that all leaching pesticides are detected in the groundwater with levels exceeding the threshold in just a small proportion of samples. For instance, the pesticide Atrazine was detected above the threshold in 66 of 603 samples. This pesticide has 0.1081 as a posterior leaching probability mean, as analysed using Model 3*, which is almost equal to the maximum likelihood estimate $66/603 = 0.1095$. However, 0.1081 is a relatively high probability in comparison with the other non-leaching pesticides such as 2.4.DCPA, which has a posterior leaching probability mean of 5×10^{-4}. The explanation for the small cut-off is because the posterior distribution is dominated by the likelihood function

Model	β_0			β_1			β_2		
	mean	sd	95% CI	mean	sd	95% CI	mean	sd	95% CI
Model 2	-16.74	11.70	(-43.67, -0.756)	-2.964	2.948	(-11.1, -0.073)	7.036	4.791	(0.540, 18.13)
Model 3	-8.893	2.877	(-15.1, -3.954)	-1.177	0.537	(-2.57, -0.434)	2.359	0.738	(1.24, 4.142)

Table 5.2: Regression parameters estimates from analysis of models 2 and 3.

Model	β_0			β_1			β_2		
	mean	sd	95% CI	mean	sd	95% CI	mean	sd	95% CI
Model 2*	-46.17	24.45	(-96.92, -5.807)	15.4	8.665	(1.798, 34.17)	-0.963	0.737	(-2.789, -0.059)
Model 3*	-17.1	5.645	(-31.86, -9.662)	4.348	1.675	(2.177, 8.747)	-0.29	0.135	(-0.632, -0.099)

Table 5.3: Regression parameters estimates from analysis of models 2* and 3*.

because the sample sizes are large.

Tables 5.2 and 5.3 show the posterior means, posterior standard deviations and 95% equitailed posterior credible intervals (CI) of regression parameters β_0, β_1 and β_2 for the different models.

5.9.1 Gustafson's contention

The signs of the estimates of the parameter coefficients of model terms are to be expected. As in Chapter 3, this can be explained using either η or *odds*. For example, for Model 2*, the *odds* can be estimated as

$$odds = \exp(-46.17 + 15.4z_2 - 0.963z_1z_2) \tag{5.37}$$

Fixing the covariate z_2 at a small value and letting the covariate z_1 vary over its range decreases both the odds and the linear predictor η and hence decreases the leaching probability, as depicted in Figure 5.33 (a) and (b). This is consistent with Gustafson's contention that non-leaching pesticides are those with low $t_{1/2}^{soil}$ and high k_{oc} values. Similarly, fixing the covariate z_1 at a small value and letting the covariate

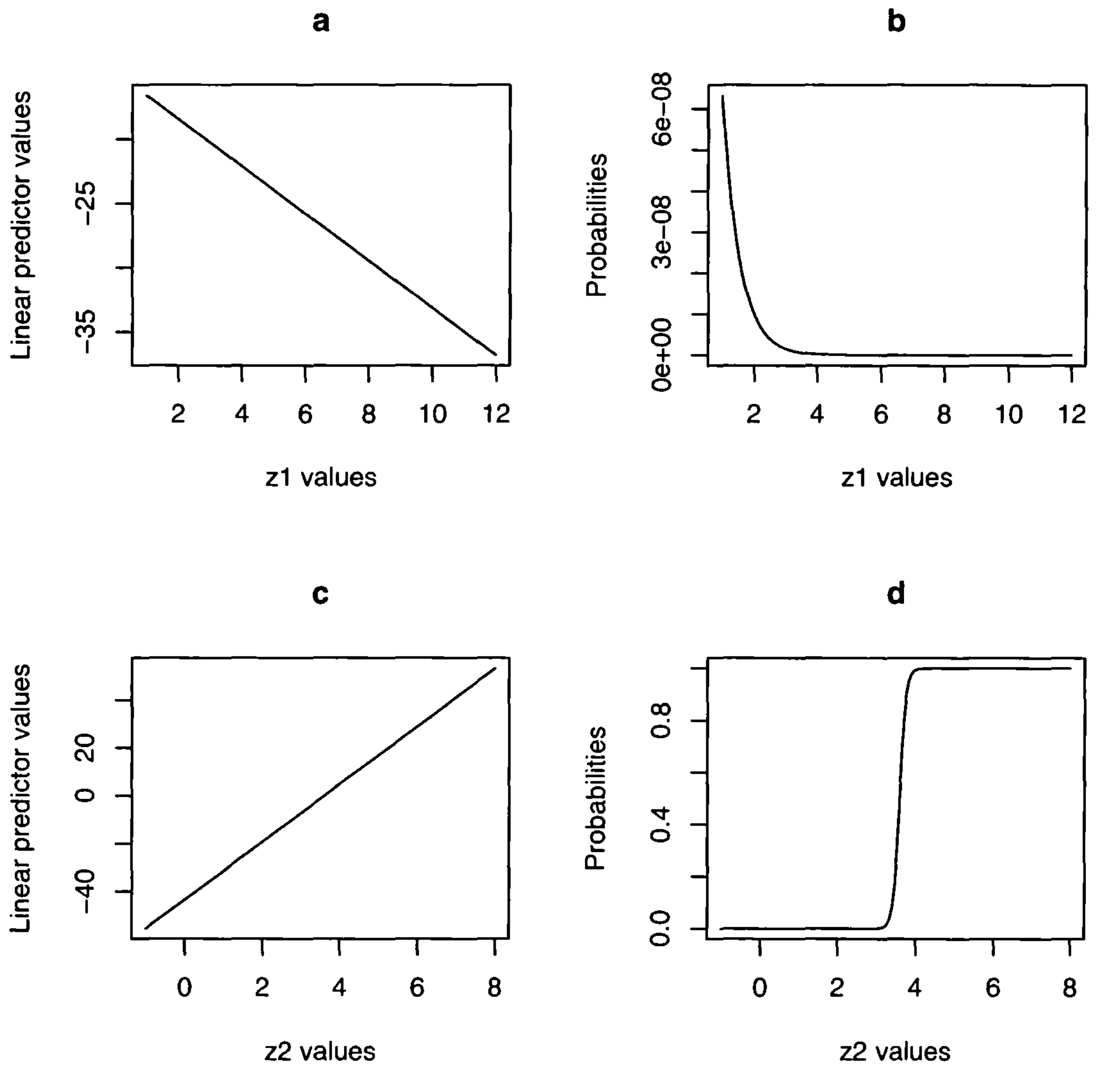

Figure 5.33: Plots of linear predictor η and probability for Model 2*: In (a) and (b) the covariate z_2 is fixed at a certain value and the covariate z_1 varies over a range of values. In (c) and (d) the covariate z_1 is fixed at a certain value and the covariate z_2 varies over a range of values.

z_2 vary over its range increases both the odds and the linear predictor η and hence increases the leaching probability, as depicted in Figure 5.33 (c) and (d). Again, this is consistent with Gustafson's contention that leaching pesticides are those with low k_{oc} and high $t^{soil}_{1/2}$ values.

5.9.2 Down-weighting prior information

Finally, down-weighting prior information derived from the analysis of Gustafson's data allow this prior specifications to be more diffuse and so give a better chance for the posterior coordinate means of k_{oc} and $t^{soil}_{1/2}$ to correspond to their leachability status.

5.9.3 Strengthening the results

The above conclusions can be supported using the multivariate runs test proposed by Friedman and Rafsky in [17] to test for the degree of separation between leaching and non-leaching pesticides in (z_1, z_2)-plane. This test was designed to test the null hypothesis of whether two groups are drawn from the same distribution. As mentioned in Chapter 2, Friedman and Rafsky proposed a statistic test $R_{m,n}$ based on the total number of edges, R, between the two groups which can be counted using a minimal spanning tree. As in [45], Friedman and Rafsky's statistic test R has expected value $2mn/(m+n)$, where m and n are the sample sizes of the two groups. In our case, the null hypothesis is whether the leaching ($m = 8$) and non-leaching ($n = 35$) pesticides are drawn from the same distribution. In this case, the expected value of R is 13.02 with standard deviation 2.15 as in [45]. For example,

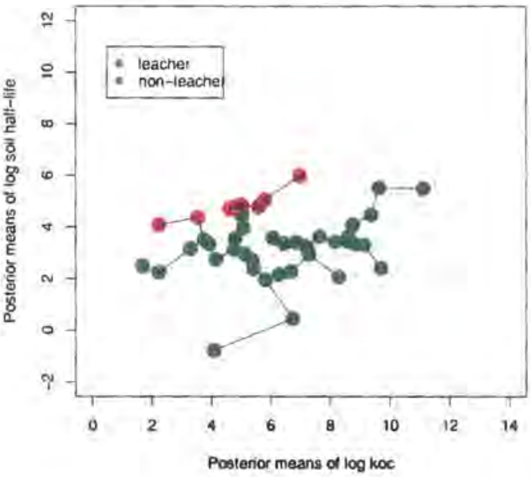

Figure 5.34: Minimal spanning tree for Model2*.

