

# **Case Study in Data Analysis: Variables related to codling moth abundance and the efficacy of the Okanagan Sterile Insect Release Program**

**Étude de cas en analyse des données :**

**Variables liées à l'abondance du  
carpocapse de la pomme et efficacité du  
programme d'épandage d'insectes stériles  
de l'Okanagan**

*Case Studies Editor / Responsable des études de cas : Sylvia R. ESTERBY*

*Key words and phrases:* Codling moth; Cox model; Fourier series; Geographic Information System; hidden Markov models; hierarchical models; integrated pest management; logistic analysis; low counts; Markov chain Monte Carlo; non-stationary Poisson process; penalized quasi-likelihood; Poisson regression; principal components; proportional hazards; random effects; spatial correlation and neighbourhoods; spatial-temporal models; sterile insect release; Tobit; Weibull regression; zero-heavy analysis.

*MSC 2000:* Primary 62P10; secondary 62J02; 62F15; 92B15.

*Abstract:* The authors consider the effect of orchard attributes and landscape in a heterogeneous area on the efficacy of a control program for the codling moth in apple orchards in British Columbia. The context is first presented, along with a set of questions of importance to the Okanagan Valley Sterile Insect Release program. Two groups of analysts then address a number of these issues using methods for spatial-temporal data including counts, proportions and Bernoulli variables. The models are then compared and the relevance of the results to this operational program is discussed.

*Résumé:* Les auteurs examinent l'influence des caractéristiques des vergers et de la topographie d'une région accidentée sur l'efficacité d'un programme de contrôle du carpocapse de la pomme dans les vergers de Colombie-Britannique. On présente d'abord le contexte, ainsi que diverses problématiques liées à un programme d'épandage d'insectes stériles dans la vallée de l'Okanagan. Deux équipes abordent ensuite plusieurs de ces questions en faisant appel à des techniques adaptées à des données spatio-temporelles se présentant sous la forme de dénombresments, de proportions et de variables de Bernoulli. Enfin, les modèles sont comparés et un jugement est porté sur la pertinence des résultats sur l'administration du programme.

## ***Introduction:***

# **A GIS application to improve codling moth management in the Okanagan Valley of British Columbia**

Bob VERNON, Howard THISTLEWOOD and Scott SMITH

### **1. INTRODUCTION**

#### *1.1. Background.*

Codling moth, *Cydia pomonella* L., is the most damaging insect pest of apples and pears in Canada and worldwide. The Okanagan-Kootenay Sterile Insect Release Board has operated an area-wide codling moth management program in British Columbia (B.C.) in the key fruit-growing areas and in neighbouring urban, native, and public lands, since 1994. As of 2004, the program now services the 200km long Okanagan Valley, the Similkameen Valley, parts of the Creston, Columbia, and Shuswap Valleys, and is divided into three Zones from North to South. It is the largest horticultural project in Canada, operating in more than 5,000ha of pome fruits and in more than 29,000 non-commercial sites such as parks and private gardens at a cost to date of approximately \$54 million (Canadian). In both commercial and non-commercial situations, it has succeeded in providing major reductions in numbers of codling moths, damage at harvest, and associated insecticide use. This has occurred as the result of the integration of many alternative control technologies against codling moth, including the globally unique operation of a Sterile Insect Release (SIR) program.

The SIR Program was created with the ambitious goal of eradicating codling moth from the target areas between 1994 and 2000 (Dyck & Gardiner 1992), and was developed from a career of research into autocidal control, later referred to as sterile insect release by M. D. Proverbs and colleagues in B.C. (Dendy 1999). The efficacy of SIR in controlling codling moth populations was demonstrated convincingly from 1976 to 1978 in 520ha of apples and pears in the geographically isolated Similkameen Valley. As of 1978, codling moth and its associated damage had been brought near to extinction in the study area (Proverbs, Newton & Campbell 1982), and no additional measures for codling moth control were required in most of the orchards until 1981. The technique, however, was about three times more expensive than conventional approaches, which at that time relied solely on the use of synthetic insecticides.

In 1994, releases of sterile moths began in the south Okanagan Valley (Osoyoos and US border to Summerland), Similkameen and Creston Valleys, collectively referred to as Zone 1. The combination of several codling moth management techniques with SIR did reduce codling moth populations in Zone 1 to unprecedented low levels, but the initial goal of area-wide eradication was not achieved in the expected time period. This was due largely to operational difficulties in effectively delivering the program on a large scale, and to greater variability in the efficacy of the SIR approach in controlling codling moth in more heterogeneous landscapes (in contrast to the smaller and more homogeneous landscape in the Similkameen Valley where SIR was first demonstrated). As a result, in 1999 the main goal of the program was shifted from eradication to area-wide suppression, and the target date for completion of the program in Zones 1–3 was extended from 2000 to 2006. Beyond 2006, the program is expected to operate at a subsistence level to maintain suppression in all Zones, and be completely funded by fruit-growers.

This case study investigates how efficacy of SIR against low levels of codling moth might vary according to differences in orchard and landscape attributes in a heterogeneous agricultural area in Zone 1 of the program (i.e., Summerland, B.C.). The effectiveness of SIR, as measured by pheromone trap catches of codling moth and damage to apples at harvest, will be correlated with

site specific orchard and landscape measurements that are contained in a Geographic Information System (GIS) compiled and maintained by Agriculture and Agri-Food Canada (A.A.F.C.) through Pacific Agri-Food Research Centre (PARC) in Summerland, and the Okanagan-Kootenay Sterile Insect Release Board.

