ANALYSIS OF AGGREGATED PLANT DISEASE INCIDENCE DATA

by

Sembakutti Samita

PhD

University of Edinburgh

1995

DECLARATION

I hereby declare that the contents of this thesis are, except where otherwise stated, my own original work.

Sembakutti Samita

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to Dr G. Hughes, University of Edinburgh, for his supervision and admirable patience that requires no further testing.

My sincere thanks also go to Dr R. O. Thattil, University of Peradeniya, for his supervision during the in-country research, conducted in Sri Lanka. Special thanks are also extended to Dr. S.C. Nelson, North Carolina State University, USA, for his helpful comments on 2DCLASS analysis.

A special acknowledgement is given to the Faculty of Agriculture, University of Peradeniya for granting me a scholarship to pursue my PhD and, Overseas Development Administration for sponsoring the scholarship through the British Council.

Sumith Senaratne is greatly appreciated for helping in carrying out the data collection survey.

Finally, I would like to thank all my friends and colleagues who helped me in numerous ways in making this study a success.

ABSTRACT

This study concerns the analysis of plant disease incidence data when the observations are made as presence and absence of disease symptoms on plants or plant units.

If diseased plants (or plant units) are randomly dispersed, the frequency distribution of diseased plants (or plant units) per sample may be described by a binomial distribution, and statistical analyses may be based on the linear logistic model. Since most disease incidence data do not have a random spatial pattern, the binomial distribution can hardly ever, in practice, be used to describe observed frequencies. In this study, the use of conditional probability distributions, such as the logistic-normal binomial distribution, for such data is illustrated. Both descriptive distribution fitting and statistical modelling are discussed.

The study evaluates several methods for analysis of incidence data which do not exhibit a random spatial pattern. Some of these methods are applied to plant disease data for the first time. A method of choosing between the different analyses is discussed. All the techniques are illustrated using examples and, as an application, survey data collected on pineapple wilt disease in Sri Lanka are extensively studied.

As an alternative method of describing disease incidence data with a non random spatial pattern, the use of two-dimensional distance class (2DCLASS) analysis was evaluated using the same survey data. 2DCLASS analysis is widely accepted in plant disease epidemiology as a method of analysing non-random spatial patterns when the observations are made as presence or absence of the disease on individual plant basis. We demonstrate the possibility of using quadrat-based data in 2DCLASS analysis. We investigate the use of 2DCLASS analysis as a methodology and find some drawbacks with this technique, which are discussed in detail. Moreover, this study introduces a new parameter in the 2DCLASS analysis called Scaled Core Cluster size, that may be more suitable to use for comparison of datasets of different sizes.

CONTENTS

1. INTRODUCTION

In an experiment in which the variable of interest is disease incidence, the observations made on an individual experimental unit (a plant, say) may take one of two possible forms, i.e., presence or absence of the disease (or its symptoms). The data in this form are said to be binary (quantal) and the two possible forms for each observation are often described generically by the terms 'success' and 'failure'.

In most circumstances, interest centres not just on the response of one particular experimental unit (individual plant) but on a group of units that have all been treated in a similar manner. Thus the individual responses from each plant in an experimental plot, in which all plants have been treated alike, may be combined to give the proportion of plants diseased. The resulting data are then referred to as grouped binary data, and represent the number of successes out of the total number of units exposed to a particular set of experimental conditions. It is well known that the data in the form of proportions are often modelled using the binomial distribution (Cox and Snell, 1989).

If every plant in the field has an equal and independent chance of becoming diseased, the resulting distribution of disease incidence over the field will be random one. However, the actual disease incidence may deviate from a random one due to diseased plants occurring together more often than would occur by chance (Cochran, 1936). The diseased plants may occur randomly over the field, as could happen if the distributing agent of the disease is an insect species, and the insects had an equal access to all the plants in the field. On the other hand, the deviation from randomness may be of a more regular type, infection being higher, for instance, near the borders of a field or plot than in the interior. However, a regular pattern of incidence seldom occurs in practice (Collett, 1991).

In plant epidemiology, dispersion of disease incidence is characterised by the spatial pattern of diseased plants (Pielou, 1977; Campbell and Madden, 1990). Occurrence of diseased plants in groups or patches is often referred to as 'aggregation' or 'patchiness'. According to Jeger (1989), aggregation is the pattern of disease commonly observed in epidemiological studies. Sometimes aggregation is referred to as 'clustering'. This is the spatial pattern observed in contrast to random pattern of diseased plants. Thus some authors refer to this situation as a non-random pattern or 'heterogeneity' of diseased plants (Campbell and Madden, 1990).

In statistical analysis, this condition is referred to as overdispersion (Collett, 1991), and data of this type are often called clustered binary data (Rao, 1992). Overdispersion is so common in practice that some would maintain that overdispersion is the norm and random dispersion the exception. The incidence and the degree of overdispersion encountered depend greatly on the field of application (McCullagh and Nelder, 1989). In the statistical literature, overdispersion has been explained as variation between binary response probabilities or correlation between binary responses (Collett, 1991). Such variability is often referred to as extra-binomial variation (Crowder, 1978; Williams, 1982; Brooks, 1984: Boos, 1993). Aggregated disease incidence exhibits overdispersion while regular disease incidence leads to underdispersion. Standard statistical procedures (section 3.2.1) are not appropriate to analyse disease incidence data when disease incidence has any spatial pattern other than random.

The first task in the statistical analysis of disease incidence data is to examine the randomness of the observations. The use of probability distributions in determining the randomness is a well-established technique in plant disease epidemiology (Madden, 1989; Hughes and Madden, 1993). In plant disease epidemiology, data are often collected in the form of incidence maps (Madden *et al.*, 1987). For incidence maps, an examination of whether the disease incidence is random or not may be done by dividing the area into sample units (quadrats) containing the same number of plants, say from 6 to 12 per quadrat, and by comparing observed and expected quadrat frequencies of number of plants diseased per quadrat (Cochran, 1936). Given certain assumptions (Cochran, 1977), the binomial distribution provides expected frequencies based on the supposition of spatial randomness for disease incidence.

When the disease incidence is not random, the binomial distribution cannot provide an adequate description of the data (Hughes and Madden, 1993). On many occasions, the observed disease frequency distribution contains higher frequencies in the upper and lower tails compared to the expected binomial frequencies. For such situations, conditional distributions such as the beta-binomial (Skellam, 1948) and logistic-normal binomial (Pierce and Sands, 1975) may provide a better description of the data.

Experiments often aim to investigate the effect of an explanatory variable on a response variable. For instance, several fungicide regimes may need to be compared for their effect on disease incidence. Or, if more than one explanatory variable is under study, an experiment may be extended to investigate the interaction between explanatory variables in their effect on a response variable of interest. On such occasions it is necessary to model the data (Collett, 1991) and the parameters based on the fitted model are used for comparison between treatments. The experimental design suggests an analysis of variance (ANOVA) (Snedecor and Cochran, 1989) approach, but data in the form of proportions present a number of problems (section 3.2) for 'normal theory' ANOVA. Instead of forcing the data into the framework of ANOVA, it may be more appropriate to use Generalised Linear Models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983; 1989) to analyse such data (section 3.2).

It is well established that the most useful analogue for incidence data of the linear model for normally distributed data is provided by the linear logistic model (Cox and Snell, 1989) in which maximum likelihood estimates can easily be obtained. However, if the incidence is overdispersed, the linear logistic model is not a suitable model for the data (Haseman and Kupper, 1979).

Several methods have been suggested for statistical analysis when incidence is overdispersed. Most of these techniques were originally developed in the area of teratology and toxicology. Thus this study attempts to investigate the applicability of these techniques in plant disease epidemiology. These methods have been discussed in chapter 3. Section 3.5 describes some guidelines for choosing the most efficient from among the available methods.

In chapter 4, the analysis of disease incidence data is illustrated using survey data collected in Sri Lanka on pineapple wilt disease.

1.1 Pineapple wilt disease

A wilt disease of pineapple (*Ananas comosus*) was first described in Hawaii in the early 1900s (Larsen, 1910) and since then has been reported as a serious problem in most areas of the world where pineapple is cultivated (Carter, 1963). The disease has been associated consistently with the presence of mealybugs (Carter, 1932; Singh and Sastry, 1974). The etiology of mealy bug wilt of pineapple has been controversial for many years. The theory that it is caused by a toxin secreted by mealybugs during feeding on pineapple was proposed by Carter (Carter, 1932; 1933). Later, biological data (Carter, 1963) suggested that mealybug wilt of pineapple was not caused solely by toxins in mealybug salivary secretions, but that an unidentified "latent transmissible factor" was also associated with the disease. The concept of a viral etiology began to emerge (Carter, 1963). Some recent work (Rohrbach *et al.*, 1988) has clearly identified the presence of a virus in mealybug-wilt-affected pineapple plants.

Several different species of mealybugs have been associated with wilt disease of pineapple in Hawaii and in other areas where pineapple is grown (Rohrbach *et al.*; 1988). Because the taxonomy of these mealybugs was not well understood until relatively recently, some confusion exists in the published literature dealing with these pests (Gunasinghe and German, 1989). Early references (Carter, 1932) to the mealybugs associated with wilt generally refer to *Pseudococcus brevipes* (Cockerell) (currently named as *Dysmicoccus brevipes*) as the pineapple mealybug. However, Ito (1938) pointed out that there were two distinct types of pineapple mealybugs associated with wilt disease in Hawaii, which he referred to as the pink form and the grey form. Beardsley (1959) demonstrated that there were valid morphological differences between pink and grey forms and recognised the grey form as a distinct new species, *Dysmicoccus neobrevipes*.

Ants are a problem in pineapple fields only because of their association with mealybugs (Rohrbach *et al.*, 1988). It is the ants' caretaking behaviour that allows the mealybug species to prosper. In the literature, there are two hypotheses that attempt to explain why mealybugs flourish as a result of ant tending (Nixon, 1951). The first is that ants protect mealybugs from any potential parasites and predators. The second hypothesis is that benefits to the mealybugs result from removal of honeydew (secreted by mealybugs) by ants. Honeydew removal prevents the accumulation of honeydew on the plant and mealybugs, and impedes sooty mold build-up, both of which may be detrimental to mealybugs.

When the plants are contaminated with the virus, the first symptoms appear in the roots, which cease growth, collapse and rot. Above-ground symptoms then follow. Carter (1956: 1963) described the sequential above-ground symptoms of pineapple wilt as follows.

1. The preliminary reddening of the leaves

2. A definite colour change from red to pink and with an inward reflexing of the leaf margins (Fig.1.1)

3. Loss of rigidity of affected leaves (wilted appearance).

Then either,

4a. A recovery state in which the centre of the plant grows out with fresh, apparently normal leaves. This stage is called "terminal wilt, normal leaves"

or

4b. According to the Department of Agriculture of Sri Lanka (1993), after loss of rigidity of leaves, leaf tips start to dry followed by death of plants.

Carter first reported the presence of pineapple wilt disease in Sri Lanka (Carter, 1956). According to this report, in Sri Lanka, the mealybug is found attended principally by fire ants (*Monomorium indicum*) often with a huge nest built around mealybug infested plants. In 1993, the Department of Agriculture of Sri Lanka (1993) reported wilt disease as the biggest phytopathological problem of the pineapple industry in Sri Lanka, and it is known to occur in all pineapple growing areas in Sri Lanka. That report also confirmed the continuous association of ants with pineapple wilt.

According to the literature (Rohrbach *et al.*, 1988), in the history of controlling pineapple wilt, the first control measure was direct control of mealybugs by spray application of chemicals. Thus the intent of research in controlling pineapple wilt had been to develop cheap and effective chemical to eliminate mealybugs. With the realisation of an active association of ants with mealybugs, researchers also started paying attention to the control of ants in pineapple fields. In addition to the use of chemicals, physical control methods have been introduced to control ants. The Department of Agriculture of Sri Lanka (Department of Agriculture of Sri Lanka, 1993) has recommended control of both ants and mealybugs simultaneously, in controlling wilt disease.

In Hawaii, currently, ants are controlled by use of mirex and heptachlor (Rohrbach *et al.*, 1988). Since mealybugs are generally controlled by controlling ants, no specific control measures for mealybugs have been implemented. In Sri Lanka, use of prophenopos or prothiopos as a prophylactive measure has been recommended (Department of Agriculture of Sri Lanka, 1993). Moreover, if diseased plants are found in the field, further application of chemicals or rogueing, depending on the extent of disease, is also being practised.

Fig. 1.2 shows a pineapple plantation affected by the pineapple wilt disease. In this figure it is easy to identify rows of affected plants and healthy plants. This is not the spatial pattern one would expect if disease incidence was random. Thus, at least in this case, incidence of the pineapple wilt disease is not random but aggregated. As mentioned earlier, an appropriate statistical analysis of disease incidence data must take account of the spatial pattern of incidence. Thus, spatial pattern analysis is an important aspect of epidemiological studies. Moreover, the use of spatial pattern in statistical analysis is required for the proper assessment of disease control strategies.

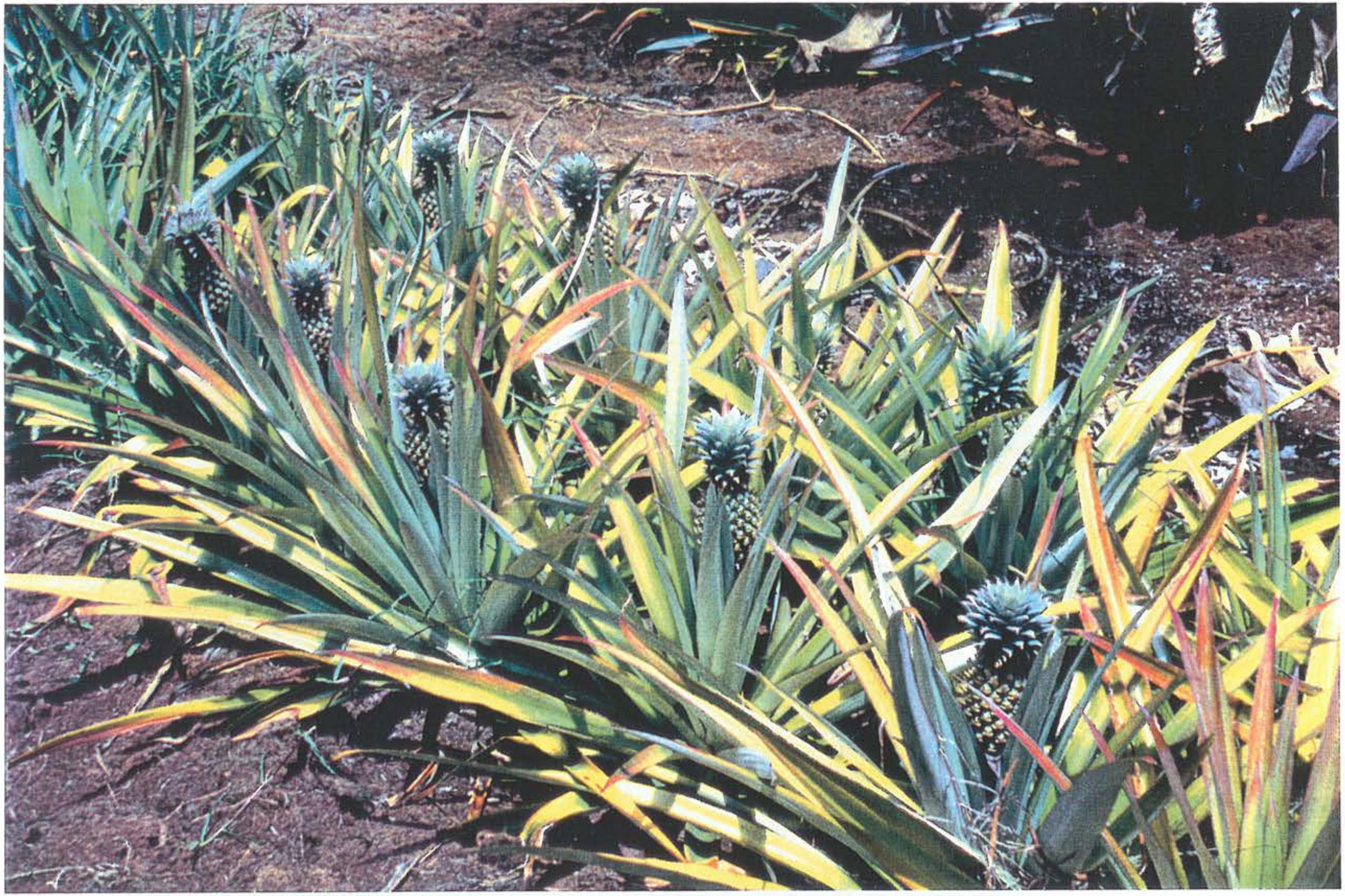


Fig 1.1 Wilt disease symptoms on pineapple plants.

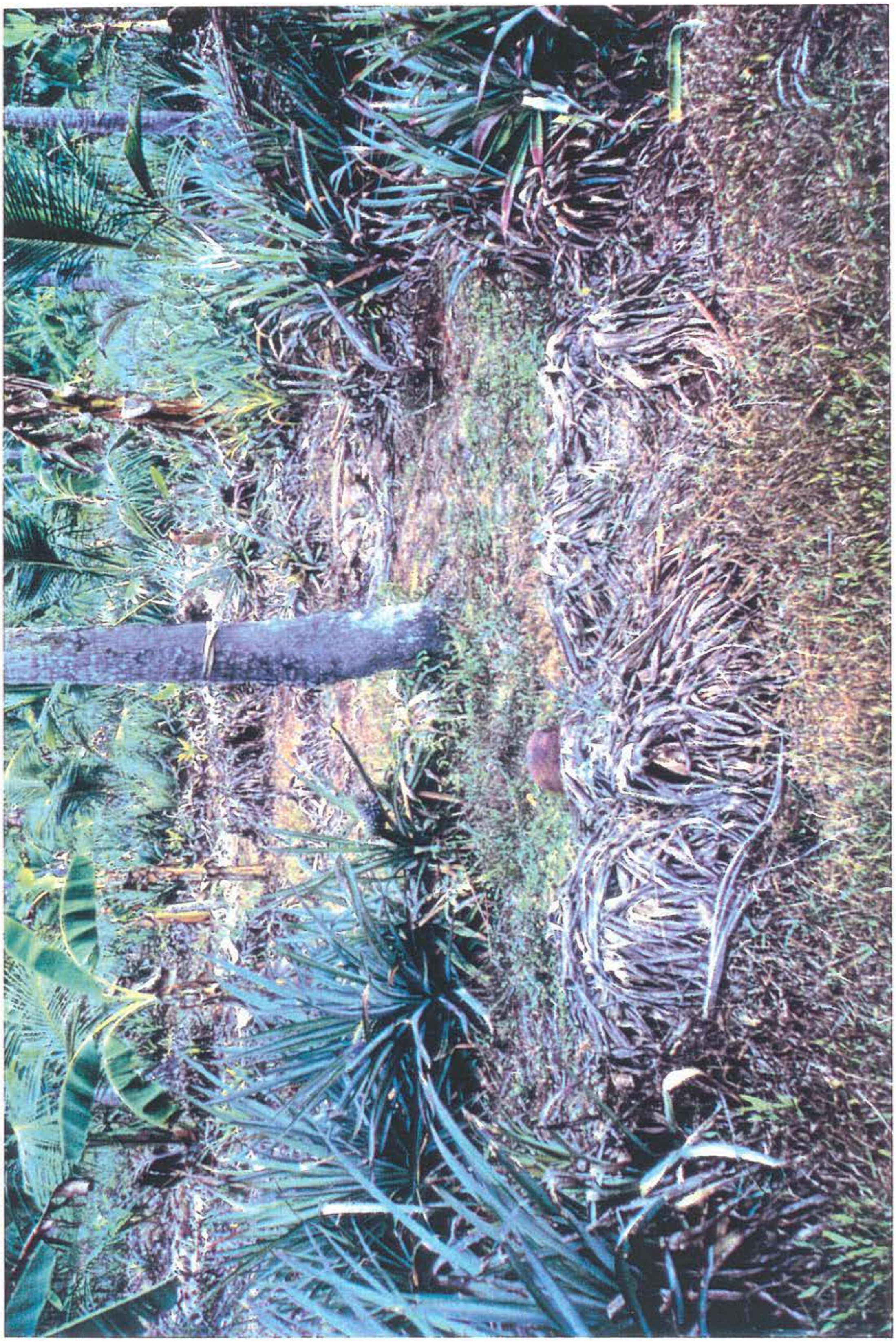


Fig. 1.2 Nature of wilt disease spread on a pineapple plantation.

1.2 Objectives of the study

This study has several objectives.

1) To illustrate possible techniques that can be used to analyse disease incidence data, specifically in the presence of extra-binomial variation. We introduce some newly applied techniques. In addition we generalise some techniques available for analysis of incidence data to facilitate wider application, specifically in the analysis of disease incidence data.

2) To identify criteria for choosing appropriate statistical procedures from among the possible procedures that can be used.

3) As an application of these techniques, detailed analysis of new epidemiological data for pineapple wilt disease is presented.

4) To further develop the use of two-dimensional distance class (2DCLASS) analysis as a tool for spatial pattern analysis in plant disease epidemiology (Nelson *et al.*, 1992).

A number of computer software programs were used in this study to implement the procedures discussed. Chapter 2 gives some uses of these software programs, and the applications of these software programs are described along with the procedures in chapter 3. Chapter 4 gives the details about the survey data collected on pineapple wilt disease, followed by the statistical analysis of these data. Chapter 5 describes the 2DCLASS analysis of these data, and gives details of the investigation of the effects quadratisation and sample size on the results of 2DCLASS analysis. Finally, chapter 6 gives an overall conclusion.

2. COMPUTER SOFTWARE FOR ANALYSIS OF DISEASE INCIDENCE DATA

Several computer software programs were used to illustrate the various methodologies described in this study. Among them GLIM and SAS are widely being used in statistical analysis and EGRET is less well known. These packages can be used to implement most of the procedures illustrated in this study, but certain other software was used for some procedures. Among these were BBD, 2DCLASS and MATHCAD. A summary of some properties and uses of these packages is given in sections 2.1-2.6, and applications are described along with the statistical analysis procedures in chapter 3. Throughout this study these applications and uses are based on the versions of each packages are as follows.

Version 3.77 of GLIM
Version 0.23.26 of EGRET
Version 1.2 of BBD
Version 1 of 2DCLASS
Version 5.0+ of MATHCAD
Release 6.04 of SAS
Version 4 of Microsoft EXCEL for Windows.

2.1 GLIM

GLIM (Generalised Linear Interactive Modelling) was written to enable fitting of the family of generalised linear models described by Nelder and Wedderburn (1972). GLIM provides powerful statistical modelling facilities which extend beyond standard normal linear models and enable many different types of response variable to be modelled in a consistent way. The command language sets GLIM apart from many other interactive statistical packages and this gives the flexibility of modelling data in different ways.

Instructions are issued to the GLIM system by means of directives. 'Directive' is the term used in GLIM syntax to refer a function command. Each directive consists of a name which begins with the directive symbol $. The format of the input file with the necessary directives are described in GLIM user guide and in Aitken *et al.* (1989). The operational commands necessary to implement procedures on GLIM are described along with the methods in chapter 3.

One important feature of GLIM is the use of 'macros'. A macro is simply a string of characters, usually a set of GLIM directives, which is stored by the program for subsequent use as often as needed. The macro begins with directive $MACRO followed by a user defined macro name. The macro is ended by the directive $ENDMAC and a created macro subsequently called by typing $USE directive followed by the macro name.

A large number of GLIM macros are distributed with the package, and additional macros are published in issues of the GLIM *Newsletter*. This has made GLIM useful for a variety of non-standard statistical analyses.

2.2 EGRET

EGRET (Epidemiological, Graphics, Estimation and Testing program) is a package specifically developed for the analysis of data from epidemiological studies. This package provides the user with menu options. EGRET can easily be used for linear logistic modelling as well as to fit models associated with conditional probability distributions. In fact this is the only package which has the ability of fitting such models as a standard option.

EGRET package consists of two modules. The first module is called DEF and is for data definition. The second module is called PECAN (Parameter Estimation through Conditional Probability Analysis) and the purpose of this part is data analysis. This package is not designed for data management and thus the facilities for data manipulation within the package are extremely limited. Details of the uses of this package are given in EGRET (1990)

2.3 BBD

BBD (Beta-Binomial Distribution Fitting Program) was developed by Madden and Hughes (1994a) to fit the beta-binomial distribution to frequency distributions of disease incidence data. The program is written in Microsoft FORTRAN and compiled by version 5.1 of Microsoft's Professional Development System (Microsoft Corporation, One Microsoft Way, Redmond, WA 98052-6399). To run the program this requires DOS 3.2 or higher. This program can be used to obtain the expected frequencies of binomial and beta-binomial distributions unless sample unit sizes vary (where it is impossible to define the expected values). The expected frequencies are computed based on the Maximum Likelihood Estimates (MLEs) and MLEs of the parameters and standard errors of the estimates are calculated using a damped Newton-Raphson Technique. Description of the software is given in Madden and Hughes (1994a) and the operational commands are given in the BBD manual (Madden and Hughes, 1994b). This software program does not need any operational commands other than specifying the input file and the output file. The required format of the input file is described in the user guide (Madden and Hughes, 1994b).

2.4 2DCLASS

2DCLASS (Two-Dimensional Distance Class Analysis Software) was developed by Nelson *et al.* (1992) to perform Gray's (Gray *et al.*, 1986) 2DCLASS analysis. This program is an adaptation of Gray's program (Gray *et al.*, 1986) and is written and compiled in the Microsoft Quick BASIC language (Microsoft Quick BASIC Version 4.5). To run the program it requires DOS 2.0 or higher. As in BBD this software program also does not need any operational commands other than specifying input and output files. Description of the software and the required format of the input file is given in Nelson *et al.* (1992).

2.5 MATHCAD

The MATHCAD software program is widely used for variety of mathematical computations. One important facility of version 5.0+ is that it can perform numerical integrations. In this study this software was used to perform a numerical integration associated with the logistic-normal probability density function (section 3.1.3 and 3.4.4). This is a 'Windows environment software program' and the details of this software package is described in the user guide (MATHCAD 5.0+, 1994).

2.6 SAS

The software program SAS (Statistical Analysis System) is extensively being used for statistical analysis in variety of study areas. As in GLIM, SAS also allows the use of macros which consist of sequences of SAS statements. But most of the non-standard analyses are programmed using a specific software module, IML (Interactive Matrix Language) included in SAS. Since most of the macros used in this study were written for GLIM, SAS is rarely used in this study. But the possibilities of the use of SAS, with necessary operational commands are discussed under relevant topics in section 3.2. All these operational commands are based on the release 6.04 of SAS. The detail use of this software program is given in the SAS user guide (SAS, 1990).

2.7 Microsoft EXCEL

In addition to the software programs in sections 2.1 to 2.6, the Microsoft EXCEL spreadsheet program was also used in this study. In data collection, automated spread sheets (chapter 4) were prepared using this software program. In section 3.1 this software was used for fitting binomial distributions to example data. Moreover, in section 3.4.4 this program was used to compute the estimates, design effect and the effective sample size, which required for the analysis described in section 3.4.4. Details of this spreadsheet program are given in the manual (Microsoft EXCEL, 1992).

3. DESCRIPTION OF THE STATISTICAL ANALYSIS OF DISEASE INCIDENCE DATA

3.1 Fitting probability distributions to incidence data

3.1.1 The binomial distribution

When the observation made is number of successes out of a total, i.e. when the response variable is the proportion of successes, the most satisfactory approximate distribution for finite populations is often the binomial distribution (Cochran, 1977). If the proportion of successes, π, remains constant, i.e. sampling is done with replacement or the original population is large relative to the sample unit n and each success is independent of other successes in the sample unit, the probability that the sample unit contains y units of successes is

$$P(Y = y) = \binom{n}{y} \pi^{y} (1 - \pi)^{n-y} \tag{3.1}$$

For instance, if the location of a diseased plant is independent of the location of other diseased plants and there is a constant probability, π, of a plant being diseased, then the probability of observing a particular number of diseased plants, Y, out of n in a sample unit, a quadrat say, has the binomial distribution and could be expressed in the form (3.1). In this case $P(.)$ represents probability and y takes the values $0,1,2,..,n$. The mean and variance of Y are then $n\pi$ and $n\pi(1-\pi)$, respectively. A χ^2 test goodness-of-fit provides a quantitative test of discrepancies between the observed and the expected frequencies in such a comparison. This gives a test of the null hypothesis: The observations are randomly drawn from a specified binomial distribution. Thus, non-rejection of null hypothesis indicates homogeneity and the rejection of null hypothesis implies non-randomness.

Table 3.1 shows data presented by Bald (1937) for the number of plants infected with tomato spotted wilt virus (TSWV) for the cultivar Early Dwarf Red irrigated by trenches with the expected frequencies based on the binomial distribution. There are 40 sample units with 9 plants per sample (quadrat).

The expected binomial frequencies were calculated from equation (3.1) using the software program Microsoft Excel. In Microsoft EXCEL, the command FREQUENCY can produce the observed frequency distribution and the command

BINDIST can produce the expected binomial probabilities from which expected frequency distribution can be derived, given the mean of disease incidence, π.

Table 3.1 Disease frequency of Bald's data for the cultivar Early Dwarf Red irrigated by Trenches.

Number of diseased plants per quadrat	Observed frequency	Expected binomial frequency
0	0	0.00
1	0	0.01
2	0	0.07
3	0	0.43
4	3	1.82
5	3	5.16
6	12	9.73
7	12	11.81
8	7	8.35
9	3	2.63
Total	40	40.01

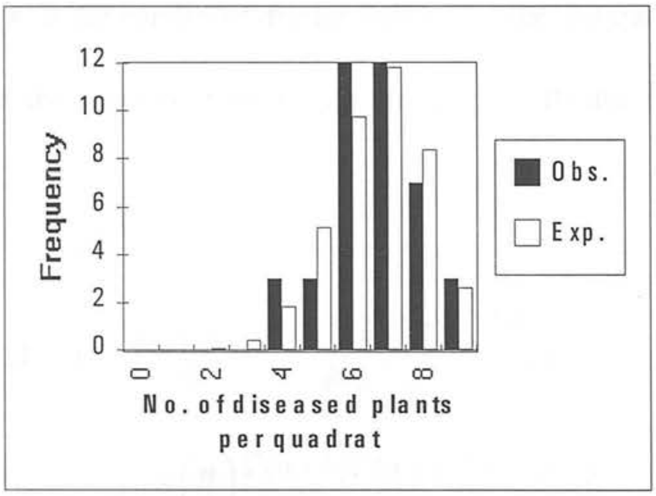


Fig. 3.1 Frequency distribution of tomato plants infected by TSWV in 40 quadrats with 9 plants per quadrat. Observed (Obs.) and expected frequencies (Exp.) based on binomial distribution are shown. Estimated parameters and goodness-of-fit test statistic are given in the text.

Fig. 3.1 shows the observed frequencies along with expected binomial frequencies. It clearly illustrates that observed frequencies agree with the expected binomial frequencies. For the binomial distribution the estimate of π is 0.74 and the goodness-of-fit χ^2=2.97, with 2 degrees of freedom (P=0.23). In applying the χ^2 test of goodness-of-fit between observed and expected frequencies classes 0-5 and 8-9 were pooled to make the smallest expectation greater than 5 (Snedecor and Cochran, 1989). The probability value, which is much greater than 0.05, indicates consistency of observed values with expected values, implying random pattern of diseased plants in the field.

3.1.2 The beta-binomial distribution

As observed in the previous section, the binomial distribution provides expected frequencies based on the supposition of spatial randomness. However, when the occurrence of disease incidence is not random, the binomial distribution cannot adequately describe the observed frequencies. Williams (1975), Crowder (1978), Paul (1982), pointed out the use of beta-binomial distribution to describe non-random incidence frequencies. Shiyomi and Takai (1979), Qu *et al.* (1993), and Hughes and Madden (1993) reported the beta-binomial distribution to be appropriate to describe aggregated disease frequencies.

The beta-binomial distribution is formed by compounding the binomial distribution with a beta density function (Skellam, 1948). Suppose there are j binomial observations made under uniform condition. In principle the exact distribution of $y = \sum_j r_j$, where r_j is the number of diseased plants for the quadrat of size n_j, can be determined under this random compounding model. In particular if u $(0 \le u \le 1)$ has the beta density

$$u^{a-1}(1-u)^{b-1} / B(a,b),$$

then

$$P(Y=y) = \binom{n}{y} \int_0^1 \frac{u^y(1-u)^{n-y}u^{a-1}(1-u)^{b-1}}{B(a,b)} du$$

$$= \binom{n}{y} \frac{\Gamma(a+b)\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(a)\Gamma(b)\Gamma(a+b+n)} \tag{3.2}$$

which is called the beta-binomial distribution.

Note that $B(a,b) = \int_0^1 u^{a-1}(1-u)^{b-1} du = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$, where

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du,$$

$$\Gamma(b) = \int_0^\infty u^{b-1} e^{-u} du$$

and

$$\Gamma(a+b) = \int_0^\infty u^{(a+b-1)} e^{-u} du.$$

Basically, the beta-binomial distribution is an extension of the binomial distribution which assumes that the probability of success varies between samples according to a two parameter beta distribution.

With a convenient reparameterization,

$$p = a(a+b)^{-1}$$

and

$$\theta = (a+b)^{-1},$$

equation (3.2) may be rewritten as (if a and b are integers)

$$P(Y=y) = \binom{n}{y} \frac{\prod_{r=0}^{y-1} (p + r\theta) \prod_{r=0}^{n-y-1} (1 - p + r\theta)}{\prod_{r=0}^{n-1} (1 + r\theta)}.$$

Here p is the mean disease incidence of the binomial parameter π and θ is a measure of the variation in π. The case of pure binomial variation corresponds to $\theta = 0$ whereas the original a, b parameterization, it corresponds to infinite parameter values. The parameters p, θ are more stable than a, b as described by Ross (1970). According to Williams (1975) stability could be further enhanced by introducing a parameter such that $\tau = \theta(1+\theta)^{-1}$ rather than θ itself but the difference between the p, θ and p, τ parameterization will be of practical significance only if θ is not small. Theoretically, p is the expectation of the underlying beta distribution, and given p, the parameter θ determines the shape of the distribution. In practice the parameter θ takes account of overdispersion or aggregation. In the literature θ is often referred to as an aggregation parameter and sometimes as random effect parameter (EGRET, 1991). The beta-binomial distribution can facilitate a wide variety of shapes such as J, L or U, depending on values of a and b (Mendenhall *et al.*, 1990). The derivation of the beta-binomial is analogous to the derivation of negative binomial distribution by compounding the Poisson with a gamma distribution (Moran, 1968). The Poisson distribution is a special case of binomial distribution when n is large and p is small (Patil and Joshi, 1968). The negative binomial is in fact a special case of beta-binomial distribution when n and $a+b$ are large (Patil and Joshi, 1968; Skellam, 1948).

For the reparameterised form (p, θ) mean and variance of beta-binomial distribution are np and $np(1-p)(1+n\theta)(1-\theta)^{-1}$ respectively (Skellam, 1948). When $\theta > 0$, the variance of the beta-binomial distribution is larger than that of binomial with the same mean and when $\theta = 0$, the beta-binomial distribution is exactly the same as the binomial distribution. The parameters can be estimated by moments. If the observed mean and the variance of y are $\overline{y}$ and s^2 respectively, then the moments estimates are (Griffith, 1983)

$$p = \overline{y} / n$$

$$\theta = \frac{s^2 - np(1-p)}{n^2 p(1-p) - s^2}$$

and maximum likelihood estimates of p and θ can be obtained using Smith's algorithm using moments estimates as the starting point for the iterative estimating procedure (Smith, 1983).

The software program BBD uses Smith's algorithm to calculate the maximum likelihood estimate of the beta-binomial distribution. Moreover, it computes the expected beta-binomial frequencies based on these maximum likelihood estimates. In using BBD there is no specific command necessary to specify, other than specifying input file (data file) and the output file. The format of input file required is described in the BBD operating manual.

Table 3.2 shows the data reported by Snedecor and Cochran (1989), in a quantitative study of tomato spotted wilt virus (TSWV). The observed number of diseased tomato plants out of 9 (quadrat size of 9) for 160 quadrats was compared with the expected frequencies based on the binomial distribution and the beta-binomial distribution.

Table 3.2 Observed and expected (binomial and beta-binomial) frequencies for the tomato TSWV data reported by Snedecor and Cochran (1989).

Number of diseased plants per quadrat	Observed frequency	Expected binomial frequency	Expected beta-binomial frequency
0	36	26.45	36.32
1	48	52.70	47.81
2	38	46.67	37.69
3	23	24.11	22.21
4	10	8.00	10.45
5	3	1.77	3.98
6	1	0.25	1.21
7	1	0.03	0.28
8	0	0.00	0.04
9	0	0.00	0.00
Total	160	159.98	160.26

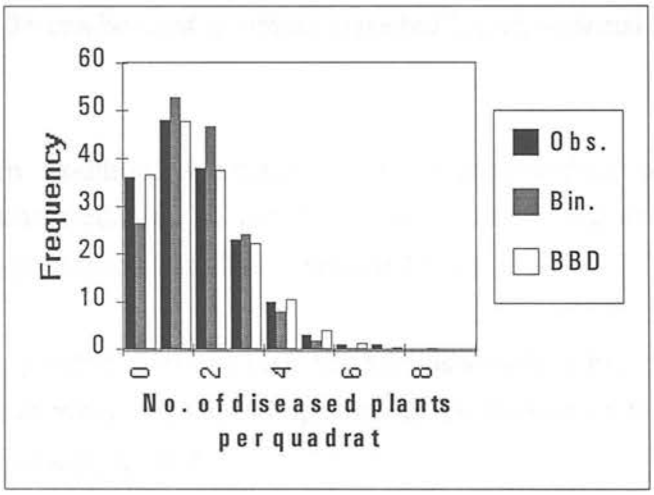


Fig. 3.2 Frequency distribution of tomato plants infected by TSWV in 160 quadrats with 9 plants per quadrat. Observed (Obs.) and expected frequencies for the binomial (Bin.) and beta-binomial (BBD) distributions are shown. Estimated parameters and goodness-of-fit test statistic are given in the text.