Figure 5.34 shows the minimal spanning tree for Model 2*, which is required to calculate R. From this figure, the total number of edges R between the pesticides in the two groups is 2. Table 5.1 shows the total number of edges R for the various models. As can be seen from this table, Models 2, 2*, 3 and 3* have $R = 5$, $R = 2$, $R = 5$ and $R = 7$, respectively, which are all small in comparison with 13.02 the expected value of R and its standard deviation 2.15, suggesting that leachers and non-leachers are indeed well-separated.

5.10 Conclusion

In this chapter, Bayesian approaches have been developed to analyse the problem of discriminating pesticides as leachers and non-leachers on the basis of two of their chemical properties, the adsorption coefficient k_{oc} and soil half-life $t^{soil}_{1/2}$ where the values of these covariates are uncertain. Prior information from two sources was used. Prior information for the covariates was based on their USDA data base means and variances and prior information for the coefficients β was based on logistic analyses of Gustafson's data. Six models were analysed; models 1, 2 and 3 each with a main effect linear predictor and models 1*, 2* and 3* with an interaction term in the linear predictor. MCMC simulation was used to draw samples from the posterior distributions using a Metropolis-Hastings algorithm implemented in the R package and the Gibbs sampler in WinBUGS. WinBUGS built-in tools were used to assess convergence of chains to their target distributions. The Deviance Information Criterion (DIC) was used to compare the proposed models. In brief, pesticides are correctly classified using these models, except that Models 3 and 3* misclassify pesticide number 32, Monuron.

These conclusions are strengthened using the multivariate runs test proposed in [17].

These models have succeeded for the first time in providing a complete separation between leaching and non-leaching pesticides, a classified in the EA database, on the basis of their chemical properties, where the values of these properties must be regarded as uncertain. They give better results than the Bayes linear method proposed in [43], which misclassified six leaching pesticides. On the other hand, the results are consistent with the data-analytic method given in [45] which chooses the combinations of data base values which best support the leacher/non-leacher status of each pesticide; namely, Gustafson's contention that leaching pesticides are those with low k_{oc} and high $t_{1/2}^{soil}$ values and non-leaching pesticides are those with low $t_{1/2}^{soil}$ and high k_{oc} values.

Pesticide	Posterior mean	Posterior sd	CI	Posterior probabilities $P[\pi_i > 0.5]$
1	0.0248	0.0777	(0, 0.245)	0.0072
2	0.1522	0.2119	(2e-04, 0.7797)	0.0915
3	0.2614	0.273	(8e-04, 0.9162)	0.2057
4	0.757	0.2559	(0.1296, 0.9987)	0.8201
5	0.0343	0.0938	(0, 0.3166)	0.0109
6	0.0606	0.1403	(0, 0.5356)	0.0303
7	0.7631	0.2587	(0.1136, 0.9991)	0.8195
8	0.0816	0.1646	(0, 0.6442)	0.043
9	0.2547	0.2702	(8e-04, 0.906)	0.1984
10	0.0437	0.1192	(0, 0.4598)	0.0204
11	0.0271	0.0963	(0, 0.3175)	0.0134
12	0.0205	0.0791	(0, 0.2458)	0.0079
13	0.8682	0.213	(0.2101, 1)	0.9119
14	0.0255	0.0986	(0, 0.3261)	0.0132
15	0.0448	0.1169	(0, 0.4294)	0.019
16	0.2076	0.2619	(1e-04, 0.8888)	0.16
17	0.1262	0.2067	(0, 0.7814)	0.0804
18	0.0458	0.1299	(0, 0.5058)	0.0259
19	0.043	0.1289	(0, 0.4914)	0.0242
20	0.0908	0.187	(0, 0.7317)	0.0598
21	0.0954	0.1567	(2e-04, 0.5841)	0.0401
22	0.7558	0.2657	(0.1005, 0.9993)	0.8128
23	0.0433	0.1045	(0, 0.3845)	0.0145
24	0.0689	0.1632	(0, 0.6513)	0.0426
25	0.0459	0.1336	(0, 0.5177)	0.0267
26	0.6961	0.2863	(0.0689, 0.9982)	0.7476
27	0.0389	0.1264	(0, 0.4831)	0.0233
28	0.1603	0.2333	(0, 0.8423)	0.113
29	0.0474	0.1164	(0, 0.4351)	0.0189
30	0.1251	0.2129	(0, 0.7966)	0.0854
31	0.1909	0.2396	(3e-04, 0.8474)	0.1326
32	0.3309	0.2905	(0.0029, 0.9445)	0.2794
33	0.1106	0.175	(2e-04, 0.6719)	0.0551
34	0.251	0.2699	(7e-04, 0.9083)	0.1944
35	0.051	0.1357	(0, 0.5374)	0.0282
36	0.8549	0.24	(0.1263, 1)	0.8884
37	0.0342	0.1039	(0, 0.3778)	0.0144
38	0.0884	0.1688	(0, 0.6584)	0.0454
39	0.7065	0.2615	(0.1178, 0.9943)	0.7711
40	0.79	0.2688	(0.0911, 1)	0.8331
41	0.0532	0.1258	(0, 0.481)	0.0233
42	0.14	0.227	(0, 0.8368)	0.0991
43	0.0396	0.1072	(0, 0.3949)	0.0156

Table 5.4: Statistical summaries for Model 1 for the leaching probability π.

Pesticide	Posterior mean	Posterior sd	CI	Posterior probabilities P[π_i > 0.5]
1	0.0092	0.0402	(0, 0.0943)	0.001
2	0.0994	0.2	(0, 0.7919)	0.0678
3	0.0964	0.2035	(0, 0.7982)	0.0703
4	0.9178	0.1841	(0.2727, 1)	0.9428
5	0.0446	0.1192	(0, 0.4492)	0.021
6	0.0552	0.1494	(0, 0.5961)	0.0347
7	0.9403	0.1599	(0.3656, 1)	0.9604
8	0.0702	0.1655	(0, 0.6504)	0.0439
9	0.114	0.215	(0, 0.8118)	0.0833
10	0.04	0.1261	(0, 0.4839)	0.024
11	0.0151	0.0685	(0, 0.1668)	0.0061
12	0.0123	0.0516	(0, 0.1288)	0.0023
13	0.9521	0.1441	(0.4184, 1)	0.9672
14	0.0283	0.1094	(0, 0.3721)	0.0169
15	0.0414	0.1151	(0, 0.4268)	0.0176
16	0.0918	0.1969	(0, 0.7708)	0.0654
17	0.0676	0.1724	(0, 0.6934)	0.0499
18	0.0259	0.11	(0, 0.3639)	0.0176
19	0.0366	0.124	(0, 0.4697)	0.022
20	0.0289	0.1107	(0, 0.4054)	0.0187
21	0.0855	0.1858	(0, 0.732)	0.057
22	0.9126	0.1874	(0.2628, 1)	0.9401
23	0.045	0.1191	(0, 0.4408)	0.02
24	0.0452	0.1401	(0, 0.5502)	0.0304
25	0.0217	0.102	(0, 0.3044)	0.0159
26	0.8812	0.2175	(0.1746, 1)	0.915
27	0.0349	0.122	(0, 0.4622)	0.0217
28	0.0766	0.1822	(0, 0.7197)	0.0562
29	0.0535	0.1424	(0, 0.5561)	0.0308
30	0.063	0.164	(0, 0.6703)	0.0423
31	0.1153	0.2157	(0, 0.8183)	0.0831
32	0.1397	0.2343	(0, 0.8552)	0.1042
33	0.0957	0.1911	(0, 0.7407)	0.0647
34	0.0732	0.1759	(0, 0.7072)	0.051
35	0.0252	0.0985	(0, 0.3197)	0.013
36	0.9284	0.1803	(0.2746, 1)	0.9461
37	0.0359	0.1195	(0, 0.4585)	0.0213
38	0.0722	0.1719	(0, 0.6921)	0.0486
39	0.9039	0.1904	(0.2671, 1)	0.9359
40	0.8798	0.2236	(0.1571, 1)	0.9101
41	0.0396	0.1097	(0, 0.4029)	0.0164
42	0.0654	0.1674	(0, 0.672)	0.0434
43	0.0291	0.1031	(0, 0.3576)	0.0153

Table 5.5: Statistical summaries for Model 1* for the leaching probability π.