### *1.2. Codling moth biology, population monitoring and management.*

Madsen & Procter (1982) summarized the biology of the codling moth in British Columbia: codling moth overwinters as a diapausing larva (first generation) in protected locations on host trees, wooden structures, or in the leaf litter at the base of trees. Pupation and emergence of adults occurs in the spring and is temperature-dependent. Adult moths appear by early May and continue to emerge until late June. Both sexes can crawl or fly within the tree and males search for females by cues including a female-produced sex pheromone. Mating and egg-laying occurs in late June and July, and the resulting larvae enter and feed in the fruit, potentially causing extensive damage. When larger, the larvae leave the fruit, search for pupation sites, and a variable proportion enter diapause for one to two years. The remainder pupate immediately and moths of the second generation appear in late July and August. Summer weather leads to more rapid development and more serious damage from larvae arising from the second generation, but owing to the long emergence period of the first flight, overlapping of generations occurs. If weather is favourable in late August or early September, a partial third generation can occur. Moth dispersal occurs via short- and long-distance flights up to several km (Thistlewood & Judd 2002), and via passive transfer in and on fruit containers (Proverbs & Newton 1975).

In the SIR program, codling moth populations are monitored weekly throughout each growing season in apple and pear orchards using codlemone-baited pheromone lures in sticky traps (1/ha), and by inspections of ca. 1,000 fruit per orchard at harvest (Madsen & Vakenti 1973). Codlemone (a synthetic copy of the female sex pheromone (E,E)-8,10-dodecadien-1-ol, 99% isomerically pure, Shin-etsu, Fine Chemicals Division, Tokyo, Japan) is selectively attractive to males of the species, and a trap baited with a 1mg lure is estimated to capture ca. 5% of a wild population within one hectare. When pheromone traps are inspected, the male moths captured are determined to be “wild” (not artificially reared and released) or “sterile” (released by the SIR program). Sterile moths are differentiated from wild moths directly in traps by a red dye acquired through their artificial diet.

When the SIR program is first introduced to an area, wild codling moth populations have to first be reduced to low levels at all target sites (i.e. orchards, gardens, etc.) before sterile insect release can begin (see below). Adequate population reduction has been achieved in some orchards by insecticide use and in other orchards by the intensive use of pheromone lures dispensed throughout the tree canopy to disrupt moth mating (Judd, Gardiner & Thomson 1996). The two types of codling moth controls (insecticide and mating disruption) require the use of different trap types and lure rates. Lures in insecticide-treated orchards are baited with 1 mg of codlemone, and in mating disruption orchards with 10 mg of codlemone. Similarly, a Pherocon 1-CP sticky wing trap (PheroTech Inc., Delta, B.C.) is deployed in insecticide-treated orchards, whereas a Delta shaped open trap with a larger interior sticky surface is used in mating disruption orchards. The pheromone traps used in the mating disruption orchards had to be more attractive than the standard traps in order to work in a pheromone-saturated environment.

### *1.3. Sterile insect release.*

For 18–22 weeks each year, some 2 million codling moths are mass reared daily, marked internally by a red dye in their diet, sterilized by gamma radiation (250Gy) from a Cobalt 60 source, and distributed into orchards. The sterile males locate and mate with local wild females, and the effect on the natural population is a major reduction in viable offspring. (Sterile females also

released may have an effect in acting as a “sink” for wild male sperm but this is regarded as much less important). Owing to the poor competitive ability of mass-reared sterile males, Proverbs, Newton & Campbell (1982) estimated that the ratio of sterile males to wild males had to be approximately 40:1 for a population effect. At this “overflooding ratio,” successful mating between fertile wild moths is rare and, over time, populations can be driven to near extinction.

Sterile moths are released from blower units mounted on a fleet of all-terrain vehicles (ATVs). The ATVs move along designated rows of trees dispensing moths in a manner designed to deliver moths uniformly over the area of an orchard and at a rate of approximately 2000 moths/ha. The period of release of sterile moths each year is determined by weather conditions and operational considerations, with the aim of overlapping the period of wild moth activity. In the study area for the years examined, releases began in the third week of April and ended in the second week of September.

#### *1.4. Development of the Geographic Information System (GIS) database.*

The use of GIS, especially when combined with the field of geostatistics, is a new and powerful tool that can be used in integrated pest management (IPM) programs to assist in better understanding the ecological relationships of pest arthropods and their surrounding landscape (e.g., Johnson 1989). With this objective in mind, a GIS database was created by A.A.F.C. and the SIR Program as a tool for managing the extensive wild versus sterile moth population data (from regular checks of more than 5,000 pheromone traps annually) and fruit damage data (from more than 30,000 orchards and gardens within the program area). All codling moth and damage data are spatially linked to trap and/or orchard locations, and all orchards are referenced with relevant attribute data.

Each of the pheromone traps was georeferenced using hand-held Global Positioning System (GPS) units, and location points entered into base maps in the GIS (ArcInfo, ©ESRI, Inc.). Trap locations within orchards were generally fixed from year to year, however the database was updated annually to reflect new, moved or deleted traps. Georeferencing provided location co-ordinates and elevation at each trap location, and 17 additional site attributes were also measured and recorded, including physical aspects of the orchard and surrounding topography (see Section 2.2).

In order to geographically locate traps and orchard blocks, a map of municipal lot boundaries was created and overlaid onto the provincial ‘TRIM’ (terrestrial resource information management) digital base map (BC Ministry of Sustainable Resource Management), which acted as the cartographic base for the application. As no digital provincial cadastral maps were available at the time of project initiation, lot boundaries had to be obtained from each of the municipal and regional jurisdictions within the SIR program area. These lot boundary files were converted to digital ArcInfo coverages, merged with adjoining coverages, linked to a unique lot identifier and then projected to fit the 1:20,000 scale TRIM base. Using the merged cadastral maps, it was then possible to link trap locations and crop damage information to legal lots which provided a spatial basis for managing the SIR data. Although GIS applications exist for a number of other agricultural pests in Canada (Johnson, Olfert, Dolinski & Harris 1996) this is the only application developed specifically to support a codling moth control program.