Using BBD, for the data in Table 3.2, estimate of p is 0.181 (standard error [SE]=0.0119) and $\hat{\theta}= 0.053$ (SE=0.0204). From Fig. 3.2, it is very clear that the expected beta-binomial frequencies are much closer to the observed frequencies than the expected binomial frequencies. For the binomial, goodness-of-fit χ^2 is 7.968, with 3 degrees of freedom ($P = 0.047$) suggesting a deviation from randomness. For the beta-binomial the goodness-of-fit χ^2 is 0.10, with 3 degrees of freedom (P >0.99) and giving a far better description of the observed frequency distribution. In applying the χ^2 test of goodness-of-fit between observed and expected binomial frequency classes 4-9 were pooled to make the smallest expectation greater than 5. Correspondingly, classes 5-9 were pooled when applying χ^2 test of goodness-of-fit to the beta-binomial distribution fitting.

3.1.3 The Logistic-normal binomial distribution

Another conditional distribution available for the description of aggregated patterns is the logistic-normal binomial distribution. In this distribution the logit of the variability in π is assumed to have normal distribution with a constant variance (Pierce and Sands, 1975). The probability density function of the logistic-normal binomial distribution is more complicated than that of beta-binomial distribution and thus details of this distribution are delayed until section 3.4.2.

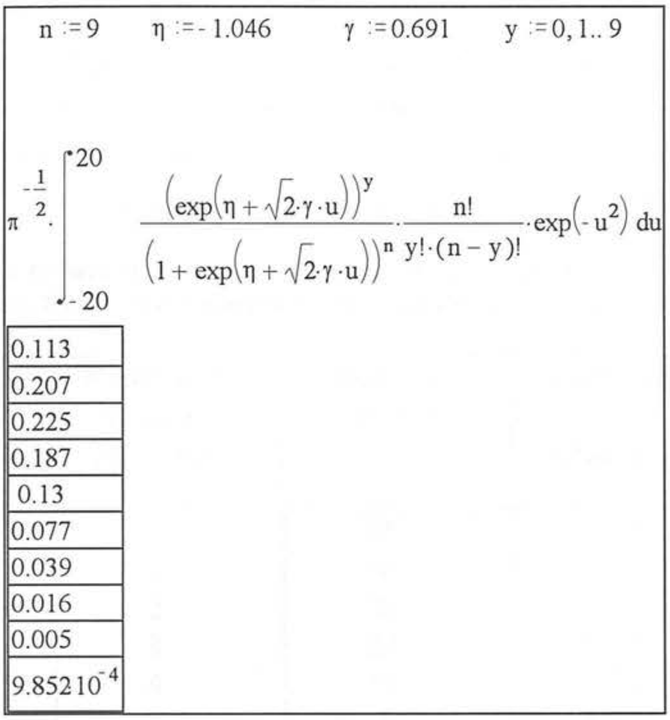
There is no particular software program which can be used to calculate the logistic-normal binomial expected frequencies. However, EGRET in association with MATHCAD 5.0+ can be used to obtain expected logistic-normal binomial frequencies as follows.

First, maximum likelihood estimates of the logistic-normal binomial distribution parameters (linear predictor, η, and the parameter associated with the random effect component, γ) (see section 3.4.2) are obtained using EGRET.

Then, define n, η, γ and y in MATHCAD 5.0+ as shown in Fig. 3.3. Note that n, η, γ take single values while y take multiple values as number of frequency classes, 0-9 say, if the quadrat size, n, is 9.

Thereafter, define the operational probability density function of logistic-normal binomial distribution function (3.3) (Collett, 1991) as in the Fig. 3.3

$$P(Y=y|n,\eta,\gamma)=\pi^{-1/2}\int_{-\infty}^{\infty}\binom{n}{y}\frac{[\exp(\eta+\sqrt{2}\gamma u)]^{y}}{[1+\exp(\eta+\sqrt{2}\gamma u)]^{n}}\exp\{-u^{2}\}du, \quad (3.3)$$



$n := 9$ $\eta := -1.046$ $\gamma := 0.691$ $y := 0, 1..9$

$$\pi^{-\frac{1}{2}} \cdot \int_{-20}^{20} \frac{\left(\exp\left(\eta+\sqrt{2}\cdot\gamma\cdot u\right)\right)^{y}}{\left(1+\exp\left(\eta+\sqrt{2}\cdot\gamma\cdot u\right)\right)^{n}} \cdot \frac{n!}{y!\cdot(n-y)!} \cdot \exp\left(-u^{2}\right)\,du$$

0.113
0.207
0.225
0.187
0.13
0.077
0.039
0.016
0.005
$9.852\,10^{-4}$

Fig. 3.3 MATHCAD 'Windows' screen view for obtaining expected logistic-normal binomial probabilities

The integral limits beyond the range ±20 do not make any difference in the computation of expected probabilities. Thus $-\infty$and ∞in the equation can be replaced as -20 and 20 respectively, as in the Fig. 3.3.

After that press the '=' sign. The outcome should appear as Fig. 3.3. The resulting column of values provides the expected probabilities for number of diseased plants per quadrat (0 to n) respectively. Finally, by multiplying these probabilities by the total number of quadrats, the corresponding expected quadrat frequencies can be obtained.

When the logistic-normal binomial distribution is fitted to the observed data in Table 3.2 using EGRET and MATHCAD 5.0+, the expected frequencies given in Table 3.3 are found. The estimate of η and the aggregation parameter γ (see section 3.4.2) were -1.617 (SE=0.0932) and 0.5989 (SE=0.1180) respectively. From Table 3.3 and Fig. 3.4, it is very clear that expected frequencies from the logistic-normal binomial (LNB) distribution are much closer to observed frequencies than the expected binomial frequencies. The goodness-of-fit χ^2 is 0.10 with 3 degrees of freedom ($P > 0.99$), and this implies logistic-normal binomial distribution can give a far better description of the observed frequencies than can the binomial distribution. Frequency classes 5-9 were pooled in applying the χ^2 goodness-of-fit between observed and expected logistic-normal binomial frequencies. For the data in Table 3.2 there is hardly any improvement of logistic-normal binomial distribution over beta-binomial distribution in describing frequency distributions. Nevertheless, it should be noted that logistic-normal binomial distribution has more liberal mathematical properties (see section 3.4.2) than beta-binomial distribution (Anderson, 1988).

Table 3.3 Observed and expected (binomial and logistic-normal binomial) frequencies for the tomato TSWV data reported by Snedecor and Cochran (1989).

Number of diseased plants per quadrat	Observed frequency	Expected binomial frequency	Expected LNB frequency
0	36	26.45	36.00
1	48	52.70	48.64
2	38	46.67	37.76
3	23	24.11	21.76
4	10	8.00	10.24
5	3	1.77	4.00
6	1	0.25	1.28
7	1	0.03	0.03
8	0	0.00	0.00
9	0	0.00	0.00
Total	160	159.98	159.71

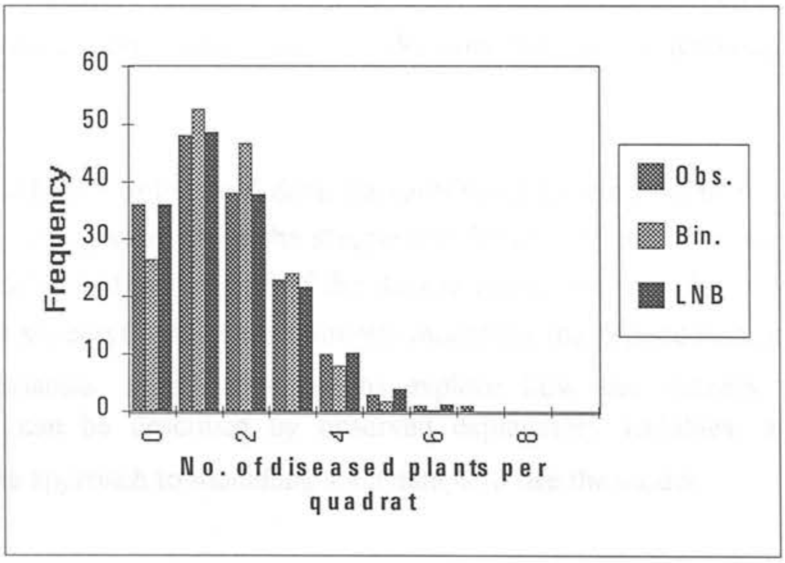


Fig. 3.4 Frequency distribution of tomato plants infected by TSWV in 160 quadrats with 9 plants per quadrat. Observed (Obs.) and expected frequencies for the binomial (Bin.) and Logistic-normal binomial (LNB) distributions are shown. Estimated parameters and goodness-of-fit test statistic are given in the text.

3.2 Fitting models to binomial data

Suppose the observations of the binary response are made as 'success' and 'failures' for each individual in a quadrat size *n*. Then the probability of y (1,2,..,*n*) being a success (diseased) can be estimated using equation (3.1) given π. This is called the distribution of success. For this distribution of success, y is symmetric when $\pi = 0.5$, positively skewed when $\pi < 0.5$ and negatively skewed when $\pi > 0.5$. An important property of binomial distribution is that as n increases, the degree of asymmetry in the distribution decreases, even when π is close to 0 or 1. From the central limit theorem it can be shown that as n increases, the binomial distribution can be approximated by normal distribution. Thus an ordinary *z*-test (Steel and Torrie, 1980) can be performed to make inferences and comparisons about the success probabilities when n is large. The χ^2 test is also often applied to compare differences in success probabilities when the data is arranged in the form of contingency tables (Snedecor and Cochran, 1989). These techniques for analysing incidence data can be useful when the structure of the data is not particularly complicated.

An alternative approach to analysing incidence data is based on the construction of a statistical model to describe the relationship between the observed response and explanatory variables. In other words, the objective of modelling is to derive a mathematical representation of the relationship between an observed response variable and a number explanatory variables, together with a measure of the inherent variability of any such relationships which might eventually helpful to understand certain systems.

Suppose that for binary or binomial data, the response from the i^{th} unit, $i = 1,2..n$, is a proportion $\tilde{\pi}_i = y_i / n_i$, where n_i is the sample size for i^{th} unit and y_i is the number of successes out of n_i in the i^{th} unit (If the data is binary $n_i = 1$ and $y_i = 0$ or 1, i.e. either failure or success). Rather than directly modelling the dependence of $E(y_i)$ on explanatory variables, it is customary to explore how the success probability $\pi_i = E(y_i / n_i)$ can be described by observed explanatory variables, x_j where, $j = 1,2..k$. One approach to modelling such data is to use the model

$$\pi_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki} \qquad (3.4)$$

and apply the method of least squares to obtain values of $\beta_0, \beta_1, \ldots \beta_k$ for which

$$\sum_i \left(\frac{y_i}{n_i} - \pi_i \right)^2 = \sum_i \left(\tilde{\pi}_i - \beta_0 - \beta_1 x_{1i} - \ldots \beta_k x_{ki} \right)^2$$

(where $\tilde{\pi}_i = y_i / n_i$) is minimised.

The first problem with this approach concerns the assumption made about the variance of $\tilde{\pi}_i$. Since y_i actually has a binomial distribution, $\text{Var}(\tilde{\pi}_i) = \pi_i(1 - \pi_i) / n_i$, which varies with π_i (not a constant) even if the binomial denominator, n_i is constant (Collett, 1991). The constant variance is a requirement in the ordinary least squares analysis in order to 'pool' the variance estimates so that *t*-test and confidence intervals can be calculated (Cox and Snell, 1989). According to the literature this is not a serious problem if the π_i lie in the range of 0.25-0.75 (Snedecor and Cochran, 1989). One early approach to overcome this problem was to transform the data in such a way that the resulting data have a constant variance. A variance stabilising transformation often adopted was the arc sine or angular transformation $\left(\sin^{-1} \sqrt{\tilde{\pi}_i} \right)$ (Snedecor and Cochran, 1989). A better procedure which is now being adopted, appropriate whether or not n_i are equal, is to use the method of weighted least squares to minimise the function

$$\sum_i w_i (\tilde{\pi}_i - \pi_i)^2$$

where the weights, w_i are reciprocals of the variance of $\tilde{\pi}_i$ given by $w_i = [\pi_i(1 - \pi_i) / n_i]^{-1}$. An iterative scheme is adopted to obtain weighted parameter estimates. Starting with the initial estimates of the π_i, $\hat{\pi}_{i0} = y_i / n_i$ and hence $w_{i0} = [\hat{\pi}_{i0}(1 - \hat{\pi}_{i0}) / n_i]^{-1}$, the iteration is continued until the parameter estimates converge to $\hat{\beta}_j (j = 1, 2..k)$ and the fitted probabilities converge to π_i. In the literature this is referred to as the iteratively weighted least squares method, although it is equivalent to maximum likelihood estimation (Aitkin *et al.*, 1989; Collett, 1991).

The second problem is that since the data are not normally distributed, the elegant distribution theory associated with fitting linear models to normal data cannot be applied. When y_i's are not normally distributed, no method of estimation that is linear in the y_i's will in general be fully efficient (Cox and Snell, 1989).

The third problem is that unless restrictions are imposed on β_j the estimates of β_j may lie in the range $(-\infty, \infty)$. Since the fitted probabilities are obtained from (3.4) there is no guarantee that the fitted values will lie in the interval $(0,1)$ if no restrictions are imposed on β_j.Then, the expression π_i as a linear combination would be inconsistent with the laws of probability. The simple and effective way of avoiding this difficulty is to use a transformation $g(\pi)$, the link function, that maps the unit interval onto the whole real line $(-\infty, \infty)$. This remedy leads to instances of generalised linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983; 1989) in which systematic part is

$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki}; i = 1,2..n. \tag{3.5}$$

The logistic, probit and complementary log-log are three transformations that are commonly used as link functions with binomial data. Of the three transformations, the logistic transformation is the most widely used. Cox and Snell (1989) compare some transformation methods and explain the usefulness of probit and logistic transformations over the angular transformation. According to them, when the probability is outside the range 0.1-0.9, the finite limits on the angular transformation usually seriously restrict its usefulness.

3.2.1 Fitting the linear logistic model

For n binomial observations of the form $\tilde{\pi}_i = y_i / n_i$, the linear logistic model, which is a generalised linear model for the dependence of π_i on the values of the k explanatory variables, $x_{1i}, x_{2i}, \ldots x_{ki}$, is given by

$$\text{logit}(\pi_i) = \log[\pi_i / (1 - \pi_i)] = \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots \beta_k x_{ki}. \tag{3.6}$$

Therefore,

$$\pi_i = \exp(\eta_i) / [1 + \exp(\eta_i)].$$

Here the $g(\pi)$ is $\log[\pi / (1 - \pi)]$ which is also called logit or logistic transformation and log is the natural logarithm, base 2.71828..

The maximum likelihood estimates of the parameters $\boldsymbol{\beta}$ are the values of the parameters that maximise the fitted log likelihood function,

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log[\pi_i / (1-\pi_i)] + n_i \log(1-\pi_i). \tag{3.7}$$

The derivative of this log likelihood function with respect to π_i is,

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - n_i \pi_i}{\pi_i (1-\pi_i)}.$$

Using the chain rule, the derivative with respect to β_j is

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - n_i \pi_i}{\pi_i (1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_j}.$$

In the case of generalised linear models, it is usually express $\frac{\partial \pi_i}{\partial \beta_j}$ as a product

$$\frac{\partial \pi_i}{\partial \beta_j} = \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{d\pi_i}{d\eta_i} x_{ij}.$$

Thus,

$$\frac{\partial l}{\partial \beta_j} = \sum_i \frac{y_i - n_i \pi_i}{\pi_i (1-\pi_i)} \frac{d\pi_i}{d\eta_i} x_{ij.}$$

For the linear logistic model, this may be expressed as,

$$\partial l / \partial \beta_j = \sum_i [y_i - n_i \pi_i] x_{ij}, \tag{3.8}$$

and in matrix notation,

$$\partial l / \partial \boldsymbol{\beta} = \mathbf{X}^T (y - \boldsymbol{\pi}), \tag{3.9}$$

where $\mathbf{X}$ is the model matrix of order $n \times p$, y is the vector of y_i and $\boldsymbol{\pi}$ is the vector of π_i. Then the Fisher information for $\boldsymbol{\beta}$ is

$$-E\left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_r} \right) = \sum_i n_i [e^{\eta_i} / (1+e^{\eta_i})^2] x_{ij} x_{ir} \tag{3.10}$$

$$=\sum_i n_i\pi_i(1-\pi_i)x_{ij}x_{ir}$$

$$=\{\ \mathbf{X}^T\mathbf{W}\mathbf{X}\}_{jr} \tag{3.11}$$

where $\mathbf{W}=\operatorname{diag}\{w_i = n_i\pi_i(1-\pi_i)\}$. The likelihood equation then amounts to equating the sufficient statistic, $\mathbf{X}^T\mathbf{Y}$, to its expectation as a function of $\boldsymbol{\beta}$. Given initial estimates $\hat{\boldsymbol{\beta}}_0$, vectors $\hat{\boldsymbol{\pi}}_0$ and $\hat{\boldsymbol{\eta}}_0$ are computed. Using these computed values an adjusted dependent variable, $\mathbf{Z}$, is defined with components

$$z_i = \hat{\eta}_i + \frac{y_i - n_i\hat{\pi}_i}{n_i}\frac{d\eta_i}{d\pi_i}, \tag{3.12}$$

all quantities being computed at the initial estimate $\boldsymbol{\beta}_0$. Maximum likelihood estimates satisfy the equation (McCullagh and Nelder, 1989)

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{W}\mathbf{Z},$$

which can be solved iteratively using standard least square methods. The revised estimate is

$$\hat{\beta}_1 = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Z}$$

where all quantities appearing on the right are computed using the initial estimate. Once $\hat{\boldsymbol{\beta}}$ has been obtained, the estimated value of the systematic component of the model is

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots \hat{\beta}_k x_{ki}, \tag{3.13}$$

which is also called the 'linear predictor'. Thus the fitted probabilities $\hat{\pi}_i$ can be obtained as

$$\hat{\pi}_i = \exp(\hat{\eta}_i)/[1+\exp(\hat{\eta}_i)]. \tag{3.14}$$

Under any given model, H_0, with fitted probabilities, $\hat{\pi}_i$, the likelihood can be obtained using (3.7). The maximum achievable log likelihood is attained at the point $\tilde{\pi}_i = y_i / n_i$, which does not usually occur in the model space under H_0. The residual deviance is defined as twice the difference between the maximum achievable log likelihood and that attained under the fitted model. The deviance function is therefore

$$D(\mathbf{y};\hat{\pi}) = 2l(\tilde{\pi},\mathbf{y}) - 2l(\hat{\pi},\mathbf{y}) \qquad (3.15)$$

$$= 2\sum\nolimits_i \{y_i \log(y_i / \hat{y}_i) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\}.$$

where $\hat{y}_i = n_i\hat{\pi}_i$. Since $D(\mathbf{y};\hat{\pi})$ is asymptotically distributed as χ^2_{n-p}, where p is number of fitted parameters under H_0, $D(\mathbf{y};\pi)$ is used as a goodness-of-fit statistic for testing the adequacy of the fitted model.

Table 3.4 shows the data presented by Bald (1937) for his 2 factor factorial experiment of 2 cultivars (Burwood Prize and Early Dwarf Red) and 2 irrigation methods (by overhead spray and by trenches). Part of the data has already been presented in Table 3.1.

Table 3.4. Bald's data for his two factor factorial experiment of 2 cultivars and two irrigation methods

Number of diseased plants per quadrat	Cv. Burwood Prize		Cv. Early Dwarf Red	
	Overhead spray	Trench	Overhead spray	Trench
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	2	0
4	0	0	6	3
5	0	0	13	3
6	3	0	6	12
7	7	6	7	12
8	20	12	6	7
9	10	22	0	3
Total	40	40	40	40

If a linear logistic model is considered, including both main effects and interaction effect, such a model can be written in subscript notation, for the observed probability for the k^{th} quadrat of the i^{th} level of cultivar (A) and the j^{th} level of irrigation method (B), π_{ijk},

$$\text{logit}(\pi_{ijk}) = \eta_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{3.16}$$

where $i = 1,2; j = 1,2; k = 1,2..40$. The parameter μ represents the overall mean, α_i and β_j represent fixed main effects of cultivar and irrigation method, respectively, and $(\alpha\beta)_{ij}$ represents the interaction between cultivar and irrigation method. The residuals ε_{ijk} are assumed to have mean 0 and a constant variance. Thus 3.16 may be written

$$\hat{\eta}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + (\widehat{\alpha\beta})_{ij} \ . \tag{3.17}$$

GLIM can be implemented to fit the linear logistic model. The directives $YVAR followed by the response variable identifier, $ERR B followed by the binomial denominator identifier and $FIT followed by the terms to be included in the model, fit the target linear logistic model in GLIM.

If EGRET is implemented, using DEF module, first the data file (which should be in ASCII form [EGRET, 1990]) should be specified. Then the format of the data file should be defined as explained in the EGRET user guide. After that variables in the data file should be defined. Finally, default analysis model and associated variables such as outcome variable and group size (binomial denominator) variables should be specified. When specifying the default analysis model, for instance, 'logistic regression' should be selected if the required model to be fitted is a linear logistic model. Then in the PECAN module, selecting 'auXcmd', factors (if there are any) can be specified. After that, selecting 'New' and specifying terms (variables or factors) to be included in the model, followed by selecting the 'Fit' fits the intended model (User guide may be consulted for details).

Table 3.5 shows the resulting deviances and corresponding degrees of freedom with different linear logistic models (with different components included in the model) for the data in the Table 3.4 using GLIM. (EGRET also produces exactly same results). The deviance difference (Δdeviance) (Table 3.5) which also has asymptotic χ^2 distribution can be used as a guideline whether to include a particular component into the model.

Table 3.5 Analysis of deviance table for Bald's data

Model	Deviance	d.f.	Δ deviance	Δ d.f.
mean only	294.60	159		
A only	181.37	158	113.23	1
B only	280.61	158	13.99	1
A+B	166.24	157	15.10 (after A)	1
			114.37 (after B)	1
A*B	166.01	156	0.23	1

The change in the deviance on adding the interaction term into the model is 0.23 on 1 degree of freedom, and $P(\chi^2 > 0.23)$ with 1 degree of freedom is 0.632, providing no evidence for the interaction effect. The deviances for main effects were significant $(P < 0.01)$ and therefore we can conclude that an adequate model comprises both main effects only. The residual deviance of 166.24 with 157 degrees of freedom $(P = 0.29)$ also indicates that the fitted model adequately describe the data and that the linear logistic model has provided the basis for valid tests of significance.

If SAS is implemented, procedures PROC LOGISTIC, PROC PROBIT with the option D=LOGISTIC or PROC CATMOD can be used to fit linear logistic models. For instance, Table 3.6 shows resulting log likelihoods for possible models when PROC PROBIT with the option D=LOGISTIC is executed for the data in the Table 3.4. Clearly, in the Table 3.6, the Δ deviance values, on which the 'significance tests' are based, closely resemble to those in Table 3.5.

Table 3.6 Analysis of log likelihood table for Bald's data

Model	Log likelihood	Δ deviance	Δ d.f.
A only	-670.844		
B only	-720.460		
A+B	-663.276	15.14 (after A)	1
		114.37 (after B)	1
A*B	-663.163	0.23	1

3.2.2 Parameterisation in the linear logistic model

By analogy with the ANOVA, a linear model for η_{ij} can be formulated as

$$\hat{\eta}_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

where

μ = intercept

α_i = effect of i^{th} level of A

β_j = effect of j^{th} level of B

$(\alpha\beta)_{ij}$= interaction of i^{th} level of A and j^{th} level of B

with symmetric constraints,

$$\sum_i \alpha_i = 0, \sum_j \beta_j = 0, \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0.$$

But GLIM uses the asymmetric "corner-point" constraints

$$\alpha_1 = 0, \beta_1 = 0, (\alpha\beta)_{i1} = 0 \ \forall_i, (\alpha\beta)_{1j} = 0 \ \forall_j .$$

Thus the parameters are

	j=1	j=2
i=1	μ	$\mu+\beta_2$
i=2	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_2+(\alpha\beta)_{22}$

Correspondingly, in the GLIM output, the parameters written as

	j=1	j=2
i=1	1	B(2)
i=2	A(2)	A(2).B(2).

For instance after fitting the full model for the data in Table 3.3 GLIM output is as follows.

```
     estimate    s.e.     parameter

1     1.998     0.1625     1
2    -1.451     0.1959     A(2)
3     0.6414    0.2664     B(2)
4    -0.1477    0.3120     A(2).B(2)
```

Thus the estimates of the equation (3.17) may be obtained as,

$$\hat{\eta}_{11} = 1$$

$$\hat{\eta}_{21} = 1 + A(2)$$

$$\hat{\eta}_{12} = 1 + B(2)$$

$$\hat{\eta}_{22} = 1 + A(2) + B(2) + A(2).B(2),$$

of which values (subjected to some rounding errors) are given in Table 3.7. The parameters and the standard errors in the GLIM output may be explained as follows.

parameter(GLIM)	actual estimate	actual S.E.
1	$\hat{\eta}_{11}$	$s.e.(\hat{\eta}_{11})$
A(2)	$\hat{\eta}_{21} - \hat{\eta}_{11}$	$s.e.(\hat{\eta}_{21} - \hat{\eta}_{11})$
B(2)	$\hat{\eta}_{12} - \hat{\eta}_{11}$	$s.e.(\hat{\eta}_{12} - \hat{\eta}_{11})$
A(2).B(2)	$\hat{\eta}_{22} - (\hat{\eta}_{21} - \hat{\eta}_{11})$ $-(\hat{\eta}_{12} - \hat{\eta}_{11}) - \hat{\eta}_{11}$	$s.e.[\hat{\eta}_{22} - (\hat{\eta}_{21} - \hat{\eta}_{11})$ $-(\hat{\eta}_{12} - \hat{\eta}_{11}) - \hat{\eta}_{11}]$

Therefore significant tests for the above estimates can be made directly from the GLIM output. For instance test for $\pi_{11} - \pi_{21}$ is equivalent to,

$$\hat{t} = \frac{\hat{\eta}_{21} - \hat{\eta}_{11}}{s.e.(\hat{\eta}_{21} - \hat{\eta}_{11})} = \frac{\text{A(2)}}{s.e.[\text{A(2)}]},$$

and numerically above $\hat{t}$ statistic is -7.41 and it gives ($P < 0.001$) significance. Note that $\hat{t}$ is compared with the standard normal distribution (GLIM, 1985; Collett, 1991). However, confidence intervals for each η_{ij} and correspondingly π_{ij} cannot be derived from the above GLIM output. Nevertheless GLIM directives $EXTRACT %VL and $PRINT %LP %VL produce the estimates and variances of estimates for each η_{ij}. The standard error for each $\hat{\eta}_{ij}$ can be obtained by taking the square root of each variance. For the data in Table 3.4, after fitting the model with all main effects and interactions, and then above directives, resulted $\hat{\eta}_{ij}$ and corresponding standard errors are given shown in Table 3.7.

Table 3.7 Parameter estimates for the linear logistic model (3.17)

cultivar	Irr. Method	$\hat{\eta}$	S.E.
Burwood Prize	Overhead spray	1.998	0.1623
Burwood Prize	Trenches	2.639	0.2112
Early Dwarf Red	Overhead spray	0.5465	0.1095
Early Dwarf Red	Trenches	1.040	0.1200

These estimates could be used to derive confidence intervals for actual proportions as follows.

The confidence interval for η_{ij} is

$$\hat{\eta}_{ij} \pm 1.96 \times s.e.\,\hat{\eta}_{ij}$$

where $s.e.\,\hat{\eta}_{ij}$ are from Table 3.7.

From equation 3.14,

$$\hat{\pi}_{ij} = \frac{e^{\hat{\eta}_{ij}}}{1+e^{\hat{\eta}_{ij}}}.$$

Therefore the confidence interval for π_{ij} is

$$\frac{e^{\hat{\eta}_{ij}-1.96\times s.e.\hat{\eta}_{ij}}}{1+e^{\hat{\eta}_{ij}-1.96\times s.e.\hat{\eta}_{ij}}} < \pi_{ij} < \frac{e^{\hat{\eta}_{ij}+1.96\times s.e.\hat{\eta}_{ij}}}{1+e^{\hat{\eta}_{ij}+1.96\times s.e.\hat{\eta}_{ij}}}.$$

3.3 Replication in epidemiological studies

In an experimental design, to test all main effects and interactions it is necessary to have replicates. These replicates, other than providing the basis for an estimate of error, generate the residual degrees of freedom necessary to test all main effects and interactions. Moreover, random allocation of treatments into replicates makes this estimate of error unbiased. In ordinary ANOVA, the errors are assumed to be normally distributed with mean zero and a constant variance which does not depend on the treatment. This leads to an estimate of a common variance and it is estimated as the unexplained variability of an experiment. Then all effects are tested against the residual, especially when the levels of the factors are fixed, not random. In contrast, if the observations are binary, the variance is related to the mean and thus the residual is not simply the unbiased estimate of unexplained variability. This makes any comparison against residual invalid. Therefore, when the data are binary, the appropriate testing procedure is to compare deviance change with the corresponding change of degrees of freedom as a guideline for whether to include a particular component into the model. The residual deviance is then used to test the adequacy of the fitted model.

Epidemiological studies in general do not have proper experimental designs. Analysis of disease incidence usually start with distribution fittings. In the distribution fitting, usually a field plot of a particular treatment combination is divided in to subplots and distributions are fitted to the subplot frequencies of number of diseased plants per subplot (quadrat). Since these subplots do not have a random allocation they do not provide a basis for an estimate of error even if one assumes normal distribution for the response variate.

When the data are binary, although the subplots do not provide any basis for test for main effects and interactions (between plot analysis), they can be used as a basis for within plot (intra plot) analysis. Thus in statistical analysis of the epidemiological studies these subplots provide the necessary degrees of freedom for testing spatial randomness. A good illustration of this aspect is given in Anderson (1988).

3.4 Modelling overdispersion in disease incidence data

When overdispersion is present, the variance of the response Y exceeds the nominal variance, $n\pi(1-\pi)$. Thus the actual variance can no longer be obtained from

$n\pi(1-\pi)$ for a given π (mean). If a linear logistic model is fitted to overdispersed data, the residual deviance exceeds its expected value (n-p), where n is the number of binomial observations and p is the number of parameters fitted in the model. If the residual deviance, after fitting the full model, exceeds its expected value, the assumption of binomial distribution for the responses is no longer valid. Then the data are said to exhibit overdispersion. Large residual deviance may occur due to number of reasons (Collett, 1991) such as outliers, inadequate terms in the model, inappropriate relationship between the response and the independent variable(s) (for instance linear relationship might have been assumed for the model when the relationship is non linear), inappropriate link function (for instance logit link function may not be appropriate). If the large residual deviance occurs even after eliminating all potential causes then the data are said to exhibit overdispersion.

Example data

Munkvold *et al.* (1993) reported these data from one of their experiments. This experiment consisted of disease assessments made in eight different sites over three different years (1989-1991). Altogether there were 22 disease assessments (22 disease incidence maps) recorded by means of presence or absence of the disease eutypa dieback on grapevines. Table 3.8 shows data extracted from this experiment for three locations (Oak Knoll-SB, Carneros-Ch and Delta-Cb) in 1990 (three field maps). In the process of extracting, each field was divided in to quadrats of 9 plants (3 rows consisting 3 plants per row) and number of diseased plants were recorded for each quadrat. A single layer of plants along the border of the field was omitted when the field map was divided into quadrats to accommodate rest of the plants fitted into quadrats of nine plants. Eventually there were 144, 256 and 304 quadrats for sites Oak Knoll-SB, Carneros-Ch and Delta-Cb respectively. Thus this layout may be described as a single factor experiment with 3 levels and, the levels one, two and three consisting sample units 144, 256 and 304 respectively.

When the model

$$\operatorname{logit}(\pi_i) = \eta_i = \mu + \alpha_i \qquad (3.18)$$

where π_i is the actual probability for the i^{th} location, μ is the overall mean and α_i is the i^{th} location effect, is fitted to the data in the Table 3.8 using GLIM, the resulting deviance is 1199.5 with 701 degrees of freedom. Parameter estimates and their standard errors associated with the model (3.18) (in GLIM convention) are given in the Table 3.9. The resulting deviance is much greater than the expected χ^2 with 701 degrees of freedom (P <0.001). Since the fitted model is the saturated model, the large residual deviance compared to its degrees of freedom is a clear indication that the data are overdispersed.

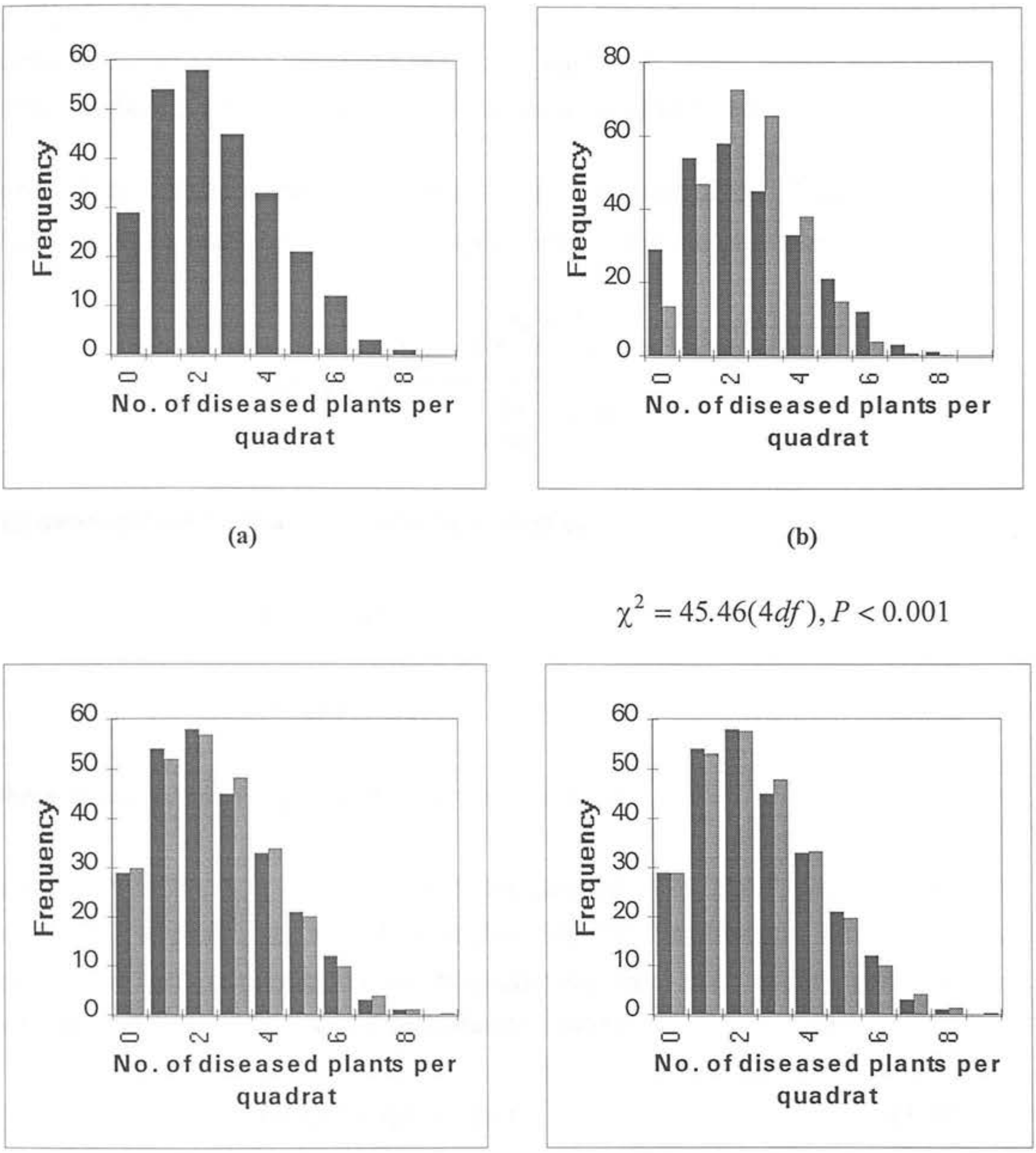
Table 3.8. Disease frequency for the data extracted from the field maps of three locations for the year 1990, presented by Munkvold *et al.* (1993) .

Number of diseased plants per quadrat	Observed frequency		
	Oak Knoll-SB	Carneros-Ch	Delta-Cb
0	0	29	15
1	3	54	49
2	4	58	75
3	17	45	69
4	21	33	51
5	27	21	23
6	32	12	14
7	25	3	4
8	10	1	2
9	5	0	2
Total	144	256	304

Table 3.9 Parameter estimates for the linear logistic model (3.18)

Parameter	Estimate	S.E.
μ	0.3907	0.0566
site 2 (Carneros-Ch)	-1.340	0.0732
site 3 (Delta-Cb)	-1.152	0.0699

As explained in section 3.1, this identification of overdispersion in the data may descriptively explained by distribution fitting. For instance, for the site Carneros-Ch, the observed and the expected frequencies of binomial and beta-binomial and logistic-normal binomial are shown in Fig. 3.5. From Fig. 3.5 it is clear that the expected beta-binomial and the logistic-normal binomial frequencies are closer to the observed frequencies than the expected binomial frequencies, indicating aggregation. The goodness-of-fit χ^2 tests (Fig. 3.5) also show that the beta-binomial and the logistic-normal binomial distributions provide far better fits to observed frequencies than the binomial distribution.