Pesticide	Posterior mean	Posterior sd	CI	Posterior probabilities P[π_i > 0.5]
1	0.0181	0.061	(0, 0.1764)	0.0026
2	0.0549	0.1116	(0, 0.3633)	0.0124
3	0.1564	0.219	(0, 0.7737)	0.0972
4	0.7353	0.2895	(0.1246, 1)	0.7541
5	0.0074	0.0278	(0, 0.0942)	0
6	0.0198	0.0691	(0, 0.2197)	0.0046
7	0.7194	0.313	(0.0632, 1)	0.7339
8	0.0141	0.0407	(0, 0.1507)	0
9	0.1956	0.2384	(0, 0.8456)	0.135
10	0.0252	0.0764	(0, 0.2311)	0.0046
11	0.0173	0.0636	(0, 0.1915)	0.0052
12	0.0078	0.0351	(0, 0.0826)	7e-04
13	0.7606	0.2954	(0.0892, 1)	0.7815
14	0.0067	0.0312	(0, 0.0768)	7e-04
15	0.0223	0.076	(0, 0.2442)	0.0052
16	0.1225	0.1995	(0, 0.7262)	0.0685
17	0.0691	0.1549	(0, 0.6186)	0.0333
18	0.0051	0.0318	(0, 0.053)	7e-04
19	0.0281	0.0917	(0, 0.3162)	0.0091
20	0.1079	0.2061	(0, 0.7701)	0.0724
21	0.0251	0.0705	(0, 0.2237)	0.0046
22	0.715	0.3015	(0.0827, 1)	0.7339
23	0.0186	0.0535	(0, 0.1812)	7e-04
24	0.0302	0.0766	(0, 0.251)	0.0039
25	0.0849	0.1762	(0, 0.6586)	0.0496
26	0.6793	0.317	(0.0561, 1)	0.6941
27	0.0163	0.0761	(0, 0.2088)	0.0065
28	0.1277	0.1947	(0, 0.6974)	0.0718
29	0.0145	0.0516	(0, 0.1489)	0.002
30	0.0658	0.1475	(0, 0.5385)	0.0313
31	0.1474	0.1781	(0, 0.6483)	0.0574
32	0.4766	0.3123	(7e-04, 0.9745)	0.4866
33	0.0588	0.1251	(0, 0.4586)	0.0215
34	0.0748	0.1562	(0, 0.6227)	0.0365
35	0.052	0.1327	(0, 0.5015)	0.0261
36	0.6191	0.3574	(0.0191, 1)	0.6132
37	0.0109	0.033	(0, 0.1183)	0
38	0.0482	0.118	(0, 0.4333)	0.0176
39	0.7307	0.2889	(0.1203, 1)	0.7593
40	0.7712	0.3011	(0.0816, 1)	0.7893
41	0.0352	0.0984	(0, 0.3391)	0.0124
42	0.1312	0.2006	(0, 0.7152)	0.0731
43	0.0322	0.0871	(0, 0.2863)	0.0085

Table 5.6: Statistical summaries for Model 2 for the leaching probability π.

Pesticide	Posterior mean	Posterior sd	CI	Posterior probabilities
				P[π_i > 0.5]
1	0.0079	0.0391	(0, 0.0854)	0.0011
2	0.0311	0.0935	(0, 0.3096)	0.0103
3	0.0924	0.1873	(0, 0.6978)	0.0608
4	0.8333	0.2557	(0.1264, 1)	0.8634
5	0.0024	0.0147	(0, 0.0274)	0
6	0.0087	0.0531	(0, 0.1073)	0.0023
7	0.8326	0.266	(0.1029, 1)	0.8576
8	0.0062	0.0338	(0, 0.0677)	0.0011
9	0.1311	0.2151	(0, 0.7902)	0.0867
10	0.0109	0.0589	(0, 0.1118)	0.004
11	0.0085	0.0559	(0, 0.0779)	0.0034
12	0.0035	0.0244	(0, 0.0235)	0
13	0.853	0.253	(0.135, 1)	0.8731
14	0.0031	0.0278	(0, 0.0222)	6e-04
15	0.0079	0.0376	(0, 0.1118)	6e-04
16	0.0435	0.1296	(0, 0.4888)	0.0247
17	0.0384	0.1179	(0, 0.438)	0.0195
18	0.0012	0.0145	(0, 0.0017)	0
19	0.0154	0.0742	(0, 0.1863)	0.0075
20	0.0586	0.1588	(0, 0.6265)	0.039
21	0.0132	0.0545	(0, 0.1656)	0.0029
22	0.8232	0.2604	(0.1262, 1)	0.8553
23	0.009	0.0445	(0, 0.1054)	0.0011
24	0.0173	0.0691	(0, 0.1943)	0.0063
25	0.0413	0.1302	(0, 0.5134)	0.0276
26	0.7752	0.2924	(0.0716, 1)	0.8071
27	0.0062	0.0413	(0, 0.0558)	0.0017
28	0.0785	0.1677	(0, 0.6541)	0.0442
29	0.0069	0.0423	(0, 0.0737)	0.0011
30	0.0392	0.1259	(0, 0.4738)	0.0207
31	0.1143	0.1823	(0, 0.659)	0.0551
32	0.4198	0.334	(3e-04, 0.9797)	0.4202
33	0.0354	0.1047	(0, 0.3654)	0.0161
34	0.0257	0.0965	(0, 0.3096)	0.0126
35	0.0283	0.0986	(0, 0.361)	0.0132
36	0.7551	0.3226	(0.0242, 1)	0.7732
37	0.0047	0.0241	(0, 0.0543)	0
38	0.0259	0.0881	(0, 0.2849)	0.0086
39	0.8477	0.2343	(0.1724, 1)	0.8915
40	0.8387	0.2661	(0.0824, 1)	0.8617
41	0.0187	0.0749	(0, 0.2008)	0.0063
42	0.0856	0.1842	(0, 0.734)	0.0557
43	0.0147	0.0606	(0, 0.1679)	0.0023

Table 5.7: Statistical summaries for Model 2* for the leaching probability π.

Pesticide	Posterior mean	Posterior sd	CI	Posterior probabilities $P[\pi_i > 0.016]$
1	5e-04	0.0024	(0, 0.0034)	0.0022
2	0.003	0.0053	(0, 0.017)	0.0301
3	0.0122	0.0164	(1e-04, 0.0609)	0.2362
4	0.1081	0.0124	(0.085, 0.1327)	1
5	1e-04	2e-04	(0, 6e-04)	0
6	8e-04	0.0033	(0, 0.0077)	0.0106
7	0.1266	0.0561	(0.0422, 0.259)	0.9991
8	5e-04	0.0015	(0, 0.0036)	0.0013
9	0.0145	0.0165	(3e-04, 0.0606)	0.3137
10	6e-04	0.0024	(0, 0.0044)	0.0035
11	3e-04	0.0016	(0, 0.0025)	0.0018
12	1e-04	5e-04	(0, 9e-04)	0
13	0.0294	0.029	(0.0011, 0.1072)	0.58
14	2e-04	0.0014	(0, 0.0014)	0.0013
15	2e-04	7e-04	(0, 0.0019)	0
16	0.0056	0.0071	(0, 0.0259)	0.0793
17	0.002	0.0039	(0, 0.0133)	0.0173
18	1e-04	4e-04	(0, 5e-04)	0
19	3e-04	9e-04	(0, 0.0025)	0
20	8e-04	0.0015	(0, 0.0056)	0
21	0.0012	0.0032	(0, 0.0085)	0.0093
22	0.0213	0.0221	(0.001, 0.0778)	0.5906
23	3e-04	8e-04	(0, 0.0026)	0
24	0.0015	0.0041	(0, 0.0109)	0.0142
25	8e-04	0.0017	(0, 0.0053)	9e-04
26	0.0249	0.0114	(0.0084, 0.052)	0.7754
27	1e-04	5e-04	(0, 0.0012)	0
28	0.0086	0.0141	(0, 0.0487)	0.1586
29	3e-04	0.0015	(0, 0.0029)	9e-04
30	0.0044	0.0097	(0, 0.0298)	0.0696
31	0.0081	0.009	(2e-04, 0.0328)	0.1232
32	0.0346	0.0269	(0.0028, 0.1049)	0.7337
33	0.0025	0.0046	(0, 0.016)	0.0257
34	0.0059	0.011	(0, 0.0368)	0.0979
35	0.0025	0.0079	(0, 0.0216)	0.0372
36	0.0289	0.0183	(0.0044, 0.0751)	0.7333
37	2e-04	4e-04	(0, 0.0013)	0
38	0.0023	0.0067	(0, 0.0175)	0.0297
39	0.0195	0.0055	(0.0102, 0.0318)	0.7222
40	0.0197	0.0117	(0.0036, 0.0488)	0.5534
41	0.0012	0.0046	(0, 0.009)	0.012
42	0.0096	0.0152	(0, 0.052)	0.1786
43	5e-04	0.0011	(0, 0.0034)	4e-04

Table 5.8: Statistical summaries for Model 3 for the leaching probability π.