#### *1.5. Linking orchard attributes to success of SIR.*

Preliminary analyses of the 1999 codling moth and orchard attribute data were conducted by ourselves (unpublished results) and Marston, Barg & Cranston (2002). The results suggested that the efficacy of SIR varies according to certain physical attributes of orchards, including tree size and orchard elevation. The catch of wild moths in pheromone traps during May and June increased in orchards with larger trees, whereas the number of released sterile moths recaptured in traps appeared to be consistent regardless of tree size. In Penticton, the region with the highest

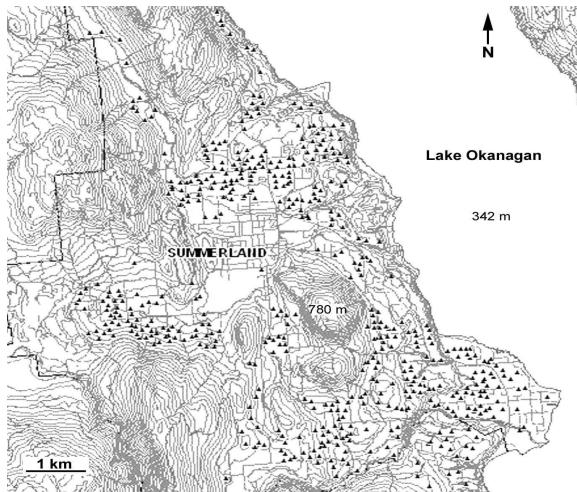


Figure 1–1: Summerland study area showing topography and the distribution of codling moth traps.

populations of wild moths in 1999, the ratio of sterile to wild moths in May and June declined as tree size increased. The preferred overflooding ratio of 40:1 was generally achieved only in orchards with smaller tree size. Catch of wild moths in May and June was typically highest in orchards at lower elevations, whereas sterile moth catches increased gradually with increased elevation.

## 2. CASE STUDY DATA AND OBJECTIVES

### 2.1. Region selection and location identifiers.

Operational questions were important in selection of the data, in order to assess the effect of trap-location or orchard attribute variables on relative wild versus sterile codling moth abundance and success of the SIR program. A dataset was required with low counts of wild codling moths in traps (so that most sites were able to rely upon SIR as the primary control method) and that provided considerable range in values of topographic and attribute variables within a relatively small but contiguous area. Following these considerations, data from the Summerland region (Figure 1–1) for the period of 1999 to 2001 were chosen for the case study.

### 2.2. Description of variables.

The database contained information for 255 trap locations organized in three data tables. A trap location table contained identification and legal lot location providing spatial resolution to: property and trap level, trap number, jurfolio (combination of legal jurisdiction, folio), trapid (combination of trap number, jurfolio making a unique field for relating to other tables), *x* co-ordinate (UTM Northing location), *y* co-ordinate (UTM Easting location), GPS date, and GPS height (elevation of trap location in m). The other trap attributes described were: trap style

(wing trap/delta trap), ground cultivation (description of ground cover surrounding the trap tree), canopy cover (% canopy coverage of sky), horizontal cover (looking across and through rows, % canopy covering open space), tree height (feet), within row tree spacing (feet), between row spacing (feet), basal diameter (cm), double planted (yes/no for an unusual type of orchard having double rows of trees planted close together), irrigation type (drip or overhead), slope (percent slope), slope type (bench/bottom of depression/bottom of gully or hill/near top of small mound or slope/near top of large hill or slope), aspect (compass bearing), orchard quality (good/poor), adjacent crop (type of orchard or if adjacent to block with trap tree), distance (from trap to urban site, if 0m to 200m only), urb200 (trap is 100.1 to 200m from an urban site, yes/no), and distance urb100 (trap is 0.1m to 100m away from an urban site, yes/no).

A trap count table included the count of sterile moths and count of wild moths at each trap site on a given date, being cumulative moth catches over a one-week period, where the week generally included at least one sterile moth release event. The variables were: trap number, trapid, jurfolio, countdate (date the moth count occurred in the field), wild moths (number in trap), sterile moths (number in trap), and program year (1999, 2000 or 2001).

An orchard damage table provided the measured area of apple and pear orchard, the counts of fruit sampled and fruit damaged, by legal lot and year. A sample of apples was taken at harvest using a consistent method and the number of damaged apples was recorded. Results were expressed as the number of damaged apples and as the percentage of apples damaged. The variables were (explanation provided for 2001 only): jurfolio, appleac01 (apple acres in 2001), pearac01 (pear acres in 2001), sample2001 (number of fruit sampled in 2001), damage2001 (number of fruit damaged by codling moth in 2001), pctdam01 (= damage2001/sample2001\*100), appleac00, pearac00, sample2000, damage2000, pctdam2000, appleac99, pearac99, sample1999, damage1999, and pctdam1999.

### *2.3. Case study objectives and suggestions for analysis.*

Prior to 1998, the aim of the SIR Program was to reduce the numbers of wild moths to zero at 95% of targeted sites. Program staff had expected this to occur in all sites within three years of program initiation, but observed that some sites took a very long time or never reached zero. Spatial relationships are very important for codling moth management; terrain and land use in the British Columbia interior is varied and fractured into small parcels, and moth movement is interrupted by many features (Dorn, Schumacher, Abivardi, Meyhofer & Brown 1999, Thistlewood & Judd 2002). Analysis of the entire SIR program trap data from Zone 1 (1994–2002) using the GIS database revealed a low level of achievement of the overflowing sterile:wild codling moth ratio (Thistlewood, Judd & Clodius 2004).

The main goal of establishing the SIR database described in section 2.2 was to determine if certain physical or spatial attributes associated with orchards positively or negatively affect the efficacy of the SIR approach. With this goal in mind, the objectives of the current study, as presented to the analyst teams, were to conduct an in-depth analysis of the orchard attribute and codling moth monitoring databases collected in the SIR program to determine:

1. Whether certain physical attributes (either singly or in combination) associated with an orchard, or trap site, have a positive or negative effect on populations of wild or sterilized male codling moths. Data relating to CM damage to apples at harvest for particular orchard sites has been included in the database and it should be included in the analyses as well as the pheromone trap data. The question is whether any of the orchard variables contribute to higher or lower levels of damage?
2. Whether variables (singly or combined), identified as being of interest in 1), above relate to their ‘relative’ effects on wild vs sterilized moth counts. For example, does a specific variable have a positive effect on both wild and sterilized moth counts, or are the effects inversely

proportional (i.e., positive effect on number of steriles equates to negative effect on numbers of wilds). Would certain orchard variables contribute to greater or less success in achieving the 40:1 overflooding ratio?