(a)

(b)

$\chi^2 = 45.46(4df), P < 0.001$

(c)

(d)

$\chi^2 = 1.11(5df), P = 0.967$

$\chi^2 = 1.16(5df), P = 0.949$

Fig. 3.5. Frequency distributions to Carneros-Ch site disease incidence. (a) Observed frequency; (b) Observed and expected binomial frequencies. ■ Observed ▨ Expected; (c) Observed and expected beta-binomial frequency ■ Observed ▨ Expected; (d) Observed and expected logistic-normal binomial frequency ■ Observed ▨ Expected.

3.4.1 Fitting the beta-binomial model

The beta-binomial model is a well known member of the general class of conditional binomials and sometimes categorised under random compounding models. Williams (1975) first demonstrated the modelling of overdispersion using the beta-binomial model. Crowder (1978) illustrated the use of the beta-binomial model for a factorial layout in which parameters are estimated by maximum likelihood.

For y_{ij} diseased plants among n_{ij} plants in the j^{th} quadrat of the i^{th} treatment group, where $1 \le j \le m_i$ and $1 \le i \le t$, equation (3.3) can be generalised as (Smith, 1983),

$$P\left(Y_{ij} = y_{ij}\right) = \binom{n_{ij}}{y_{ij}} \frac{\prod_{r=0}^{y-1} p_i + r\theta_i \prod_{r=0}^{n_{ij}-y_{ij}-1} (1 - p_i + r\theta_i)}{\prod_{r=0}^{n_{ij}-1} (1 + r\theta_i)}. \qquad (3.19)$$

The mean and the variance of y_i may be written as

$$E(y_i) = n_i p_i \qquad (3.20)$$

$$Var(y_i) = n_i p_i (1 - p_i) \tau_i \qquad (3.21)$$

where $\tau_i = 1/(1 + a_i + b_i)$ and a and b are as in section 3.1.2.

In equation (3.21), τ_i is sometimes called the 'heterogeneity factor', and it depends not only n_i but a_i and b_i as well. This implies that the beta-binomial model can take account of different amounts of heterogeneity that may occur with different treatments. But if τ_i is constant over different treatments (3.21) becomes

$$Var(y_i) = n_i p_i (1 - p_i) \tau. \qquad (3.22)$$

If there is a reason to believe that values of π_i near zero or unity are unlikely, the density function of the beta distribution must be unimodal, and zero at both zero and unity. Then, $a_i > 1$ and $b_i > 1$, and so the variance of π_i cannot exceed $p_i(1 - p_i)/3$. This could be rather restrictive (Collett, 1991).

Ignoring the constants involving only the observations, the log likelihood of (3.19) may be written as (Williams, 1975)

$$L=\sum_{i=1}^{t}\sum_{j=1}^{m_i}\left[\sum_{r=0}^{y_{ij}-1}\log(p_i+r\theta_i)+\sum_{r=0}^{n_{ij}-y_{ij}-1}\log(1-p_i-r\theta_i)-\sum_{r=0}^{n_{ij}-1}\log(1+r\theta_i)\right]. \quad (3.23)$$

EGRET can easily be implemented to obtain the maximum likelihood estimates of the beta-binomial model parameters. A GLIM macro published by Brooks (1984) can also be used if the model-fitting is done using GLIM, but this produces only approximate estimates. Table 3.10 shows the results of six different beta-binomial models to the data in Table 3.8 using EGRET. The equivalent models for each of the fits in Table 3.10 are shown in Table 3.11. Table 3.12 uses likelihood ratio statistics based on these fits (Table 3.10) to analyse the data.

Table 3.10 Six fits to the Munkvold *et al.* data, using the beta-binomial model with and without random effects

Model	Fixed effect parameters	Random effect parameters	Deviance	d.f.
A	%GM		1586.79	703
B	%GM, SITE		1199.56	701
C	%GM	%SCL	1335.81	702
D	%GM,SITE	%SCL	1116.97	700
E	%GM	%SCL,SITE	1301.20	700
F	%GM,SITE	%SCL,SITE	1114.59	698

Table 3.8 Equivalent models specified in Table 3.7

Model	Fixed effect component	Random effect component
A	μ	
B	$\mu+\alpha_i$	
C	μ	θ
D	$\mu+\alpha_i$	θ
E	μ	θ_i
F	$\mu+\alpha_i$	θ_i

In the Table 3.10, %GM refers to overall mean, μ, and %SCL refers to the aggregation parameter, θ. This is the notation adopted by EGRET in model fitting.

Model A is fitting a common mean (ignoring the site effect and aggregation) to all the data. Model B is fitting common mean including site effect, α_i, ignoring aggregation. Model C is fitting common mean and common θ for all the data. Model D is fitting common mean and site effect including an aggregation parameter (common). Model E is fitting common mean and (different) aggregation parameters for each level of the site factor. Model F is fitting common mean, site effect and different aggregation parameters to three different sites.

In Table 3.12, test 1 is the traditional test for site (treatment) effect when there is no extra-binomial variation, and it is seen that disease incidence varies between sites. Test 2 is the test for extra-binomial variation, given that there is site effect. The results of test 2 indicate that there is an excess variation. In this test, the testing hypothesis is $H_0\colon \theta = 0$ versus $H_1\colon \theta > 0$. Since θ is not allowed to be negative, the test compares the likelihood ratio statistic against the square of a one-tailed normal distribution, rather than a χ^2 as is usually done, i.e. test 2 does not necessarily have 1 d.f. since likelihood ratio statistic not necessarily have χ^2 distribution. Test 3 tests for site effect again, but this time accounting for the presence of excess variation. From the results of this test, it is clear that the site effect is still substantial. Test 4 goes on to test whether excess variation differs across three different sites, given that the site effect is present. According to the results, there is no evidence to say that excess variation differs between sites. As in the case with test 2, test 4 also does not necessarily have 2 d.f. Test 5 again tests for treatment effect, but this time accounting for different levels of excess variation in three sites. This test also suggests that there is a site effect. Thus the most appropriate model for the data seems to be Model D. Parameter estimates and their standard errors associated with Model D are given in Table 3.13. The parameterisation used in EGRET is same as the parameterisation used in GLIM.

The estimates in the Table (3.13), except θ, are very similar to those in Table 3.9, without allowing for overdispersion. But the standard errors have been inflated, in this particular case by nearly 23%. This in turn means that quantities derived from these estimates, such as fitted probabilities, will have larger standard errors than they would have had in the absence of overdispersion. The corresponding confidence intervals for these quantities will then be wider than they would have been if no adjustment were made for overdispersion.

It is important to note that the deviance of the saturated model is still much larger than the residual degrees freedom (698). This is an apparent indication that the data set could still be fit better. In addition, though the beta-binomial model is an attractive option from a theoretical point of view compared to standard linear logistic model, in practice it seems there is no reason to rely on a specific form of overdispersion.

Table 3.12 Analysis of the fits of the Munkvold *et al.* (1993) data reported in Table 3.8

Test	Compare s models	Likelihood ratio statistic	d.f.	P-value
1. Test for site effect ignoring excess variation	A vs. B	387.23	2	<0.001
2. Test for excess variation	B vs. D	82.59	1	<0.001
3. Test for site effect in the presence excess variation	C vs. D	218.84	2	<0.001
4. Test for different levels excess variation in the sites	D vs. F	2.38	2	0.304
5. Test for site effect in the presence of differing levels of excess variation	E vs. F	186.61	2	<0.001

Table 3.13 Parameter estimates for the beta-binomial model (D)

Parameter	Estimate	S.E.
μ	0.3900	0.0696
site 2 (Carneros-Ch)	-1.346	0.0901
site 3 (Delta-Cb)	-1.146	0.0859
θ	0.0688	0.0102

3.4.2 Fitting the logistic-normal binomial model

Luning, Sheridan, Ytterborn and Gullberg (1966) suggested an alternative to beta-binomial in which they assumed the variability in the response probability follows a normal distribution rather than beta distribution. Then the response probability is not restricted to the $(0,1)$ range. Since in practice the response probability could not occur outside $(0,1)$ range, their proposed model hardly became known. Pierce and Sands (1975) adjusted this procedure, incorporating the assumption that the variability in the logit of the response probability rather than response probability itself, varies normally with the expectation $\eta_i = \sum \beta_j x_{ji}$. Then the variability in the response probability has a logistic-normal distribution, which is equivalent to the beta-distribution in the beta-binomial model. When the logistic-normal distribution is compounded with the binomial distribution, the response probability has a logit scale and the resulting distribution is logistic-normal binomial distribution. The logistic-normal binomial distribution does not violate the restriction imposed by the fact that probabilities must lie in the (0,1) range (Aitchison and Shen, 1980). The rationale behind logistic-normal binomial model is that if the fixed effects η_i, exert their influence linearly on the logit scale, it is reasonable to assume that the random effect z_i would also act similarly.

The logistic-normal binomial model may be formulated as

$$\operatorname{logit}(\pi_{\mathrm{i}}) = \eta_i + \gamma z_i, \tag{3.24}$$

where π_i is the true response probability, i.e. the expected response probability and the variability in the response probability together, $\eta_i = \sum \beta_j x_{ji}$, γ is the coefficient of the random term z_i. The z_i is the standardised variable (zero mean and unit variance) for the random effect (variability in the response probability) and thus γ is the standard deviation of the random effect variable.

Thus

$$E[\operatorname{logit}(\pi_i)] = \eta_i$$

and

$$Var[\operatorname{logit}(\pi_i)] = \gamma^2.$$

The likelihood function of the logistic-normal binomial model may be written as

$$\begin{aligned} L(\beta, \gamma, z_i) &= \prod_{i=1}^{n} \binom{n_i}{y_i} \pi^{y_i} (1-\pi_i)^{n_i - y_i} \\ &= \prod_{i=1}^{n} \binom{n_i}{y_i} \frac{[\exp(\eta_i + \gamma z_i)]^{y_i}}{[1+\exp(\eta_i + \gamma z_i)]^{n_i}}. \end{aligned} \tag{3.25}$$

Likelihood estimations of the models that involve random variables with specified distributions need integration of the likelihood function with respect to the distribution of these variables. The resulting function is a marginal likelihood function,

$$L(\beta, \gamma) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \binom{n_i}{y_i} \frac{[\exp(\eta_i + \gamma z_i)]^{y_i}}{[1+\exp(\eta_i + \gamma z_i)]^{n_i}} \frac{\exp[-z_i^2/2]}{\sqrt{(2\pi)}} dz_i \tag{3.26}$$

This function involves only β and γ, and not z as in (3.25) and maximum likelihood estimates are the estimates of these parameters that maximise this likelihood function. (Note that in equation (3.26) π is the mathematical constant 3.141..).

By taking $u = \frac{z}{\sqrt{2}}$, we may write

$$dz = \sqrt{2}du.$$

Then equation (3.26) becomes

$$L(\beta,\gamma) = \pi^{-n/2}\prod_{i=1}^{n}\int_{-\infty}^{\infty}\binom{n_i}{y_i}\frac{[\exp(\eta_i+\sqrt{2}\gamma u_i)]^{y_i}}{[1+\exp(\eta_i+\sqrt{2}\gamma u_i)]^{n_i}}\exp(-u_i^2)du_i \tag{3.27}$$

This integral can only be integrated numerically. The technique EGRET employs for numerical integration is use of Gauss-Hermite quadrature formula:

$$\int_{-\infty}^{\infty} f(u)\exp(-u^2)du \approx \sum_{j=i}^{m} c_j f(s_j),$$

where values of c_j and s_j are given in standard tables (Abramowitz and Stegun, 1972). Pierce and Sands (1975) reported $m = 10$ is acceptable but recommended that m to be approximately equal to 20 for assured precision. However, using c_j and s_j equation (3.27) can be written as a summation and thus equation (3.27) may be written as (Collett, 1991)

$$L(\beta,\gamma) = \pi^{-n/2}\prod_{i=1}^{n}\binom{n_i}{y_i}\sum_{j=1}^{m} c_j\left\{\frac{[\exp(\eta_i+\gamma s_j\sqrt{2})]^{y_i}}{[1+\exp(\eta_i+\gamma s_j\sqrt{2})]^{n_i}}\right\}. \tag{3.28}$$

In equation (3.28)π is the mathematical constant 3.141.. The values of $\hat{\beta}$ and $\hat{\gamma}$ which maximise the equation (3.28) can then be determined numerically.

EGRET can be used to fit logistic-normal binomial models. After fitting a model, the deviance for the model can be computed and the deviance difference from two nested models with random effects will have an approximate χ^2 distribution. This can be used to test the significance of the nesting component in the usual way. But the deviance difference of two models, one with a systematic component and a random effect and the other with the same systematic component alone, can not be used to test for the excess variation as explained in the section 3.4.1. The appropriate testing procedure was also discussed in the section 3.4.1.

Table 3.14 shows the results of fitting six different logistic-normal binomial models (in EGRET notation) to the data in Table 3.8, using EGRET. The description of the models and model fitting procedures are exactly the same as for fitting beta-binomial models. Table 3.15 shows the results of the analysis of the possible likelihood ratio tests.

Table 3.14 Six fits to the Munkvold *et al.* data, using the logistic-normal binomial model with and without random effects

Mo del	Fixed effect parameters	Random effect parameters	Deviance	d.f.
A	%GM		1586.79	703
B	%GM,SITE		1199.56	701
C	%GM	%SCL	1333.20	702
D	%GM,SITE	%SCL	1117.02	700
E	%GM	%SCL,SITE	1277.56	700
F	%GM,SITE	%SCL,SITE	1113.84	698

Table 3.15 Analysis of the fits of the Munkvold *et al.* data reported in Table 3.8

Test	Compares models	Likelihood ratio statistic	d.f.	P-value
1. Test for site effect ignoring excess variation	C vs. D	387.23	2	<0.001
2. Test for excess variation	B vs. D	82.54	1	<0.001
3. Test for site effect in the presence of excess variation	C vs. D	216.18	2	<0.001
4. Test for different levels excess variation in the sites	D vs. F	3.18	2	<0.204
5. Test for site effect in the presence of differing levels of excess variation	E vs. F	163.72	2	<0.001

According to Table 3.15, the site effect is significant in the presence of random effect but there is no evidence for different random effects over different sites. It is important to note that the conclusions made using this method are the same as the conclusions made when the data were analysed using the beta-binomial model. The parameter estimates and their standard errors, obtained for the best fit (Model D) using EGRET are shown in Table 3.16.

Table 3.16. Parameter estimates for the logistic-normal binomial model (D)

Parameter	Estimate	S.E.
μ	0.4204	0.0762
site 2 (Carneros-Ch)	-1.442	0.0982
site 3 (Delta-Cb)	-1.237	0.0940
γ	0.5804	0.0453

The estimates and standard errors in Table 3.16 can be used to test the difference between disease intensities of three different sites. However, since little is known about the properties of the parameter estimates of the logistic-normal binomial model it would be prudent to use percentage points of *t*-distribution rather than the standard normal distribution in performing such tests.

In comparison with Table 3.13, standard errors produced with logistic-normal binomial model are larger than those produced with beta-binomial model.

According to the Table 3.14 the residual deviance of the most satisfactory model is 1117.02 with 700 degrees of freedom. The deviance is much larger than expected value, the residual degrees of freedom. Even after fitting the saturated model (Fit F) the residual deviance is still larger than its expected value. The distribution of the residual deviance for a logistic-normal binomial model is not known (Collett, 1991). Thus the residual deviance does not necessarily have a χ^2 distribution. Therefore the deviance of the satisfactory model need not be approximately equal to its degrees of freedom. Hence, after fitting a logistic-normal binomial model it is not possible to evaluate whether there is any additional source of overdispersion in the data. In other words, there is no proper goodness-of-fit test for the model.

3.4.3 Fitting overdispersed data using Williams procedures

3.4.3.1 Williams' Model II procedure

Williams (1982) suggested this method of taking extra-binomial variability into account in modelling binomial data. To allow for extra-binomial variation he introduced an unobserved continuous random variable ϑ_i independently distributed on (0,1) with $E(\vartheta_i)=\pi_i$, $Var(\vartheta_i)=\phi\pi_i(1-\pi_i)$, and assumed that, conditional on ϑ_i, y_i is distributed binomially with (n_i, ϑ_i), i.e.,

$$E(y_i|\vartheta_i)=n_i\vartheta_i$$

and

$$Var(y_i|\vartheta_i)=n_i\vartheta_i(1-\vartheta_i)$$

Thus unconditionally,

$$E(y_i)=E\{E(y_i|\vartheta_i)\}=E(n_i\vartheta_i)=n_i\pi_i \tag{3.29}$$

and

$$Var(y_i)=E\{Var(y_i|\vartheta_i)\}+Var\{E(y_i|\vartheta_i)\}$$

$$=E\{n_i\vartheta_i(1-\vartheta_i)\}+Var(n_i\vartheta_i)$$

$$=n_i\pi_i(1-\pi_i)[1-\phi]+n_i^2\phi\pi_i(1-\pi_i).$$

Therefore

$$Var(y_i)=n_i\pi_i(1-\pi_i)[1+(n_i-1)\phi]. \tag{3.30}$$

Here $\phi(\geq 0)$ is an unknown scale parameter. The equation (3.30) is the Model II of Williams (1982). In equation (3.30) the quantity $[1+(n_i-1)\phi]$ is referred to as the heterogeneity factor. If equation (3.30) is written in the form

$$Var(y_i)=n_i\pi_i(1-\pi_i)\sigma_i^2 \tag{3.31}$$

where $\sigma_i^2=1+(n_i-1)\phi$, then σ_i^2, the heterogeneity factor, depends only on n_i, the binomial denominator. Equation (3.31) is equivalent to equation (3.21) except that σ_i^2 does not take account of possible variable response probabilities between different treatments and thus is exactly equivalent to equation (3.22). In equation (3.30), if $\phi=0$, i.e. in the absence of random variation in the response probabilities, the variance of y_i is exactly the binomial variance. Maximum likelihood cannot be used for parameter estimation because the distribution of the y_i is not fully specified, but if ϕ is known the relationship between the expectation and variance of y_i allows the definition of a quasi-likelihood (Wedderburn, 1974) which is maximised with respect to the parameters $\boldsymbol{\beta}$ by the iteratively weighted least square equations

$$\mathbf{X}^T\mathbf{WVX}\hat{\beta} = \mathbf{X}^T\mathbf{WVY}, \qquad (3.32)$$

where $\mathbf{W}$ =diag (w_i), $\mathbf{V}$ =diag (v_i) which is $[n_i\pi_i(1-\pi_i)]$ and all quantities are computed using the initial estimates. Since ϕ is not usually known, Williams (1982) suggested the estimate of ϕ to be obtained by equating the value of Pearson's X^2 statistic for the model to its approximate expected value.

For n binomial observations the X^2 is given by

$$X^2 = \sum\nolimits_{i=1}^{n}\left\{\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + \frac{[(n_i - y_i) - (1-\hat{\pi}_i)n_i]^2}{n_i(1-\hat{\pi}_i)}\right\}$$

$$= \sum\nolimits_{i=1}^{n}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

If the observations have associated weights w_i, X^2 is then given by

$$X^2 = \sum\nolimits_{i=1}^{n}\frac{w_i(y_i - n_i\hat{\pi}_i)}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

and the approximate expected value of this statistic is

$$\sum\nolimits_{i=1}^{n} w_i(1 - w_i v_i d_i)\{1+\phi(n_i - 1)\} \qquad (3.33)$$

where $v_i = n_i p_i(1-p_i)$ and d_i is the i^{th} diagonal element of the variance-covariance matrix of the linear predictor, $\hat{\eta}_i = \sum \hat{\beta}_j x_{ji}$ (i.e. $p \times p$ matrix where p is the number of parameters). Thus to obtain ϕ, X^2 should be obtained first. But X^2 depends on ϕ. Therefore an iterative procedure is required to estimate ϕ.

In the iteration procedure, first X^2 is estimated at w_i=1 for the full model. Then equation (3.33) becomes

$$\sum_i (1-v_i d_i)\{1+\phi(n_i-1)\}$$

$$=\sum_i \{1-v_i d_i+(1-v_i d_i)[\phi(n_i-1)]\}$$

$$=\sum_i \{1-v_i d_i+\phi(n_i-1)-v_i d_i n_i \phi+v_i d_i \phi\}$$

$$=\sum_i \{1-v_i d_i+\phi[n_i-1-v_i d_i(n_i-1)]\}$$

$$=n-p+\phi\sum_i \{(n_i-1)(1-v_i d_i)\}.$$

Therefore the initial estimate of ϕ is,

$$\hat{\phi}_0=\frac{\{X^2-(n-p)\}}{\sum\{(n_i-1)(1-v_i d_i)\}},$$

from which the initial estimates of the weights are obtained as

$$w_{i0}=[1+(n_i-1)\hat{\phi}_0]^{-1}.$$

Then the model is refitted using weights as w_{i0} and new X^2 is calculated. Then the new $\hat{\phi}(=\hat{\phi}_1)$ becomes

$$\hat{\phi}_1=\frac{[X^2-\sum_i\{w_i(1-w_i v_i d_i)\}]}{\{w_i(n_i-1)(1-w_i v_i d_i)\}}.$$

If X^2 still remain large relative to its degrees of freedom (*n-p*) an additional cycle of this iterative procedure is carried out and this iteration should be continued until X^2 becomes very close to its degrees of freedom.

In the Williams' Model II procedure, the iterative scheme forces a fit and thus the Pearson X^2 statistic is forced down to be close to the residual degrees freedom. Hence, the residual deviance cannot be used as a measure of the extent to which the model that incorporates overdispersion fits the observed data. Instead, diagnostics for model checking have to be adopted (Collett, 1991). One might argue that ϕ is overestimated if some of the overdispersion occurs for other reasons such as unmeasured covariates. It is true that some believe (as noted in EGRET, 1990) that variability in the response probability is may be due to unmeasured covariates and that if they could be measured, they would provide enough information to account for the differences between similarly treated experimental units. But even if only most important variables were taken into account, there would quickly be more variables than observations. If the overdispersion problem is seen in this perspective, the Williams' Model II procedure may be a good solution. The model is appealing since it scales up the variance of the binomial variate without altering the mean. However, the Williams' Model II procedure is not applicable for binary (where $n_i = 1$) data. One other problem with Williams' Model II procedure is that it does not produce an estimate for the standard error of $\hat{\phi}$. Moore (1986) suggested a method to obtain an estimate for this standard error, but one could easily assume that if it is necessary to estimate $\hat{\phi}$, it is obviously different from zero.

Williams (1982) gives a GLIM listing which can implement this procedure. After fitting the full model (all main effects and interactions) in the usual way, a GLIM listing in the form of a macro is activated by the directive $USE followed by the identifier of the macro. This results a residual deviance followed by quasi-likelihood estimates for the saturated model. Then by eliminating terms from the model, the residual deviance of reduced model can be obtained. Two nested models can then be compared by examining the change in the deviance with the change in the degrees of freedom using a χ^2 test. The GLIM macro which implements the Williams' Model II procedure is given in Collett (1991).

When the Williams' Model II procedure is used to analyse the data in Table 3.8, after fitting the full model (3.18), $\hat{\phi}$ was found to be 0.0665. The resulting deviances for different models using this value for $\hat{\phi}$ are given in the Table 3.17.

Table 3.17 Deviances on fitting weighted linear logistic models to the data in Table 3.8

Terms fitted in the model	Deviance	d.f.	Δ deviance	Δ d.f.
$\mu + \alpha_i$	782.88	701	-	-
μ	1035.6	703	252.7	2

The deviance after fitting full model, in this case the model including a site effect, is quite close to its degrees of freedom, as it must be. Indeed, the only reason for the difference observed is that the iterative scheme for estimating ϕ is continued until Pearson's X^2 statistic, rather than the deviance, is close to its degrees of freedom. The increment in deviance on removing the term corresponding to the main effect of site is 252.7 on 1 degrees of freedom ($P<0.001$). This suggests different disease incidence in three different sites. Parameter estimates and their standard errors shown in Table 3.18 can be used to compare disease incidence between the three sites.

Table 3.18 Parameter estimates and standard errors for the model (3.18) using Williams' Model II procedure

Parameter	Estimate	S.E.
μ	0.3907	0.0701
site 2 (Carneros-Ch)	-1.340	0.0906
site 3 (Delta-Cb)	-1.152	0.0865

The estimates in the Table (3.18) also are very similar to those in Table 3.9, without allowing for overdispersion. But the standard errors have been inflated. These standard errors are larger than those obtained by the beta-binomial method (Table 3.13) but smaller than those obtained by the logistic-normal binomial model (Table 3.16). However, in most situations, Williams procedures produces larger standard errors compared to all other procedures Collett (1991).

3.4.3.2 Williams' Model II as a generalisation of Finney's procedure

This general method of adjusting the analysis to take account of overdispersion was first proposed by Finney (Finney, 1947; 1971). If Y given π has a binomial distribution, and π has a distribution with mean p and variance σ^2, then for $n>1$,

$$E(Y)=E[E(Y|\pi)]=nE(\pi)=np$$

$$Var(Y)=Var[E(Y|\pi)]+E[Var(Y|\pi)]$$

$$=n^2\sigma^2+nE[\pi(1-\pi)]$$

$$=n^2\sigma^2+n[p-(p^2+\sigma^2)]$$

$$=np(1-p)+n(n-1)\sigma^2. \tag{3.34}$$

The equation (3.34) clearly shows the effect of varying π is to increase the variance of y (more than the binomial variance), leading to large residual deviances for models which would fit well if the random variation were correctly specified. When $n_i = n$ for all i, σ^2, the heterogeneity factor is an unknown constant. Then the expected value of Pearson's X^2 statistic for the model that includes all main effects and interactions (full model) can be approximated by $(n-p)\sigma^2$ (McCullagh and Nelder, 1989). The parameter σ^2 may therefore be estimated by $X^2/(n-p)$, where p is the number of parameters and an iterative estimation procedure, which was used in the Williams' Model II procedure, is no longer required. Thus Williams' Model II procedure is a generalisation of Finney's procedure.

If linear logistic models are fitted in the usual manner, two nested models can be compared by examining the ratio of the difference in deviances divided by the change in the degrees of freedom to the mean deviance (which is the residual deviance divided by its degrees of freedom) for the full model. This quantity has F distribution and is also independent of σ^2. So the models can be compared using an F test as in ANOVA for continuous response variables. Thus an analysis of deviance table can be formed.

For the data in Table 3.8 the deviances for fitting possible linear logistic models using GLIM are given in Table 3.19. Table 3.20 is the analysis of deviance table that can be formed Table 3.19.

Table 3.19 Deviances for the possible linear logistic models for the data in Table 3.8

Terms fitted in the model	Deviance	d.f.
μ	1586.8	703
$\mu+\alpha_i$	1199.5	701

Table 3.20 Analysis of deviance table obtained from the Table 3.13

Source of variation	d.f.	Deviance	Mean deviance	F-ratio
Site	2	387.2	193.60	113.22
Residual	701	1199.5	1.71	
Total	703	1586.8		

The observed value of F-statistic for the site effect is significant ($P < 0.001$) and thus suggests different disease intensities in the three different sites. GLIM can easily be implemented to perform this analysis to obtain F ratios. In the terminology of the statistical package GLIM, σ^2 is referred to as the scale parameter. When there is no overdispersion, that is, when standard linear logistic model is fitted, the scale parameter, σ^2 is taken as 1 for all i. But when overdispersion is present and when n_i are equal, we can set the scale parameter to be the mean deviance of the full model. That is, after fitting the full model of the standard linear logistic model we can declare the scale parameter to be taken as the mean deviance of the fitted full model. In GLIM, scale parameter settings may be done by the directives, $CAL %S=%DV/%DF and %$SCALE %S. Then two nested models can be compared by examining the resulting mean deviance difference with the percentage points of the F distribution. The appropriate degrees of freedom for percentage points of the F distribution are the degrees of freedom corresponding to deviance difference and the degrees of freedom corresponding to residual deviance with full model. For this particular example the scale parameter is set to 2.257.

A consequence of this method is that the variance of y_i being inflated by a σ^2. But the parameter estimates (Table 3.21) are same as obtained with the standard linear logistic model. The standard errors of the parameter estimates (Table 3.21) are the standard errors after fitting corresponding standard linear logistic model multiplied by a factor of $\sqrt{\sigma^2}$. In addition, confidence intervals based on parameter estimates will need to be constructed from percentage points of t-distribution instead of z-distribution (Finney, 1971; Collett, 1991). Finney's method gives same result as the Williams' Model II procedure when n_i are all equal. The small numerical differences of SEs in the Table 3.21 from Table 3.18 result because in the Williams' Model II procedure ϕ is based on the residual χ^2 and the Finney's method the scale parameter is based on the residual deviance.

Table 3.21 Parameter estimates for the model (3.18) using Finney's method implemented in GLIM

Parameter	Estimate	S.E.
μ	0.3907	0.0715
site 2 (Carneros-Ch)	-1.340	0.0913
site 3 (Delta-Cb)	-1.152	0.0877

The major limitation of this procedure is that it is not possible to test the goodness-of-fit of the fitted model. In addition, practical situations in which each proportion is based on same number of binary observations are comparatively rare. However, Finney's method is not very sensitive to differences in n_i and thus this method could be used to adjust overdispersion even when n_i are not all equal. Nevertheless, Williams' Model II procedure has the advantage that correct standard errors can be obtained directly. Finney's method is therefore recommended only when approximate estimates are sufficient (Aitken *et al.*, 1989; Collett, 1991).

SAS can also be used to implement this procedure. In SAS, PROC PROBIT with LACKFIT option produces similar estimates and standard errors of estimates. It is important to note that the output produced by SAS will not be a duplicate of the output of GLIM because SAS adopts the parameterisation which sets the parameter of the last level of the factor to be zero in contrast to parameter of the first level is zero in GLIM. Moreover, PROC PROBIT does not identify different observations with same treatment combinations as separate individual observations. Therefore heterogeneity cannot be taken account for a model which contains all main effect and interaction terms.

3.4.3.3 Williams' Model III procedure

Williams (1982) suggested an approximate method of obtaining parameters ($\boldsymbol{\beta}$) and γ of the logistic-normal binomial model. The assumption in the logistic-normal binomial model is that $E[\text{logit}(\pi_i)] = \eta_i$ and $Var[\text{logit}(\pi_i)] = \gamma^2$ (section 3.4.2). If the γ is sufficiently small, the approximate variance of a function of a random variable can be derived for the variance of $\text{logit}(\pi_i)$. That is (as cited in Collett [1991]),

$$Var[\text{logit}(\pi_i)] \approx \left\{ \frac{d[\text{logit}(\pi_i)]}{d\pi_i} \right\}^2 |_{p_i} \, Var(\pi_i)$$

$$= \frac{1}{p_i^2(1-p_i)^2} Var(\pi_i), \qquad (3.35)$$

where p_i is the expected response probability $[E(\pi_i)]$ for the i^{th} observation and $|_{p_i}$ stands for 'evaluated at p_i'. Then,

$$Var(\pi_i) \approx p_i^2(1-p_i)^2 Var[\text{logit}(\pi_i)]$$

$$= \gamma^2 p_i^2 (1-p_i)^2 \qquad (3.36)$$

which is the Model III of Williams (1982). The expected value and the variance of y_i then may be written as

$$E(y_i) \approx n_i p_i \qquad (3.37)$$

$$Var(y_i) \approx n_i p_i (1-p_i)[1+\gamma(n_i-1)p_i(1-p_i)]. \qquad (3.38)$$

The difference between (3.38) and (3.30) is that ϕ in equation (3.30) has been replaced by $\gamma p_i(1-p_i)$. A GLIM macro used to implement Williams' Model II procedure can easily be modified to implement this approximation procedure. The only change that has to be made to the algorithm is that the quantity $[1+\phi(n_i-1)]$ is replaced by $[1+\gamma(n_i-1)p_i(1-p_i)]$. Then γ can be estimated iteratively by equating the value of Pearson's X^2-statistic to its approximated expected value as in the section 3.4.3.1. The complete GLIM listing required to implement this procedure is given in Collett (1991).

When this approximation procedure is used to analyse the data in Table 3.8 after fitting model with main effects (3.18), $\hat{\gamma}$ was found to be 0.3094. The resulting deviances for the different models using this value for $\hat{\gamma}$ are given in the Table 3.22.

Table 3.22. Deviances on fitting models using Williams' Model III procedure to the data in Table 3.8

Terms fitted in model	Deviance	d.f.	Δ deviance	Δ d.f.
$\mu+\alpha_i$	782.95	701	-	-
μ	1028.9	703	245.9	2

The models with different linear systematic components can be compared by examining the difference in deviance between two models with percentage points of the χ^2 distribution, after fitting these models with weights based on full model with the random effect. Thus, according to Table 3.22, the deviance difference for fitting the site effect component is 245.9 with 2 degrees of freedom ($P<0.001$) and this confirms different disease incidence in the three sites. The parameter estimates for the best fit [model (3.18)] are shown in Table 3.23.

Table 3.23 Parameter estimates for the best fit (full model) using Williams' Model III procedure

Parameter	Estimate	S.E.
μ	0.3907	0.0715
site 2 (Carneros-Ch)	-1.340	0.0913
site 3 (Delta-Cb)	-1.152	0.0877

In this approximation method, as in the Williams' Model II procedure, the resulting deviance after fitting the model cannot be used to compare the goodness-of-fit of the model, for the very same reason mentioned in section 3.4.3.1. On the other hand, since the residual deviance after fitting the full model in logistic-normal binomial will not necessarily be as small as its number of degrees of freedom, the estimated variance of the random effect, $\hat{\gamma}$, obtained using approximate procedure, can be considerably smaller than that found using logistic-normal binomial model. When $0.2 \le \hat{p}_i \le 0.8$ this approximate procedure for fitting logistic-normal binomial model produces similar results to the procedure described in the section 3.4.3.1.

3.4.3.4 A generalised Williams' Model II procedure for modelling overdispersion

Moore (1987) suggested that the moment and beta-binomial likelihood methods could be generalised to allow modelling the extraneous variance. Thus he expressed the τ_i of equation (3.21) or equivalently ϕ of equation (3.30) as a function of p_i where p_i is $E(\pi_i)$, that is,

$$\phi(p_i) = \rho[p_i(1-p_i)]^{\xi-1},$$

where ρ and ξ are unknown scale parameters. Thus,

$$Var(\pi_i) = \rho[p_i(1-p_i)]^{\xi}.$$

This yields the generalisation of the unconditional variance of y_i,

$$Var(y_i) = n_i p_i(1-p_i)\{1+(n_i-1)\rho[p_i(1-p_i)]^{\xi-1}\} \qquad (3.39)$$

Equation (3.39) is same as the equation 12 of Moore (1987) and is a generalisation of Williams' model II (equation 3.30). Moreover, when $\xi = 1$, equation (3.39) reduces to equation (3.22) and (3.30), and when $\xi = 2$, equation (3.39) reduces to equation

(3.38), which corresponds approximately to a constant variance on logit scale (Williams' model III).

As Moore (1987) suggested, if the beta-binomial likelihood method is generalised, maximum likelihood estimates of the parameters may be found by an iterative function maximisation routine. If equation (3.30) is generalised without assuming any distribution for π_i, quasi-likelihood estimates can be obtained for the parameters. The estimate of ρ may be obtained by equating the Pearson X^2 statistic to its expected value as in the Williams' Model II and Model III procedures, which Moore refers to as moment method. He further suggested that an estimate for ξ to be made by minimising the function,

$$Q(\xi)=\sum_i^n\left[\hat{e}_i^2-\{[\hat{p}_i(1-\hat{p}_i)/n_i]+\hat{\rho}[\hat{p}_i(1-p_i)]^{\xi}\}\right]^2, \qquad (3.40)$$

where $\hat{e}_i=[(y_i/n_i)-\hat{p}_i]$ and $Q(\xi)$ is the sum of squares of the residuals about their approximate expected values.

The GLIM listing of Williams (1982) can be easily modified to implement this procedure. This listing in association with equation (3.40), can be used to obtain the estimates of the parameters in equation (3.39). The optimum estimate of ξ can be obtained by trial and error. The complete GLIM macro, including the listing required to implement the equation (3.40), is given in the appendix I. First, a value for ξ should be introduced to the system. This may be done by the GLIM directive $CALC followed by the identifier for ξ, followed by the value for ξ with the '=' sign. Then after fitting the full model (separate mean for each assessment) the macro associated with equation (3.39) is activated by the directive $USE followed by the identifier of this macro. After that the macro associated with the equation (3.40) is activated. Thereafter, by trial and error with different values of ξ, the estimate ξ with minimum $Q(\xi)$ is obtained. At this $Q(\xi)$, two nested models can be compared by examining the change in the deviance with the change in the degrees of freedom using χ^2 test.

According to the algorithm, the estimate of ξ is based on the p_is. Specifically, since the estimate of ξ is obtained by trial and error, the estimate may not be reliable with only a few p_is. Thus this method is more appropriate when the number of p_i values is large. Because of this, the method is illustrated with the complete data reported by Munkvold *et al.* (1993), described in the section 3.4.1. When the linear logistic model, with a separate mean for each assessment, was fitted using GLIM to the complete data, the residual deviance was 7881.8 with 5113 degrees of freedom, which clearly indicates overdispersion.

Using the GLIM macro in appendix I after fitting the model with a separate mean for each assessment, the estimates of ρ and ξ were found to be 0.0432 and 0.9 respectively. The resulting deviances for possible models using $\hat{\rho}=0.0432$ and $\hat{\xi}=0.9$ are given in Table 3.24.

Table 3.24 Deviances on fitting model using generalised Williams' Model II procedure to the data of Munkvold *et al.* (1993) (all assessments)

Terms fitted in the model	Deviance	d.f.	Δ deviance	Δ d.f.
$\mu+\alpha_i$	5589.3	5113	-	-
μ	12240	5134	6650	21

The increase in deviance on excluding the term corresponding to main effect from the model that contains the main effect is 6650 with 21 degrees of freedom ($P<0.001$), indicating differences between assessments. The parameter estimates for the best fit, i.e. the model with main effects are shown in Table 3.25.