Pesticide	Posterior mean	Posterior sd	CI	Posterior probabilities $P[\pi_i > 0.016]$
1	5e-04	0.0115	(0, 0.0023)	0.0015
2	0.0024	0.0047	(0, 0.015)	0.0224
3	0.0086	0.0144	(0, 0.0481)	0.1624
4	0.1081	0.0125	(0.0848, 0.1339)	1
5	1e-04	2e-04	(0, 8e-04)	0
6	8e-04	0.0041	(0, 0.0071)	0.0097
7	0.1308	0.0569	(0.0431, 0.259)	1
8	4e-04	0.0014	(0, 0.0034)	0.001
9	0.0118	0.016	(1e-04, 0.056)	0.2336
10	5e-04	0.0019	(0, 0.0041)	0.0027
11	3e-04	0.0014	(0, 0.0022)	0.0015
12	1e-04	9e-04	(0, 8e-04)	5e-04
13	0.027	0.0279	(8e-04, 0.1037)	0.5388
14	2e-04	0.0039	(0, 9e-04)	0.0012
15	3e-04	6e-04	(0, 0.0017)	0
16	0.0026	0.0048	(0, 0.017)	0.0281
17	0.0017	0.0035	(0, 0.0116)	0.0117
18	0	2e-04	(0, 2e-04)	0
19	3e-04	8e-04	(0, 0.0021)	0
20	7e-04	0.0014	(0, 0.0049)	0
21	0.0011	0.003	(0, 0.0088)	0.0062
22	0.0222	0.0231	(0.001, 0.0873)	0.5945
23	3e-04	7e-04	(0, 0.0022)	0
24	0.0013	0.0043	(0, 0.0094)	0.0119
25	4e-04	0.0011	(0, 0.0037)	0
26	0.025	0.0115	(0.0078, 0.0521)	0.7644
27	1e-04	5e-04	(0, 0.001)	0
28	0.0077	0.014	(0, 0.0477)	0.141
29	4e-04	0.002	(0, 0.0032)	0.0027
30	0.0032	0.0087	(0, 0.0237)	0.0473
31	0.0077	0.0096	(1e-04, 0.0353)	0.1211
32	0.0331	0.0286	(0.0019, 0.1088)	0.6769
33	0.0028	0.0064	(0, 0.0189)	0.0336
34	0.0025	0.0071	(0, 0.02)	0.0351
35	0.0014	0.0051	(0, 0.0123)	0.0177
36	0.03	0.0192	(0.0048, 0.0781)	0.7527
37	2e-04	4e-04	(0, 0.0014)	0
38	0.002	0.0054	(0, 0.0146)	0.0197
39	0.0198	0.0055	(0.0107, 0.0315)	0.7418
40	0.0189	0.0114	(0.0035, 0.0474)	0.5294
41	0.001	0.0033	(0, 0.0076)	0.0087
42	0.007	0.012	(0, 0.0406)	0.1264
43	4e-04	8e-04	(0, 0.0024)	2e-04

Table 5.9: Statistical summaries for Model 3* for the leaching probability π.

Chapter 6

Conclusions and further studies

6.1 Introduction

The aim of this thesis has been to develop Bayesian methods to discriminate between leaching and non-leaching pesticides on the basis of two of their chemical properties: the adsorption coefficient k_{oc} and soil half-life $t_{1/2}^{soil}$. The problem was that these covariates (k_{oc} and $t_{1/2}^{soil}$) are uncertain in the sense that there are a range of values reported, in the USDA database, for both of them for each pesticide. The study was limited to 43 pesticides extracted from the UK Environment Agency (EA) where complete information was available regarding these pesticides. In addition, data from 22 pesticides, known as "Gustafson's data", with a single value reported for k_{oc} and $t_{1/2}^{soil}$ was analysed. The information derived from analysis of Gustafson's data together with the USDA database values was chosen as prior information in the analysis of the EA data.

In Chapter 1, the aims of this thesis and its objectives are stated. In addition, a detailed description of the data and its sources and deficiencies have been given.

Furthermore, there was discussion of related literature. The general methodology used throughout the thesis is outlined.

Chapter 2 reviews the statistical concepts, methods and tools used in the thesis. In particular, a description has been given of the logistic regression model and one of its deficiencies arising from this research. Also, there is a brief outline of Bayes linear methods and application to the general linear model with random covariates. Furthermore, the chapter contains discussion of some aspects of Bayesian inference and implementation using MCMC simulation techniques.

Chapters 3 and 4 extend the Bayesian method proposed in [44] and the Bayes linear approach proposed in [43], respectively.

Three Bayesian models to analyse the EA data are proposed in Chapter 6. These models use logistic regression with random covariates and prior information derives from both available data base values of k_{oc} and $t_{1/2}^{soil}$ for the covariates and Gustafson's data for the regression parameters. Each model has two stages. In all three models, the first stage uses logistic regression model with a logit link, while in the second stage different prior information for the unknown quantities is chosen for the three models. For each of the proposed models, combining the data with the prior information yielded complex joint posterior distributions where high dimensional integrations would be required to calculate marginal posterior distributions analytically. Consequently, MCMC simulation techniques were used to draw samples from the marginal posterior distributions. These techniques were implemented both via the WinBUGS software and the Metropolis-Hastings algorithm in R. Convergence of MCMC algorithms to their target distribution were assessed via various diagnostic tests such as tracing or the history of the chains, autocorrelation plots, posterior density plots and the modified Gelman-Rubin diagnostic test.

Half of the models use a linear function of the two covariates in the linear predictor, leading to linear discrimination, whereas the other half include an interaction term between the two covariates, leading to non-linear discrimination analogous to that proposed by Gustafson in [25]. However, a deficiency arises when fitting the latter logistic regression model to Gustafson's data; namely, the maximum likelihood of the regression parameters estimates (MLE) do not exist since there is complete separation between leachers and non-leachers in the space of the covariates relative to this interaction model. To remedy this, maximum estimated likelihood (MEL) is used instead.

For the Bayes linear models (Chapter 4) we used some Bayes linear diagnostics, such as the system resolution and the size ratio, for analysing the observed adjustments and examining any conflict between data and prior specification.

For the Bayes models, in Chapter 5, two statistical tools, the Deviance Information Criterion (DIC) and misclassification statistics, were used to compare and measure the ability of the proposed models to discriminate between leaching and non-leaching pesticides. The conclusions from the proposed models were supported using the multivariate runs test proposed in [17] to test for the degree of separation between leaching and non-leaching pesticides in the plane of the covariates. The next section summarizes the findings of this thesis.

6.2 Findings of the thesis

This study leads to satisfactory findings which can be summarized as follows.

1. This thesis documents the literature studies which are concerned with developing methods to help in predicting the potential of pesticides to leach into the soil and pollute the groundwater.

2. In the review of the literature studies, in particular [43], an error was noted in the plotting of the posterior discriminant, which was caused by using inappropriate covariates. The correct analysis suggests that the Bayes linear approach still gives good prediction.

3. A general formula to represent joint posterior distribution for logistic regression with uncertain covariates using a DAG is provided; see Section 2.4.6.

4. Formulae were derived to make Bayes linear computations for any general linear model with random covariates. These formulae were used in Chapter 4 to improve the Bayes linear approach proposed in [43]. They were also used to derive the prior information regarding Models 1 and 1* in Chapter 6.

5. A USDA database published several chemical and physical properties for each pesticides. These published values vary with soil type and climate. However, as discussed in [25] and from the analyses in Chapter 3, the adsorption coefficient (k_{oc}) and the estimated half-life of pesticide in the soil ($t_{1/2}^{soil}$) appear to have the most influence on the leaching potential of a pesticide.

6. The logistic regression model proposed in [44] to fit Gustafson's data was improved using logistic regression with an interaction term in the linear predictor.

The idea was suggested in [44], was formulated in [37] and implemented in this thesis using weighted maximum estimated likelihood as proposed in [8]. The WEMEL analysis led to perfect separation between leaching and non-leaching pesticides, while three pesticides are misclassified using logistic regression with the linear predictor proposed in [44].

7. The Bayesian method proposed in [44] and the the Bayes linear approach proposed in [43] were improved by introducing an interaction term of k_{oc} and $t_{1/2}^{soil}$. This led to slightly better results than the original models.

8. Three models were studied to analyse the EA data. These use logistic regression and prior information derived from the available values of the k_{oc} and $t_{1/2}^{soil}$ and from Gustafson's data. These models were improved using logistic regression with an interaction term. MCMC simulation techniques were used to draw samples from posterior distributions. These techniques were implemented using the WinBUGS software and the R package. The ability of these models to predict the potential of pesticides to leach varied from model to model. However, it was apparent that logistic regression with an interaction term (starred models) were better in fitting the EA data than the original models, where the predictor is linear in the two covariates.