3. Whether certain variables can be correlated with earlier or later dates of first wild moth catch, or whether catch of released steriles takes longer to increase to maximum levels in orchards with certain characteristics.

#### 2.3.1. Tree size.

In a preliminary analysis, a new variable called “Tree Area” was generated by multiplying together two attributes, “Within Row” (spacing between trees down a row) and “Between Row” (spacing between tree rows). We assumed that the “Tree Area” variable would usually reflect the size of trees in an orchard, since wide spacing between trees and rows permit trees that are larger in area (common in traditional plantings with large trees), and small spacings only permit small trees planted at higher densities (common in newer, trellised style plantings). Another variable related to tree size, “Basal Diameter” was found to correlate fairly well with the “Tree Area” variable, whereas the “Tree Height” variable did not. Two other orchard attributes, “Horizontal Cover” and “Canopy Cover” were deemed too subjective to be of use in defining tree area. We suggest that any analysis should include a “Tree Area” variable derived from the “Within Row” and the “Between Row” variables, perhaps with the “Basal Diameter” variable also included. A final derived “Tree Area” variable will likely be of importance, since microclimate (temperature, sunlight, wind, etc.) can be quite different in more sheltered traditional orchards versus more open trellised orchards.

#### 2.3.2. Temporal concerns.

Pheromone trap data were collected on a weekly basis from May until September in all years. The numbers of wild and sterile moths often increased or decreased in orchards in a reciprocal fashion with time, being higher for wild moths in the first generation (May to June) than later in summer, and vice versa for sterile moths. Therefore, the time of sampling should be taken into account in the analysis to detect changes in the relative sterile versus wild codling moth population trends over the growing season.

#### 2.3.3. Dealing with zeros.

As a consequence of mass rearing and successful management, many more sterile moths occur in traps than wild moths. In fact, the vast majority of trap sites in Summerland for the 1999 season contained zero wild moths, which makes certain analyses challenging (i.e., when attempting to study the success or failure in achieving the 40:1 overflooding ratio). Generally, however, there were ample sterile moths for analysis at any time, which at least provides some information on the relative competitiveness of sterile moths in various orchard habitats.

### ACKNOWLEDGEMENTS

We thank Robert Fugger, Tiiia Meere, and Brenda Moreira of the Okanagan-Kootenay Sterile Insect Release Board for providing the case study data and Figure 1-1; Chris Marston, Kevin Barg, and Jerome Cranston of British Columbia Institute of Technology for initial GIS analyses and discussions; Grace Frank, GIS analyst at PARC, for the design and establishment of the SIR GIS application, Todd Kabaluk of PARC, Agassiz for determining the orchard attributes collected in the GPS surveys and for preliminary analyses, Rob Batenburg and Markus Clodius of PARC for data collection and management assistance. Howard Thistlewood was Director and Technical

Director of the SIR Program from 1998–2001. Research funding and support was provided in part by the Okanagan-Kootenay Sterile Insect Release Board and by the Matching Investment Initiative of A.A.F.C.

---

Bob VERNON: [vernonbs@agr.gc.ca](mailto:vernonbs@agr.gc.ca)  
*Pacific Agri-Food Research Centre  
Agriculture and Agri-Food Canada  
Agassiz, British Columbia  
Canada V0M 1A0*

Howard THISTLEWOOD: [thistlewoodh@agr.gc.ca](mailto:thistlewoodh@agr.gc.ca)  
Scott SMITH: [smithcas@agr.gc.ca](mailto:smithcas@agr.gc.ca)  
*Pacific Agri-Food Research Centre  
Agriculture and Agri-Food Canada  
Summerland, British Columbia  
Canada V0H 1Z0*

## **Analysis/Analyse 1: Codling moth incidence in Okanagan orchards**

Farouk NATHOO, Laurie AINSWORTH, Paramjit GILL and Charmaine B. DEAN

### 1. INTRODUCTION

The SIR case study addresses important concerns for the British Columbia fruit industry. On the statistical side, it is rich with challenges, including the analysis of small counts and proportions, the analysis of temporal data which may be spatially correlated, hierarchical random effects modeling, and handling count data with a high proportion of zeros. In the subsequent discussion we address a few questions of interest in some preliminary analyses.

Recently there have been substantial developments both in methods for the analysis of count data and in methods for spatial analyses. Balshaw & Dean (2002), Lawless & Nadeau (1995) and Sun & Wei (2000) consider methods for regression analysis of panel count data. With panel data, the number of events which have occurred within specific time intervals are available rather than the exact times of occurrence of events. This is the case for the SIR study where events refer to the arrival of moths in traps. Methodological developments for spatial analyses of rates have also seen tremendous development (Besag, York & Mollié 1991; Yasui & Lele 1997, for example). MacNab & Dean (2001) and Waller, Carlin, Xia & Gelfand (1997) discuss methods for spatio-temporal analyses. Essentially, hierarchical or multilevel models accommodate multiple sources of variation using random effects in a generalized linear mixed model analysis. Spatially, temporally or spatio-temporally correlated random effects may be included. Autoregressive random effects may handle both temporal and spatial correlations, for example, and flexible forms such as splines may be used for incorporating temporal trends. Methods for analysis include approximations of the likelihood common in generalized linear mixed models, for example quasi-likelihood techniques, and Bayesian approaches using Markov chain Monte Carlo (MCMC) methods.

The objective we considered for the case study was to determine the geophysical attributes associated with trap sites which have an effect on the population of wild moths. We also considered which geophysical attributes were related to the proportion of wild moths out of the total collected.