Compared to standard errors obtained using the method in section 3.4.3.1, the standard errors in the Table 3.25 have been subjected to possible variability in the heterogeneity parameter due to different groups or treatments. Hence one can argue that the procedure discussed in this section is more acceptable compared to the method discussed in section 3.4.3.1.

If either the logistic-normal binomial or the beta-binomial model is fitted using EGRET, for variable aggregation, a separate aggregation parameter is fitted for each treatment. But in generalised Williams' Model II procedure there are only two parameters for the similar situation (variable aggregation). Thus, there is an advantage of a simpler parameterisation in generalised Williams' Model II procedure.

The properties of the moment estimates for fixed ξ have been studied by Kleinman (1973) and Moore (1986), and they have shown that the estimates are consistent and asymptotically normal under reasonable condition, although asymptotic properties of ξ would be difficult to establish. One of the advantage of the method described in this section is that it does not require the assumption of a particular distribution for π_i and hence, this method may also have some robustness properties.

Table 3.25 Parameter estimates for the best fit (full model) in Table 3.24

Parameter	Estimate	S.E.	Estimate	S.E.
	(with generalised Williams' Model II procedure)		(with linear logistic model)	
μ	-1.663	0.0593	-1.663	0.0497
Assessment 2	0.714	0.0809	0.714	0.0680
Assessment 3	1.235	0.0768	1.235	0.0646
Assessment 4	0.507	0.0779	0.507	0.0655
Assessment 5	0.901	0.0766	0.901	0.0644
Assessment 6	1.132	0.0741	1.132	0.0623
Assessment 7	-1.736	0.1884	-1.736	0.1548
Assessment 8	-1.088	0.1283	-1.088	0.1064
Assessment 9	0.122	0.0991	0.122	0.0832
Assessment 10	0.439	0.0850	0.439	0.0714
Assessment 11	1.174	0.0791	1.174	0.0666
Assessment 12	2.019	0.0766	2.019	0.0645
Assessment 13	2.053	0.0894	2.053	0.0753
Assessment 14	3.074	0.1021	3.074	0.0858
Assessment 15	1.126	0.0782	1.126	0.0658
Assessment 16	2.445	0.0796	2.445	0.0670
Assessment 17	3.161	0.0873	3.161	0.0733
Assessment 18	1.802	0.0916	1.802	0.0772
Assessment 19	2.604	0.0983	2.604	0.0827
Assessment 20	0.053	0.0927	0.053	0.0777
Assessment 21	0.906	0.0819	0.906	0.0689
Assessment 22	1.621	0.0793	1.621	0.0667

3.4.4 Modelling overdispersion based on the concepts of design effect and effective sample size

Rao and Scott (1992) suggested a method for comparing independent experimental treatments. This method is based on the concepts of design effect and effective sample size widely used in sample surveys (Kish, 1965). According to Rao and Scott (1992) this method gives asymptotically correct results as the number of clusters (quadrats, say) in each treatment tends to infinity and can be implemented using any standard computer program for the analysis of independent binomial data after an adjustment to the data.

Let y_{ij} be the number of responses in n_{ij} units of the j^{th} replicate (quadrat) $(j = 1,2..m_i)$ of the i^{th} treatment group $(i = 1,2,..t)$, where m_i is the number of replicates in the i^{th} treatment group and t is the number of treatment groups. Then the proportion of the responses in the i^{th} treatment group, $\hat{\pi}_i$, is given as

$$\hat{\pi}_i = \sum_j y_{ij} / \sum_j n_{ij}. \tag{3.41}$$

An estimate of the variance of $\hat{\pi}_i$ for large m_i may be obtained as (Rao and Scott, 1992)

$$v_i = m_i (m_i - 1)^{-1} n_i^{-2} \sum_{j=1}^{m_i} (y_{ij} - n_{ij}\hat{\pi}_i)^2 , \tag{3.42}$$

and under mild regularity conditions on the population variances in the n_{ij} and r_{ij}, where $r_{ij} = y_{ij} - n_{ij}\hat{\pi}_i$, it follows that $[(\hat{\pi}_i - \pi_i)/v_i^{1/2}]$ is asymptotically N(0,1) as m_i increases (Scott and Wu, 1981).

Thus the ratio of observed variance, v_i, to the estimated binomial variance denoted by,

$$d_i = n_i v_i / [\hat{\pi}_i (1 - \hat{\pi}_i)]. \tag{3.43}$$

The d_i represents the inflation of variance due to overdispersion and referred to as the 'design effect'. The sample size adjusted for the inflation, known as the 'effective sample size', is obtained as $\tilde{n}_i = n_i / d_i$ and effective response is obtained as $\tilde{y}_i = y_i / d_i$. Thus the estimate $\tilde{\pi}_i$ is given as $\tilde{\pi}_i = \tilde{y}_i / \tilde{n}_i$ and the adjusted estimated binomial variance is given by $\tilde{v}_i = \tilde{\pi}_i (1 - \tilde{\pi})_i / \tilde{n}_i$. Therefore,

$$(\tilde{\pi}_i - \pi_i) / \sqrt{\tilde{v}_i}$$

is asymptotically N(0,1). Thus the $\tilde{y}_i$ and $\tilde{n}_i$ may be used to model the effect of factors on response variables using the standard linear logistic model.

For the data in Table 3.8, using equations (3.41), (3.42) and (3.43), the inflation factors can be computed as $d_1 = 1.2, d_2 = 2.3$ and $d_3 = 3.7$ for Oak Knoll-Sb, Carneros-Ch and Delta-Cb sites respectively. Using these estimates, values of $\tilde{y}_{ij}$ and

$\tilde{n}_{ij}$ can be computed as above. Then the standard linear logistic models can fitted to the adjusted values as described in section 3.2. The deviance on fitting possible linear logistic models using GLIM for the adjusted data are shown in Table 3.26.

In the Table 3.26, μ and α_i represent the mean and site effect respectively. The deviance difference for including the term for a site effect into the model is 257.7 with 2 degrees of freedom ($P < 0.001$), indicating differences in disease incidence at different sites. The residual deviance after fitting full model, 782.04, is almost equal to the residual deviance obtained using Williams' Model II procedure after fitting the full model. However, in this method, no iteration scheme forces the residual deviance to be close to its degrees of freedom and this method is mathematically much simpler than all other procedures. One other advantage of this method is that this method can be applied to a variety of statistical procedures involving independent treatment groups of clustered binary data such as testing homogeneity of proportions, testing trend of proportions, and computing Mantel-Haenszel χ^2 statistic for independence. Moreover this method do not assume any dependence structure among binomial observations and thus the estimates are robust. However, Rao and Scott (1992) themselves observed some loss of the power of significant testing in this method compared to optimal tests under a specified model for overdispersion, provided the assumed model fits the data adequately. Nevertheless, according to Table 3.23 the deviance difference due to including the site effect term is 257.7, which is slightly higher than the corresponding deviance difference observed under some procedures such as the Williams procedures.

Table 3.26 Deviances on fitting possible linear logistic models for the data in Table 3.8 after adjusting for overdispersion using Rao and Scott (1992) suggestions

Terms fitted in the model	Deviance	d.f.	Δ Deviance	Δ d.f.
$\mu + \alpha_i$	782.04	701	-	-
μ	1039.7	703	257.7	2

A major limitation one may find with this procedure is that it is not very clear how design effect is computed for factorial situations. We suggest, for instance, that for a two factor factorial layout one may compute design effects and effective sample size as follows. First, design effects, d_{ij} for each factorial combination are computed separately and, using d_{ij}, effective sample size, $\tilde{n}_{ij}$ and $\tilde{y}_{ij}$ may be computed as $\tilde{n}_{ij} = n_{ij} / d_{ij}$ and $\tilde{y}_{ij} = y_{ij} / d_{ij}$ respectively. Then, using the adjusted data, linear logistic model can be fitted using GLIM in the usual way.

However, if interactions are non-significant it is not very sensible that the adjustment is carried out for each factorial combination separately. Thus, if interaction is non-significant a design effect for each factor, d_i and d_j, may be computed separately and $\tilde{n}_{ij}$ and $\tilde{y}_{ij}$ may be computed as $\tilde{n}_{ij} = [n_{ij} / (d_i d_j)^{1/2}]$ and $\tilde{y}_{ij} = [y_{ij} / (d_i d_j)^{1/2}]$ respectively. Then the linear logistic model can be fitted in the usual way and test for the main effects. If only one main effect is detected the data have to be adjusted only for the significant main factor only and the linear model can then be fitted for the significant main effect. The asymptotic properties of this proposed method have not been studied yet and we hope to investigate it in the near future.

3.5 Choosing an appropriate statistical procedure

Sections 3.1 to 3.4 described statistical techniques that can be used to analyse epidemiological data expressed as disease incidence. It is clear that if there is no overdispersion, the most appropriate statistical procedure is to model the data using a linear logistic model. When overdispersion is present there are several methods available to analyse the data. Thus it is necessary to choose the most appropriate procedure to analyse a particular set of data.

Hughes and Madden (1992) investigated the relationship between observed variance and binomial variance for aggregated disease incidence data and reported that a good description of epidemiological data were provided by this relationship:

$$\log(v_o) = \log(c) + m[\log(v_b)], \tag{3.44}$$

where v_o is the observed variance, v_b is the expected binomial variance and c and m are parameters to be estimated.

If the mean disease incidence is p, where $p = E(\pi)$, v_b is obtained as $p(1-p)/n$ and v_o is obtained as variance of π in the usual way. For instance, let there be N quadrats for the i^{th} treatment. First, disease incidence for each quadrat $(\tilde{\pi})$ is obtained by dividing number of diseased plants per each quadrat by the quadrat size, n. Then the estimate of p for i^{th} treatment is obtained as the sum of $(\tilde{\pi})$ divided by N. The v_o for the i^{th} treatment may be obtained as the variance of $\tilde{\pi}$ of the i^{th} treatment in the usual way.

As an illustration data reported Munkvold *et al.* (1993) consists 22 assessments (experiment details were given in the section 3.4) and for these 22 assessments variance-variance plot is shown in Fig. 3.5.

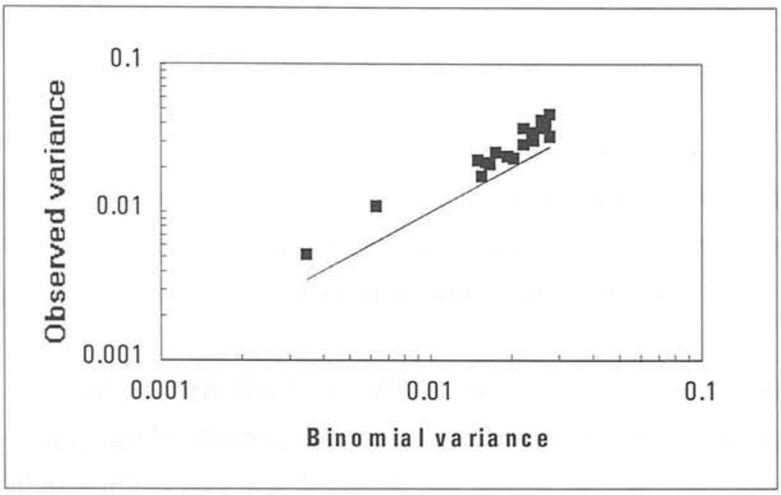


Fig. 3.5 Variance-variance plot on a log-log scale for the Munkvold *et al.* (1993) data (all 22 sites; years 1990-1992)

In Fig. 3.5, the continuous line represents the expected binomial variance. In a variance-variance plot, a scatter point lying above the binomial line indicates that the observed variance is greater than the expected binomial variance while a point below the binomial line indicates observed variance is less than the expected binomial variance. If the disease incidence is random all data points should lie on the binomial line. Thus the points above the binomial line represent overdispersion and points below the binomial line represent underdispersion.

Since all of the points in the Fig. 3.5 lie above the binomial line it is clear that disease incidence is aggregated in this case. Moreover, since all the points lie more or less parallel to the binomial line, this suggest that aggregation does not vary between treatments. The straight line regression fit for observed variance against the binomial variance gave a slope of 0.97, of which the 95% confidence interval was 0.86 and 1.09. The slope equal to unity confirms that aggregation does not differ between treatments. Thus the variance-variance plot gives an indication of spatial pattern of a diseased plants. In addition, since the variance-variance plot indicates the nature of the dispersion, this plot may be used as a tool of choosing an appropriate statistical technique to analyse epidemiological data. The link between variance-variance relationship and aggregation parameter θ (section 3.1.2) is established (Madden and Hughes, 1995). According to them, this link is given as

$$\theta = [cn^{-m} - \tfrac{f(p)}{n}] / [f(p) - cn^{-m}],$$

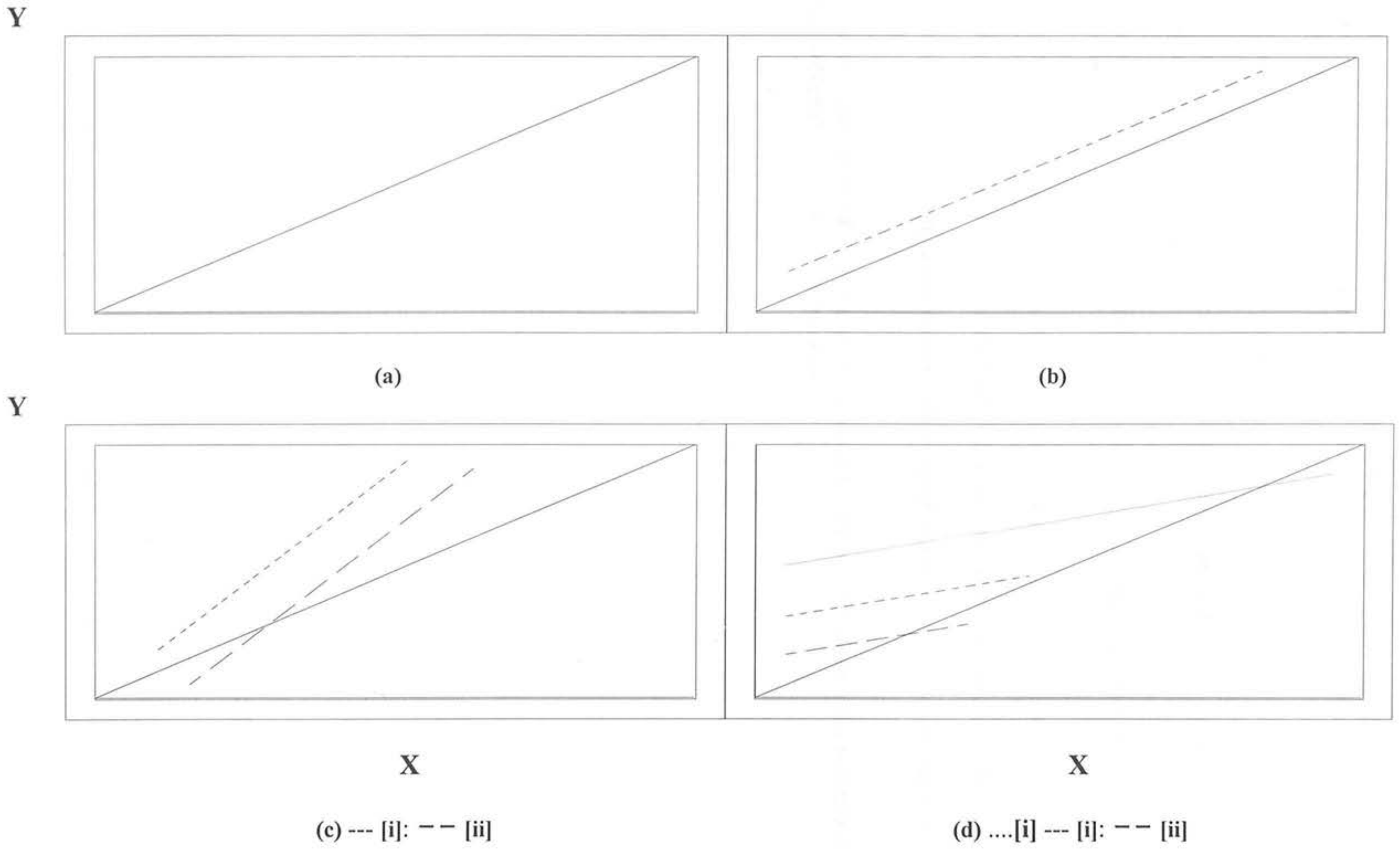
where θ is the aggregation parameter, $f(p) = [p(1-p)]^{1-m}$ and c, p and m are as in equation (3.44) and n is the sample unit size.

It is clear if there is no overdispersion all the points lie more or less on the binomial line of the variance-variance plot. Then, as described earlier, the most appropriate statistical procedure is to model the data using linear logistic model.

When overdispersion is present, the most appropriate procedure for the analysis of a particular set of data can be chosen using the nature of aggregation as described by the variance-variance plot as a guideline. If the data set has only a few groups (treatment combinations), for instance if it is a 2×2 factorial, then there will be only four points in the variance-variance plot. With a few scatter points on the plot it is often difficult to decide on the nature of aggregation. Under these circumstances, fitting either the beta-binomial model or the logistic-normal binomial model using EGRET would be more appropriate. The beta-binomial has a limitation that for positive a_i and b_i, variance of π_i cannot exceed $p_i(1-p_i)/3$ as described in the section 3.4.1. Moreover, there is no particular reason to justify the assumption in the beta-binomial model that π_i is to have a beta distribution. In contrast logistic-normal binomial model has an intuitive rationale that the random effects and fixed effects are added together on the same logit scale. These reasons have made us prefer the logistic-normal binomial model rather than the beta-binomial model.

If there is a substantial number of scatter points on the variance-variance plot, some possible relationships between observed variance and the binomial variance are shown in Fig. 3.6. Fig 3.6(a) is a situation where there is random occurrence of incidence. In this case all the scatter points in the graph lie along the binomial line. The data in the Table 3.4 are an example for such situations, and the linear logistic model gives an adequate fit. Fig. 3.6(b) illustrate a situation where the data points lie above and parallel to the binomial line. Since all the scatter points lie above the binomial line, for all the assessments, observed variances are larger than the binomial variance. In other words this is the situation of overdispersion. All the scatter points lying parallel to the binomial line implies the slope of the graph is equal to one, and this suggests constant aggregation for all treatments. Fig. 3.5 is an example for this situation. For analysis, Williams' Model II procedure would be the most efficient, though beta-binomial, logistic-normal binomial on EGRET, generalised Williams' Model II procedure and the method associated with design effect and effective sample size are also appropriate. Fig. 3.6c is the situation where slope of the graph is larger than one. As shown in Fig. 3.6c this situation may be of two types. The first is where all the scatter points lie along a straight line with all points above the binomial line (Fig. 3.6c[i]). The data of Madden *et al.*, (1987) are an example of this situation. The generalised Williams' Model II procedure would be the most efficient for this situation, as explained in section 3.4.3.2, and the method associated with design effect and effective sample size is also appropriate for the analysis. The second type is where the data lie along a straight line but the line crosses the binomial line with in the range of the data (Fig. 3.6c[ii]), i.e. data points exhibit both overdispersion and underdispersion. Data on grape downy mildew (Madden *et al.*, 1994) provide an example for this type. EGRET may not be suitable for this situation because the algorithm adopted in EGRET does not converge under underdispersion (EGRET, 1990), and the generalised Williams' Model II procedure and the method associated with design effect and effective sample size are appropriate for the analysis. Between both methods, the generalised Williams' Model II procedure may be more efficient because of the convenience in implementation.

It is important to note that this second type (Fig. 3.6c[ii]) could be quite crucial. If first half of the data points lie below that binomial line the other half lie above the binomial line, fitting the binomial model might apparently give an adequate fit. The reason for this is that when the scatter points are below the binomial line the residual deviance is less than the expected under binomial and this could compensate for the large deviance due to overdispersion. Thus, fitting an apparently adequate linear logistic model might produce misleading conclusions. Hence, examination of the variance-variance plot, where possible, may be a useful precursor to the analysis of epidemiological data in the form of incidence.



(a)

(b)

(c) --- [i]: — — [ii]

(d) ....[i] --- [i]: — — [ii]

(See next page for the figure caption)

Fig. 3.6 Some possible variance-variance plots on a log-log scale for epidemiological data. X axes - Binomial variance; Y axes - Observed variance. In all four plots the continuous line represent the binomial line.

(a). **All the scatter points lie on the binomial line**

(b). **All the scatter points lie more or less parallel to the binomial line and above the binomial line**

(c)[i]. **The slope of the scatter line greater than unity and all the scatter points above the binomial line**

(c)[ii]. **The slope of the scatter line greater than unity and the line crosses the binomial line.**

(d)[i]. **The slope of the scatter line less than unity and the line crosses the binomial line.**

(d)[ii]. **The plot consists of only a few scatter points and all the points lie above the binomial line.**

(d)[iii]. **The plot consists of only a few scatter points and some points lie below the binomial line, but still exhibit overall overdispersion**

Fig. 3.6d is the situation where the slope of the line is less than the slope of the binomial line, i.e. slope is less than unity. According to our experience this situation is not very common. Nevertheless, if the scatter points occurs along a line but the line crosses the binomial line within the range of the data (Fig. 3.6d[i]) generalised Williams' Model II procedure may be suitable to take account of deviation from the randomness. Even a negative slope would not be a problem with the generalised Williams' Model II procedure. If the number of scatter points is low and the scatter points lie above the binomial line (Fig. 3.6d[ii]) the method associated with EGRET would be the most efficient. However, if some of these scatter points lie below the binomial line (Fig. 3.6d[iii]), but still showing overall overdispersion, the method associated with design effect and effective sample size may be the most appropriate method. Table 3.27 summarises the analyses appropriate for data following each of the observed-binomial variance relationships shown in Fig. 3.6.

Table 3.27 Summary table of the choice of the statistical procedure

Variance-variance relationship (Fig. 3.6)	Suggested analysis	Section reference
(a)	Linear logistic model	3.2.1
(b)	Williams' model II procedure	3.4.3.1
[(c)i]	Generalised Williams' Model II procedure	3.4.3.4
[(c)ii]	Generalised Williams' Model II procedure	3.4.3.4
[(d)i]	Generalised Williams' Model II procedure	3.4.3.4
[(d)ii]	Logistic-normal binomial model	3.4.2
[(d)iii]	Method based on design effects and effective sample size	3.4.5

4 ANALYSIS OF PINEAPPLE WILT DISEASE INCIDENCE DATA

4.1 Data collection

In order to investigate the effect of certain factors, such as cultivation practices, on the occurrence of pineapple wilt disease and the spatial pattern of this disease, a survey of commercial pineapple plantations was conducted during November 1993 to May 1994 in Gampaha and Kurunegala districts of Sri Lanka (Fig 4.1). Pineapple is a popular fruit crop in Sri Lanka, often grown under coconut in plantations in Gampaha and Kurunegala districts of the low-country (0-300 m altitude) wet-zone (150-250 cm rainfall per annum). Pineapple is a perennial crop, for which the economic life span lasts four years.

Ageing is one possible factor among the many factors that may affect wilt disease incidence on pineapple. Ageing might cause the plants more susceptible to the disease, so that disease incidence might vary on plantations in different years of stand. So we decided to include age ('year of stand') of the crop as a factor in our study.

In an investigation with years of stand of the crop as a factor on disease incidence, we may want to examine contrasts between different years of stand for disease incidence. Pineapple being a perennial crop (with a productive cycle of four years), in order to fully investigate year of stand of the crop as a factor on wilt disease incidence, we would need to examine the crop in four different years of stand. Thus if we want to establish a designed field experiment (field trial) to investigate year of stand of the crop as a factor on disease incidence, we need to establish field plots with year of stand one, year of stand two, year of stand three and year of stand four. Therefore, for establishment of the field trial alone, it takes four years.

On the other hand, if we conduct a survey of established commercial plantations, we have the opportunity of observing disease incidence on plantations in different years of stand. This enables the collection of disease incidence data for various years of stand of the crop in a single season without having to establish plots and waiting for them to mature.

In a purposely established field experiment the field plots are normally substantially smaller than commercial plantations. The size of plots in field experiments very often

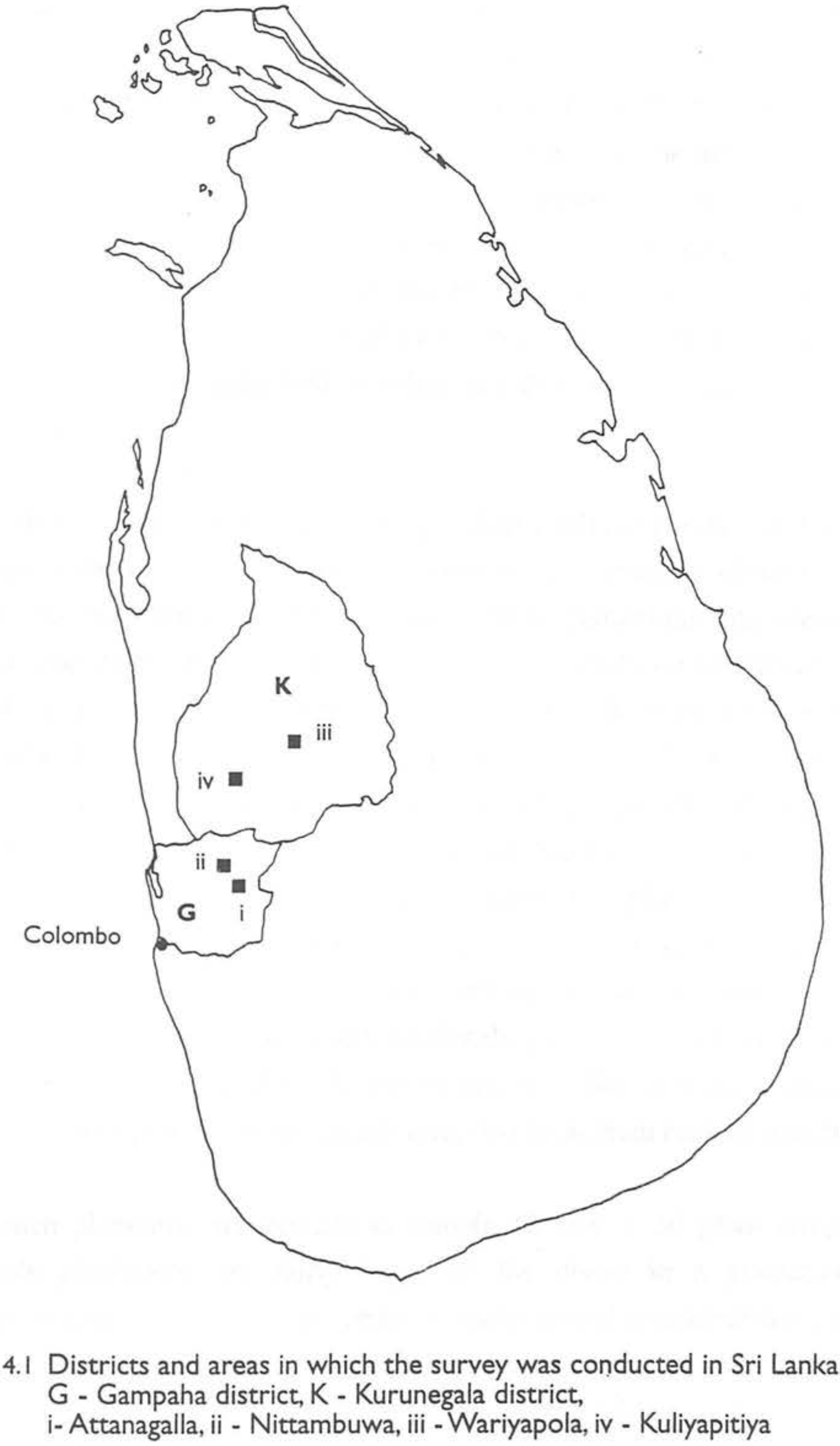


Fig. 4.1 Districts and areas in which the survey was conducted in Sri Lanka.
G - Gampaha district, K - Kurunegala district,
i- Attanagalla, ii - Nittambuwa, iii - Wariyapola, iv - Kuliyapitiya

is restricted by the cost and by practical reasons, such as availability of resources and time, involved in establishing the field plots. The use of relatively small plots could disrupt the spatial patterns of disease incidence that may occur in large plantations. Small plots are more prone to edge effects (per unit area) than large plantations. These edge effects may mask the spatial patterns that we actually observe in large plantations. Thus the spatial patterns we observe in small plots in a field trial may not be the spatial pattern that may occur in large commercial plantations. Moreover, in a field trial it is quite difficult to prevent the field plots being affected by treatments applied in adjacent plots. Because of this interference, unless we control the experiment carefully, the spatial pattern and the disease incidence we observe under particular treatment conditions may not be fully representative of that particular condition alone. After considering all these reasons, we decided to conduct a survey rather than establish a field trial to collect the data on the occurrence of pineapple wilt disease.

In the survey, after visiting district agricultural advice centres of Gampaha and Kurunegala districts, a list was made of commercial plantations where pineapple wilt disease had been reported. We categorised these plantations into areas (villages). Categorising the plantations in this way allows us to make an assumption that all the sites in a given area are subject to reasonably homogeneous environmental conditions, so that sites within the area can be compared for the differences in occurrence of disease under different 'treatment' conditions. From the list of areas, we selected two areas randomly for each district (see Fig 4.1). Since the aim of this study is to apply statistical tests for factorial experiments to plant pathological data, we wanted to obtain data that have a 'factorial structure'. So we chose four plantations from each selected area, from which we could collect the data under such a factorial framework. The details of the conditions that the pineapple crops on these plantations have been subjected are given in the section 4.2. We sampled a total of sixteen plantations, four plantations from each area, two areas from each of two districts.

From each plantation we decided to sample 12 row × 30 plant array. Although pineapple plantations are fairly large, all the plants in a plantations are not necessarily arranged in one large array. Usually several smaller arrays can be found in a plantation. These arrays can be found in various shapes. The 12 row × 30 plants array size was a reasonable size for which we could take random samples from all selected plantations. This made us to decide on this array size for our samples. Moreover, in a similar study on tomato spotted wilt virus disease, Bald (1937) also used same sample size. The reason for choosing a fixed array size is that, then we can avoid possible difficulties concerned with scale-dependence of spatial pattern of disease incidence.

Samples were obtained randomly, one from each plantation. Random number tables were used to select the initial position of the sample in the plantation. Depending on number of rows and plants per row in a chosen plantation, the initial position was selected, giving equal chance to each position (individual plant). If it was impossible to cover an array size of 12 rows × 30 plants per row from the chosen initial position, the initial position was reselected.

The plant arrangement in a plantation is illustrated in the Fig. 4.2. The distance between rows was 1.5 m. Each row consisted of plants arranged in a triangular manner, spaced 45 cm apart (Fig. 4.2). In a well-established plantation the 'double row' is hardly distinguishable.

From each plantation, as explained earlier, a 12 rows × 30 plant array was selected randomly, and the presence or absence of pineapple wilt disease symptoms on each plant was recorded. All the recording (mapping) was done on automated pre-prepared spreadsheets implemented in Microsoft Excel. Thus, a separate map was made for each sample from every plantation examined. No guard rows were considered when collecting data because the randomly selected areas were within established plantations.

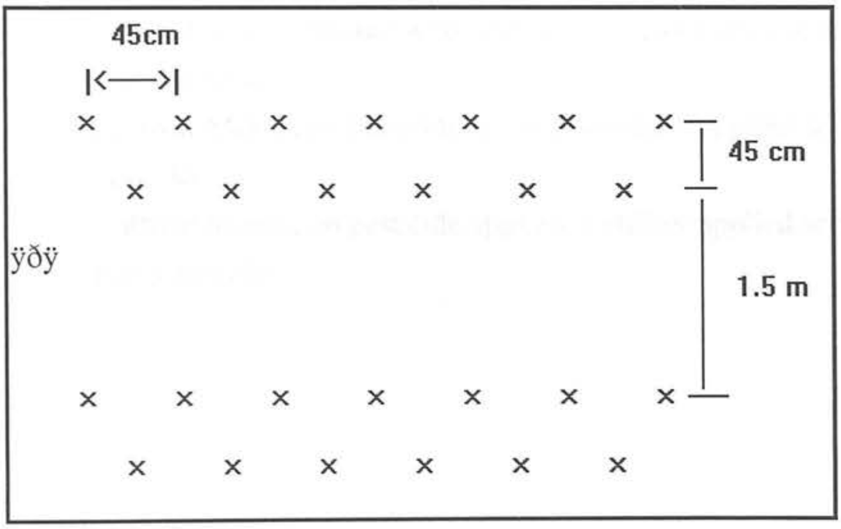


Fig. 4.2 Field plant arrangement of pineapple

4.2 Structure of the collected data

Two different areas in each district were surveyed. For each area, four different plantations were examined. Apart from different cultivation practices in given areas, all the fields had been subjected to similar conditions. The data collected can be summarised as follows.

District - Gampaha; Area - Attanagalla

Plantation	*Conditions*
Kattota	Cultivar Murici, treated with pesticide, fertiliser applied and two years old
Kalagedihena	Cultivar Murici, treated with pesticide, fertiliser applied and four years old
Navadiga	Cultivar Kew, treated with pesticide, fertiliser applied and two years old
Urapola	Cultivar Kew, treated with pesticide, fertiliser applied and four years old

District - Gampaha; Area - Nittambuwa

Plantation	*Conditions*
Walgammana	Cultivar Murici, treated with pesticide, fertiliser applied and one year old
Welhena	Cultivar Murici, treated with pesticide, fertiliser applied and four years old
JEDB	Cultivar Murici, no pesticide applied, fertiliser applied and one year old
Aluwala	Cultivar Murici, no pesticide applied, fertiliser applied and four years old

District - Kurunegala; Area - Wariyapola

Plantation	*Conditions*
Hettipola	Cultivar Murici, no pesticide applied, fertiliser applied and one year old
Panduwasnuwara	Cultivar Murici, no pesticide applied, no fertiliser applied and one year old
Bingiriya	Cultivar Murici, treated with pesticide, fertiliser applied and one year old
Kobeigane	Cultivar Murici, treated with pesticide, no fertiliser applied and one year old

District - Kurunegala; Area - Kuliyapitiya

Plantation	*Conditions*
Devasarana	Cultivar Murici, no pesticide applied, fertiliser applied and one year old
Akkarawatta	Cultivar Murici, treated with pesticide, fertiliser applied and one year old
Munamaldeniya	Cultivar Murici, no pesticide applied, fertiliser applied and four years old
Mukalanyaya	Cultivar Murici, treated with pesticide, fertiliser applied and four years old

Here the application of pesticide refers to application of either profenofos or prothiofos at rates recommended by the Department of Agriculture of Sri Lanka (Department of Agriculture of Sri Lanka, 1993) and application of fertiliser refers to mixture of fertiliser applied at rates recommended by the Department of Agriculture of Sri Lanka (Department of Agriculture of Sri Lanka, 1993). Maps of disease incidence made for each plantation is shown in figures 4.3-4.6.

These data were not from designed experiments. However, since the structure of the data from each area is similar to that of a factorial experiment, and one of the objectives of this study is to illustrate the statistical analysis of disease incidence data, the structure in each area are considered as a 'factorial experiment'. In this study the set-ups in each area is henceforth referred to as factorial experiments. Thus the factorial set-up of each experiment can be summarised as follows.

Factorial experiment 1: District - Gampaha; Area - Attanagalla

2×2 factorial set-up, two types of cultivars, Murici and Kew, and two different years of stand, two year old and four years old.

Factorial experiment 2: District - Gampaha; Area - Nittambuwa

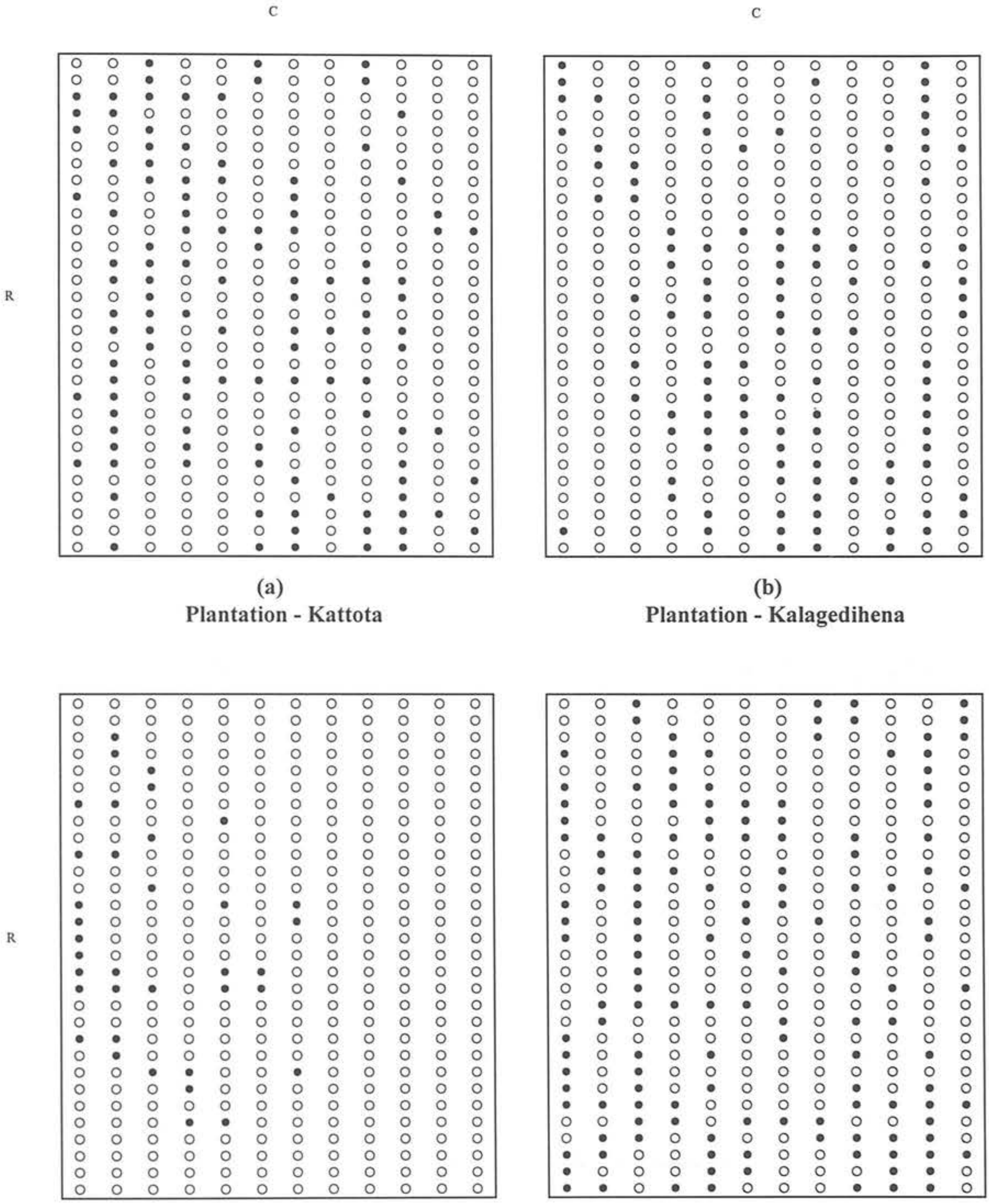
2×2 factorial set-up, application of pesticides, with or without, and two different years of stand, one year old and two years old. Here, pesticides were either profenofos or prothiofos, with the quantities recommended by the Department of Agriculture (Department of Agriculture, 1993).