9. The analyses of the Bayes and Bayes linear models led us to believe that the prior information derived from Gustafson's data should be down-weighted. A general form to down-weight prior information by modifying the prior variance-covariance matrix of regression parameters was analysed. The modification gives better results.

6.3 Suggestions for future work

6.3.1 Accounting for other uncertainties

We noticed in Chapter 1 that EA and CDFA use a different classification basis. In our analyses in this thesis, we assumed, as in [43], that the classification is secure. However, we believe that a further work is needed to investigate whether the classification is secure and account for any possible uncertainty in the classification in any future work.

Also, further investigations are needed to study the reasons behind the absence of a pesticide in a sample, whether this because it has not been used in that locality or it has been used but not yet reached the groundwater in a detectable amount, and account for any possible uncertainty in this regard; see the example given in Section 1.2.4.

6.3.2 Predictive and discrete models

The predictive and discrete models are fully described in Section 5.5.7. However, some convergence difficulties in implementing these models arose, and hence further work is needed to overcome these difficulties.

6.3.3 Leachability prediction for pesticides with uncertain chemical properties

In the USDA database there are an additional 17 pesticides with at least two values for each of k_{oc} and $t_{1/2}^{soil}$, but these pesticides are not part of the EA database. This raises the question of how we might predict the leachability status of these 17 pesticides given the data for the 43 EA pesticides and Gustafson's data.

6.3.4 Likelihood for hidden logistic regression

A proper likelihood approach is available under the hidden logistic regression model, proposed by Christmann and Rousseeuw in [8] (discussed here in Chapter 2) instead of the maximum estimated likelihood (MEL) method they propose. The structure of the hidden logistic regression model is described in Section 2.2.2 and depicted in Figure 2.1. Consider a single Bernoulli observation y. Then it is straightforward to show

$$\mathrm{P}\left[y \,|\, \boldsymbol{\beta}\right] = \sum_{t=0,1} \mathrm{P}\left[y \,|\, t\right] \mathrm{P}\left[t \,|\, \boldsymbol{\beta}\right] = \theta^{y}(1-\theta)^{1-y} \qquad y = 0, 1 \tag{6.1}$$

where y is related to t as described in Section 2.2.2 and

$$\theta = \delta_0 \mathrm{P}\left[t = 0 \,|\, \boldsymbol{\beta}\right] + \delta_1 \mathrm{P}\left[t = 1 \,|\, \boldsymbol{\beta}\right] \tag{6.2}$$

where $\mathrm{P}\left[t = 0 \,|\, \boldsymbol{\beta}\right] = \frac{1}{1+e^{\eta}}$ and $\eta = \boldsymbol{x}^T\boldsymbol{\beta}$. For n observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, 6.1 becomes

$$\mathrm{P}\left[\boldsymbol{y} \,|\, \boldsymbol{\beta}\right] = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i} \tag{6.3}$$

where

$$\begin{aligned} \theta_i &= \delta_0 \mathrm{P}\left[t_i = 0 \,|\, \boldsymbol{\beta}\right] + \delta_1 \mathrm{P}\left[t_i = 1 \,|\, \boldsymbol{\beta}\right] \\ &= \delta_0 + (\delta_1 - \delta_0) \cdot \tfrac{e^{\eta_i}}{1+e^{\eta_i}} \end{aligned} \qquad (6.4)$$

where $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$. Hence, provided the error probabilities δ_0 and δ_1 are specified (known values), we can maximize 6.3, with respect to $\boldsymbol{\beta}$, to obtain the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$.

Bayesian analysis for the above hidden logistic regression model is possible, provided prior distributions can be assessed for the $\boldsymbol{\beta}$, δ_0 and δ_1. Since $0 < \delta_0 < \frac{1}{2} < \delta_1 < 1$, we can put, for example, $0 < P_0 = 2\delta_0 < 1$ and $0 < P_1 = 2\delta_1 - 1 < 1$ and then choose independent beta distributions for P_0 and P_1. This analysis can be implemented using, for example, MCMC simulation with the prior assessment and the likelihood function in 6.3.

6.3.5 Bayes linear methods with likelihood

The Bayes linear method proposed in [43] may be extended in conjunction with "likelihood" function as follows. Let $\boldsymbol{\theta}$ denote a vector of unknown parameters. Then, as in Chapter 2, the adjusted mean and the adjusted variance for $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$ are given by

$$\mathrm{E}_{\boldsymbol{y}}[\boldsymbol{\theta}] = \mathrm{E}[\boldsymbol{\theta}] + \mathrm{Cov}[\boldsymbol{\theta}, \boldsymbol{y}]\mathrm{Var}[\boldsymbol{y}]^{-1}[\boldsymbol{y} - \mathrm{E}[\boldsymbol{y}]] \qquad (6.5)$$

$$\mathrm{Var}_{\boldsymbol{y}}[\boldsymbol{\theta}] = \mathrm{Var}[\boldsymbol{\theta}] - \mathrm{Cov}[\boldsymbol{\theta}, \boldsymbol{y}]\mathrm{Var}[\boldsymbol{y}]^{-1}\mathrm{Cov}[\boldsymbol{y}, \boldsymbol{\theta}] \qquad (6.6)$$

where the terms $\mathrm{Cov}[\boldsymbol{\theta},\ \boldsymbol{y}], \mathrm{Var}[\boldsymbol{y}]$ and $\mathrm{E}[\boldsymbol{y}]$ can be calculated as follows.

$$\mathrm{Cov}[\boldsymbol{\theta},\ \boldsymbol{y}] \ = \mathrm{Cov}[\boldsymbol{\theta},\ \mathrm{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]] \tag{6.7}$$

$$\mathrm{Var}[\boldsymbol{y}] = \mathrm{Var}[\mathrm{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]] + \mathrm{E}[\mathrm{Var}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]] \tag{6.8}$$

$$\mathrm{E}[\boldsymbol{y}] = \mathrm{E}[\mathrm{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]] \tag{6.9}$$

Hence, if we have expressions for $\mathrm{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]$ and $\mathrm{Var}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]$ for each $\boldsymbol{\theta}$, then the calculation can proceed, either exactly or approximately.

Example 1 Let $y \sim \mathrm{Binomial}(n, \theta)$, then $\mathrm{E}[y \,|\, \theta] = n\theta$ and $\mathrm{Var}[y \,|\, \theta] = n\theta(1 - \theta)$ and straightforward calculations give

$$\mathrm{Cov}[\theta,\ y] = n\mathrm{Var}[\theta] \tag{6.10}$$

$$\mathrm{Var}[y] = n(n-1)\mathrm{Var}[\theta] + n\mathrm{E}[\theta]\,(1 - \mathrm{E}[\theta]) \tag{6.11}$$

$$\mathrm{E}[y] \ = \ n\mathrm{E}[\theta] \tag{6.12}$$

So, provided we can assess the prior mean $\mathrm{E}[\theta]$ and prior variance $\mathrm{Var}[\theta]$ of θ, the posterior expectation $\mathrm{E}_{\boldsymbol{y}}[\theta]$ and the posterior variance $\mathrm{Var}_{\boldsymbol{y}}[\theta]$ are readily obtained. Note that we use exact expressions for $\mathrm{E}[y \,|\, \theta]$ and $\mathrm{Var}[y \,|\, \theta]$. Generalization to a vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ with independent binomial components is straightforward.

The above approach can be generalized to cases where exact expressions for $\mathrm{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]$ and $\mathrm{Var}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]$ are not available in closed form or are difficult to calculate. In such cases, we can derive simple approximations to these quantities using, for example, a Taylor series expansion for $\mathrm{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}]$.