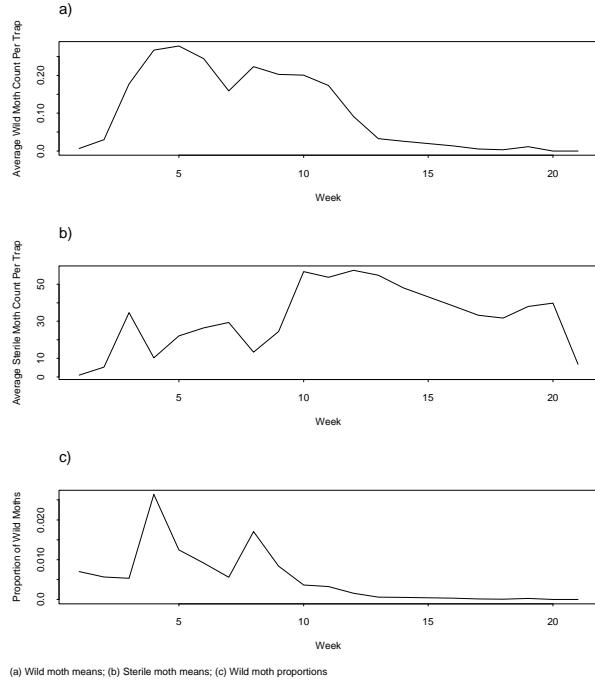


Figure 2-1: Wild and sterile moth means and wild moth proportions over time.

The SIR program seems very much in control in 2000 and 2001 with the wild moth occurrence over the whole area being extremely sparse. Only 290 and 263 wild moths were captured in 2000 and 2001 respectively. On the other hand, in 1999, 1069 wild moths were captured. We focus throughout on analyses of moth counts in 1999. Plots of the mean number of wild moths caught per trap over time and the proportions of wild moths out of the total caught are provided in Figure 2-1. These overall means and proportions are quite low. Although not displayed here, note, however, that there is considerable variability in these quantities over traps and there are some very large counts for a few individual traps.

Principal component analysis was employed to investigate reducing the dimensionality of the four tree size/space variables: Tree Height, Diameter, Between Row spacing and Within Row spacing. The first component reflected overall tree size; an average of tree size (Height and Diameter) and space available for growth (Between and Within Row spacing). The second component was a contrast between tree size and available space: Diameter and Height versus Between and Within row spacing, a reflection of whether a tree is small in a large space or overcrowded in a small space. In terms of model building, direct use of tree size/spacing variables was comparable to the use of principal components. In the spirit of simplicity we chose to use the most parsimonious model based directly on variables which are considered important by the scientists: Tree Area (a product of the Within Row and Between Row spacings), Height and Diameter, rather than on principal components. One severely outlying value for Tree Area was reassigned to the second largest value.

The variable Aspect in combination with Slope had several special cases with few traps in each. We simplified these by considering five cases: flat land, on a slope facing broadly north, south, east or west. For example, a slope facing south had Aspect between south-east and south-west. There were three missing values for trap type. These were imputed as the more common type, Wing traps; sensitivity analyses indicated no striking effect of the imputation for these particular traps. Table 2-1 provides descriptive statistics for these variables.

Table 2–1: Mean elevation, tree area and diameter by trap type and direction

	Mean (S.D.)			Proportion
	Elevation in meters	Tree Area in sq feet	Diameter in centimeters	
Wing/Unknown Trap ( $n = 448, 3$ )	480.3 (61.01)	177.5 (99.63)	18.9 (9.04)	0.89
Delta Trap ( $n = 55$ )	446.4 (74.68)	182.7 (96.45)	20.6 (9.78)	0.11
North ( $n = 96$ )	497.5 (58.11)	172.3 (91.42)	19.4 (9.20)	0.19
Other direction ( $n = 410$ )	471.7 (63.71)	179.5 (101.01)	19.0 (9.12)	0.81
Overall ( $n = 503$ )	476.6 (63.44)	178.1 (99.21)	19.1 (9.13)	

Spatial correlation at different scales seems evident and a thorough analysis of this phenomenon would be helpful. For our purposes, it seemed important to consider that counts from traps located close together may be correlated. Several definitions of neighbourhoods were considered including the  $j$  closest traps,  $j$  being some small number, or traps within distance  $L$ , where  $L$  is fixed. Another formulation might use orchards to define neighbourhoods, but several orchards are quite large and it did not seem apparent that correlation existed between trap counts at far reaches of large orchards. We chose a simple definition. Traps within 400 meters in the longitude-latitude plane form a neighbourhood; eight special cases have neighbours slightly farther away. Our definition evolved through inspection of the map, a study of the resulting distribution of the number of neighbours in the corresponding neighbourhoods and an intention to keep neighbourhoods small. The distribution of the number of such neighbours has mean, median and mode around 12.

Likelihood and quasi-likelihood analyses reported here were programmed in R 1.8.1. Bayesian analyses using MCMC methodology were implemented in WinBUGS (to be found at [www.mrc-bsu.cam.ac.uk/bugs](http://www.mrc-bsu.cam.ac.uk/bugs)). For each Bayesian analysis, though not reported here, five different choices of priors were considered. Results show a fair robustness to choice of prior. For each parameter we examined convergence of the associated Markov chain to the posterior distribution using the univariate Gelman–Rubin statistic as well as trace-plots.

Section 2 considers analysis of the wild moth counts aggregated over time. Section 3 focusses on logistic analyses of the proportions of wild moths over time, specifically considering whether the ratio of trapped sterile to wild moths was at least 40:1. If the number of released sterile males to naturally occurring wild males is maintained at 40:1, scientists expect that within a few years populations of wild moths can be driven to extinction. Zero-heavy analyses are discussed in Section 4. First, a zero-inflated mixed Poisson model is considered for analysis of the time-aggregated counts. This is extended to a hidden Markov model in a spatio-temporal analysis where the allocation of traps to the zero-inflated mixture components changes dynamically over time. The report concludes with a summary of the results.