Factorial experiment 3: District - Kurunegala; Area - Wariyapola

2×2 factorial set-up, application of pesticide, with or without, and application of fertiliser, with or without. Here application of fertiliser refers to the mixture of fertiliser applied in quantities recommended by the Department of Agriculture (Department of Agriculture, 1993).

Factorial experiment 4: District - Kurunegala; Area - Kuliyapitiya

2×2 factorial set-up, application of pesticides, with or without, and two different years of stand, one year old and two years old. Here also, pesticides were as in the second set-up.



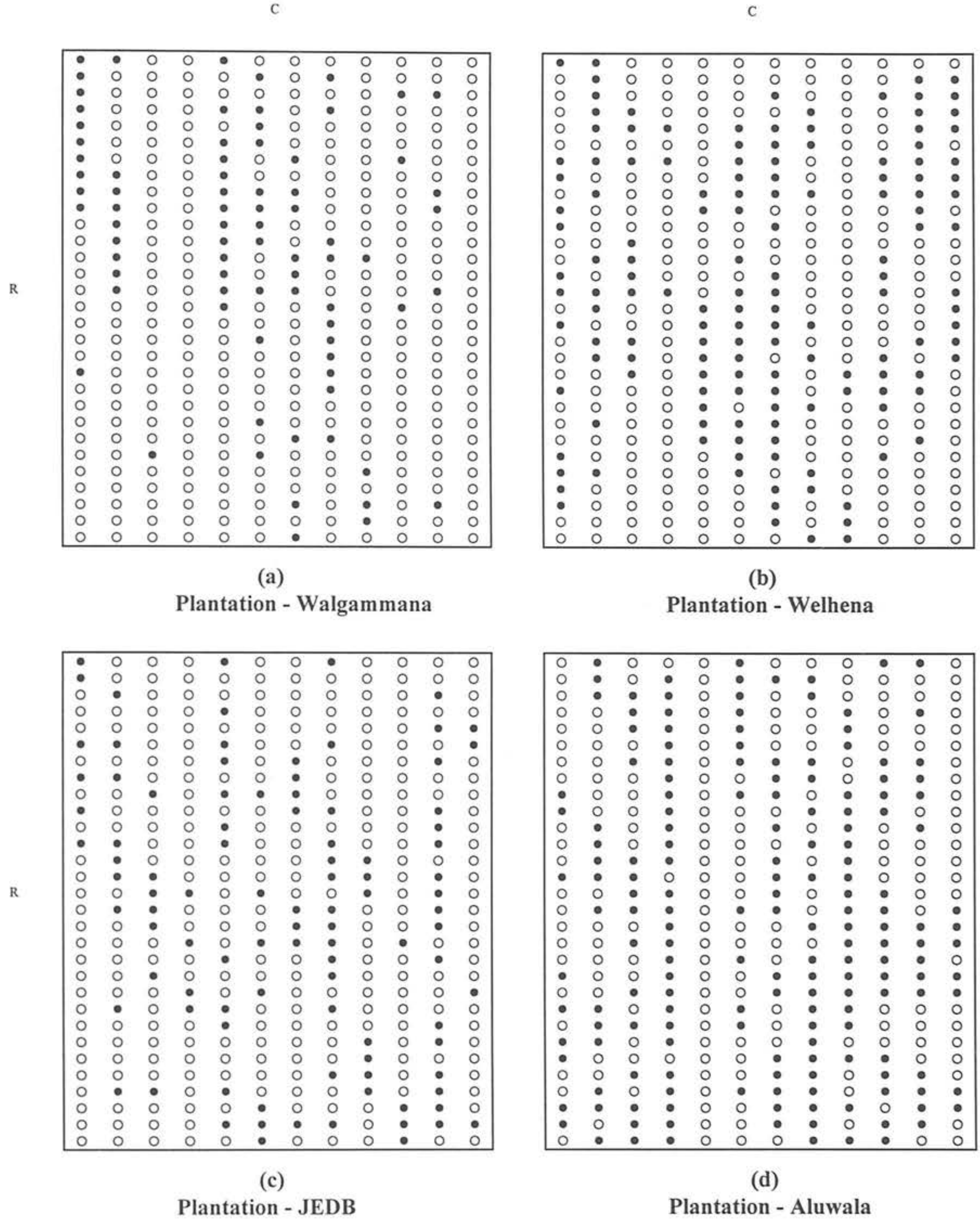
(a)
Plantation - Kattota

(b)
Plantation - Kalagedihena

(c)
Plantation - Navadiga

(d)
Plantation - Urapola

Fig. 4.3 Diseased incidence maps. District - Gampaha, Area - Attanagalla. ● - Infected plant; ○ - Healthy plant; C - Agronomic rows runs vertically; R - Plants within agronomic rows.



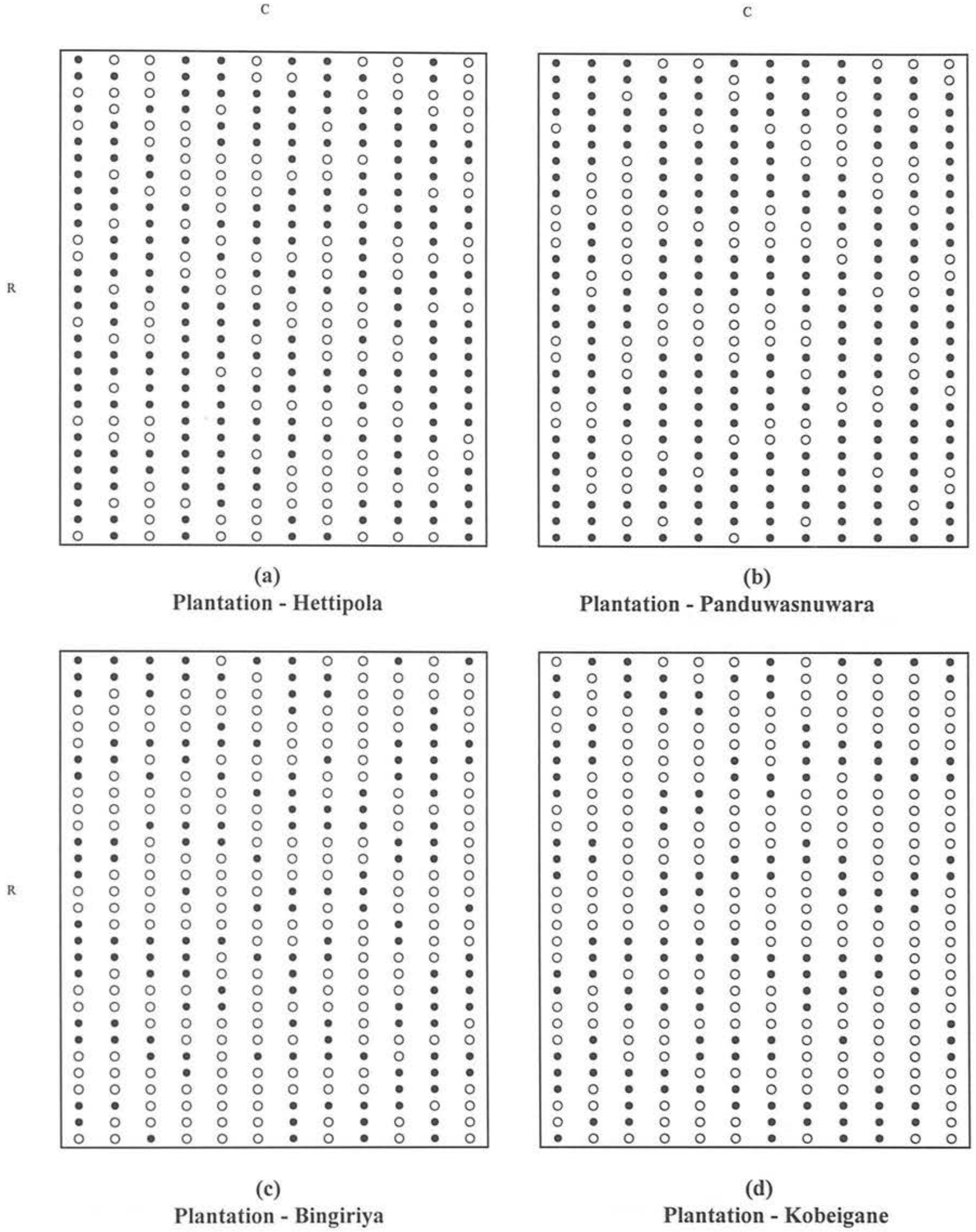
(a)
Plantation - Walgammana

(b)
Plantation - Welhena

(c)
Plantation - JEDB

(d)
Plantation - Aluwala

Fig. 4.4 Diseased incidence maps. District - Gampaha, Area - Nittambuwa. • - Infected plant; o - Healthy plant; C - Agronomic rows runs vertically; R - Plants within rows.



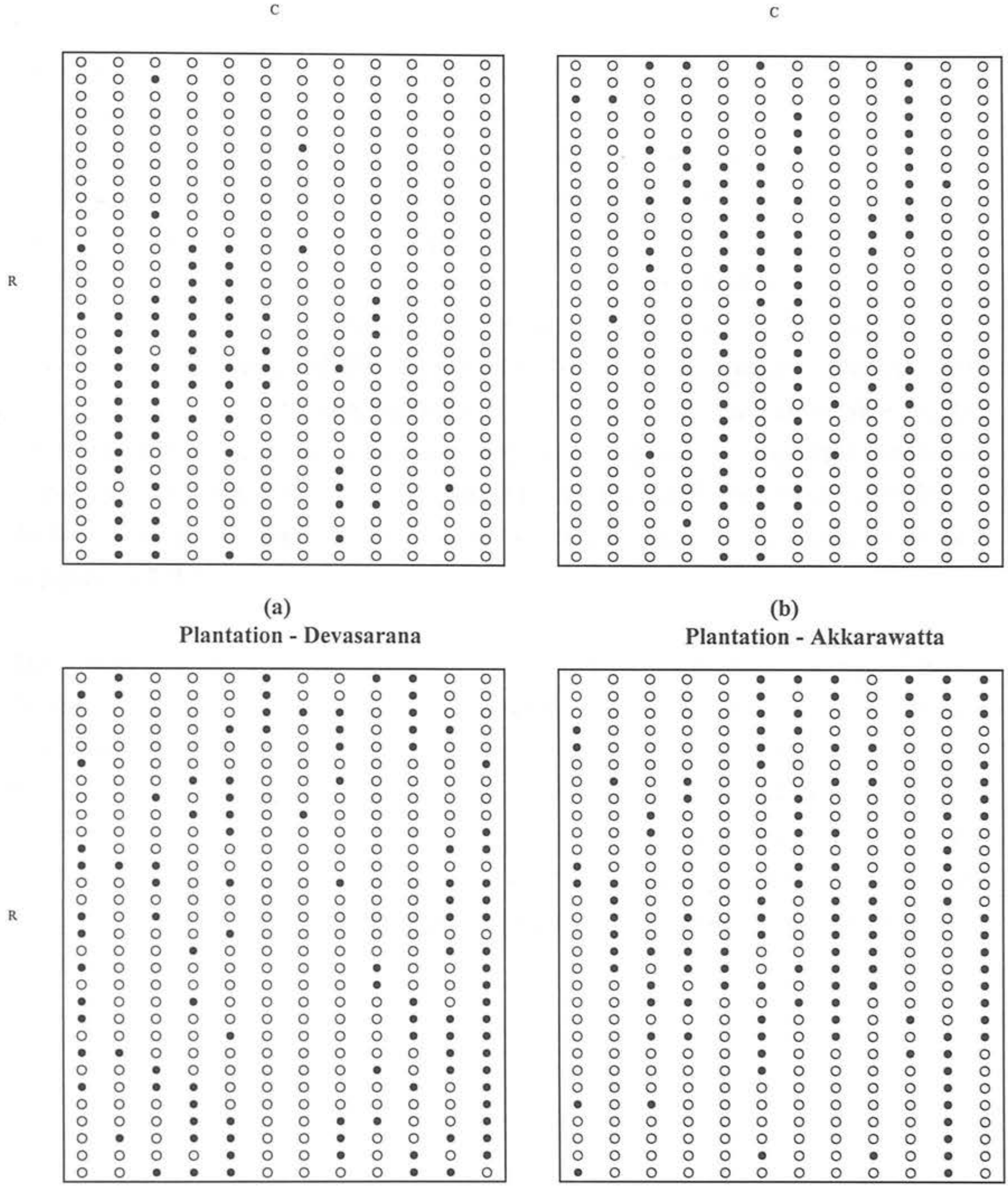
(a)
Plantation - Hettipola

(b)
Plantation - Panduwasnuwara

(c)
Plantation - Bingiriya

(d)
Plantation - Kobeigane

Fig. 4.5 Diseased incidence maps. District - Kurunegala, Area - Wariyapola. • - Infected plant; ○ - Healthy plant; C - Agronomic rows runs vertically; R - Plants within rows.



(a)
Plantation - Devasarana

(b)
Plantation - Akkarawatta

(c)
Plantation - Munamaldeniya

(d)
Plantation - Mukalanyaya

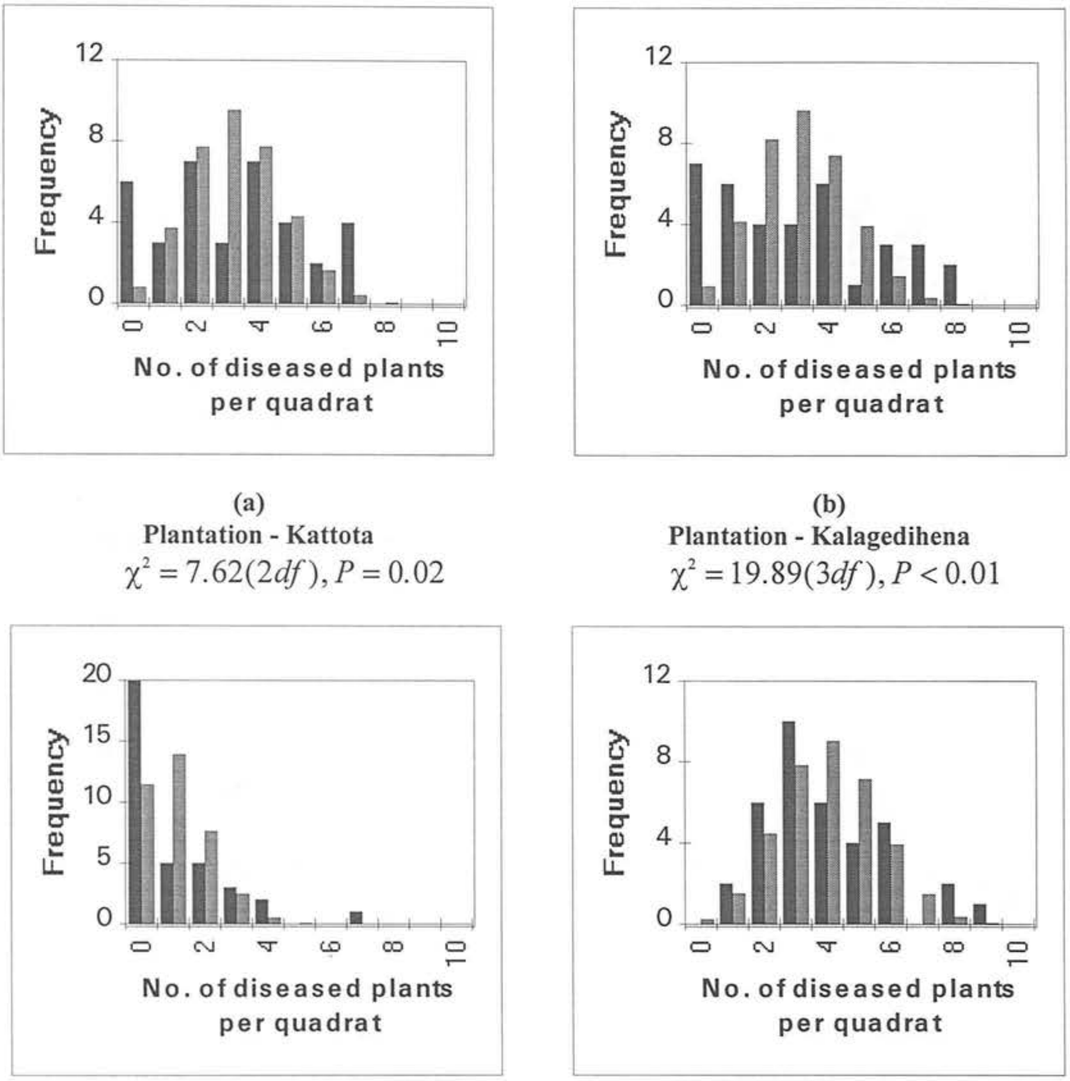
Fig. 4.6 Diseased incidence maps. District - Kurunegala, Area - Kuliyapitiya. ● - Infected plant; ○ - Healthy plant; C - Agronomic rows runs vertically; R - Plants within rows.

4.3 Examination of randomness of the disease incidence data

4.3.1 Fitting the binomial distribution

As described in the chapters 1 and 2, the starting point for the statistical analysis of experiments of this nature is to examine the pattern of disease incidence. In order to investigate pattern, each incidence map was divided into sample units (quadrats) and then observed quadrat frequencies were compared with the expected binomial frequencies. When the maps were divided into quadrats, each map was divided into 36 quadrats, each consisting of 10 plants along a single row. From the incidence maps (Fig. 4.3-4.6) it is noticeable that disease clusters occur more often along the columns (agronomic rows) than across columns. For this reason the shape of the quadrats was made long and thin. As Cochran (1936) reported, quadrat size is usually taken to be 6 to 12 plants. After dividing the field maps into quadrats of size 10, the binomial distribution was fitted to each map separately. Expected binomial frequencies were calculated using Microsoft Excel. Expected frequencies after fitting the binomial distribution, along with observed frequencies for each map, are shown in figures 4.7-4.10.

The goodness-of-fit χ^2 values (Fig. 4.7-4.10) were computed after pooling adjacent frequency classes to make the smallest expectation at least 5. For most of the disease maps the computed χ^2 values were significant ($P < 0.05$), indicating departure from the random occurrence, specifically aggregation of diseased plants in the field. However, for Fig. 4.9(a)-4.9(c) computed χ^2 values were not greater than expected ($P > 0.05$), and indicated a random occurrence of diseased plants in this particular case. Thus, it may be concluded that, in general, spatial pattern of pineapple wilt disease is aggregated.



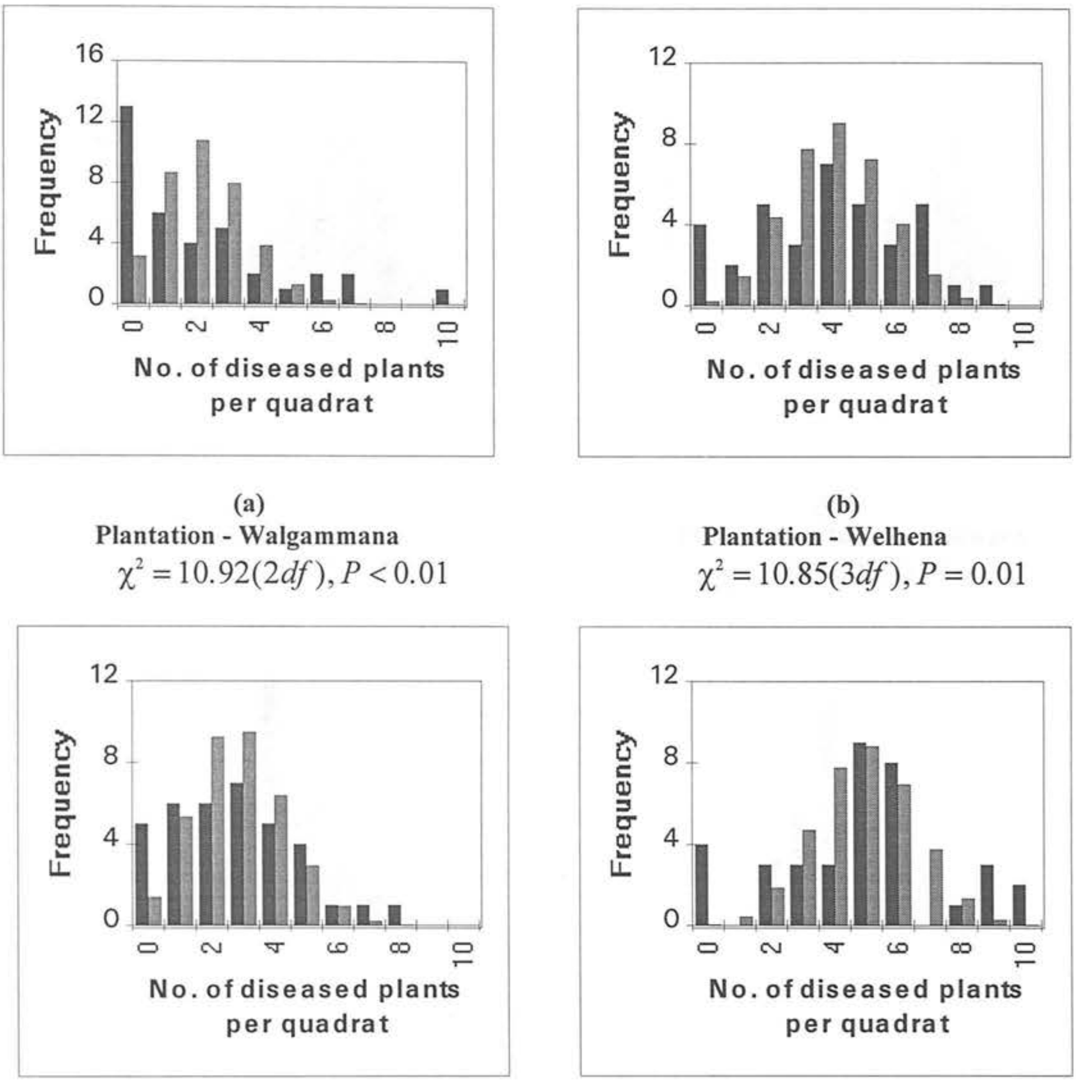
(a)
Plantation - Kattota
$\chi^2 = 7.62(2df), P = 0.02$

(b)
Plantation - Kalagedihena
$\chi^2 = 19.89(3df), P < 0.01$

(c)
Plantation - Navadiga
$\chi^2 = 12.12(1df), P < 0.01$

(d)
Plantation - Urapola
$\chi^2 = 4.35(3df), P = 0.23$

Fig. 4.7 Expected (after fitting binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.3. ■- Observed ▨- Expected.



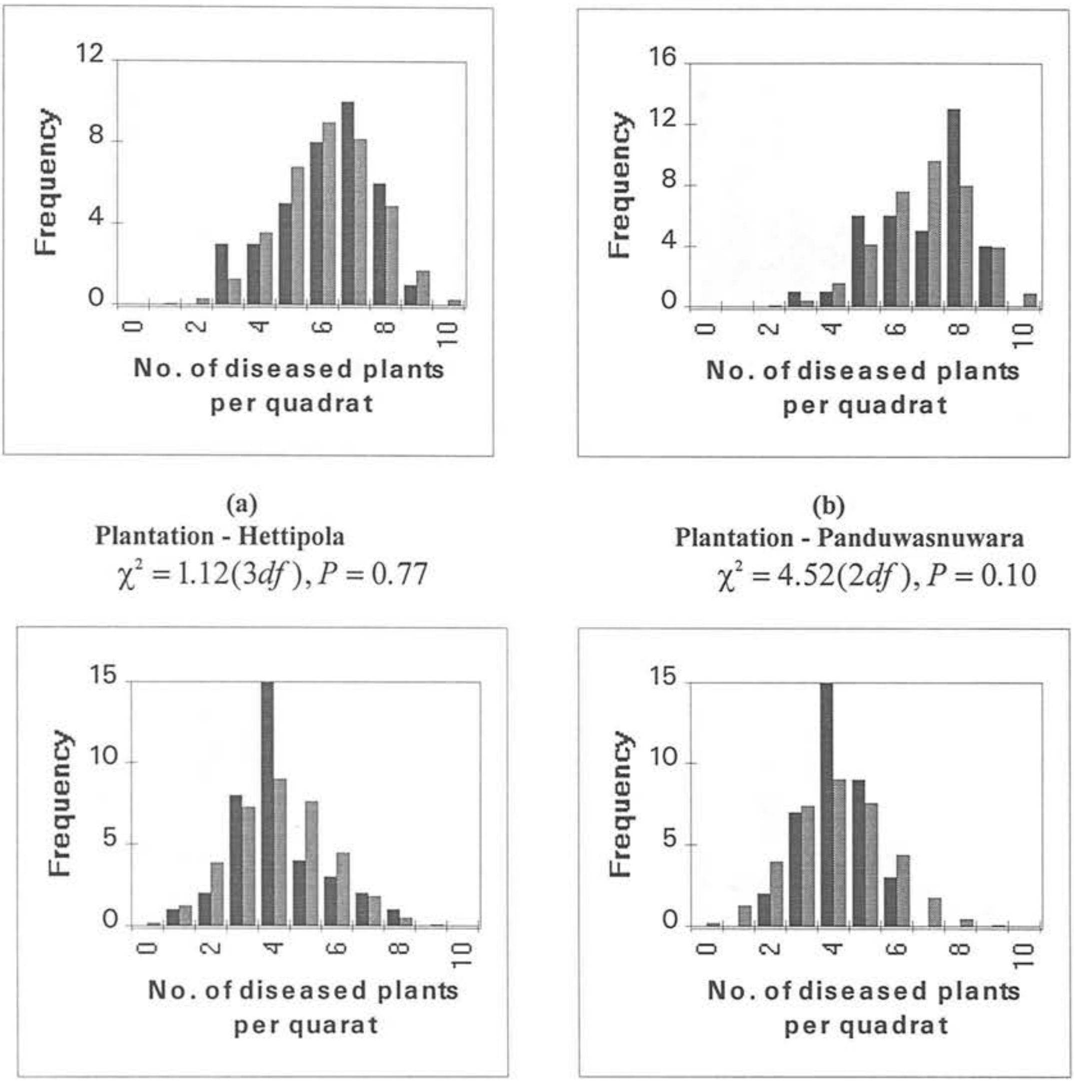
(a)
Plantation - Walgammana
$\chi^2 = 10.92(2df), P < 0.01$

(b)
Plantation - Welhena
$\chi^2 = 10.85(3df), P = 0.01$

(c)
Plantation - JEDB
$\chi^2 = 4.70(2df), P = 0.10$

(d)
Plantation - Aluwala
$\chi^2 = 4.4(3df), P = 0.22$

Fig. 4.8 Expected (after fitting binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.4. ■ - Observed ▨ - Expected.



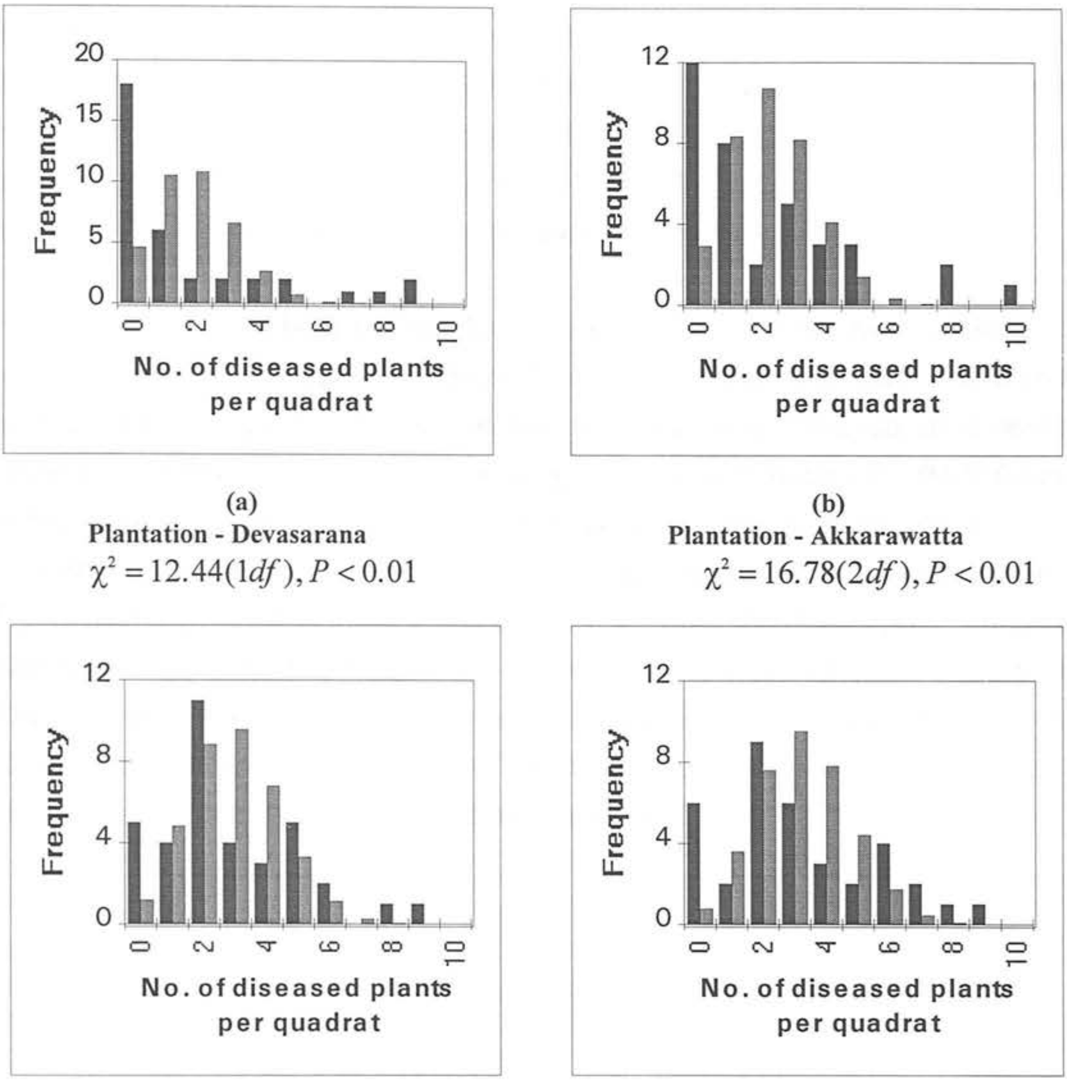
(a)
Plantation - Hettipola
$\chi^2 = 1.12(3df), P = 0.77$

(b)
Plantation - Panduwasnuwara
$\chi^2 = 4.52(2df), P = 0.10$

(c)
Plantation - Bingiriya
$\chi^2 = 7.06(3df), P = 0.07$

(d)
Plantation - Kobeigane
$\chi^2 = 8.45(3df), P = 0.04$

Fig. 4.9 Expected (after fitting binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.5. ■- Observed ▨- Expected.



(a)
Plantation - Devasarana
$\chi^2 = 12.44(1df), P < 0.01$

(b)
Plantation - Akkarawatta
$\chi^2 = 16.78(2df), P < 0.01$

(c)
Plantation - Munamaldeniya
$\chi^2 = 5.26(2df), P = 0.07$

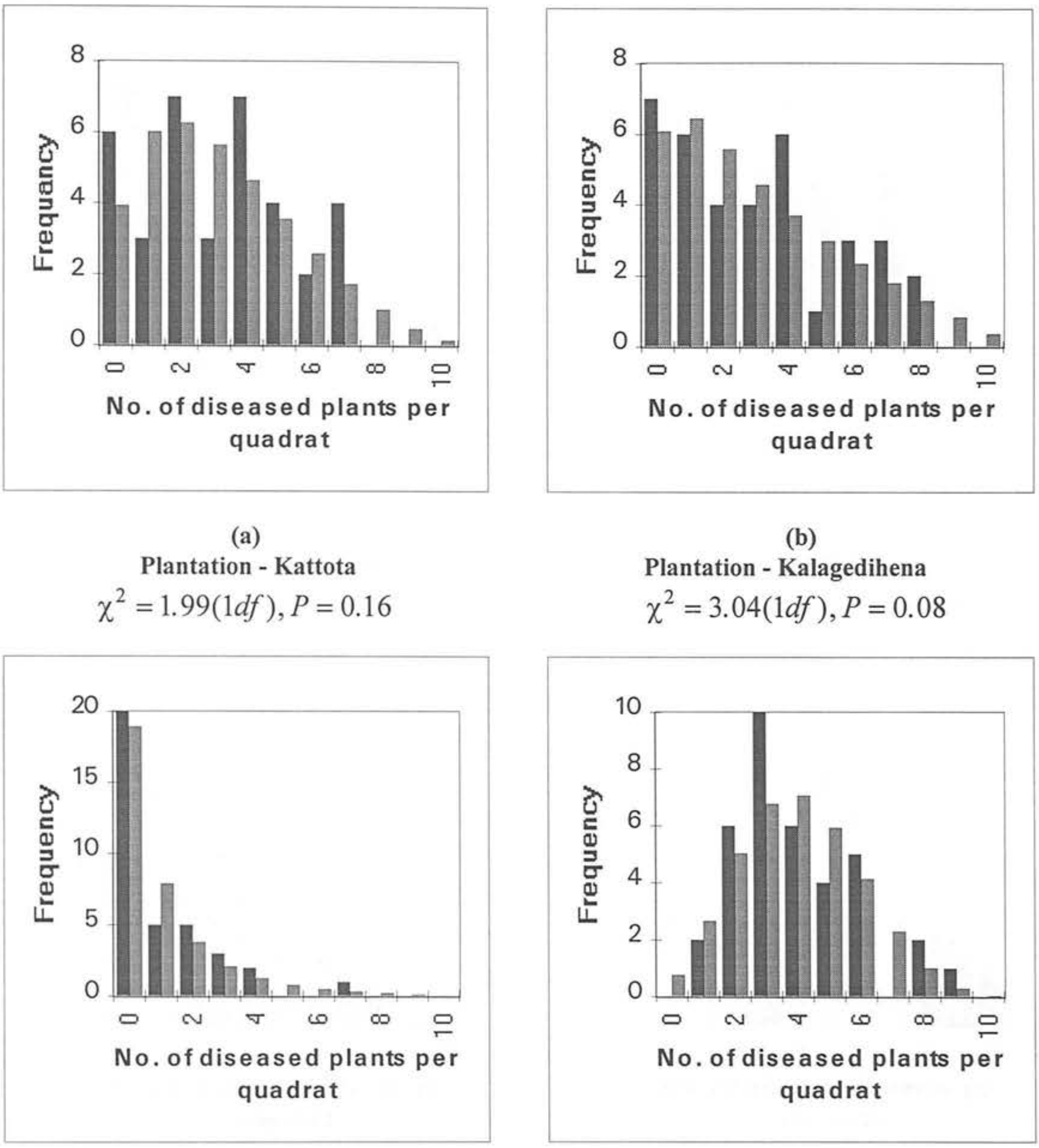
(d)
Plantation - Munamaldeniya
$\chi^2 = 8.03(2df), P = 0.02$

Fig. 4.10 Expected (after fitting binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.6. ■- Observed ▨- Expected.

4.3.2 Fitting the logistic-normal binomial distribution

As explained in chapter 3, when the disease incidence is not random, the binomial distribution cannot adequately describe the observed frequencies. For aggregated disease incidence, an adequate description of observed frequencies may be obtained by fitting either the beta-binomial distribution or the logistic-normal binomial distribution. The logistic-normal binomial distribution was chosen for this study for the reasons explained in section 3.5. When the logistic-normal binomial distribution is fitted to the incidence maps (Fig. 4.3-4.6), the expected frequencies, along with observed frequencies, are shown in Fig. 4.11-4.14. EGRET in association with MATHCAD 5.0+ was used to obtain expected frequencies.