First, put

$$\begin{aligned}\boldsymbol{\mu} &= \boldsymbol{\mu}(\boldsymbol{\theta}) = \mathrm{E}[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \qquad (6.13)\\ \Sigma &= \Sigma(\boldsymbol{\theta}) = \mathrm{Var}[\boldsymbol{y} \,|\, \boldsymbol{\theta}] \qquad (6.14)\end{aligned}$$

The aim now is to find a simple linear approximation to $\boldsymbol{\mu}(\boldsymbol{\theta})$. Let $\boldsymbol{t} = \mathrm{E}[\boldsymbol{\theta}]$ and $\mathrm{V} = \mathrm{Var}[\boldsymbol{\theta}]$ be the prior mean and prior variance of $\boldsymbol{\theta}$, respectively. A Taylor series expansion around $\boldsymbol{t}$ for $\boldsymbol{\mu}(\boldsymbol{\theta})$ is

$$\boldsymbol{\mu}(\boldsymbol{\theta}) \approx \boldsymbol{\mu}(\boldsymbol{t}) + \boldsymbol{\mu}'(\boldsymbol{t})(\boldsymbol{\theta} - \boldsymbol{t}) \qquad (6.15)$$

where $\boldsymbol{\theta}$ is a $p \times 1$ vector, $\boldsymbol{\mu}(\boldsymbol{\theta})$ is an $n \times 1$ vector and $\boldsymbol{\mu}'(\boldsymbol{t})$ is the $n \times p$ matrix of partial derivatives $\left[\frac{\partial \mu_i}{\partial \theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{t}}$. Hence,

$$\mathrm{Cov}[\boldsymbol{\theta}, \boldsymbol{y}] = \mathrm{Cov}[\boldsymbol{\theta}, \boldsymbol{\mu}(\boldsymbol{\theta})] \approx \mathrm{Cov}[\boldsymbol{\theta}, \boldsymbol{\mu}(\boldsymbol{t}) + \boldsymbol{\mu}'(\boldsymbol{t})(\boldsymbol{\theta} - \boldsymbol{t})] = \mathrm{V}\boldsymbol{\mu}'(\boldsymbol{t})^T \qquad (6.16)$$

$$\begin{aligned}\mathrm{Var}[\boldsymbol{y}] &= \mathrm{Var}[\boldsymbol{\mu}(\boldsymbol{\theta})] + \mathrm{E}[\Sigma(\boldsymbol{\theta})]\\ &\approx \boldsymbol{\mu}'(\boldsymbol{t})\mathrm{V}\boldsymbol{\mu}'(\boldsymbol{t})^T + \Sigma(\boldsymbol{t})\end{aligned} \qquad (6.17)$$

$$\begin{aligned}\mathrm{E}[\boldsymbol{y}] &= \mathrm{E}[\boldsymbol{\mu}(\boldsymbol{\theta})]\\ &\approx \mathrm{E}[\boldsymbol{\mu}(\boldsymbol{t}) + \boldsymbol{\mu}'(\boldsymbol{t})(\boldsymbol{\theta} - \boldsymbol{t})] = \boldsymbol{\mu}(\boldsymbol{t})\end{aligned} \qquad (6.18)$$

Therefore, the Bayes linear estimate and adjusted variance can be approximated using these expressions for $\mathrm{Cov}[\boldsymbol{\theta}, \boldsymbol{y}]$, $\mathrm{Var}[\boldsymbol{y}]$ and $\mathrm{E}[\boldsymbol{y}]$.

Example 2 Consider a logistic regression model with a single Bernoulli observation; i.e. $n = 1, \boldsymbol{\theta} = \boldsymbol{\beta}$ and $\boldsymbol{t} = \boldsymbol{b}$. Then,

$$\mu(\boldsymbol{\beta}) = \mathrm{E}[y \,|\, \boldsymbol{\beta}] = \frac{e^{\boldsymbol{x}^T\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}^T\boldsymbol{\beta}}} \qquad (6.19)$$

So,

$$\mu'(\boldsymbol{\beta}) = \mu(\boldsymbol{\beta})(1 - \mu(\boldsymbol{\beta}))\boldsymbol{x}^T \tag{6.20}$$

Hence,

$$\mu(\boldsymbol{\beta}) = \mu(\boldsymbol{b}) + \mu(\boldsymbol{b})(1 - \mu(\boldsymbol{b}))\boldsymbol{x}^T(\boldsymbol{\beta} - \boldsymbol{b}) \tag{6.21}$$

Extending this example to n independent observations is straightforward.

Application of these ideas to logistic regression with random covariates, the problem considered in this thesis, is presently under investigation.

Bibliography

[1] Adini, A. 2005. *Environmental Enlightenment.* Access via: `http://www.amiadini.com`.

[2] Albert, A., Anderson, J. A., 1984. *On the existence of maximum likelihood estimates in logistic regression models.* Biometrika, **71**, 1-10

[3] Bernardo, J. M., & Smith, A. F. M., 1994. *Bayesian Theory.* New York: Wiley.

[4] Besag, J., 1989. *A candidate's formula: A curious result in Bayesian prediction.* Biometrika, **76**, 1, p.183.

[5] Birkes, D., Dodge, Y., 1993. *Alternative methods of regression.* New York: Wiley- Interscience publication. itemBranham Branham, B., Milnert, E., Rieke, P., 1995. *Potential Groundwater Contamination from Pesticides and Fertilizers Used on Golf Courses.* USGA Green Section Record, **33(1)**, 33-37.

[6] Brooks, S. P. and Gelman, A., 1998. *Convergence Assessment Techniques for Markov Chain Monte Carlo.* Statistics and Computing, **8**, 319-335.

[7] Christmann, A., Rousseeuw, P.J., 2001. *Measuring Overlap in Logistic Regression.* Computational Statistics & Data Analysis, **37**, 65-75.

[8] Christmann, A., Rousseeuw, P.J., 2003. *Robustness Against Separation and Outliers in Logistic Regression.* Computational Statistics & Data Analysis, **43**, 315-332.

[9] Cohen, S.Z., Creeger, S. M., Carsel, R. F. & Enfield, C. G., *Potential pesticide contamination of groundwater from agricultural uses.* In Treatment and disposal of pesticide waste, ed. R. F. Kruger & J. N. Seiber. ACS Symposium Series no. 259, American Chemical Society, Washington DC, 1984, pp. 297-325.

[10] Collett, D., 1991. *Modelling binary data.* London: Chapman & Hall.

[11] Congdon, P., 2001. *Bayesian Statistical Modelling.* Chichester: Wiley.

[12] Copas, J.B., 1988. *Binary Regression Models for Contaminated Data.* With discussion. J. Roy. Statist. Soc. B **50**, 225-265.

[13] Craig, P.S., 2004. *Bayesian Statistics 4* . Department of Mathematical sciences, Durham University. Lecture notes.

[14] Dobson, Annette J. 2002. *An Introduction to Generalized Linear Models.* USA: Chapman and Hall.

[15] Draper, N., Smith, H., 1998. *Applied Regression Analysis.* New York: Wiley & sons.

[16] Ekholm, A., Palmgren, J., 1982. *A Model for Binary Response with Misclassification.* In: Gil-christ, R. (Ed.), GLIM-82, Proceedings of the International Conference on Generalized Linear Models. Springer, Heidelberg, pp. 128-143.

[17] Friedman, J., and Rafsky, L. C., 1979. *Multivariate generalizations of the Wolfowitz and Smirnov two-sample tests.* Annals of Statistics **7** 697

[18] Gelman, A., Carlin, J., Stern, H. and Rubin, D.B., 2004: *Bayesian Data Analysis.* London:Chapman and Hall.

[19] Gelman, A., Rubin, D. B., 1992. *Inference from Iterative Simulation Using Multiple Sequences.* Statistical Science, **7**, 457-511.

[20] Gelman, A., Sturtz, S., Ligges, U., 2005. *R2WinBUGS: A Package for Running WinBUGS from R.* Journal of Statistical Software, **12**, issue 3.

[21] Gilks, W., Richardson, S. and Spiegelhalter, D. (1996) *Markov Chain Monte carlo in Practice.* London: Chapman & Hall.

[22] Gill, J., 2002. *Bayesian Methods: A Social and Behavioral Sciences Approach.* Chapman and Hall.

[23] Goldstein, M., 1998. *Bayes Linear Analysis*, Encyclopedia of Statistical Sciences, **3**.

[24] Goldstein, M. & Wooff, D., 2007. *Bayes Linear Statistics: Theory and Methods.* Chichester: Wiley.

[25] Gustafson, D.I., 1989. *Groundwater ubiquity score: a simple method for assessing pesticide leaching.* Environmental Toxicology and Chemistry, **8**, 339-357.

[26] Henze, N., and Penrose, M. D., 1998. *On the multivariate runs test* Annals of Statistics **27**, 290

[27] Hastie, T.J., Tibshirani, R.J. and Friedman, J. (2001)*The elements of statistical learning. Data mining inference and prediction.* New York: Springer-Verlag.

[28] Jury, W. A., Focht, D. D. & Farmer, W. J. 1987. *Evaluation of pesticide groundwater pollution potential from standard indices of soil chemical adsorption and biodegradation.* J. Envir. Qual., **16**, 422-8.

[29] Kendall, M., Stuart, A.,1952. *The advanced theory of statistics.* London: Charles Griffin.

[30] Lanchenbruch, P. A. (1975)*Discriminant analysis.* New York: Hafner Press.

[31] Lee, P.M., 2004. *Bayesian Statistics: An Introduction.* Hodder Arnold.