## 2. ANALYSIS OF WILD MOTH COUNTS

Let  $Y_i$  denote the number of wild moths collected in trap  $i$ ,  $i = 1, \dots, 506$  aggregated over  $m_i$  weeks of collection in 1999. We set  $m_i$  to be 14 as the aggregated counts to week 14 are available for 503 out of the 506 traps. There is a substantial amount of missing data after week 14, and simultaneously, almost no wild moth counts. For the three traps without counts for week 14, we use data available from the closest time, week 13. All three of these traps collected no wild moths at week 13. Note that there are many sites with zero aggregated counts. Thus, it may be more appropriate to use the zero-heavy analyses discussed later, which account for excess zeros not accommodated by a simple mixed Poisson model. As an exploratory tool, it seems that the

aggregated analysis, though lacking, still identifies sites with large counts and pinpoints covariates of interest.

Conditional on random trap effects, we assume independent Poisson distributions for  $Y_i$  with a loglinear relationship between the conditional mean and the covariates. Two types of random effects are assumed: one to represent heterogeneity and a second to account for spatial correlation. A simple local conditional autoregressive model is adopted here for the spatial site-specific random effects  $\{\alpha_i, i = 1, \dots, 506\}$ , so

$$\alpha_i | \alpha_{-i} \sim \mathcal{N}(\bar{\alpha}_i, \sigma_\alpha^2 / \#\delta_i),$$

where  $\alpha_{-i} = \{\alpha_j : j \neq i\}$ ,  $\bar{\alpha}_i$  denotes the mean of the random effects corresponding to traps in the neighbourhood of the  $i$ th trap, and  $\#\delta_i$  denotes the cardinality of the neighbourhood. A more general formulation would set the conditional mean of  $\alpha_i$  to be a weighted mean of the other trap effects, where the weights could depend on the distance between traps. The spatial random effect reflects both moth behaviour and environmental covariates which are similar in neighbourhoods. Hence we have

$$Y_i | b_i, \beta \sim \mathcal{P}(\mu_i), \quad \log \mu_i = \mathbf{x}_i^\top \beta + b_i,$$

where  $\mathbf{x}_i$  denotes the covariates associated with site  $i$ ,  $\beta$  the corresponding parameters,

$$b_i = \alpha_i + \gamma_i, \quad \gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2).$$

Let  $\sigma^2 = \sigma_\alpha^2 + \sigma_\gamma^2$  and  $\lambda = \sigma_\alpha^2 / \sigma^2$ . Then the vector of random effects  $\mathbf{b}$  is distributed as  $\mathcal{N}(0, \Sigma)$ , where  $\Sigma = \sigma^2 \{\lambda \mathbf{Q}^{-1} + (1 - \lambda) \mathbf{I}\}$ ;  $\mathbf{I}$  is an identity matrix, and  $\mathbf{Q}$  is the spatial connectivity matrix defined by the conditional distribution  $\alpha_i | \alpha_{-i}$ . If  $\lambda = 1$ , the distribution of the random effects  $\mathbf{b}$  reduces to the intrinsic CAR model. Note Bernardinelli, Clayton & Montomoli (1995) discuss complications regarding the comparability of  $\sigma_\alpha$  and  $\sigma_\gamma$  arising because of the marginal nature of  $\gamma$  and the conditional nature of  $\alpha$ .

Table 2–2: Estimates from the PQL analysis of aggregated wild moth counts

	Estimate (S.E.)	95% Confidence interval
Intercept	2.04 (0.83)	(0.42, 3.66)
Tree Area (sq feet)	0.0022 (0.0008)	(0.0007, 0.0037)
Elevation (meters)	-0.0053 (0.0017)	(-0.0087, -0.0019)
Trap Type: Delta	0.446 (0.233)	(-0.011, 0.904)
North	-0.383 (0.209)	(-0.794, 0.027)
$\sigma^2$	4.63 (1.04)	(2.59, 6.68)
$\lambda$	0.81 (0.072)	(0.67, 0.95)

Penalized quasi-likelihood (PQL; Breslow & Clayton 1993) was used for estimation. PQL relies on a series of approximations to the mixture model. Estimation proceeds in a computationally straightforward manner using an appropriate working vector and the restricted maximum likelihood (REML) equations under the normal theory linear model. PQL has some severe shortcomings, such as bias in the analysis of mixed binomial models with low denominators (Breslow & Clayton 1993; Sutradhar & Qu 1998). However, it seems to offer a viable method for the analysis of count data, even for low order counts (Breslow & Lin 1995). We complemented PQL with a Bayesian analysis where the inverse of each of the variance components  $\sigma_\alpha^2, \sigma_\gamma^2$  was given a weakly informative gamma distribution. These are proper, conjugate priors.

The PQL and Bayesian analyses yielded similar results in terms of variables included and point/interval estimates. However, the variance estimate was twice as large from the Bayesian analysis and the estimated coefficient for north-facing slope was twice as large in the PQL analysis. Table 2–2 gives results from the PQL analysis. The variable North is an indicator for North-facing

slope with all other values of Aspect included in the intercept; Type indicates Delta traps, with Wing Traps as the baseline included in the intercept. Most of the residual variation, represented by the random effects, is accounted for by spatial correlation. Wild moth prevalence has considerable variation and increases were associated with lower elevation and larger tree area. There was also a tendency for fewer wild moths on north-facing slopes though this effect was only marginally significant. As expected, based on documentation provided with the study, there was a tendency for more wild moths to be caught in Delta traps than in Wing Traps. Although not significant, Trap Type is included for correspondence with the zero-heavy analyses to follow. There was high agreement in all the estimates of the residual relative risks,  $\exp(b_i)$ , from both analyses. Large residual relative risks indicate traps with high wild moth counts after adjusting for covariate effects discussed above: Elevation, North aspect, Trap Type and Tree Area. Figure 2–2 identifies the traps which have the ten largest and ten smallest values of the posterior estimates of the residual relative risk. There is a small cluster of large values in the south-eastern part of the region. Note also the cluster of small risks in the south.