From the Fig. 4.11-4.14 it is clear that expected frequencies are much closer to the observed frequencies than the expected binomial frequencies. Thus the logistic-normal binomial distribution is superior to binomial distribution in describing observed frequencies. The goodness-of-fit χ^2 was non-significant ($P > 0.05$) for most of the incidence maps, confirming these results. The probabilities for goodness-of-fit χ^2 were computed following the same principles as with binomial fitting except for Fig. 4.12(d), in which pooling was done only until the smallest expectation greater than 4. The reason for this is that none of the frequency classes had a frequency of 5. The goodness-of-fit χ^2 tests for each incidence map are given in Table 4.1. This table clearly shows the overall superiority of logistic-normal distribution over the binomial distribution for describing aggregated incidence data.



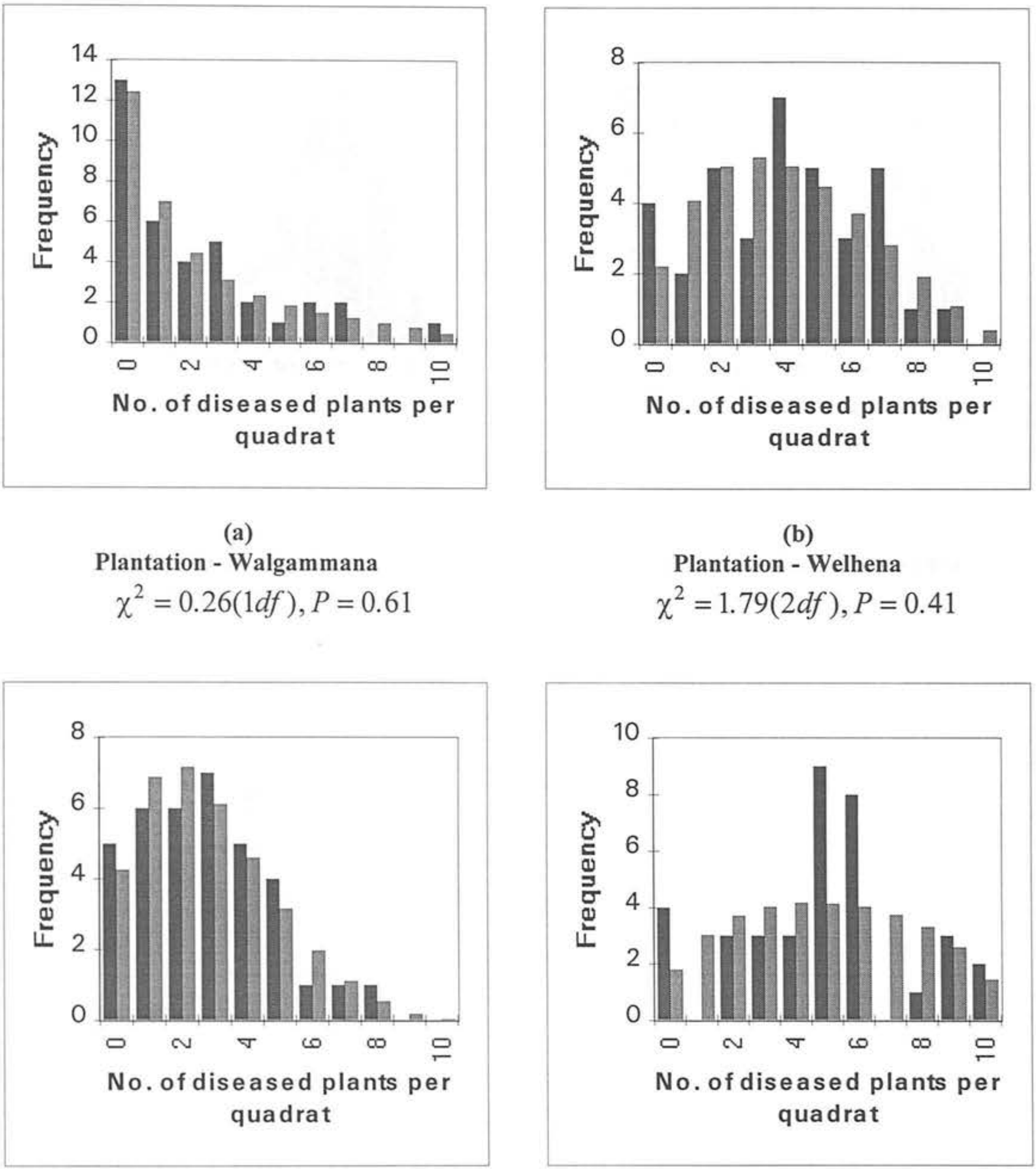
(a)
Plantation - Kattota
$\chi^2 = 1.99(1df), P = 0.16$

(b)
Plantation - Kalagedihena
$\chi^2 = 3.04(1df), P = 0.08$

(c)
Plantation - Navadiga
$\chi^2 = 1.57(1df), P = 0.21$

(d)
Plantation - Urapola
$\chi^2 = 2.39(2df), P = 0.30$

Fig. 4.11 Expected (after fitting logistic-normal binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.3. ■ - Observed ■ - Expected.



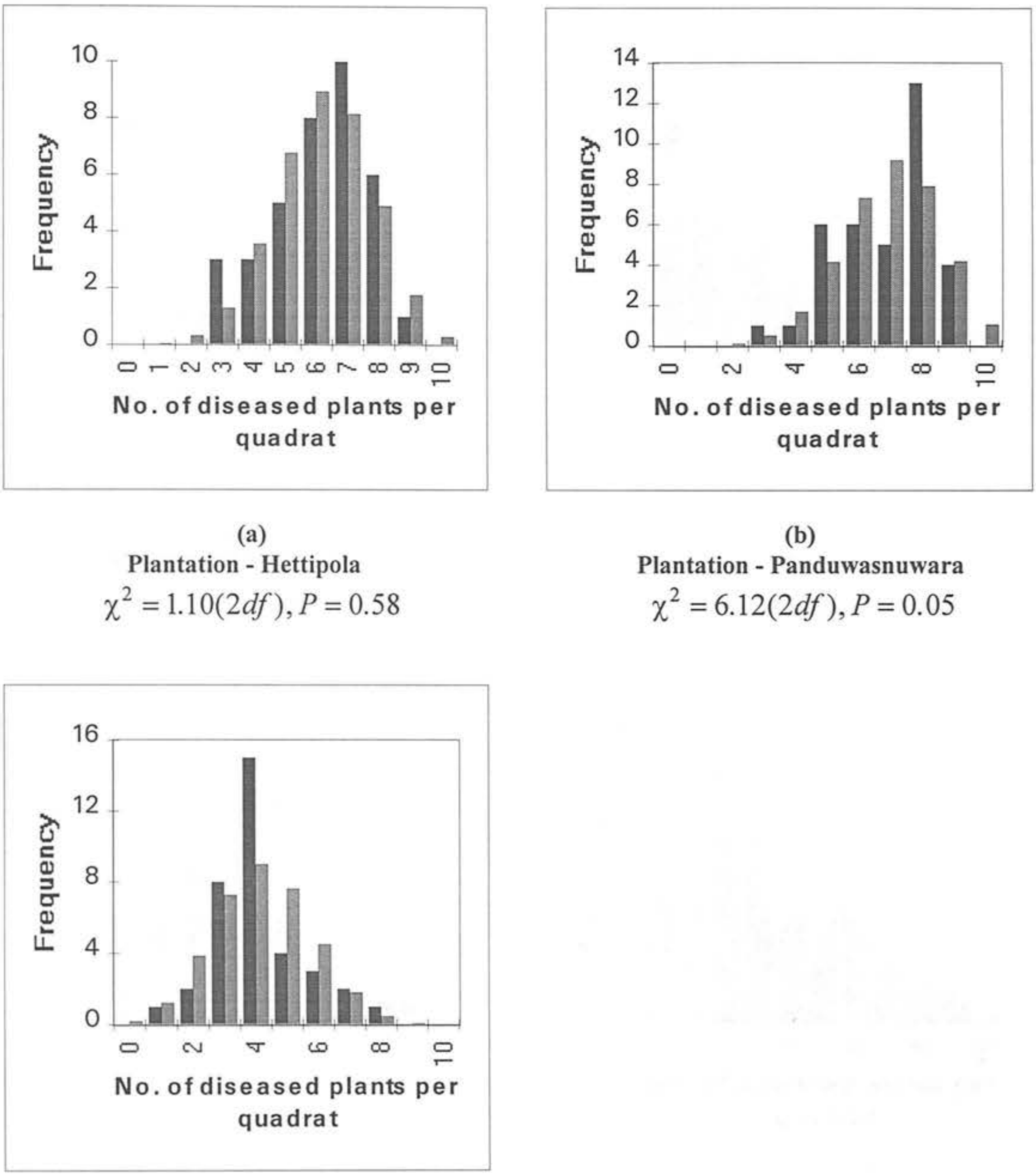
(a)
Plantation - Walgammana
$\chi^2 = 0.26(1df), P = 0.61$

(b)
Plantation - Welhena
$\chi^2 = 1.79(2df), P = 0.41$

(c)
Plantation - JEDB
$\chi^2 = 0.37(1df), P = 0.55$

(d)
Plantation - Aluwala
$\chi^2 = 6.64(1df), P = 0.01$

Fig. 4.12 Expected (after fitting logistic-normal binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.4. ■- Observed ■- Expected.

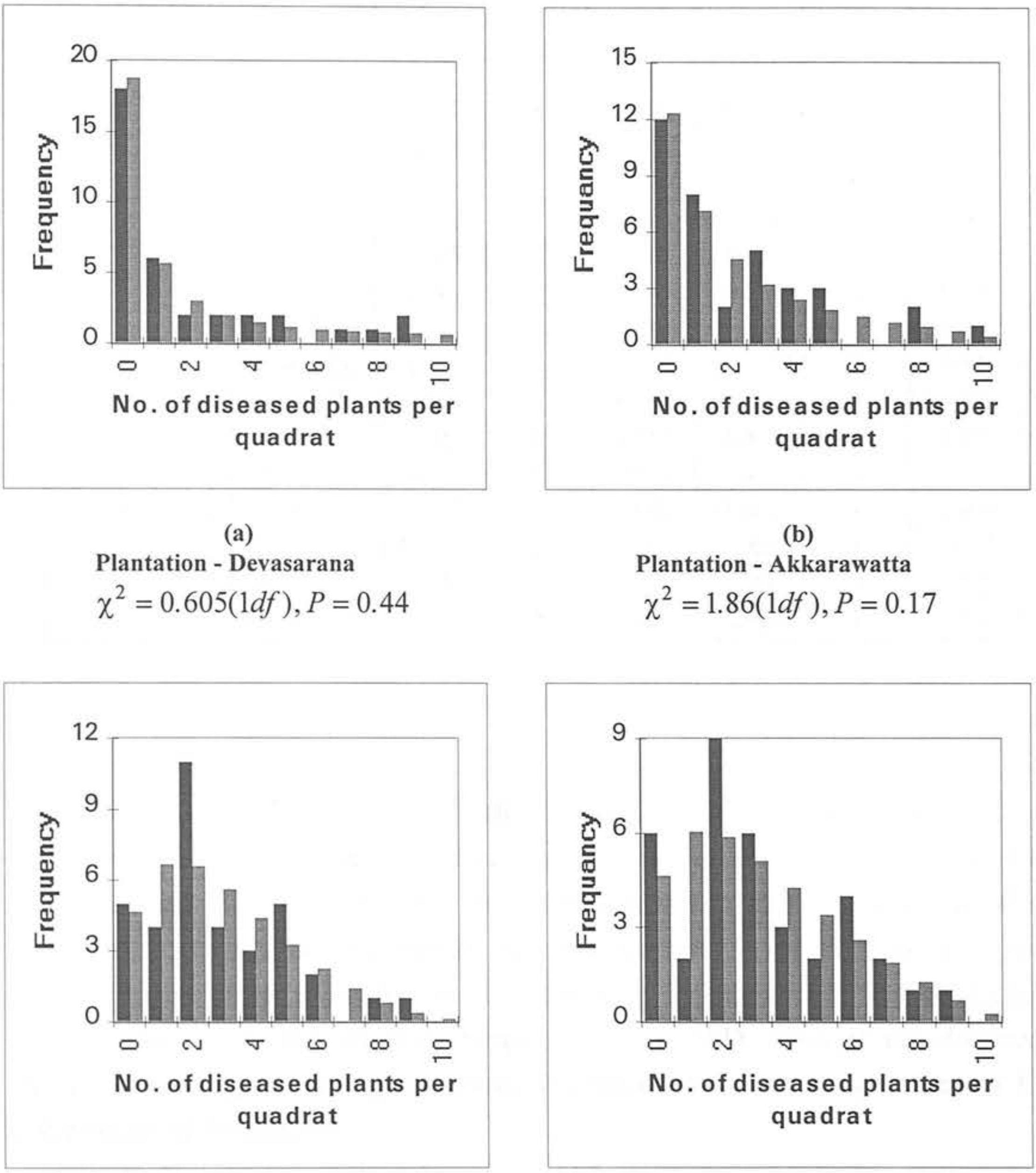


(a)
Plantation - Hettipola
$\chi^2 = 1.10(2df), P = 0.58$

(b)
Plantation - Panduwasnuwara
$\chi^2 = 6.12(2df), P = 0.05$

(c)
Plantation - Bingiriya
$\chi^2 = 6.87(2df), P = 0.03$

Fig. 4.13 Expected (after fitting logistic-normal binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.5. ■- Observed ▨- Expected.



(a)
Plantation - Devasarana
$\chi^2 = 0.605(1df), P = 0.44$

(b)
Plantation - Akkarawatta
$\chi^2 = 1.86(1df), P = 0.17$

(c)
Plantation - Munamaldeniya
$\chi^2 = 3.91(1df), P = 0.05$

(d)
Plantation - Mukalanyaya
$\chi^2 = 2.62(1df), P = 0.11$

Fig. 4.14 Expected (after fitting logistic-normal binomial distribution) and the observed disease frequencies for the diseased maps in the Fig. 4.6. ■- Observed ▨- Expected.

Table 4.1 Goodness-of-fit χ^2 tests for disease incidence maps (Fig. 4.3-4.6)

Factorial experiment No.	Incidence map	Goodness-of-fit (binomial)			Goodness-of-fit (logistic-normal binomial)		
		χ^2	d.f.	$P > \chi^2$	χ^2	d.f.	$P > \chi^2$
1	Fig. 4.3 (a)	7.62	2	0.02	1.99	1	0.16
	Fig. 4.3 (b)	19.89	3	<0.01	3.04	1	0.08
	Fig. 4.3 (c)	12.12	1	<0.01	1.57	1	0.21
	Fig. 4.3 (d)	4.35	3	0.23	2.39	2	0.30
2	Fig. 4.4 (a)	10.92	2	<0.01	0.26	1	0.61
	Fig. 4.4 (b)	10.85	3	0.01	1.79	2	0.41
	Fig. 4.4 (c)	4.70	2	0.10	0.37	1	0.55
	Fig. 4.4 (d)	4.40	3	0.22	6.64	1	0.01
3	Fig. 4.5 (a)	1.12	3	0.77	1.10	2	0.58
	Fig. 4.5 (b)	4.52	2	0.10	6.12	2	0.05
	Fig. 4.5 (c)	7.06	3	0.07	6.87	2	0.03
	Fig. 4.5 (d)	8.45	3	0.04	-	-	-
4	Fig. 4.6 (a)	12.44	1	<0.01	0.605	1	0.44
	Fig. 4.6 (b)	16.78	2	<0.01	1.86	1	0.17
	Fig. 4.6 (c)	5.26	2	0.07	3.91	1	0.05
	Fig. 4.6 (d)	8.03	2	0.02	2.62	1	0.11

The incidence map for plantation Walgammana (Fig. 4.4a) is a typical case of overdispersion. The clusters of diseased plants can easily be identified in the incidence map. The corresponding binomial distribution fit (Fig. 4.8a) for this incidence map clearly indicates that the observed frequency is very different from the expected binomial fit ($P < 0.01$). The corresponding logistic-normal fit (Fig. 4.12a) gave a better fit to the observed frequencies ($P = 0.61$). Thus if the diseased incidence is aggregated the logistic-normal binomial distribution can give a better fit to the observed frequencies.

The incidence map for plantation JEDB (Fig. 4.4c) is a typical incidence map for a random pattern of disease incidence. In the incidence map it is easy to identify individual diseased plants, as well as diseased plants in groups, with varying group sizes, all throughout the map. The corresponding binomial distribution fit (Fig. 4.8c) shows that the expected binomial frequencies are similar to observed frequencies

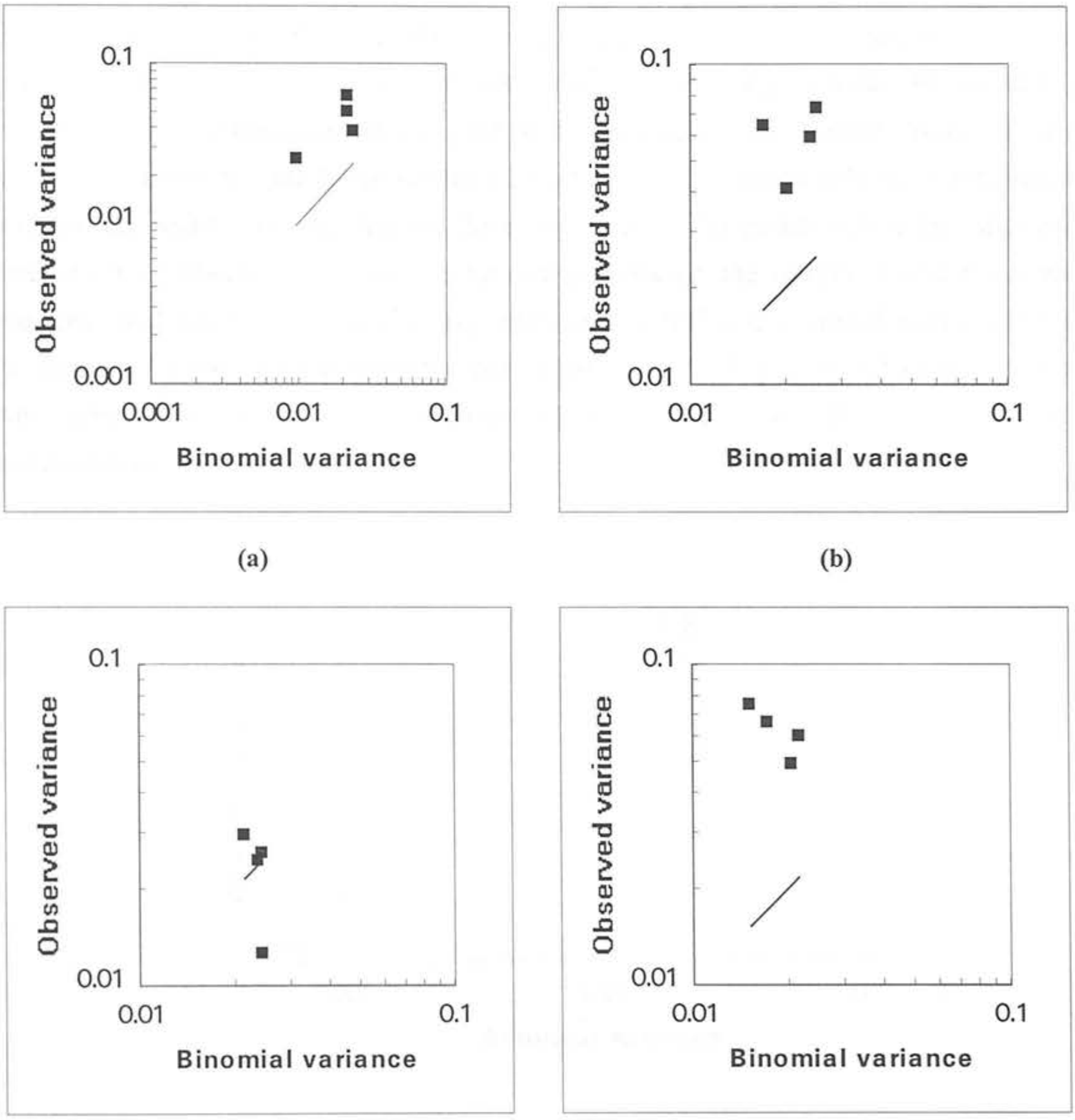
($P = 0.10$). This confirms that if the disease incidence is random, the observed frequencies can adequately be described by the binomial distribution. However, the corresponding logistic-normal binomial distribution (Fig. 4.12c) also produced an adequate fit ($P = 0.55$) for the observed frequencies.

The incidence map for plantation Aluwala (Fig. 4.4d) is also an example for a situation where the disease incidence is random. The binomial fit (4.8d) clearly showed that expected binomial frequencies are similar equal to observed frequencies ($P = 0.22$). The corresponding logistic-normal distribution (Fig. 4.12d) however failed to provide an adequate fit ($P = 0.01$) indicating that binomial distribution sometimes can give better fits to random patterns.

The incidence map for plantation Kobeigane (Fig. 4.5d) is a situation for a regular pattern of disease incidence. Regular disease incidence results underdispersion. In the event of an underdispersion, the goodness-of-fit χ^2 for a binomial distribution is less than its expected value (d.f.). However, in the Table 4.1 and Fig. 4.9d, the corresponding χ^2 value is larger than its d.f. ($P = 0.04$). The reason for this is that the χ^2 values in the Table 4.1 were obtained after pooling the frequency classes in order to have the expected frequency of at least 5. The logistic-normal binomial distribution cannot be produced when the disease incidence is underdispersed because the algorithm adopted in EGRET does not converge under underdispersion (EGRET, 1990). This is the reason that in Fig. 4.13 the logistic-normal binomial distribution fit corresponding to Fig. 4.5d has not been made.

4.3.3 Variance-variance plots.

In section 3.5 we showed that variance-variance plots can be used to describe spatial pattern of disease incidence and that this could be used as a guideline from which to choose the appropriate statistical technique in the analysis of incidence data. The variance-variance plots for each factorial experiment (section 4.1) are shown in Fig. 4.15. Each point in a plot represents a field map that belongs to that particular factorial experiment (section 4.1). These plots were made using Microsoft EXCEL, as explained in section 3.5.



(a)

(b)

(c)

(d)

Fig. 4.15 Variance-variance plots for the diseased maps in the Fig. 4.3-4.6. Plot (a) corresponding to Fig. 4.3; plot (b) corresponding to Fig. 4.4; plot (c) corresponding to Fig. 4.5; plot (d) corresponding to Fig. 4.6 (All graphs are in log-log scale).

In the variance-variance plots, the continuous line represents the expected binomial variance. In Fig. 4.15 all the scatter points except one lie above the binomial line. Thus some degree of overdispersion is clear with all maps except one. In Fig. 4.15(c), the point below the binomial line indicates that the observed variance is smaller than the expected binomial variance. This indicates underdispersion of disease incidence in this particular map. This scatter point represents the incidence map Fig. 4.5(d), and this is the map for which EGRET fails to compute logistic-normal distribution parameters.

Fig. 4.16 summarises the four plots in Fig. 4.15. This is a useful summary which gives an overview of dispersion of disease incidence. Thus Fig. 4.6 may be useful for making general comments on the pattern of pineapple wilt disease. Since all the scatter points except one lie above the binomial line, the spatial pattern of pineapple wilt disease seems to be aggregated. Moreover, the scatter points above the binomial line, do not indicate a consistent relationship between the observed and binomial variance, and this implies variable aggregation over different factorial combinations. In contrast, in the variance-variance plot in section 3.5 (Fig. 3.5), all scatter points were parallel to the binomial line indicating consistent aggregation in all factorial combinations.

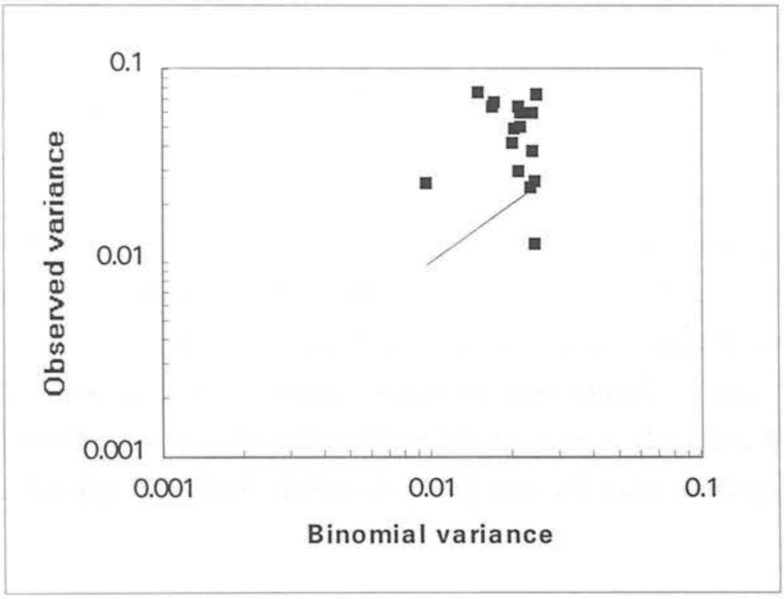


Fig. 4.16 Variance-variance plot (in log-log scale) for the diseased maps in Fig. 5.3-5.6 (pooled)

4.4 Investigating factorial effects.

To investigate the effect of the factors and their interactions on disease incidence, models were fitted to the data from four factorial experiments. Since the data collection for each of the four experiments was done separately, the analyses were carried out separately. As the initial step, linear logistic models were fitted to the data for each factorial experiment. The residual deviances, after fitting the models with all main effects and the interactions were much greater than expected, except for the third factorial experiment (Table 4.1-4.4). A large residual deviance indicates overdispersion, and that the linear logistic model is not suitable as a basis to analyse such data. Thus one of the procedures discussed in section 3.4 has to be used to model the data. As discussed in section 3.5, since each factorial experiment has 2×2 treatment combinations, i.e. there were only four points in the variance-variance plot (Fig. 4.15), and all the points in the plots (except in the third experiment) lay above the binomial line, EGRET was chosen to perform the analysis. Among the possible two types of models on EGRET, the beta-binomial model and the logistic-normal model, the logistic-normal model was chosen to perform the analysis, for reasons noted in the section 3.5.

4.4.1 Statistical model fitting for the first experiment

This experiment consisted of a 2×2 factorial. Factor one (X1), was the cultivar with 2 cultivars, Murici and Kew. Factor two (X2), is the year of stand with two different years of stand, year two and year four.

The deviances after fitting possible models to the data from this experiment using EGRET are shown in Table 4.2(a). In Table 4.2(a), the models which do not include random effect parameters refer to fitting linear logistic models and the models with random effect parameters refer to fitting logistic-normal models. Table 4.2(b) uses respective likelihood ratio tests based on these fits to analyse the data. The models equivalent to all the fits and interpretation of all the tests are same as described in the section 3.4.1.

From Table 4.2(b) it is evidently the saturated model (fit K) that is the most appropriate model to describe the data. That is, all main effects (difference between cultivars, and difference between year of stands) and the interaction between cultivars and year of stand are present with respect to disease incidence. In addition, varying aggregation among different factorial combinations were also apparent. The conclusion from this analysis is that disease incidence varies depending on the cultivar and year of stand, and the differences between two cultivars is influenced by the year of stand. Furthermore, the aggregation varies depending on the treatment combination applied. The parameter estimates and their standard errors for the best fit, along with the corresponding linear logistic model parameters and their standard errors are given in the Table 4.3. As noted in section 3.4.1, the parameter estimates of the best fit are similar to those of the corresponding linear logistic model but the standard errors have been inflated. Since these inflated standard errors take into account of aggregation, these standard errors should be used for significant testing and deriving confidence intervals as illustrated in section 3.2.2.

Table 4.2 Possible models and likelihood ratio tests for the data in the first factorial experiment.

(a) Possible models

Fit	Fixed effects parameters	Random effects parameters	d.f.	Deviance
A	%GM		143	458.113
B	%GM,X1		142	452.054
C	%GM,X2		142	423.541
D	%GM,X1,X2		141	417.333
E	%GM,X1,X2,X1.X2		140	368.540
F	%GM	%SCL	142	332.975
G	%GM,X1	%SCL	141	330.714
H	%GM,X2	%SCL	141	318.322
I	%GM,X1,X2	%SCL	140	315.079
J	%GM,X1,X2,X1.X2	%SCL	139	292.589
K	%GM,X1,X2,X1.X2	%SCL,X1,X2,X1.X2	136	267.975

(b) Possible likelihood ratio tests

Test	Description of the test	Compare fits	Likelihood ratio statistic	d.f.	*P* value
1	Test for effect of factor1 adjusted for the other factor	C vs. D	6.207	1	0.01
2	Test for effect of factor2 adjusted for the other factor	B vs. D	34.720	1	<0.01
3	Test for interaction adjusted for both factors	D vs. E	48.793	1	<0.01
4	Test for excess variation	E vs. J	75.951	1	<0.01
5	Test for factor 1 (adjusted) in the presence of excess variation	H vs. I	3.243	1	0.07
6	Test for factor 2 (adjusted) in the presence of excess variation	G vs. I	15.635	1	<0.01
7	Test for interaction (adjusted) in the presence of excess variation	I vs. J	22.491	1	<0.01
8	Test for differing levels of excess variation	J vs. K	24.614	3	<0.01

Table 4.3 Estimates and standard error of estimates of the best fit (K) with corresponding linear logistic model parameters and their standard errors.

Factorial combination	Under logistic-normal binomial model		Under linear logistic model	
	Estimate (η)	S.E	Estimate (η)	S.E
μ	-0.9266	0.211	-0.7691	0.113
Cv. Kew	-1.953	0.481	-0.0518	0.161
Yr. 2	-0.1669	0.338	-1.339	0.204
Cv. Kew*Yr. 2	2.605	0.567	1.743	0.257

4.4.2 Statistical model fitting for the second experiment

This experiment also consisted of a 2×2 factorial. Factor one (X1) was the application of pesticide, with 2 levels, with pesticide or without pesticide. The pesticide had contained either prothiofos or profenofos at recommended dosage (Department of Agriculture of Sri Lanka, 1993). Factor two (X2) was the year of stand, with two different years of stand, year one and year two.

As with the experiment one, deviances after fitting possible models are shown in Table 4.4(a), and Table 4.4(b) uses respective likelihood ratio tests based on these tests to analyse the data.

From Table 4.4(b), the model with both main effects and common aggregation parameter (fit I) seems to be the most appropriate model to describe the data.

It is important to note that if one takes the critical probability limit as 0.10 instead of 0.05, test 8 in the Table 4.4(b) might be interpreted as aggregation varying with treatment combination including interaction. However, since interaction between factors is not evident it is not very sensible to include the varying aggregation (including interaction) component into the model. Since only main effects are significant one may attempt to fit model K. But test 9 suggests that varying aggregation (without interaction) is not significant and therefore the most appropriate model is the model with both main effects and common aggregation parameter (fit I). Thus the conclusion is that both application of pesticide and year of stand affect the disease incidence, but the effect of application of pesticide does not depend on year of stand. Moreover, aggregation does not vary with the treatment combination imposed.

The parameter estimates and standard error of estimates for the best fit along with the corresponding linear logistic model parameters and their standard errors are given in the Table 4.5. As explained with the experiment 1, the estimates and standard errors of the best fit have to be used for significant testing and deriving confidence intervals.

Table 4.4 Possible models and likelihood ratio tests for the data in the second factorial experiment.

(a) Possible models

Fit	Fixed effects parameters	Random effects parameters	d.f.	Deviance
A	%GM		143	530.319
B	%GM,X1		142	521.676
C	%GM,X2		142	468.600
D	%GM,X1,X2		141	459.573
E	%GM,X1,X2,X1.X2		140	459.565
F	%GM	%SCL	142	360.354
G	%GM,X1	%SCL	141	357.037
H	%GM,X2	%SCL	141	339.755
I	%GM,X1,X2	%SCL	140	335.787
J	%GM,X1,X2,X1.X2	%SCL	139	335.769
K	%GM,X1,X2	%SCL,X1,X2	138	334.338
L	%GM,X1,X2,X1.X2	%SCL,X1,X2,X1.X2	136	328.836

(b) Possible likelihood ratio tests

Test	Description of the test	Compare fits	Likelihood ratio statistic	d.f.	*P* value
1	Test for effect of factor1 adjusted for the other factor	C vs. D	9.028	1	<0.01
2	Test for effect of factor2 adjusted for the other factor	B vs. D	62.10	1	<0.01
3	Test for interaction adjusted for both factors	D vs. E	0.008	1	0.93
4	Test for excess variation	E vs. J	123.796	1	<0.01
5	Test for factor 1 (adjusted) in the presence of excess variation	H vs. I	3.968	1	<0.05
6	Test for factor 2 (adjusted) in the presence of excess variation	G vs. I	21.251	1	<0.01
7	Test for interaction (adjusted) in the presence of excess variation	I vs. J	0.018	1	0.89
8	Test for differing levels of excess variation	J vs. L	6.933	3	0.07
9	Test for differing levels of excess variation for factor 1 and 2 when only main effects of both factors are significant	I vs. K	1.449	2	0.49

Table 4.5 Estimates and standard error of estimates of the best fit (K) with corresponding linear logistic model parameters and their standard errors.

Factorial combination	Under logistic-normal binomial model		Under linear logistic model	
	Estimate (η)	S.E	Estimate (η)	S.E
μ	-1.662	0.220	-1.291	0.106
Untreated	0.4740	0.238	0.3412	0.114
Yr. 2	1.125	0.240	0.8904	0.115

4.4.3 Statistical model fitting for the third experiment

As in the previous experiments, this experiment also consists of a 2×2 factorial. Factor one (X1) is the application of pesticide with 2 levels, with pesticide or without pesticide as in the first experiment. Factor two (X2) is the application of fertiliser, with two levels; with or without. The applied fertiliser had contained the fertiliser mixture at recommended dosage (Department of Agriculture of Sri Lanka, 1993).

As with the previous experiments, deviances after fitting possible models are shown in Table 4.6(a) and Table 4.6(b) uses respective likelihood ratio tests based on these tests to analyse the data.

Test 4 in table 4.6(b) suggests that there is no overdispersion associated with these data and the linear logistic model can accommodate the observed variability. In fact, no overdispersion with this experiment was reflected in the distribution fitting with these data (section 4.3.1). In this experiment, observed frequencies with all the disease incidence maps were adequately described by fitting the binomial distribution (Fig. 4.9).

Test 3 suggests no interaction between application of pesticide and application of fertiliser on disease incidence, and only the application of pesticide seems to have an influence on disease incidence. Therefore, the most appropriate model to describe the data is model B. So the conclusion may be made that application of pesticide affects the incidence of disease but application of fertiliser does not. The difference in the response at two levels of pesticide does not depend on application of fertiliser. Moreover, in this case, the spatial pattern does not vary depending on whether pesticide is applied or not. The parameter estimates and their standard errors for the best fit are given in Table 4.7 and these standard errors can be used for significant testing and deriving confidence intervals.

Table 4.6 Possible models and likelihood ratio tests for the data in the third factorial experiment.

(a) Possible models

Fit	Fixed effects parameters	Random effects parameters	d.f.	Deviance
A	%GM		143	211.795
B	%GM,X1		142	128.460
C	%GM,X2		142	209.909
D	%GM,X1,X2		141	126.459
E	%GM,X1,X2,X1.X2		140	123.989
F	%GM,X1,X2,X1.X2	%SCL	139	123.989

(b) Possible likelihood ratio tests

Test	Description of the test	Compare fits	Likelihood ratio statistic	d.f.	*P* value
1	Test for effect of factor1 adjusted for the other factor	C vs. D	83.450	1	<0.01
2	Test for effect of factor2 adjusted for the other factor	B vs. D	2.001	1	0.16
3	Test for interaction adjusted for both factors	D vs. E	2.470	1	0.12
4	Test for excess variation	E vs. F	0.00	1	0.50

Table 4.7 Estimates and standard error of estimates of the best fit (K) with corresponding linear logistic model parameters and their standard errors.

Factorial combination	Under linear logistic model	
	Estimate (η)	S.E
μ	0.6251	0.0782
Treated	-0.9788	0.109

4.4.4 Statistical model fitting for the fourth experiment

As all the other experiment this experiment also consisted of a 2×2 factorial. Factor one (X1) was application of pesticide with 2 levels, with pesticide or without pesticide (as in the first experiment). Factor two (X2) was the year of stand, with two different years of stand, year one and year two.

As with the other experiments, deviances after fitting possible models are shown in Table 4.8(a) and Table 4.8(b) uses respective likelihood ratio tests based on these tests to analyse the data.

From Table 4.8(b) it is evident that there is no interaction between application of pesticide and year of stand, and that only the effect of application of pesticide is significant. From test 9 it is clear that aggregation varies between levels of application of pesticide. Test 8 suggests differing levels of aggregation, but as explained earlier, since the interaction and main effect of factor 2 are not significant, test 8 becomes obsolete. Thus the most appropriate model to describe the data is model K. One may therefore conclude that application of pesticide affects disease incidence but that the year of stand does not, and the spatial pattern varies between pesticide treatments.

The parameter estimates and their standard errors for the best fit, along with the corresponding linear logistic model parameters and their standard errors are given in the Table 4.9. As explained with the experiment 1 and 2, the estimates and standard errors of the best fit have to be used for significant testing and deriving confidence intervals.

Table 4.8 Possible models and likelihood ratio tests for the data in the fourth factorial experiment.