[32] Lindley, D. V. & Scott, W. F., 1995. *New Cambridge Statistical Tables.* Second Edition. Cambridge: Cambridge University Press.

[33] McCullagh, P. and Nelder, J. A., 1989. *Generalized Linear Models.* Second Edition. New York: Chapman & Hall.

[34] Pawitan, Y., 2001. *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* New York: Oxford University Press.

[35] R Development Core Team (2006). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

[36] Santner, T.J., Duffy, D.E., 1986. *A Note on A. Albert and Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models.* Biometrika **73**, 755-758.

[37] Seheult, A. H., 2001. *Discriminant analysis.* In: El-Shaarawi, A. & Piegorsch, W. W. (eds). Encyclopedia of Environmentrics. J Wiley & sons, Ltd. Chichester, **1**, 523-530.

[38] Seheult, A. H., 2004. *Statistical Mthods III.* Department of Mathematical sciences, Durham University. Lecture notes.

[39] Spiegelhalter, D., N. Best, B. Carlin, and A. Van der Linde, 2002. *Bayesian Measures of Model Complexity and Fit (with Discussion).* Journal of Royal Statistic Society B **64**, 583-640.

[40] Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D., 2003. *WinBUGS User Manual. Version 1.4 (http://www.mrc-bsu.cam.ac.uk/bugs.)* Technical report, Medical Research Council Biostatistics Unit. Cambridge.

[41] Venables, W. N. and Ripley, B. D., 2002: *Modern Applied Statistics with S.* Fourth Edition. New York: Springer-Verlag.

[42] Wilkerson, M.R. & Kim, K.D. 1986. *The Pesticide Contamination Prevention Act: Setting Specific Numerical Values.* California Department of Food and Agriculture, Environmental Monitoring and Pest Management, Sacramento, CA.

[43] Wooff, D.A., Seheult, A.H., Coolen, F.P.A. & Worrall, F., 1998. *Bayesian Discrimination with Uncertain Covariates for Pesticide Contamination.* In:

Barnett, V., Stein, A.& Feridun Turkman, K. (eds.) Statistics for the Environment **4**. Wiley, Chichester, 337-354.

[44] Worrall, F., Wooff, D.A., Seheult, A.H. & Coolen, F.P.A. 1998. *A Bayesian Approach to the Analysis of Environmental Fate and Behaviour Data for Pesticide Registration*, Pesticide Science, **54**, 99-112.

[45] Worrall, F., Wooff, D.A., Seheult, A.H. & Coolen, F.P.A, 2000. *New approaches to assessing the risk of groundwater contamination by pesticides*, Journal of the Geological Society, London, **157**, 877-884.