### 3. LOGISTIC ANALYSIS OF THE PROPORTION OF WILD MOTH COUNTS

Let  $Y_i(t)$  and  $U_i(t)$  denote the number of wild and sterile moths trapped at trap  $i$ ,  $i = 1, \dots, 506$  in week  $t$ ,  $t = 1, \dots, m_i$ , respectively. For each  $i$  and  $t$  we condition on the total number of moths caught,  $N_i(t) = Y_i(t) + U_i(t)$ , and examine the proportion of these which are wild. There are cases of missing data where information on  $Y_i(t)$  and  $N_i(t)$  are not available. Fortunately, for periods of highest moth abundance when we require high precision in estimates, such as weeks 4 and 5, 98% of the traps were visited. When  $Y_i(t)$  and  $N_i(t)$  are unavailable, some information on these values is still accessible; counts accumulate until a trap is visited and data are recorded. No counts are available for any trap in the 15th week. We applied a naive imputation technique to estimate missing values. The total obtained from the week following a missing week was divided equally to obtain estimates for each week. In view of the time trends in the abundance of the moths which manifested in the subsequent analysis, this imputation technique needs refinement. On the other hand, gains in using a more sophisticated technique are likely to be small as missing data is least prevalent for the period of greatest moth abundance.

The mixed binomial model isolates factors associated with the relative presence of wild moths, given  $N_i(t)$ . The conditional distribution of  $Y_i(t)$  is assumed to be  $\text{BIN}(p_i(t), N_i(t))$  with the logit of  $p_i(t)$  driven by covariates and site-specific random effects and modulated by a function of time in an additive model:

$$\text{logit } p_i(t) = \mathbf{x}_i^\top \boldsymbol{\beta} + g(t) + b_i,$$

where  $b_i = \alpha_i + \gamma_i$  as in the previous section. Several forms for  $g(t)$  could be considered; we used a cubic B-spline with two boundary and three inner knots (at 6, 10, and 14 weeks; see Figure 2–1).

Table 2–3: Estimates from the logistic analysis of the proportion of wild moths

	Posterior mean (SD)	95% Credible interval
Tree Area	0.0021 (0.00089)	(0.00039, 0.0039)
Elevation	-0.0072 (0.0034)	(-0.014, -0.00045)
Trap Type	1.69 (0.29)	(1.13, 2.25)
$\sigma^2$	13.75 (2.94)	(9.02, 20.35)
$\lambda$	0.95 (0.030)	(0.88, 1.00)

A Bayesian analysis used MCMC sampling based on two chains of length 50,000 run in parallel. All regression coefficients, except the intercept, were assigned  $N(0, 10^6)$  priors; the intercept was assigned a  $\mathcal{U}(-\infty, \infty)$  prior, while  $\mathcal{G}(0.01, 0.01)$  prior distributions were used for both precision

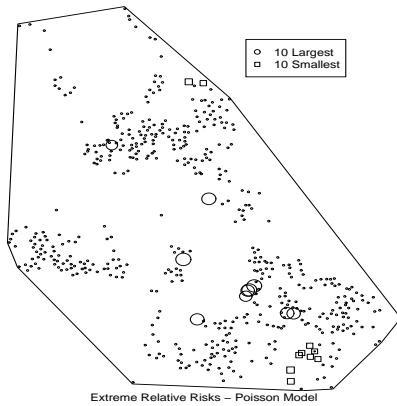


Figure 2–2: Extreme residual relative risks as identified in the analysis of aggregated counts. The figure identifies the positions of all 506 traps. The positions associated with the ten largest (circles) and ten smallest (squares) posterior estimates of the residual relative risks are also indicated.

parameters. The  $\mathcal{G}(\alpha, \beta)$  distribution has mean  $\alpha/\beta$  and variance  $\alpha/\beta^2$ . A burn-in of 15,000 for each chain seemed more than sufficient to ensure convergence and good mixing.

Tree Area, Elevation and Trap Type emerged as important covariates for explaining wild moth variation. Table 2–3 summarizes results regarding the covariates. The estimated odds of catching a wild moth is over five times greater (95% c. i. 3.1 – 9.5) with Delta traps than with Wing traps. This is in general agreement with expectations of the scientists running the study. Increases in the proportion of wild moths are also observed with greater tree size and lower elevation.

Figure 2–3a displays the time trend in the proportion of wild moths. Posterior estimates of the random effects for each trap and site-specific proportions of wild moths from the logistic analysis of  $Y_i(t)$  are displayed in Figures 2–3b and 2–3c. Circles correspond to positive random effects, and squares, to negative random effects with the size of the symbol reflecting the magnitude of the estimated random effect. Circles correspond to traps where the proportion of wild moths is greater than accounted for by the covariates while squares correspond to traps where the proportion is correspondingly less. The central and south-central regions have a higher “residual risk” of wild moths, while corresponding risks for the southern, western and northern regions are quite low. Site-specific proportions of wild moth are also of interest and posterior estimates of these are displayed in Figure 2–3c for  $t = 5$ , the time of greatest moth abundance.

The apparent interaction between trap type and moth type makes it difficult to discuss the effectiveness of the program in achieving the overflooding ratio, even if we were to condition on the pattern of abundance of the sterile moths over time and space as observed here. If such patterns reflect the schedule of release of the sterile moths, we would then be postulating on the overflooding ratio based on a similar sort of release in the future. We consider the scenario as observed, and calculate the posterior probability that the proportion of wild moths exceeds the overflooding level. Those twenty-eight traps for which this posterior probability is greater than 0.95 are identified in Figure 2–3d.