(a) Possible models

Fit	Fixed effects parameters	Random effects parameters	d.f.	Deviance
A	%GM		143	546.778
B	%GM,X1		142	527.691
C	%GM,X2		142	544.667
D	%GM,X1,X2		141	525.551
E	%GM,X1,X2,X1.X2		140	525.450
F	%GM	%SCL	142	349.023
G	%GM,X1	%SCL	141	340.474
H	%GM,X2	%SCL	141	348.165
I	%GM,X1,X2	%SCL	140	339.600
J	%GM,X1,X2,X1.X2	%SCL	139	339.391
K	%GM,X1	%SCL,X1	140	328.984
L	%GM,X1,X2,X1.X2	%SCL,X1,X2,X1.X2	136	326.056

(b) Possible likelihood ratio tests

Test	Description of the test	Compare fits	Likelihood ratio statistic	d.f.	*P* value
1	Test for effect of factor1 adjusted for the other factor	C vs. D	19.115	1	<0.01
2	Test for effect of factor2 adjusted for the other factor	B vs. D	2.140	1	0.14
3	Test for interaction adjusted for both factors	D vs. E	0.101	1	0.75
4	Test for excess variation	E vs. J	186.05	1	<0.01
5	Test for factor 1 (adjusted) in the presence of excess variation	H vs. I	8.565	1	<0.01
6	Test for factor 2 (adjusted) in the presence of excess variation	G vs. I	0.874	1	0.35
7	Test for interaction (adjusted) in the presence of excess variation	I vs. J	0.209	1	0.65
8	Test for differing levels of excess variation	J vs. L	13.335	3	<0.01
9	Test for differing levels of excess variation for factor 1 when only main effect of factor 1 is significant	G vs. K	11.490	1	<0.01

Table 4.9 Estimates and standard error of estimates of the best fit (K) with corresponding linear logistic model parameters and their standard errors.

Factorial combination	Under logistic-normal binomial model		Under linear logistic model	
	Estimate (η)	S.E	Estimate (η)	S.E
μ	-2.358	0.361	-1.360	0.0925
Untreated	1.350	0.394	0.5329	0.123

4.5 Conclusions of the analysis

It is important to note that conclusions made here were separately for each experiments and that it is difficult to make general conclusions for the pineapple wilt disease pathosystem based on these results. Since four experiments were from four different situations, comparison among experiments are not valid. This is why different conclusions have been made with different experiments, even with same factors (Table 4.4 and 4.8). This suggests that factors other than those recorded here may also influence disease incidence.

However, influence of each of the factors on disease incidence, in each experiment can be summarised as in the Table 4.10.

Table 4.10 Summary of the factors investigated and their influence

Exp. No.	Factors investigated and their effect			
	Cultivar	Year of stand	Application of Pesticide	Application of Fertiliser
1	Significant	Significant	Not investigated	Not investigated
2	Not investigated	Significant	Significant	Not investigated
3	Not investigated	Not investigated	Significant	Not Significant
4	Not investigated	Not Significant	Significant	Not investigated

From Table 4.10 some general comment could be made on the influence of the factors investigated. All experiments in which pesticide application was investigated clearly showed the effect of pesticide on controlling the disease. The year of stand seems to have an influence on the disease incidence. The experiment where the effect of fertiliser application was investigated failed to show any effect of fertiliser on disease incidence. The type of cultivar also seems to have some effect on disease incidence although choice between them is based largely on fruit quality characteristics (Department of Agriculture of Sri Lanka, 1993).

5 TWO-DIMENSIONAL DISTANCE CLASS ANALYSIS

In chapter 3 we discussed ways of examining spatial patterns, by means of statistical distribution fitting, model fitting and variance-variance plots. For all these methods, the form of data required was the number of plants diseased in each quadrat of a particular size. Although it is well established (Cochran, 1936) that a quadrat size of 6-12 plants is usually used in epidemiological studies, one may still argue that all these methods are size (scale) dependent, i.e. conclusions based on these methods depend on the size of the quadrat.

Methods such as two-dimensional distance class (2DCLASS) analysis, which can be used to describe spatial patterns, utilise the data on an individual plant basis. Thus one may suggest that methods such as 2DCLASS analysis are scale independent.

In the literature, several methods which utilise incidence data on an individual plant basis have been employed to describe the spatial pattern of plant disease. Among them, the ordinary runs test and doublet analysis are quite common in plant disease epidemiology (Madden *et al.*, 1982; Converse *et al.*, 1979). These methods only considered the spread of disease between adjacent plants in one direction (within rows) and were intolerant of missing data. (Gray *et al.*, 1986).

Gray *et al.* (1986) first described the method of analysing the two-dimensional spread of incidence of a virus disease within row crops (which may be regarded as plant distribution lattices) known as two-dimensional distance class (2DCLASS) analysis. One of the advantages of this 2DCLASS analysis is that missing data within the lattice neither limit the analysis nor affect the interpretation of the data. This method is a development of the method described by Proctor (1984). 2DCLASS analysis, other than being useful for the detection of non-random spatial pattern, can also be used for quantification of average cluster size, distance between clusters, relative cluster location within the lattice, within and across row aggregation, and edge effects (Nelson *et al.*, 1992).

2DCLASS analysis is based on the concept of grouping pairs of infected plants into distance-orientation classes. According to Proctor (1984) a distance orientation class is defined as the number of horizontal (X) and vertical (Y) unit moves that separate a pair of infected plants from each other in a lattice. The horizontal distance (number of agronomic rows) is the abscissa and the vertical distance (number of plants within the agronomic row) is the ordinate. Thus the [X,Y] distance classes do not identify the position of the infected plants in the lattice, but refer to the absolute distance between plants by means of horizontal and vertical distances between a pair. The number of distance-orientation classes is defined by the overall dimensions of the lattice. For instance, 5 × 4 lattice (4 agronomic rows and 5 plants within an agronomic row) (Fig. 5.1) contains 19 different distance-orientation classes (Table 5.1). For a 5 × 4 lattice, number of plant pairs for each of these 19 different distance-orientation classes are given in Table 5.1.

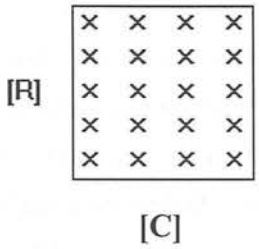


Fig. 5.1 5 × 4 **lattice. [C] Agronomic rows, [R] Plants within agronomic rows**

Table 5.1 Distance-orientation classes and number of possible plant pairs for each distance-orientation class

Distance orientation class [X,Y]	Number of possible pairs
[0,1]	16
[0,2]	12
[0,3]	8
[0,4]	4
[1,0]	15
[1,1]	12
[1,2]	9
[1,3]	6
[1,4]	3
[2,0]	10
[2,1]	8
[2,2]	6
[2,3]	4
[2,4]	2
[3,0]	5
[3,1]	4
[3,2]	3
[3,3]	2
[3,4]	1
Total	130

The specific software 2DCLASS (Nelson *et al.*, 1992) is capable of performing 2DCLASS analysis. This program allows for rapid and concise analysis of the spatial pattern of binary data within a two-dimensional matrix (where each plant has a specific location in the field; row number and plant number with in row). In the analysis, first, the pairs of infected plants are grouped into two-dimensional [X,Y] distance classes. Since the total possible number of pairs vary among [X,Y] distance classes, the number of pairs of diseased plants in each [X,Y] distance class is standardised by dividing the total possible number of pairs of living plants within the same [X,Y] distance class. The standardised number of pairs of diseased plants is generally referred to as 'standardised count frequency' [SCF]. The program calculates the expected SCFs by simulation. Expected SCFs are determined when the same number of infected plants are randomly assigned to locations within a lattice of the same dimensions. In this process, the location of the infected plants in the lattice are generated by a pseudo-random function and the missing plants observed in the field are assigned the same fixed X and Y co-ordinates (Nelson *et al.*, 1992). The expected SCFs are computed for all [X,Y] distance classes for each of user-specified number of sets of simulator data. Nelson *et al.* (1992) recommend the use of 400 sets of simulated data.

Comparison of observed and expected standardised counts in each [X,Y] distance class is used to define and quantify the randomness of diseased pairs of plants and their orientation within the lattice. The mean SCF and the standard deviation of the expected SCF for each [X,Y] distance class is computed in the program. The significance level on the observed SCF for each [X,Y] distance class is computed directly by counting the number of times the simulated expected SCF exceeds the observed SCF during the specified number of simulations. In addition, the program calculates 95% lower and upper confidence limits on the significance level using the formula;

$$p \pm \sqrt{1.96 \times p \times (1-p)/n}$$

where p is the expected SCF based on specified number of sets (n) of simulations.

Moreover, the program calculates the number of [X,Y] distance classes with an observed SCF significantly higher (upper confidence limit on significance ≤ 0.05) than expected (indicated by a '+' in 2DCLASS matrix [Fig. 5.2]) and significantly lower (lower confidence limit on level of significance ≥ 0.95) than expected (indicated by a '$' in 2DCLASS matrix). In the 2DCLASS matrix non-significant classes, i.e. observed SCF is neither greater or less than expected SCF, are indicated by '0'.

According to the program, a pattern is considered to be non-random if more than 5% of the total number of [X,Y] classes have SCFs that are significantly greater ($P \leq 0.05$) than expected with a random pattern of diseased plants (Gray *et al.*, 1986; Nelson and Campbell, 1993; Hughes and Nelson, 1995). Strength of non-randomness is interpreted as being directly proportional to the number of significant SCFs (Hughes and Nelson, 1995). It is very important to note that in 2DCLASS analysis, non-randomness does not mean aggregation. It could either be aggregation or regular spatial pattern.

The property of minimum 'core cluster size' is defined as the number of significant ($P \leq 0.05$) and adjacent distance classes (including the [0,0] class) that constitute a discrete, contiguous group in the [0,0] corner. For instance, Fig. 5.2a represents a minimum core cluster size of four and Fig. 5.2c represents a minimum core cluster size of fifteen. Resolution of the core cluster size is limited to a range of values, because the significance level calculated for classes, for example, is based on both unidirectional and bi-directional comparison of plants within the incidence map. Bi-directional refers to each plant being compared with all other plants in that distance class, into both sides of the row (within row) from a reference plant and only into one side of the row in the case of unidirectional. For instance the core cluster size of four in Fig. 5.2a consists of distance classes [0,0], [0,1], [0,2] and [0,3].

All distance classes having [X]=0 in the core cluster implies that the cluster of diseased plants does not involve more than one agronomic row, i.e. the cluster occurs along a single row. The class [0,0] indicates the reference plant itself. For unidirectional resolution, the class [0,1] implies that the adjacent plant (in one direction, i.e. either up or down along the agronomic row) of the reference plant is diseased. The class [0,2] implies second plant next to the reference plant (in the same direction) is diseased. The class [0,3] implies the third plant next to the reference plant (in the same direction) is diseased. Thus according to unidirectional resolution, four plants are diseased (including the reference plant). If the other direction was considered, a similar conclusion would be made. Thus if both directions were considered it would be concluded that seven plants are diseased (since the reference plant is common to both directions). Thus the core cluster size is four to seven. An intuitive formula may be made, that if the number of adjacent distance classes with SCFs significantly greater ($P \leq 0.05$) than expected SCF in the [0,0] corner is u excluding [0,0], the core cluster size is $(u+1)$ to $(2u+1)$.

Elsewhere, other than the core cluster in the 2DCLASS matrix, the occurrence proximal distance classes with SCFs significantly greater than expected ($P \leq 0.05$) reflects similar arrangements of diseased plants in the corresponding plot map (Nelson and Campbell, 1993). Similar arrangements of healthy plants are likewise reflected by proximal distance classes with SCFs significantly less than expected ($P \geq 0.95$). For instance, in Fig. 5.3b, elsewhere other than the core cluster, classes [4,0-13] and [5,0-16] having observed SCFs greater ($P \leq 0.05$) than expected SCFs implies that clusters are found at a distance of 4 to 5 rows and 0-16 plants within row, in the incidence map. The corresponding incidence map (Fig. 4.4b) reflects this finding. Thus this information reflects the relative location of the clusters in an incidence map.

According Nelson *et al.* (1992), if more than 12.5% of the [X,Y] classes in the outermost row and column (other than in the core cluster) of the distance class analysis are significant ($P \leq 0.05$), an 'edge effect' is indicated. For instance 2DCLASS matrix of Fig. 5.6c, seventeen such significant classes out of forty-one in the outermost row and column (which is more than 12.5%) can be identified thus an edge effect is detected in this particular case. However, all the incidence maps made in this study (Fig. 4.3-4.6) are not complete crop fields or plots, but part of a field within a site. Since an edge effect cannot really exist without a proper crop boundary, Fig. 5.6(c) does not represent a true edge effect. For this reason, edge effects will not be discussed in relation to 2DCLASS analysis of the incidence maps (Fig. 4.3-4.6) in section 5.1.

5.1 2DCLASS analysis of pineapple wilt disease incidence data

Experiment 1

The 2DCLASS matrix of the incidence maps of experiment 1 (Fig. 4.3) are shown in Fig. 5.2. Fig. 5.2a represents four distance classes having observed SCFs significantly greater ($P \leq 0.05$) than expected SCFs in the [0,0] corner, implying a core cluster size of four to seven. As explained earlier, since the classes in the core cluster are [0,0-3], clustering along the agronomic rows is evident. The proportion of significant SCFs is greater than 5% and thus a non-random spatial pattern is detected. No clusters of significant classes can be identified in the distance class matrix other than the core cluster.

The distance class matrix of Fig. 5.2b represent six distance classes having observed SCFs significantly greater ($P \leq 0.05$) than expected SCFs in the [0,0] corner, indicating a core cluster size of six to eleven. The classes in the core cluster are [0,0-5] and this implies that clusters occur along the agronomic rows. The proportion of significant SCFs is greater than 5% and thus a non-random spatial pattern is detected. As in Fig. 5.2a, other than the core cluster no specific clusters of significant classes can be identified.

The distance class matrix of Fig. 5.2c reveals a core cluster size of fifteen to twenty-nine. The significant ($P \leq 0.05$) distance classes in the core cluster are [0,0-7] and [1,0-6] and this implies that a clusters of diseased plants involve two agronomic rows. Other than the core cluster, clusters of distance classes [2,1-2], [2,4-6] and [2,8-9], of which observed SCFs are greater than expected ($P \leq 0.05$), can be identified. This implies that the clusters occur more or less two rows apart. The proportion of distance classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

The distance class matrix of Fig. 5.2d represents a core cluster size of three to five. Since the core cluster consists of significant ($P \leq 0.05$) distance classes [0,0-2], the clusters of diseased plants occur along the agronomic rows. Other than the core cluster some clusters of significant distance classes ($P \leq 0.05$) can be identified in the lower right quarter of the distance class matrix. This may be interpreted as clusters occurring more-or-less six to ten rows apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

Thus the 2DCLASS analysis of experiment 1 may be summarised as follows. A non-random spatial pattern is evident and average core cluster size is seven to eleven. In general, clusters occur along the agronomic rows and it is not clear that the different clusters have similar relative locations in the field.

X

Y	0	1	2	3	4	5	6	7	8	9	10	11
0	+	0	+	+	0	0	0	0	$	0	$	$
1	+	0	0	0	0	0	0	0	0	$	$	0
2	+	+	0	0	0	0	0	0	0	$	0	$
3	+	0	+	0	0	0	0	0	0	$	$	$
4	0	0	0	0	0	0	0	0	0	0	$	0
5	0	+	0	0	0	0	0	0	0	$	$	$
6	+	+	+	0	0	0	0	0	0	0	0	$
7	0	0	0	0	0	+	0	0	0	0	0	$
8	0	+	0	0	0	0	0	0	0	0	$	0
9	0	0	+	0	0	0	0	0	0	0	$	$
10	+	0	0	0	0	0	0	0	0	0	$	$
11	0	0	0	0	0	0	0	0	0	0	$	$
12	+	0	+	0	0	0	0	0	0	0	$	$
13	+	+	0	0	0	0	0	+	0	0	0	$
14	0	+	0	0	0	0	0	0	0	0	$	0
15	0	0	0	0	0	0	0	0	0	$	0	0
16	0	+	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	$	0
18	0	0	0	0	0	0	$	0	0	0	0	0
19	0	0	0	0	0	+	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	$	0
22	0	0	0	0	0	0	0	0	0	0	0	0
23	0	$	$	0	0	0	0	0	0	0	0	0
24	0	0	$	0	0	0	0	0	0	0	0	0
25	$	0	0	0	0	0	0	0	0	0	0	0
26	0	0	$	0	0	0	0	0	+	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(a)

Plantation - Kattota

X

0	1	2	3	4	5	6	7	8	9	10	11
+	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	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	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	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	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	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	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	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

(b)

Plantation - Kalagedihena

0	+	+	0	0	+	0	0	0	0	0	0	0
1	+	+	+	0	+	0	0	$	0	0	0	0
2	+	+	+	0	0	0	0	$	0	0	0	0
3	+	+	0	0	0	0	0	$	0	0	0	0
4	+	+	+	0	+	0	0	0	0	0	0	0
5	+	+	+	+	+	0	0	0	0	0	0	0
6	+	+	+	0	0	0	0	0	0	0	0	0
7	+	0	0	0	0	0	0	0	0	0	0	0
8	0	+	+	0	0	0	0	0	0	0	0	0
9	+	+	+	+	0	0	0	0	0	0	0	0
10	+	0	0	+	0	0	0	0	0	0	0	0
11	+	+	0	0	0	0	0	0	0	0	0	0
12	0	+	0	0	0	0	0	0	0	0	0	0
13	0	0	+	0	0	0	0	0	0	0	0	0
14	+	+	0	0	0	$	0	0	0	0	0	0
15	0	+	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0
18	+	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(c)

Plantation - Navadiga

+	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	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	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	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	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	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	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	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	0	0	0	0	+	+	0	0	+

(d)

Plantation - Urapola

Fig. 5.2 The 2-D Class matrices for the experiment 1. X - No. of row distance; Y- No. of plants distance within the row.

Experiment 2

The 2DCLASS matrices of the incidence maps of experiment 2 (Fig. 4.4) are shown in Fig. 5.3. According to Fig. 5.3a, the core cluster size is nine to seventeen and clusters occur along the agronomic rows. Other than the core cluster, clusters of significant ($P \leq 0.05$) distance classes [1,3-11], [4,0-6], [5,0-3] and [5,5-6] may be interpreted as relative locations of clusters that could be either one row apart or four to five rows apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

The 2DCLASS matrix of Fig. 5.3b reveals a core cluster size of nine to seventeen, and clusters involve two agronomic rows. As in Fig. 5.3a, clusters of significant ($P \leq 0.05$) distance classes [4,0-13] and [5,0-16] may be interpreted as relative locations of clusters four to five rows apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

The 2DCLASS matrix of Fig. 5.3c represents a core cluster size of four to seven plants, and the clusters occur along the agronomic rows. Other than the core cluster, the clusters of significant ($P \leq 0.05$) distance classes can be identified at the row distance of three in the 2DCLASS matrix, indicating relative locations of clusters three rows apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

The 2DCLASS matrix of Fig. 5.3d reveals a core cluster size of four to seven and occurrence of clusters along the rows. Other than the core cluster, significant ($P \leq 0.05$) distance classes [0,7-14] and [1,8-13] indicate relative location of clusters as seven to fourteen plants apart in the same row, or one row apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

Thus the conclusion from experiment 2 is that a non-random spatial pattern is evident and that the average core cluster size is about seven to twelve. Moreover, aggregation is predominantly along the rows and the relative location of clusters varies from several plants in the same row to five rows apart.

X

Y	0	1	2	3	4	5	6	7	8	9	10	11
0	+	0	0	0	+	+	0	0	$	0	0	0
1	+	0	0	0	+	+	0	0	$	0	0	0
2	+	0	0	+	+	+	0	$	$	0	0	0
3	+	+	0	+	+	+	0	0	$	0	0	0
4	+	+	0	0	+	0	0	0	$	0	0	0
5	+	+	0	0	+	+	+	$	$	0	0	0
6	+	+	0	0	+	+	+	$	0	0	0	0
7	+	+	0	0	0	0	0	0	0	0	0	0
8	+	+	0	0	0	0	+	0	0	0	0	0
9	0	+	0	0	0	0	0	0	$	0	0	0
10	0	+	0	0	0	0	0	0	0	0	0	0
11	0	+	0	0	0	0	0	0	$	0	0	0
12	0	0	0	0	0	0	0	0	$	0	0	0
13	0	0	0	0	0	0	0	0	$	0	0	0
14	0	0	0	0	0	0	0	0	$	0	0	0
15	0	$	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	$	0	0	0	0	0
18	0	$	0	0	0	0	0	+	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	$	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	$	0	0	0	0	0	0	0	0
23	0	$	0	0	$	0	0	0	0	0	0	0
24	$	0	0	0	0	0	0	$	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(a)

Plantation - Walgammana

X

0	1	2	3	4	5	6	7	8	9	10	11
+	+	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
+	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
+	+	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	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	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
$	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	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0

(b)

Plantation - Welhena

0	+	0	0	0	0	0	+	0	0	0	0	0
1	+	0	0	0	+	0	0	0	0	0	0	0
2	+	0	0	0	0	0	0	0	0	0	0	0
3	+	0	0	+	0	0	0	$	0	0	0	0
4	0	0	0	0	0	0	+	0	0	0	0	0
5	0	0	0	+	0	0	0	0	0	0	0	0
6	0	0	0	+	0	0	+	0	0	0	0	0
7	0	0	0	+	0	0	+	0	0	0	0	0
8	+	0	0	0	0	0	+	0	0	0	0	$
9	0	0	0	+	0	0	0	0	0	0	0	0
10	+	0	0	+	0	0	+	0	0	0	0	0
11	+	0	0	+	0	0	0	0	0	0	0	0
12	+	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	+	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	+	0	0	0	0	0	0	0	0
16	0	$	0	0	0	$	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	$	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	$	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	+	0
27	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	+
29	0	0	0	0	0	0	0	0	0	0	0	0

(c)

Plantation - JEDB

+	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	$	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	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	$	$	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	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	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	$	$	$
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	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
0	0	0	0	0	0	0	0	0	0	0	0

(d)

Plantation - Aluwala

Fig. 5.4 The 2-D Class matrices for the experiment 2. X - No. of row distance; Y- No. of plants distance within the row.

Experiment 3

The 2DCLASS matrices of the incidence maps of experiment 3 (Fig. 4.5) are shown in Fig. 5.4. According to the 2DCLASS matrix of Fig. 5.4a, core cluster size is two to three and clusters occur along the rows. Other than the core cluster some significant ($P \leq 0.05$) distance classes can be identified at the row distance of ten in the distance class matrix. This situation is quite common (Nelson *et al.*, 1992) when the edge effect is significant. However, in this case, the criteria for identification of an edge effect were not met. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is evident.

The 2DCLASS matrix of Fig. 5.4b reveals a core cluster size of four to five, and the clusters involve two agronomic rows. A special feature of this 2DCLASS matrix is that large number of significant ($P \leq 0.05$) distance classes in the bottom half of the matrix. This may be interpreted as clusters in any of the agronomic rows twenty to twenty-nine plants apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is evident.

The 2DCLASS matrix of Fig. 5.4c reveals a core cluster size of three to five, and clusters involve two agronomic rows. No specific cluster of significant distance classes can be identified other than the core cluster. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is evident.

The 2DCLASS matrix of Fig. 5.4d indicates a core cluster size of four to seven, and clusters involve two agronomic rows. The group of significant distance classes at the bottom of the distance class matrix suggests that the distance between clusters is about four to eight rows and twenty-six to twenty-eight plants apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

Thus the results of the experiment 3 may be summarised as follows: aggregation is evident and average core cluster size is about three to five. Moreover, clusters predominantly involve two agronomic rows. The relative location of clusters is mostly at twenty to twenty-eight plants apart along the agronomic rows.

X

Y	0	1	2	3	4	5	6	7	8	9	10	11
0	+	0	0	0	0	0	0	0	0	0	+	+
1	+	0	0	0	0	0	0	0	0	0	+	0
2	0	0	0	0	0	0	0	0	+	0	0	0
3	0	0	0	0	0	0	0	0	+	0	+	0
4	+	0	0	0	0	0	+	0	+	0	+	0
5	+	0	0	0	0	+	+	+	0	0	+	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	+	+
9	0	0	0	0	0	0	0	0	0	0	+	0
10	0	0	0	0	0	0	0	0	0	0	+	0
11	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	+	0
14	0	0	0	0	0	0	0	0	0	0	+	0
15	0	0	0	0	0	0	0	0	0	+	0	0
16	0	0	0	0	0	0	+	0	0	0	0	0
17	0	$	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	$	0	0	$	0	0	0	0	0	0
22	0	0	$	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0
25	$	$	0	0	0	0	0	0	0	0	0	0
26	$	$	$	0	0	0	0	0	0	$	0	0
27	0	$	0	0	0	$	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	$	0	0	0	0	0	0	0	0	0

(a)
Plantation - Hettipola

X

Y	0	1	2	3	4	5	6	7	8	9	10	11
0	+	+	0	0	$	0	0	0	0	0	0	0
1	+	0	0	0	0	0	0	0	0	0	0	0
2	$	0	0	0	0	0	0	0	0	0	0	0
3	$	$	$	0	0	0	0	0	0	0	0	0
4	$	$	$	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	$	$	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	$	0	0	0	0	$	0	0	0	0	0
9	$	$	$	0	0	0	0	0	0	0	0	0
10	$	$	$	0	0	0	0	0	0	0	0	0
11	0	$	$	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	$	0	0	0	0	0
13	+	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0
18	+	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	+	+	+	0	0	0	0	0	0
21	0	0	0	+	+	+	+	0	0	0	0	+
22	0	0	0	+	+	+	+	0	0	0	0	+
23	0	0	0	+	+	0	+	0	0	+	+	0
24	+	+	+	+	+	+	+	+	0	+	+	0
25	0	0	0	0	0	+	0	0	+	0	0	0
26	+	+	+	+	+	+	+	+	+	0	+	0
27	+	0	0	+	0	0	+	+	0	+	0	0
28	0	0	0	0	0	0	0	+	+	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(b)
Plantation - Panduwasnuwara

Y	0	1	2	3	4	5	6	7	8	9	10	11
0	+	+	0	$	0	0	+	0	0	0	0	0
1	+	0	0	0	0	0	0	0	0	0	0	0
2	$	$	0	0	0	$	$	0	0	0	0	0
3	$	$	$	0	0	0	0	0	0	0	+	0
4	0	0	$	0	0	0	0	0	0	0	0	0
5	+	0	0	0	0	0	0	0	0	+	0	0
6	+	0	0	0	0	$	0	0	0	+	+	0
7	0	0	0	0	0	0	0	0	0	0	+	0
8	$	0	$	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	+	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	+	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	$	0	0	0	0	0	0	0	0	0
15	0	0	$	0	0	0	0	0	0	0	0	0
16	+	0	0	0	0	0	0	0	0	0	0	0
17	+	+	0	0	+	0	0	0	0	0	0	0
18	+	0	0	0	+	0	0	0	0	0	+	+
19	+	0	0	0	0	0	0	0	0	+	+	0
20	0	0	0	$	0	0	0	0	0	0	0	0
21	+	0	0	$	0	0	0	0	0	0	0	0
22	0	0	0	0	+	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	+	+	0
25	0	$	0	$	0	$	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	+	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(c)
Plantation - Bingiriya

Y	0	1	2	3	4	5	6	7	8	9	10	11
0	+	+	0	0	0	0	+	0	0	0	0	0
1	+	+	0	0	0	0	0	0	0	0	0	0
2	$	$	0	0	0	0	0	0	0	0	0	0
3	$	$	0	0	0	0	$	$	0	0	$	0
4	$	$	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	+	+	0	0	+	+	+	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	$	$	0	0	0	0	0
9	$	0	0	0	0	$	$	0	0	0	$	$
10	$	$	0	0	0	0	0	0	0	0	0	0
11	0	0	$	0	0	0	+	0	0	0	0	0
12	0	0	0	0	0	+	+	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	+	0	$	0	0	0	0	0	$	0
15	0	0	+	0	0	0	$	0	0	0	$	0
16	0	0	0	0	0	0	0	$	$	0	0	0
17	0	0	0	0	0	0	+	0	0	0	0	0
18	0	0	0	0	+	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	+	0	0	0	0	0	0	0	0	0
22	0	+	+	0	0	0	0	0	0	0	0	0
23	+	+	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	+	+	+	+	0	+	0	0	0
27	0	0	0	0	+	+	+	0	+	0	0	0
28	+	0	0	0	0	0	+	0	+	+	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(d)
Plantation - Kobeigane

Fig. 5.5 The 2-D Class matrices for the experiment 1. X - No. of row distance; Y- No. of plants distance within the row.

Experiment 4

The 2DCLASS matrix of the incidence maps of experiment 4 (Fig. 4.6) are shown in Fig. 5.5. The 2DCLASS matrix of Fig. 5.5a reveals a core cluster size of thirty-five to sixty-nine and clusters involve four agronomic rows. Apart from core cluster, no other groups of clusters of distance classes could be identified in the 2DCLASS matrix. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is evident.

The 2DCLASS matrix of Fig. 5.5b represents a core cluster size of fifteen to twenty-nine, and clusters involve two agronomic rows. Apart from the core cluster, distance classes [0,11-18], [1,15-19], [2,1-6] and [3,1-6] can be identified in the 2DCLASS matrix and this may be interpreted as distance between clusters is eleven to eighteen plants apart in the same row, to three rows apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is detected.

The 2DCLASS matrix of Fig. 5.5c reveals a core cluster size of four to seven and clusters occur along the agronomic rows. Other than the core cluster, a group of significant distance classes can be identified in the bottom of the 2DCLASS matrix and, as explained under experiment 3, this may also be interpreted as the relative location of clusters being twenty-three to twenty-seven plants apart. In addition, clusters of distance classes [7,4-17], [8,1-3] and [9,13-15] may be interpreted as clusters also at seven to nine rows apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is evident.

The 2DCLASS matrix of Fig. 5.5d reveals a core cluster of four to seven, and clusters occur along the rows. Other than the core cluster, a group of significant distance classes can be identified at the row distance of four to six in the 2DCLASS matrix and this may be interpreted as clusters at four to six rows apart. The proportion of significant classes with significant SCFs is greater than 5% and thus a non-random spatial pattern is evident.

The results of experiment 4 may be summarised as follows: aggregation is evident and average core cluster size is fifteen to twenty eight. Moreover, clusters occur predominantly occur along the rows, but could involve several rows.

X

Y	0	1	2	3	4	5	6	7	8	9	10	11
0	+	+	+	+	0	0	0	0	0	$	0	0
1	+	+	+	+	0	0	0	0	$	0	$	0
2	+	+	+	+	0	0	0	0	$	0	$	0
3	+	+	+	+	0	0	0	0	0	0	$	0
4	+	+	+	+	0	0	0	0	0	0	$	0
5	+	+	+	+	0	0	0	0	$	0	0	0
6	+	+	+	+	0	0	0	0	$	0	0	0
7	+	+	+	+	0	0	0	0	$	0	0	0
8	+	+	+	0	0	0	0	0	$	0	0	0
9	0	+	+	0	0	0	0	0	$	0	0	0
10	+	+	+	+	0	0	0	0	$	0	0	0
11	0	+	+	0	0	0	0	0	0	$	0	0
12	0	+	0	0	0	0	0	0	$	$	0	0
13	0	0	0	0	0	0	0	0	$	$	0	0
14	0	0	0	0	0	0	0	0	$	$	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	$	$	0	0	0	0
17	0	0	0	0	0	0	$	0	$	0	0	0
18	0	0	0	$	$	0	$	$	$	0	0	0
19	0	$	$	$	$	0	$	$	0	0	0	0
20	0	0	$	$	0	$	$	$	0	0	0	0
21	0	$	$	$	$	0	$	$	0	0	0	0
22	$	$	$	$	0	0	$	$	0	0	0	0
23	0	0	$	$	0	0	0	0	0	0	0	0
24	0	$	0	$	0	0	0	0	0	0	0	0
25	0	0	$	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(a)

Plantation - Devasara

X

0	1	2	3	4	5	6	7	8	9	10	11
+	+	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
+	+	+	+	+	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	$	$	$	$	$	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	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
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	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	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

(b)

Plantation - Akkarawatta

0	+	0	0	0	0	$	0	0	0	0	0	+
1	+	0	$	0	0	$	$	0	+	0	+	+
2	+	0	0	$	0	$	$	0	+	+	+	+
3	+	0	0	$	$	$	0	0	+	0	0	+
4	0	0	0	$	$	$	$	+	0	0	0	+
5	+	0	0	0	$	$	0	+	0	0	+	+
6	0	0	0	0	$	$	0	+	+	0	0	+
7	0	0	$	$	$	$	0	+	0	0	0	+
8	0	0	0	$	0	0	0	+	0	0	+	+
9	+	0	$	$	$	0	0	+	+	+	+	+
10	0	0	$	$	0	$	$	+	0	+	0	+
11	0	0	$	$	0	$	0	+	0	0	0	+
12	0	0	$	0	0	0	0	+	0	0	+	+
13	0	0	$	$	$	$	0	+	0	+	0	+
14	0	$	0	0	0	0	0	+	0	+	0	+
15	0	0	0	0	0	0	0	+	0	+	0	+
16	0	0	0	0	0	0	0	+	0	0	0	0
17	0	0	0	0	0	0	0	+	0	0	+	+
18	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	+	0	0
20	0	0	0	0	0	+	0	0	+	0	0	0
21	0	0	0	0	+	0	0	+	0	0	0	0
22	0	0	0	0	0	0	0	+	0	0	+	0
23	0	0	0	0	0	0	+	0	$	0	0	0
24	0	0	+	+	0	0	+	0	0	0	0	0
25	0	0	+	0	0	+	+	0	0	0	0	0
26	+	+	+	0	0	+	+	0	0	0	+	0
27	0	0	+	+	+	+	+	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	+	0
29	0	0	0	0	0	0	0	0	0	0	0	0

(c)

Plantation - Munamaldeniya

+	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	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	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	+	$	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	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
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	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
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
0	0	0	0	0	0	0	0	0	0	0	+
0	0	0	0	0	0	0	0	0	0	0	+

(d)

Plantation - Mukalanyaya

Fig. 5.6 The 2-D Class matrices for the experiment 1. X - No. of row distance; Y- No. of plants distance within the row.

It is important to note that the incidence maps in Fig. 4.3-4.6 are not complete crop fields but part of the fields within the sites. Moreover, each field had been subjected to different conditions, such as control measures applied and year of stand. Thus it is impossible to make solid conclusions about the pineapple wilt pathosystem from these results. However, one common feature with all 2DCLASS matrices is the aggregation of disease incidence. In addition, all incidence maps reveal more clusters along rows than across rows and this is a good indication that the spread of the disease occur along the rows compared to across rows. The conventional conclusion from this analysis would be that average core cluster size of 8-14 plants within the row may be the general cluster size of pineapple wilt disease. This assumes the analysis to be scale-independent, a quality discussed further in section 5.3.

5.2 Effect of quadratization in 2DCLASS analysis

In section 5.1 we investigated complete field maps in order to understand the spatial pattern of the disease. But very often epidemiological data are collected on the basis of quadrats. Moreover, if the field under study is large and 2DCLASS analysis is done on the basis of individual plants, the problem of software limitation may also arise. If a single value could be used for the information in a quadrat, i.e. if quadrat information is converted into a single value and then used in the 2DCLASS analysis, software limitation might be easily overcome. In addition, if this is possible, data collected on the quadrat basis may be directly used in the 2DCLASS analysis.

In order to investigate the effect of this quadratization we analysed a 32×32 array map of pineapple wilt disease (Fig. 5.6). When the 2DCLASS analysis is performed on the basis of individual plants, the resulting 2DCLASS matrix is shown in Fig. 5.7. From the 2DCLASS matrix a core cluster in the top left corner with distance class having observed SCFs significantly greater than expected SCFs ($P \leq 0.05$) can easily be recognised, and it identifies an average core cluster size of approximately 26 to 53 plants. In addition, large 'reflected cluster' with distance classes having observed SCFs significantly greater than expected ($P \leq 0.05$) can be identified at the right one-third of the matrix. A cluster of distance classes having observed SCFs significantly less than expected ($P \geq 0.95$) occurs in lower left of the 2DCLASS matrix. These reflect consistent periodicities between pairs of diseased (or healthy) plants over the array map.

Then the map was divided in to quadrats of size two (two plants). The quadratization was made along the rows, since the observed pattern suggests that the disease spread was predominantly in this direction. For the 2DCLASS analysis, quadrats with mean disease incidence greater than the overall mean incidence were coded as 'diseased' and the rest as 'healthy'. When 2DCLASS analysis is performed for this quadratized map, the resulting 2DCLASS matrix is shown in Fig. 5.8. Fig. 5.9 is the equivalent map when quadrat size is taken as four plants (also along the rows).