Appendix A

MCMC codes

A.1 Fun.Model6.int

```
#This function is used to implement the discrete model,

#as a general application, in R.

function(num.iters,sigma.beta0,sigma.beta1,sigma.beta2,

         prec.beta0,prec.beta1,prec.beta2){

p11<-sort(c(4.99721227376411, 5.66296048013595, ...))#

     Koc values of the first leacher pesticide,

     which is the 4-th pesticide in Table 1.1.

p21<-sort(c(3.52636052461616, 3.55534806148941))

.

.

.

p91<-sort(c(8.77276520994979, 8.29404964010203, 8.76405326934776))#

    Koc values of the first non-leacher pesticide,
    which is the 1st pesticide in Table 1.1.

.

.

.

p431<-sort(c(8.89562962713648, 8.43381158247719, ...))

p12<-sort(c(5.15329159449778, 3.73766961828337, ...))#

    Soil half-life values of the first leacher pesticide,

     which is the 4-th pesticide in Table 1.1.

p22<-sort(c(2.30258509299405, 2.63905732961526, ...))

.

.

.

p92<-sort(c(3.40119738166216, 4.0943445622221, ...))

    Soil half-life values of the first non-leacher pesticide,

    which is the 1st pesticide in Table 1.1.

.

.

.

p432<-sort(c(4.0943445622221, 4.0943445622221, ...))

p1<-list(p11,p21, ...,p431)

p2<-list(p12,p22,...,p432)

y<-c(1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0,

     0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
    0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

x1.star<-x2.star<-matrix(NA,43,num.iters)

beta0.star<-beta1.star<-beta2.star<-rep(NA,num.iters)

prop1.star<-prop2.star<-matrix(NA,1,num.iters)

x1.star[,1]<-koc.c.group.logs

x2.star[,1]<-t.c.group.logs

prop1.star[,1]<-0.5

prop2.star[,1]<-0.5

beta0.star[1]<-3

beta1.star[1]<--1

beta2.star[1]<-1

u.x1.set<-runif(num.iters)

u.x2.set<-runif(num.iters)

u.beta0.set<-runif(num.iters)

u.beta1.set<-runif(num.iters)

u.beta2.set<-runif(num.iters)

counter.x1<-counter.x2<-counter.beta0<-counter.beta1<-counter.beta2<-0

x1.cand<-matrix(NA,43,1)

x2.cand<-matrix(NA,43,1)

prop1.cand<-matrix(NA,43,1)

prop2.cand<-matrix(NA,43,1)

for(i in 2:num.iters){

for(j in 1:8){
```

```
x1.cand[j,]<-sample(p1[[j]],size=1,prob=c(2^((length(p1[[j]])-1):0)))
}
for(k in 9:43){
x1.cand[k,]<-sample(p1[[k]],size=1,prob=c(2^(0:(length(p1[[k]])-1))))
}
for(l in 1:8){
x2.cand[l,]<-sample(p2[[l]],size=1,prob=c(2^(0:(length(p2[[l]])-1))))
}
for(m in 9:43){
x2.cand[m,]<-sample(p2[[m]],size=1,prob=c(2^((length(p2[[m]])-1):0)))
}
for(n in 1:8){
prop1.cand[n,]<-2^((length(p1[[n]])-1):0)[which(p1[[n]]
               ==x1.cand[n,])[1]]/sum(2^((length(p1[[n]])-1):0))
prop2.cand[n,]<-2^(0:(length(p2[[n]])-1))[which(p2[[n]]
              ==x2.cand[n,])[1]]/sum(2^(0:(length(p2[[n]])-1)))
}
for(w in 9:43){
prop1.cand[w,]<-2^(0:(length(p1[[w]])-1))[which(p1[[w]]
              ==x1.cand[w,])[1]]/sum(2^(0:(length(p1[[w]])-1)))
prop2.cand[w,]<-2^((length(p2[[w]])-1):0)[which(p2[[w]]
              ==x2.cand[w,])[1]]/sum(2^((length(p2[[w]])-1):0))
}
```

```
prop1.new<-sum(log(prop1.cand))

eta<-beta0.star[i-1]+beta1.star[i-1]*x2.star[,i-1]

    +beta2.star[i-1]*x1.cand*x2.star[,i-1]

p.cand<-exp(eta)/(1+exp(eta))

f.cand<-sum(y*eta-log(1+exp(eta)))

eta<-beta0.star[i-1]+beta1.star[i-1]*x2.star[,i-1]

    +beta2.star[i-1]*x1.star[,i-1]*x2.star[,i-1]

p.old<-exp(eta)/(1+exp(eta))

f.old<-sum(y*eta-log(1+exp(eta)))

x1.part<-prop1.star[,i-1]+f.cand-prop1.new-f.old

if(log(u.x1.set[i])<min(0,x1.part)){

x1.star[,i]<-x1.cand

 prop1.star[,i]<-prop1.new

 counter.x1<-counter.x1+1

}else{

 x1.star[,i]<-x1.star[,i-1]


prop1.star[,i]<-prop1.star[,i-1]

 }

prop2.new<-sum(log(prop2.cand))

eta<-beta0.star[i-1]+beta1.star[i-1]*x2.cand

    +beta2.star[i-1]*x1.star[,i]*x2.cand

p.cand<-exp(eta)/(1+exp(eta))
```

```
f.cand<-sum(y*eta-log(1+exp(eta)))

eta<-beta0.star[i-1]+beta1.star[i-1]*x2.star[,i-1]

    +beta2.star[i-1]*x1.star[,i]*x2.star[,i-1]

p.old<-exp(eta)/(1+exp(eta))

f.old<-sum(y*eta-log(1+exp(eta)))

x2.part<-prop2.star[,i-1]+f.cand-prop2.new-f.old


if(log(u.x2.set[i])<min(0,x2.part)){

x2.star[,i]<-x2.cand

 prop2.star[,i]<-prop2.new

 counter.x2<-counter.x2+1

 }else{

 x2.star[,i]<-x2.star[,i-1]

prop2.star[,i]<-prop2.star[,i-1]

}

beta0.cand<-rnorm(1,beta0.star[i-1],sigma.beta0)

eta<-beta0.cand+beta1.star[i-1]*x2.star[,i]

    +beta2.star[i-1]*x1.star[,i]*x2.star[,i]

p.cand<-exp(eta)/(1+exp(eta))

f.cand<-sum(y*eta-log(1+exp(eta)))

prior.cand<--0.5*prec.beta0*((beta0.cand+20.384)^2)

eta<-beta0.star[i-1]+beta1.star[i-1]*x2.star[,i]

    +beta2.star[i-1]*x1.star[,i]*x2.star[,i]
```

```
p.old<-exp(eta)/(1+exp(eta))

f.old<-sum(y*eta-log(1+exp(eta)))

prior.old<--0.5*prec.beta0*((beta0.star[i-1]+20.384)^2)

beta0.part<-f.cand+prior.cand-f.old-prior.old

if(log(u.beta0.set[i])<min(0,beta0.part)){

beta0.star[i]<-beta0.cand

 counter.beta0<-counter.beta0+1

 }else{

 beta0.star[i]<-beta0.star[i-1]

 }

beta1.cand<-rnorm(1,beta1.star[i-1],sigma.beta1)

eta<-beta0.star[i]+beta1.cand*x2.star[,i]

    +beta2.star[i-1]*x1.star[,i]*x2.star[,i]

p.cand<-exp(eta)/(1+exp(eta))

f.cand<-sum(y*eta-log(1+exp(eta)))

prior.cand<--0.5*prec.beta1*((beta1.cand-17.380)^2)

eta<-beta0.star[i]+beta1.star[i-1]*x2.star[,i]

    +beta2.star[i-1]*x1.star[,i]*x2.star[,i]

p.old<-exp(eta)/(1+exp(eta))

f.old<-sum(y*eta-log(1+exp(eta)))

prior.old<--0.5*prec.beta1*((beta1.star[i-1]-17.380)^2)

beta1.part<-f.cand+prior.cand-f.old-prior.old

if(log(u.beta1.set[i])<min(0,beta1.part)){
```

```
beta1.star[i]<-beta1.cand
  counter.beta1<-counter.beta1+1
 }else{
 beta1.star[i]<-beta1.star[i-1]
 }
beta2.cand<-rnorm(1,beta2.star[i-1],sigma.beta2)
eta<-beta0.star[i]+beta1.star[i]*x2.star[,i]
    +beta2.cand*x1.star[,i]*x2.star[,i]
p.cand<-exp(eta)/(1+exp(eta))
f.cand<-sum(y*eta-log(1+exp(eta)))
prior.cand<--0.5*prec.beta2*((beta2.cand+1.947)^2)
eta<-beta0.star[i]+beta1.star[i]*x2.star[,i]
    +beta2.star[i-1]*x1.star[,i]*x2.star[,i]
p.old<-exp(eta)/(1+exp(eta))
f.old<-sum(y*eta-log(1+exp(eta)))
prior.old<--0.5*prec.beta2*((beta2.star[i-1]+1.947)^2)
beta2.part<-f.cand+prior.cand-f.old-prior.old
if(log(u.beta2.set[i])<min(0,beta2.part)){
beta2.star[i]<-beta2.cand
counter.beta2<-counter.beta2+1
 }else{
 beta2.star[i]<-beta2.star[i-1]
 }
 }

ac.rate.x1<-counter.x1/num.iters

ac.rate.x2<-counter.x2/num.iters

ac.rate.beta0<-counter.beta0/num.iters

ac.rate.beta1<-counter.beta1/num.iters

ac.rate.beta2<-counter.beta2/num.iters

x1.sim<-apply(x1.star,1,mean)

x2.sim<-apply(x2.star,1,mean)

x1.sd<-apply(x1.star,1,sd)

x2.sd<-apply(x2.star,1,sd)

beta0.mean<-mean(beta0.star)

beta1.mean<-mean(beta1.star)

beta2.mean<-mean(beta2.star)

beta0.sd<-sd(beta0.star)

beta1.sd<-sd(beta1.star)

beta2.sd<-sd(beta2.star)

CI.x1<-matrix(NA,43,2)

CI.x2<-matrix(NA,43,2)

CI.beta<-matrix(NA,3,2)

for(i in 1:43){

CI.x1[i,]<-quantile(x1.star[i,],probs=c(0.025,0.975))

CI.x2[i,]<-quantile(x2.star[i,],probs=c(0.025,0.975))

}
```

```
CI.beta[1,]<-quantile(beta0.star,probs=c(0.025,0.975))

CI.beta[2,]<-quantile(beta1.star,probs=c(0.025,0.975))

CI.beta[3,]<-quantile(beta2.star,probs=c(0.025,0.975))

DIC.STAT<-Dic.fun(num.iters,t(x1.star),t(x2.star),

        beta0.star,beta1.star,beta2.star)

TABLE.X1<-cbind(x1.sim,x1.sd,x1.x2.mode[,1],

          CI.x1[,1],CI.x1[,2])

colnames(TABLE.X1)=c("mean","sd","mode","lower","upper")

TABLE.X2<-cbind(x2.sim,x2.sd,x1.x2.mode[,2],

          CI.x2[,1],CI.x2[,2])

colnames(TABLE.X2)=c("mean","sd","mode","lower","upper")

TABLE.beta0<-cbind(beta0.mean,beta0.sd,

           CI.beta[1,1],CI.beta[1,2])

colnames(TABLE.beta0)=c("mean","sd","lower","upper")

TABLE.beta1<-cbind(beta1.mean,beta1.sd,

           CI.beta[2,1],CI.beta[2,2])

colnames(TABLE.beta1)=c("mean","sd","lower","upper")

TABLE.beta2<-cbind(beta2.mean,beta2.sd,

           CI.beta[3,1],CI.beta[3,2])

colnames(TABLE.beta2)=c("mean","sd","lower","upper")

list(x1.star=x1.star,x2.star=x2.star,beta0.star=beta0.star,

beta1.star=beta1.star,beta2.star=beta2.star,TABLE.X1=TABLE.X1,

TABLE.X2=TABLE.X2,x1.x2.mode=x1.x2.mode,TABLE.beta0=TABLE.beta0,
TABLE.beta1=TABLE.beta1,TABLE.beta2=TABLE.beta2,DIC.STAT=DIC.STAT,

beta0.mean=beta0.mean,beta1.mean=beta1.mean,beta2.mean=beta2.mean,

plot(x1.sim,x2.sim,xlab="Simulated values of x1=Log Koc",

     ylab="Simulated values of x2=Log soil half-life",cex=2,

     pch=as.numeric(y.cir.group),col="red"),

text(x1.sim,x2.sim,1:43,cex=0.6,adj=0.5),

curve(-beta0.mean/(beta1.mean+beta2.mean*x),add=TRUE)

)
```

A.1.1 Dic.fun

```
#This function is used in "Fun.Model6.int" to calculate the DIC

statstics.

function(N,x1,x2,beta0,beta1,beta2){

dbar<-matrix(NA,N,1)

beta0.mean<-mean(beta0)

beta1.mean<-mean(beta1)

beta2.mean<-mean(beta2)

X1.mean<-X2.mean<-matrix(NA,43,1)

for(i in 1:43){

X1.mean[i,]<-mean(x1[,i])

X2.mean[i,]<-mean(x2[,i])

}

Dhat<--2*Log.Lik(X1.mean,X2.mean,beta0.mean,beta1.mean,beta2.mean)
for (i in 1:N){

dbar[i,]<-Log.Lik(x1[i,],x2[i,],beta0[i],beta1[i],beta2[i])

}

Dbar<--2*mean(dbar)

pD=Dbar-Dhat

DIC=Dbar+pD

list(Dbar=Dbar,Dhat=Dhat,DIC=DIC,pD=pD)

}
```

A.1.2 Log.Lik

```
#This function is used in "Dic.fun".

function(x1,x2,beta0,beta1,beta2)

{

sum((beta0+beta1*x2+beta2*x1*x2)*y.group

     -log(1+exp(beta0+beta1*x2+beta2*x1*x2)))

}
```

A.2 WinBUGS code

```
model{

  for( i in 1 : 43 ) {

       y[i] ~ dbern(pi[i])

       logit(pi[i])<-eta[i]

       }
```

```
     eta[1:43]~dmnorm(mean[],precision[,])

     }

data ;

list(y=c(0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,

         0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0,

         0, 0, 0, 0, 0, 0, 1, 0, 0,1, 1, 0, 0, 0),

mean=c(-18.1602550369174, 6.18955405790843, ...),

precision=structure(.Data=c(0.161852929395662, 0.000687084927984936,...

),.Dim=c(43,43)))

inits;

list(eta=c(-1.23692607507110, 1.16338457912207, ...))

list(eta=c(-9.29418446263298, 0.196954105049372, ...))

list(eta=c(-17.5244785845280, -4.65396864339709, ...))
```