A careful examination of Fig. 5.8 and 5.10 reveals that almost all the properties observed in Fig. 5.7 have been reflected in Fig. 5.8 and 5.10. For instance, more or less proportional reduction of core cluster size may be identified along with the increase of quadrat size. In addition, clusters of distance classes having observed SCFs significantly greater ($P \leq 0.05$) and less ($P \geq 0.95$) can be observed in same locations, with proportional scale. In other words output of the 2DCLASS analysis has been consistent irrespective of quadratization. This suggests that data based on quadrats may directly be used in 2DCLASS analysis with little if any loss of information in return.

```
                              C
  * H * * * * * * * H H H * * * * * * * * H H * H * * * * * * H *
  H * * * * * * * * H H * * * H H * * * * H * * * * * * * * * H *
  * * * * * * H H H * * * * * * * * * H H H H H * * * * * * * * *
  H H H * * * H H * H H * H H * * * * * * * H * * * * * * * * * *
  H H H * * * * H H H H H H * * * * * H * * H * * * * * * * H * *
  * H H * H * H H H * * H H * * * * H H * * H H * * * * * * * H *
  H H * * * * H H H H H H H H H * * * H H * H H H * * * * * * * *
  H H H H * H H * H H H H H H H H * H H H H H * * * * * * * * * *
  * H H H * H H * * H H H * H H * * * * H H H * * * * * * * * * *
  H H H H H H H * H H H H * H H H H H H * H H * H * * * * * * * *
  H * * H * H H H * * * H * H * H H * H * H * * H H * * H * * * *
  H * * H * H H H * H H * H H * H * * * * * H * H H * * * * * * *
  H * H H * H H H * * H * * H * * * H H * * H * * H * * * * * * *
  H * H * * * H * * H H * H H H H H H * * * * * * H * * * * * * *
R * H * * H * * * H H * * * H H * * * * * * * * * H * * H * H * *
  H H H * H * H * * H * * * H * * * H H H * * H * H * * * * * * *
  H * H H * H * * * * H * * * H * * H H H * * * * * * * * * * * *
  H H H * * * H * H * * * * H * * * H H * H * * * * * * * * H * *
  H * * * H H * * H * * H * * H * H * * * * * * * H * * H * H * *
  H * * * * H H H H H * * * * H * * H H * H * * H H * * * * * * *
  * * * H * * H * H H * H * * * * * * * H H H H H * * * H * * * *
  * * * * * * * * * H * * * H * * * H * H * H H * * * * H * * * *
  H * * * * * * * * H * H * H * * H H H H * * H H * * * H * * * H
  * * * * H H H * * H H * * * * H * * H H * H H H * * * H * * * *
  * * * * * * * H * H * * * * H * H * H H * H H * * H * H * * * H
  * * * H H * * H * H H H * * * * * H H H * H H H * * * H H * * *
  * * * H * H * H * H H H * H * * H H H H H H H H H * * H H * * *
  * * * * * H H H * * H H * H * H H H H * H H * H H * * H H * H *
  * * * H H H H H H * H H * * H H H H H H H H H H * H * H H H H H
  * * * H H H H * H * H * * H * H * H H H * H H H H H * H * * * H
  H * H H H H H H H H * * H * H H H H * H * H * H H * * H H * * H
  H * * H H H H H * H * * H * H H H H H H * H * H H H H H H * * H
```

Fig. 5.6. Pineapple wilt disease incidence map for the large array (32×32), C- Agronomic rows run vertically; R-Plants within rows; H-healthy plant; *-diseased plant.

X

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	0	+	+	+	0	0	0	$	$	$	0	0	0	+	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	0	0
	1	+	+	+	+	0	0	0	$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	0	0
	2	+	+	+	+	0	0	0	$	0	0	0	0	0	0	0	0	$	0	0	0	0	0	+	+	+	+	+	+	+	+	0	0
	3	+	+	+	0	0	0	$	$	0	0	0	0	0	0	+	0	0	$	0	0	0	0	+	+	+	+	+	+	+	+	0	0
	4	+	+	+	0	0	0	$	$	0	0	$	0	0	0	0	0	0	$	0	0	0	0	+	+	+	+	+	+	+	+	+	0
	5	+	+	0	0	0	0	$	$	$	$	0	0	0	0	0	0	0	$	0	+	0	0	+	+	+	+	+	+	+	+	+	0
	6	+	+	+	0	0	0	0	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+	0
	7	+	+	0	0	0	0	0	$	$	0	0	0	0	0	0	0	0	0	0	+	0	0	+	+	+	+	+	+	+	+	0	0
	8	+	+	0	0	0	0	0	$	0	0	0	0	0	0	0	+	0	0	0	+	0	0	+	+	+	+	+	+	+	+	0	0
	9	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0	0	+	0	0	+	+	+	+	+	+	+	+	+	0
	10	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	0	+	+	0	0	+	+	+	+	+	+	+	+	+	+	0
	11	0	+	0	0	0	0	0	$	0	0	0	0	0	0	0	+	0	0	0	+	0	+	+	+	+	+	+	+	+	+	+	0
	12	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	0	0	+	+	+	+	+	+	+	+	+	0
	13	0	+	0	0	0	0	0	0	0	$	0	0	0	0	+	0	0	0	0	+	0	+	+	+	+	+	+	+	+	+	+	0
	14	0	0	0	0	0	0	0	$	$	$	$	0	0	0	0	0	0	0	0	+	0	+	+	+	+	+	+	+	+	+	+	0
	15	0	0	0	0	0	0	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+	+	0
Y	16	0	0	0	0	0	$	$	$	$	$	0	0	$	0	0	0	$	0	0	0	0	0	+	+	+	+	+	+	+	+	+	0
	17	0	0	0	$	$	$	$	$	$	$	0	$	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+	0
	18	0	0	0	$	$	$	$	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0
	19	0	0	0	0	$	$	$	$	$	$	0	$	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0
	20	0	0	0	$	$	$	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	0	0
	21	0	0	0	$	$	$	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0
	22	$	$	$	$	$	$	$	$	$	$	$	$	$	$	$	0	0	0	$	$	$	0	0	+	+	+	+	+	+	+	0	0
	23	$	$	$	$	$	$	$	$	$	$	$	0	0	0	0	0	0	$	0	$	$	0	0	0	0	+	+	+	+	+	0	0
	24	$	$	$	$	$	$	$	$	$	$	$	$	$	0	0	0	$	$	$	$	$	0	0	0	0	+	+	+	+	+	+	0
	25	$	$	$	$	$	$	$	$	$	$	$	$	$	0	$	$	$	$	$	$	$	$	0	0	0	+	+	+	+	+	0	0
	26	$	$	$	$	$	$	$	$	$	$	$	$	$	$	$	0	$	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0
	27	$	$	$	$	$	$	$	$	$	$	$	$	$	$	0	0	0	0	$	$	$	$	0	0	0	0	0	0	0	0	0	0
	28	$	$	$	$	$	$	$	$	$	$	$	$	$	$	$	0	0	0	0	$	$	$	0	$	0	0	0	0	0	0	0	0
	29	$	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
	30	$	$	$	$	0	0	0	0	0	$	$	0	$	0	0	0	0	0	0	$	0	0	0	0	0	0	0	0	0	0	0	0
	31	$	$	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

Fig. 5.7 2DCLASS matrix of Fig. 5.6 (individual plant basis). X-No. of row distance, Y-No. of plant distance with in row.

X

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	0	+	+	+	+	0	0	0	$	$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	0	0
	1	+	+	+	+	0	0	0	$	0	$	0	0	0	0	0	0	$	$	0	0	0	0	0	+	+	+	+	+	+	+	0	0
	2	+	+	+	+	0	0	0	$	$	$	0	0	0	0	0	0	0	$	0	0	0	0	0	+	+	+	+	+	+	+	0	0
	3	+	+	+	0	0	0	$	$	$	$	0	0	$	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	0	+	0	0
	4	+	+	0	0	0	0	0	$	$	0	$	$	$	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0	0
	5	+	+	0	+	0	0	0	0	0	$	0	$	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+	0
	6	+	+	0	0	0	0	0	0	$	0	$	0	0	$	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+	0	0
Y	7	+	+	0	0	0	0	0	0	$	$	$	0	0	0	0	0	0	0	0	0	0	+	0	+	+	+	+	0	+	+	0	0
	8	0	+	0	0	0	0	$	$	$	$	0	$	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0	0
	9	0	0	0	0	0	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	0
	10	0	0	0	0	0	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	+	+	0
	11	0	0	0	0	$	0	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	+	0	0
	12	0	0	0	0	0	$	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	0	0	0
	13	$	$	0	0	0	0	0	$	$	$	$	$	0	0	0	0	0	0	0	0	$	0	0	0	0	0	0	+	0	0	0	0
	14	$	$	$	0	0	0	0	0	$	$	$	$	$	0	0	0	0	0	0	0	0	$	0	0	0	0	0	0	0	0	0	0
	15	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

Fig. 5.8 2DCLASS matrix of Fig. 5.6 (for the quadrat size 2), X-No. of rows distance; Y-No. of plant distance within row.

X

		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
	0	+	+	+	0	+	0	0	$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	0	0	0	0
	1	+	+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$	0	0	0	0	+	+	+	+	+	+	+	0	0	0
	2	+	+	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
Y	3	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
	4	+	0	0	0	0	0	$	$	0	$	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	+	0	0	0
	5	0	0	0	0	$	$	$	$	$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	+	+	+	+	+	0	0
	6	0	$	0	0	0	$	$	$	$	$	0	0	0	0	0	0	0	0	0	$	$	0	0	0	0	0	+	+	0	0	0	0
	7	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

Fig. 5.9 2DCLASS matrix of Fig. 5.6 (for the quadrat size 4), X-No. of rows distance; Y-No. of plant distance within row.

5.3 Sampling from a large array

In section 5.1 we performed 2DCLASS analysis for arrays of 12×30 plants. In section 5.2 we performed the same analysis for an array of 32×32 plants scoring the same pineapple wilt disease. The properties observed with array 32×32 were different to those observed with the array of 12×30. For instance, 2DCLASS analysis for all disease incidence maps of size 12×30 identified an average minimum core cluster size of eight plants in contrast to twenty-six with the 32×32 array. There is no published information about how array size could effect on properties observed in 2DCLASS analysis although researchers seem to make the tacit assumption that array size has no effect on the analysis. We investigated the effect of array size on the property of minimum core cluster size, which is perhaps the most important information obtained from 2DCLASS analysis.

From the 32×32 array (Fig. 5.6), we obtained random subsamples of different size arrays (Table 5.2). The smallest size (5×15) was sufficient to include the core cluster identified from the analysis of the full 32×32 array. Random number tables were used to identify the initial positions of these arrays in the original map (32×32 array). In addition, in selecting the co-ordinates of the sample arrays, number of plants per each row was always taken to be greater than number of rows since aggregation was more dominant along the rows compared to across rows. For each sample array, 2DCLASS analysis were performed. From each analysis, the average minimum core cluster size was determined (Table 5.2).

According to Table 5.2, the average minimum core cluster size corresponding to array size 12×30 is eight, and this tallies well with the average minimum core cluster identified with all 12×30 arrays in section 5.1.

Fig. 5.10 shows the relationship between array size as a proportion of the total area and average minimum core cluster size. This clearly shows a linear relationship between minimum core cluster size and the array size (proportion). A regression fit of a linear relationship showed a coefficient of determination (R^2) of 92%, and the fitted model with slope equal to 23.3 and intercept equal to 0.24. This shows that the average minimum core cluster size is directly proportional to the array size of the incidence map. In other words, the average minimum core cluster seems to be a scale-dependent characteristic.

Table 5.2 Average minimum core cluster size for different sample arrays

Sample size*	Number of grid points for the sample	Average minimum core cluster size
5×15 (15)	75	2.07
10×20 (3)	200	3.67
10×30 (3)	300	9.33
12×30 (3)	360	8.00
15×30 (3)	450	9.67
20×30 (3)	600	16.00
25×30 (3)	750	13.67
32×32 (1)	1024	25.00

* The values in the parenthesis are the number of array samples obtained for each array size and the largest array corresponding to the original incidence map of 32×32 array.

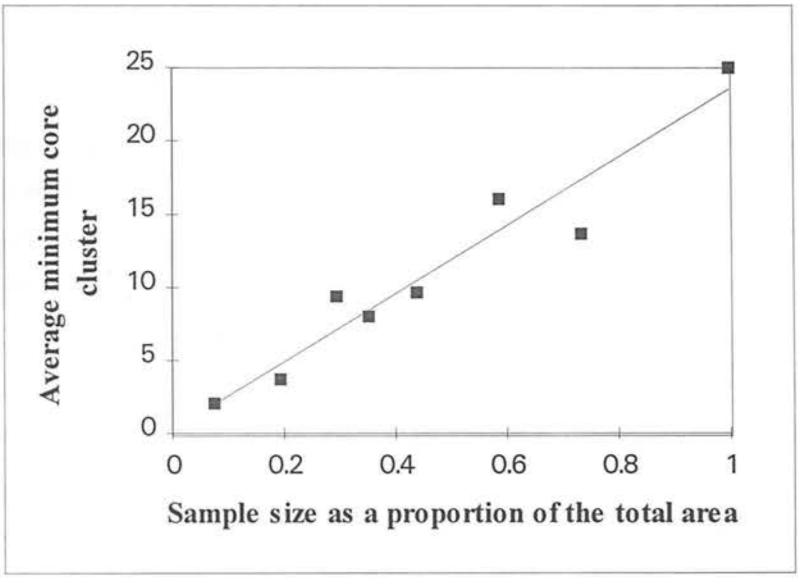


Fig. 5.10 Relationship between average minimum core cluster size and sample size as a proportion of the total area

Dr. S.C. Nelson (personal communication) suggested examining the relationship between minimum core cluster size as a proportion of the size of the distance class matrix (referred to as the Scaled Core Cluster [SCC]), and the proportion of the array size to the total area. This relationship is shown in Fig. 5.11.

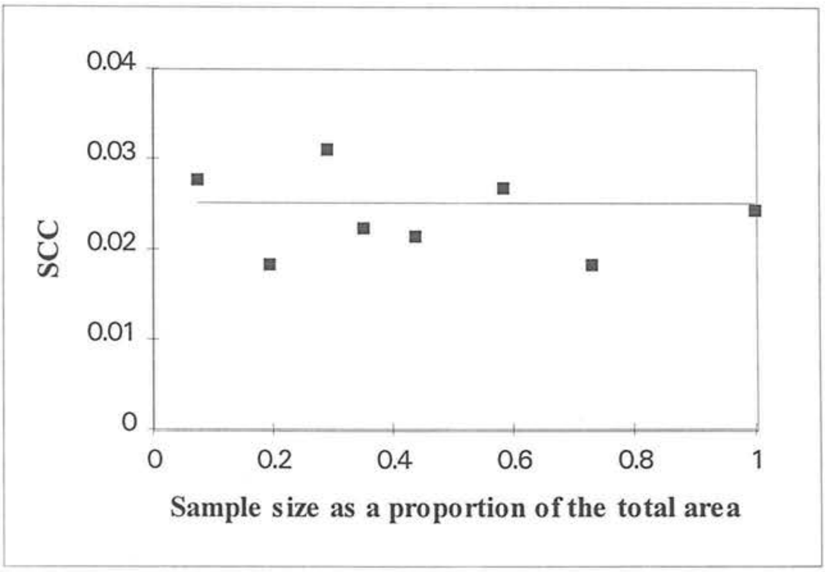


Fig. 5.11 Relationship between scaled core cluster (SCC) and sample size as a proportion of the total area

The fitted line is parallel to the *x*-axis in Fig. 5.11 confirming the findings of Fig. 5.10 that minimum core cluster as identified by 2DCLASS analysis is directly proportional to the array size. Since SCC is scale independent it may be a better parameter to be used in comparisons than the core cluster. However, SCC gives no information about the cluster size in a pathosystem. These findings suggest that caution should be exercised in drawing general conclusions with respect to the epidemiology of a disease on the basis of 2DCLASS analysis (Samita and Hughes, 1995) especially when array size has not been held constant. For instance, in Munkvold *et al.* (1993), 2DCLASS analysis has been carried out for different array sizes, but conclusions have been made with out taking account of the array size. However, conclusions in this section were based only on the data used in this study and further investigations will be carried out to find out whether these findings are repeated with other data sets.

6 CONCLUSION

Fitting the binomial distribution, or equivalently the linear logistic model, is the conventional method to analyse incidence data. It is true that if the incidence data satisfy the assumptions underlying the binomial distribution, this procedure is the most appropriate method to model the such data. However, in practical circumstances, sometimes, incidence data do not satisfy the assumptions underlying the binomial distribution. Thus, fitting the binomial distribution (or linear logistic model) is often inadequate to analyse such data. Use of binomial models in practical situations, such as toxicological testing, has been criticised (Haseman and Kupper, 1979) on the grounds that they generally provide a poor fit to actual experimental data.

Epidemiological studies of aggregated spatial patterns of disease incidence are very common in practice (Jeger, 1989). Fitting the binomial distribution (or linear logistic model) does not adequately describe disease incidence when the disease incidence is aggregated. The use of binomial models when unsuitable gives misleading results and thus should not be used at all in such situations. Clearly, routine use of χ^2 or Fisher's exact test for comparing the overall proportion of successes (diseased), which continues to be widespread in the biological literature, such as germination and epidemiological experiments, should be avoided. These procedures ignore the aggregation effect, and as a result may greatly inflate the type I error rate (Kruger, 1970; Schardein *et al.*, 1973; Haseman and Hogan, 1975; Haseman and Soares, 1976). One possible approach would be to carry out preliminary homogeneity tests, i.e. testing goodness-of-fit to the binomial distribution or linear logistic model and identifying any departure from random spatial pattern. If aggregation is present, an appropriate method for analysing such data could be chosen based on variance-variance plots (section 3.5). Since a variance-variance plot gives an insight into the spatial pattern of disease incidence, in model fitting, it is advisable to investigate the variance-variance plot in the first place, even before fitting the linear logistic model.

From chapter 4 we found that the logistic-normal binomial distribution (section 3.4.2) is far superior to the binomial distribution in describing disease frequencies for the data we analysed in this study. In the literature, the beta-binomial distribution (section 3.4.1) has been used as an alternative to binomial distribution in describing disease frequencies when the disease incidence is aggregated. However, the logistic-normal binomial distribution is much more appealing than the beta-binomial distribution in describing aggregated disease incidence, because the underlying assumptions of the logistic-normal binomial distribution agree with the intuitive notions (section 3.5) of what is happening in such data. This is the first study to illustrate the use of the logistic-normal binomial distribution in describing disease incidence data. However, our extensive investigation on the possibility of use of this distribution in describing disease incidence revealed that this distribution can effectively be used to describe and characterise aggregated spatial patterns in epidemiological studies. Thus we recommend to use the logistic-normal as an alternative to the beta-binomial distribution to describe aggregated disease incidence data.

As a basis for hypothesis testing, the logistic-normal binomial model is appealing intuitively since it provides a quantitative measure of the heterogeneity and usually gives a much better fit to real data than the simple binomial model. Because of the nature of the aggregation of the pineapple wilt disease incidence data (section 4.4.1) we chose the logistic-normal binomial model to analyse those data and we found that the logistic-normal binomial model is an effective statistical tool to analyse aggregated disease incidence data.

From the results of the statistical analysis of pineapple wilt data we found that, among the cultivars used in commercial cultivation, the cultivar Kew is more susceptible to the wilt disease than the cultivar Murici. The pesticides profenofos and prothiofos are effective in controlling pineapple wilt diseases. Moreover, the disease incidence can substantially increased with time and therefore control of the disease at the early stage is crucial.

As an alternative method to distribution fitting, 2DCLASS analysis, provides a quantitative assessment of the spatial pattern of infected plants. The dispersal from the initial loci may lead to random or clustered patterns of disease. Presence of small core cluster sizes may be due to the dispersal by a single vector. Although this information may be obtained by 2DCLASS analysis, summary statistics cannot be obtained by 2DCLASS analysis. Thus no quantitative comparisons can be made among different systems. Hence, 2DCLASS analysis may have to be carried out in association with statistical modelling in order to get better understanding of epidemiological data.

The 2DCLASS analysis was originally developed for the spatial evaluation of disease whose incidence is measured on individual plant basis using a presense or absense classification. But from this study we found that information based on quadrats can be used in 2DCLASS analysis with little loss of information. With the understanding of disease symptoms a single code could be made to a quadrat either as diseased or healthy. This may overcome the limitation of 2DCLASS software to accommodate large number of data points. Moreover, it saves time spent on data collection. Sampling being a part of a disease management program, this information could eventually help to economise the cost on such program.

Section 5.3 clearly demonstrated that 2DCLASS analysis is array size dependent (scale dependent). One may argue that aggregation may be dependent on field size. But section 5.3 confirms that the properties of 2DCLASS analysis, especially the core cluster size, are scale dependent. Thus the properties should strictly be discussed with reference to the size of the sample.

One other problem we found with 2DCLASS analysis is the criterion for identifying non-random spatial patterns. For instance, according to 2DCLASS software (Nelson *et al.*, 1992), if the proportion of distance classes with observed SCFs greater ($P < 0.05$) than 5%, then the incidence is said to have a non-random spatial pattern. But there is no objective basis for that 5% criterion. In section 4.3.1, we found certain disease incidence maps to have random spatial pattern. But in 2DCLASS analysis, all these maps appear to have non-random spatial pattern. As explained in chapter 5, in 2DCLASS analysis, strength of non-randomness is interpreted as being directly proportional to number of significant ($P < 0.05$) SCFs (Nelson *et al.*, 1992) and, in distribution fitting, strength of aggregation is accounted by means of the value of the aggregation parameter after fitting the logistic-normal binomial distribution (section 3.4.2). Fig. 6.1 shows the scatter plot of proportion of significant ($P < 0.05$) distance classes in 2DCLASS analysis versus the aggregation parameter obtained in logistic-normal binomial distribution fitting for the disease incidence maps (Fig. 4.3-4.6). Fig. 6.1 clearly shows that there is no relationship (rank correlation coefficient [Spearman's rho] estimated to be 0.16 with 14 degrees of freedom [$P = 0.55$]) between these two supposed indices of aggregation.

In distribution-fitting there is a definite statistical basis on which to decide whether a particular spatial pattern is random or not. Therefore it is well understood and established that distribution-fitting can be used to characterise spatial patterns. In contrast, in 2DCLASS analysis, the only statistical criterion used is to decide whether a SCF is significantly different from that expected on the basis of randomness. There is no agreed formal statistical basis to distinguish spatial patterns. It seems that the two methods measure different aspects of the data. More work is required on 2DCLASS analysis, especially on the development of 2DCLASS 'guidelines'.

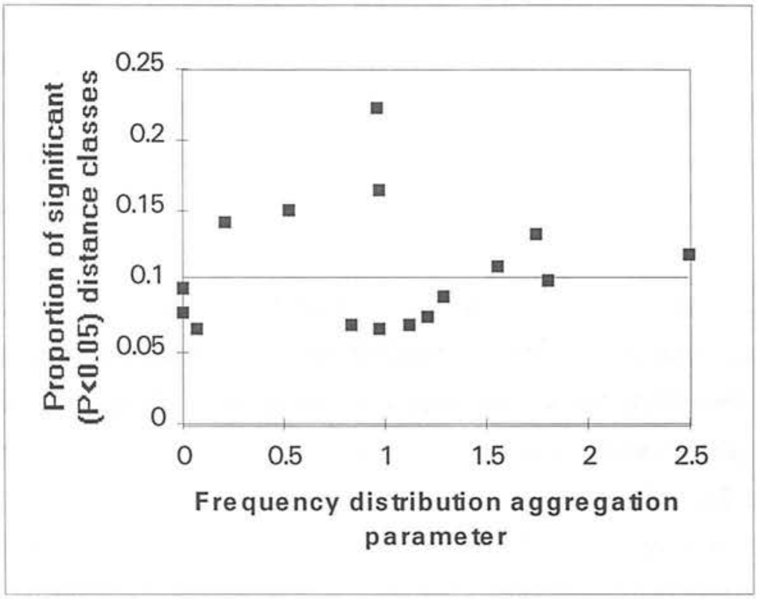


Fig. 6.1 The plot of proportion of significant ($P < 0.05$) distance classes verses frequency distribution aggregation parameter

Based on this study, we recommend use of 2DCLASS analysis only to investigate characteristics of spatial patterns (such as core cluster size, edge effect and relative location of clusters) with respect to the array size. The aggregation parameter in the logistic-normal distribution can be used to determine the spatial pattern of the disease incidence. If the logistic-normal binomial distribution is a significantly better fit to the data than the binomial, the data can be considered to be aggregated, and the aggregation parameter of the distribution is then a measure of the extent of aggregation. If the binomial distribution is an adequate description of the data, disease incidence can be considered as randomly dispersed. This is equivalent to assuming that the aggregation parameter of the logistic-normal binomial distribution is equal to zero. When the spatial pattern is regular (i.e., the data are underdispersed) the algorithm adopted in EGRET does not converge. In this scheme the extent of underdispersion cannot be evaluated by estimation of parameter, but in practice this is not much of a drawback.

From the available software, EGRET has of most of the required facilities to analyse disease incidence data. In fact EGRET is the only software which has facilities to fit models with random effects as a standard option. MATHCAD has the facilities for numerical integration and thus it can be used to fit logistic-normal binomial distribution. Spreadsheet software programs play an important role in data management in epidemiological studies. As reported earlier (Chapter 4) automated prepared spreadsheets (using spreadsheet software) were used in data collection in this study. With the availability of portable computers, these spreadsheets software programs can save time spend on data collection and thus can be used efficiently in epidemiological studies.

A number of different analysis techniques are available to analyse epidemiological data. Different techniques may illustrate different aspects of the data, and techniques can be misinterpreted easily. The properties described in all published analyses are not always well-understood. Careful selection of an appropriate technique plays a major role in reaching correct decisions. But the responsibility of interpretation should lie with the investigator, not with authors of software or analyses.

References

Abramowitz, M. and Stegun, I.A. (1972) *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. US Government Printing Office, Washington.

Aitchison, J. and Shen, S. M. (1980) Logistic-normal distributions: some properties and uses. *Biometrica*, 67, 261-272.

Aitken, M., Anderson, D.A., Fransis, B. and Hinde, J.P. (1989) *Statistical Modelling in GLIM*. Clarendon Press, Oxford.

Anderson, D.A. (1988) Some models for overdispersed binomial data. *Australian Journal of Statistics*, 30, 125-148.

Bald, J.G. (1937) Investigations on "Spotted Wilt" of tomatoes. III. Infection field plots. *Bulletin* 106, Council for Scientific and Industrial Research, Melbourne, Australia.

Beardsley, J.W. (1959) On the taxonomy of pineapple mealybugs in Hawaii, with a description of a previously unnamed species (Homoptera: Pseudococcidae). *Proc. Hawaii Entomol. Soc.* 17, 29-37.

Boos, D.D. (1993) Analysis of dose-response data in the presence of extra-binomial variation. *Applied Statistics*, 42, 173-183.

Brooks, R.J. (1984) Approximate likelihood ratio tests in the analysis of beta-binomial data. *Applied Statistics*, 33, 38-44.

Campbell, C. L. and Madden, L. V. (1990) *Introduction to Plant Disease Epidemiology*. Wiley-Interscience NY

Carter, W. (1932) Studies of populations of *Pseudococcus brevipes* (Ckl.) occuring on pineapple plants. *Ecology*, 13, 296-304.

Carter, W. (1933) The pineapple mealybug, *Pseudococcus brevipes*, and wilt of pineapples. *Phytopathology*, 23, 207-242.

Carter, W. (1956) Notes on some mealybugs (*Coccidae*) of economic importance in Ceylon. *FAO Plant protection Bulletin*, 4(4), 49-52.

Carter, W. (1963) Mealybug wilt of pineapple; A reappraisal. *Ann. N.Y. Acad. Sci.* 105, 741-764.

Cochran, W.G. (1937) The statistical analysis of field counts of diseased plants. *Journal of Royal Statistical Society*, Supplement 4, 49-67.

Cochran, W.G. (1977) W.G. *Sampling Techniques*. (3rd ed.). John Wiley and Sons, New York.

Collett, D. (1991) *Modelling Binary Data*. Chapman and Hall, London.

Converse, R.H., Seely, J. and Martin, L.W. (1979) Evidence for random local spread of aphid-borne mild yellow-edge virus in strawberries. *Phytopathology*, 69, 142-144.

Cox, D.R. and Snell, E.J. (1989) *Analysis of Binary Data*. (2nd ed.). Chapman and Hall, London.

Crowder, M.J. (1978) Beta-binomial ANOVA for proportions. *Applied Statistics*, 27, 34-37.

Department of Agriculture of Sri Lanka. (1993) *Pineapple Cultivation*. Booklet published by the Department of Agriculture of Sri Lanka.

EGRET (1991) *The EGRET System, Revision 3 Manual*. Statistics and Epidemiological Research Corporation, Washington.

Finney, D.J. (1947) *Probit Analysis*. (1st ed.), Cambridge University Press, Cambridge, UK.

Finney, D.J. (1971) *Probit Analysis*. (3rd ed.). Cambridge University Press, Cambridge, UK.

GLIM (1985) *The GLIM System, Release 3.77 Manual*. Numerical Algorithms Group: Oxford.

Gray, S.M., Moyer, J.W. and Bloomfield, P. (1986) Two-dimensional distance class model for quantitative description of virus-infected plant distribution lattices. *Phytopathology*, 76, 243-248.

Griffith, D.A. (1983) Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics*, 29, 637-648.

Gunasinghe, U.B. and German, T.L. (1989) Purification and partial characterization of a virus from pineapple. *Phytopathology*, 79: 1337-1341.

Haseman, J.K. and Hogan, M.D. (1975) Selection of the experimental unit in teratology studies. *Teratology*, 12, 165-172.

Haseman, J.K. and Kupper, L.L. (1979) Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, 35, 281-293.

Haseman, J.K. and Soares, E.R. (1976) The distribution of fetal death in control death and its implication on statistical tests for dominant lethal effects. *Mutation research*, 41, 277-288.

Hughes, G. and Madden, L.V. (1993) Using the beta-binomial distribution to describe aggregated patterns of disease incidence. *Phytopathology*, 83, 759-763.

Hughes, G. and Nelson, S.C. (1995) Spatial pattern analysis of plant virus diseases and its applications. *In*: *Epidemiology and Management of plant virus diseases*, Madden, L.V., Raccah, B. and Thresh, M.J. (ed.) Wiley-Interscience, NY (in press).

Hughes, H. and Madden, L.V. (1992) Aggregation and incidence of disease. Letter to the editor, *Plant Pathology*, 41, 657-660.

Ito, K. (1938) Studies on the life history of the pineapple mealybug, *Pseudococcus brevipes* (Ckl). J. Econ. Entomol. 31, 291-298.

Jeger, M.J. (1989) The spatial component of plant disease epidemics, *In*: *Spatial Component of Plant Disease Epidemics*, Jeger, M.J. (ed.), Prentice Hall, London, 1-13.

Kish, L. (1965) *Survey Sampling*. Wiley, New York.

Kleinman, J.C. (1973) Proportions with extraneous variance: single independent samples. *Journal of American Statistical Association*, 68, 46-54.

Kruger, J. (1970) Statistical research in mutation research. *In: Chemical Mutagenesis in Mammals and Man*. Vogel, F. and Rohrborn, G. (eds.) Springer-Verlag, Heidelberg, 460-502.

Larsen, L.D. (1910) Diseases of pineapple. *Hawaii Sugar Plant. Assoc. Pathol. Physiol. Ser. Exp.* Stn. Bull. 10: 1-72.

Luning, K.G., Sheridan, W., Ytterborn, K.H. and Gullberg, U. (1966) The relationship between the number of implantation and the rate of intra-uterine death in mice. *Mutation Research*, 3, 444-451.

Madden, L.V. (1989) Dynamic nature of with-in field disease and pathogen distributions. *In: Spatial Components of Plant Disease Epidemics*. Jeger, M.J. (ed.), Prentice Hall, London, 96-126.

Madden, L.V. and Hughes, G. (1994a) BBD-Computer software for fitting the beta-binomial distribution to disease incidence data. *Plant disease*, 78, 536-540.

Madden, L.V. and Hughes, G. (1994b) *The BBD Operating Manual, version 1.2.*

Madden, L.V. and Hughes, G. (1995) Plant disease incidence: Distributions, heterogeneity, and temporal analysis. Annu. Rev. *Phytopathology* (in press)

Madden, L.V., Hughes, G. and Ellis, M.A. (1994) Spatial pattern of the incidence of grape downy mildew. *Phytopathology*, 84, [Abstract] (in press).

Madden, L.V., Louis, R., Abt, J.J. and Knoke, J.K. (1982) Evaluation of test of randomness of infected plants. *Phytopathology*, 72, 195-198.

Madden, L.V., Pirone, T.P. and Raccah, B. (1987) Analysis of spatial patterns of virus-diseased tobacco plants. *Phytopathology*, 77, 1409-1417.

MATHCAD +5 (1994) *MATHCAD plus 5 user guide*, Mathsoft Inc, Cambridge,USA.

McCullagh, P. and Nelder, F.A. (1983) *Generalized Linear Models*. (1st ed.). Chapman and Hall, London.

McCullagh, P. and Nelder, F.A. (1989) *Generalized Linear Models*. (2nd ed.). Chapman and Hall, London.

Mendenhall, W., Wackerly, D.D. and Scheaffer, R.L. (1990) *Mathematical Statistics with Applications*. (4th ed.). PWS-KENT Publishing Company, Boston, USA.

Microsoft EXCEL (1992) Microsoft EXCEL user guide, Microsoft Corporation.

Moore, D.F. (1986) Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika*, 73, 583-588.

Moore, D.F. (1987) Modelling the extraneous variance in the presence of extra-binomial variation. *Applied Statistics*, 36, 8-14.

Moran, P.A.P. (1968) *An Introduction to Probability Theory*. Oxford University Press, London.

Munkvold, G.P., Duthie, J.A. and Marois, J.J. (1993) Spatial patterns of grapevines with Eutypa dieback in vineyards with or without perithecia. *Phytopathology*, 83, 1440-1448.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized linear models. *Journal of Royal Statistical Society,* A 135, 370-384.

Nelson, S.C. and Campbell, C.L. (1993) Comparative spatial analysis of foliar epidemicson white clover caused by viruses, fungi, and a bacterium. *Phytopathology*, 83, 288-301.

Nelson, S.C., Marsh, P.L. and Campbell, C.L. (1992) 2DCLASS, a two-dimensional distance class analysis software for the personal computer. *Plant Disease*, 76, 427-432.

Nixon, G.E.J. (1951) The association of ants with aphids and coccids. *Commonwealth Institute of Entomology*, London. 36.

Patil, G.P. and Joshi, S.W., (eds.). (1968) *A Dictionary and Bibliography of Discrete Distributions*. Oliver and Boyd, Edinburgh.

Paul, S.R. (1982) Analysis of proportions of affected foetuses in teratological experiments. *Biometrics*, 38, 361-370.

Pielou, E.C. (1977) *Mathematical Ecology*. John Wiley and Sons, New York.

Pierce, D.A. and Sands, B.R. (1975) Extra-bernoulli variation in binary data. *Technical Report*, 46, Department of Statistics, Oregon State University.

Proctor, C.H. (1984) On the detection of clustering and anisotrophy using binary data from a lattice patch. *Commun. Statist.-Theor. Meth.*, 13, 617-638.

Qu, Y., Green, T. and Piedmonte, R. (1993) Symmetric Bernoulli distribution and generalized binomial distributions, *Biometrics*, 32, 229-242.

Rao, J.N.K and Scott, A.J. (1992) A simple method for the analysis of clustered binary data. Shorter communications, *Biometrics*, 48, 577-585.

Rohrbach, K.G., Beardsley, J.W., German, T.L., Reimer, N.J. and Sanford, W.G. (1988) Mealybug wilt, mealybugs and ants on pineapple, *Plant Disease*, 72, 558-565.

Ross, G.J.S. (1970) The efficient use of function minimization in non-linear maximum-likelihood estimation. *Applied Statistics*, 19, 205-221.

Samita, S. and Hughes, G. (1995) Pineapple wilt disease of Sri Lanka. *Abstracts.* The 6th International Plant Virus Epidemiology Symposium, The International Society of Plant Pathology, 35.

SAS (1990) *The Statistical Analysis System, Release 6.04 Manual*. SAS Institute Inc. North Carolina, USA.

Schardein, J.L., Dresner, A.J., Hentz, D.L., Petrere, J.A., Fitzgerald, J.E. and Kurtz, S.M. (1973) The modifying effect of folinic acid on diphenylhydantoin-induced teratogeneicty in mice. *Toxicology and Applied Pharmacology*, 24, 150-158.

Shiyomi, M. and Takai, A. (1979) The spatial pattern of infected or infested plants and negative hypergeometric series. *Japanese Journal of Applied Entomology and Zoology*, 23, 224-229 (in japanese).

Singh, S.J. and Sastry, K.S.M. (1974) Wilt of pineapple-a new virus disease in India. *Indian Phytopathology*, 27: 298-303.

Skellam, J.G. (1948) A probability distribution derived from the binomial distribution by regarding probability of success as variable between sets of trials. *Journal of Royal Statistical Society*, B 10, 257-261.

Smith, D.M. (1983) Maximum likelihood estimation of the parameters of the beta-binomial distribution. *Applied Statistics*, 32, 192-204.

Snedecor, G.W. and Cochran, W.G. (1989) *Statistical Methods*. (8th ed.). Iowa State University Press, Ames, Iowa.

Steel, R.G.D. and Torrie, J.H. (1980) *Principles and Procedures of Statistics*. (2nd ed.), McGraw-Hill, New York.

Wedderburn, R.W.M. (1974) Quasilikelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 61, 439-447.

Williams, D.A. (1982) Extra-binomial variation in logistic linear models. *Applied Statistics*, 31, 144-148.

Williams, D.A. (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, 31, 949-952.

Appendix I

```
$macro disp
$cal %z1=%coc $out $warn $cal %z2=1 : %z3=%z4=0 : pw_=1
$swi %pwf pwt_ $cal w_=pw_ $while %z2 est_ $out %z1
$pr : 'estimate of phi = ' %z3 ' after '*i %z4 ' iterations' :
: 'deviance for full model, parameter estimates and standard errors'
: 'using a heterogeneity factor are as follows:' : $wei w_ $f + $dis e
$pr : 'note: weight directive with weights in w_ is operative' :
$warn $del wv_ n1_
$$endmac

$macro est_
$wei w_ $f + $ext %vl $cal wv_=%wt*%vl*%pw : %z5=%z3
$cal n1_=(%bd-1)*(%wt/%bd)**(%b-1)
: %z3=(%x2-%cu(%pw*(1-wv_)))/%cu(n1_*%pw*(1-wv_)) : w_=(1+%z3*n1_)
: w_=pw_/w_ : %z2=%z5-%z3 : %z2=%sqrt(%z2*%z2)>=0.0001 : %z4=%z4+1
$$endmac

$macro pwt_ $cal pw_=%pw $$endmac

$m fn
$ca %f=%cu(((%yv/%bd-%fv/%bd)**2
   -(%fv/%bd*(1-%fv/%bd)/%bd
   +%z3*(%fv/%bd*(1-%fv/%bd))**%b))**2)
$pr %z3 : %b $
$ac 8 $pr %f $$e

$return
```