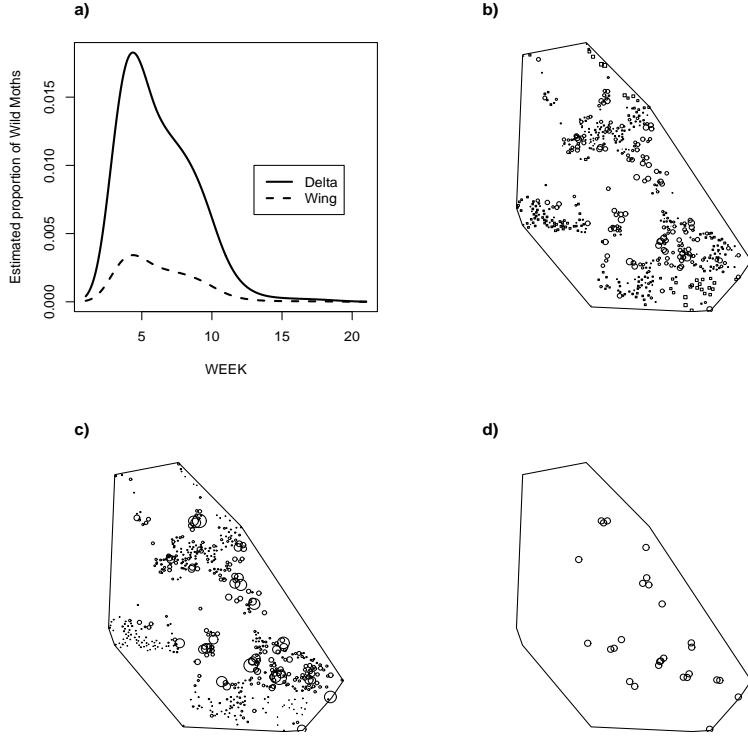


Figure 2–3: Logistic analysis of the proportion of wild moths. (a) Estimated proportion of wild moths over time (b) Estimated random effects (area of each circle or square is proportional to the estimate) (c) Week 5 wild moth proportions (d) Traps with posterior probability greater than 0.95 of exceeding a 40:1 ratio of sterile to wild.

#### 4. ZERO-HEAVY ANALYSES OF COUNTS OF WILD MOTHS

The large number of zeros in the data is certainly striking, but it is possible for such a large number of zero counts to arise under a mixed Poisson model with a low mean. In order to check this and to provide a mechanism to account for extra zeros, we consider two alternative analyses below. The first uses a zero-heavy mixture model operating on the time-aggregated counts. The second extends this to analyze counts over time where sites may dynamically move into and out of two states: one generates no wild counts, while the second generates counts according to a log-linear mixed Poisson model.

##### 4.1. Two-component model for aggregated counts.

Let  $Y_i$  denote the number of wild moths collected in trap  $i$  as in Section 2. The zero-heavy model assumes that, conditional on site-specific random effects, the counts  $Y_i$  are independent such that  $Y_i = 0$ , with probability  $p_i$ , and, with probability  $1 - p_i$ ,  $Y_i \sim \mathcal{P}(\mu_i)$ . The conditional mean  $\mu_i$  is modeled as in Section 2, with site-specific independent and spatially correlated random effects.

However, it now represents the mean for those sites whose counts are not identically zero. Here, we consider the simple case where  $p_i = p$  for all  $i$ .

Zero-heavy Poisson models have been discussed in the literature. Lambert (1992), for example, presents methods for their analyses. Such models are not uncommon in the econometrics field (see, e.g., Mullahy 1986). Dobbie & Welsh (2001) consider time-series correlated zero-heavy count data and develop generalized estimating equations for inference. Our analysis is similar to these in that we consider two components where a logistic model determines assignment to the components and a Poisson-type model is used for the “non-zero” component. However, here we incorporate spatial correlation in the conditional Poisson component. The Bayesian model used a beta prior,  $\mathcal{B}(1, 1)$  for  $p$ , weakly informative gamma hyperpriors,  $\mathcal{G}(0.5, 0.0005)$  for  $\sigma_\alpha^{-2}$  and  $\sigma_\tau^{-2}$ , and a  $\mathcal{N}(0, 10^6)$  prior for the intercept and covariate effects. Markov chain Monte Carlo sampling was run generating 50,000 samples. The first 15,000 iterations were used as burn-in.

Table 2–4: Estimates from the zero-heavy analysis of aggregated wild moth counts

	Posterior mean (SD)	95% Credible interval
Intercept	1.05 (2.22)	(−3.12, 5.88)
Tree Area (sq feet)	0.0028 (0.0009)	(0.0010, 0.0045)
Elevation (meters)	−0.0043 (0.0046)	(−0.0144, 0.0044)
Trap Type: Delta	0.290 (0.292)	(−0.259, 0.855)
North	−0.246 (0.251)	(−0.728, 0.250)
$\sigma^2$	8.25 (2.55)	(4.34, 13.83)
$\lambda$	0.88 (0.073)	(0.71, 0.98)
$p$	0.034 (0.028)	(0.0012, 0.1029)

Estimates of the parameters are provided in Table 2–4. The covariate estimates are quite similar to those from Section 2. There is about a 3.4% probability of belonging to the zero state. The posterior probability of belonging to the zero state is mapped in Figure 2–4a. The top ten largest residual relative risks in the mixed Poisson component (those 474 sites with posterior probability greater than 0.90 of not belonging to the zero component) are identified in Figure 2–4b. This analysis also identifies a cluster of high risks in the south-central region.

#### 4.2. Hidden Markov analysis.

The zero-heavy analysis of the aggregated counts can be extended to a temporal analysis of counts which permits allocation of the traps to the mixture components to change dynamically over time. The mixture model is then Markov-dependent and postulates that there are two *states* in the process: counts are either zero, with probability 1, or derived from a conditional Poisson log-linear model with spatially correlated random effects. It also postulates that movement between these two states occurs through a Markov dependence. Then, the state occupied in any week would depend strongly on the state occupied in the previous week. Markov-dependent mixture models generalize mixture distributions by introducing serial correlation in the movement over time between states.

If specific sites had no wild moths throughout the study period, while others had wild moths in varying abundances over time, then the estimate of the probability of moving out of the zero state would be close to zero. For the non-zero state, the Poisson mean gives an indication of the average wild moth abundance, given that wild moths tend to occur at that location. The analysis aims to separate out the sites and times of zero wild moth counts and allows a focus of the mean wild moth counts over time elsewhere.

The model is more easily expressed in terms of a “hidden” or “latent” variable  $Z_i(t) \in \{0, 1\}$  which expresses the state of site  $i$  in week  $t$ . Let  $Y_i(t)$  denote the number of wild moths trapped

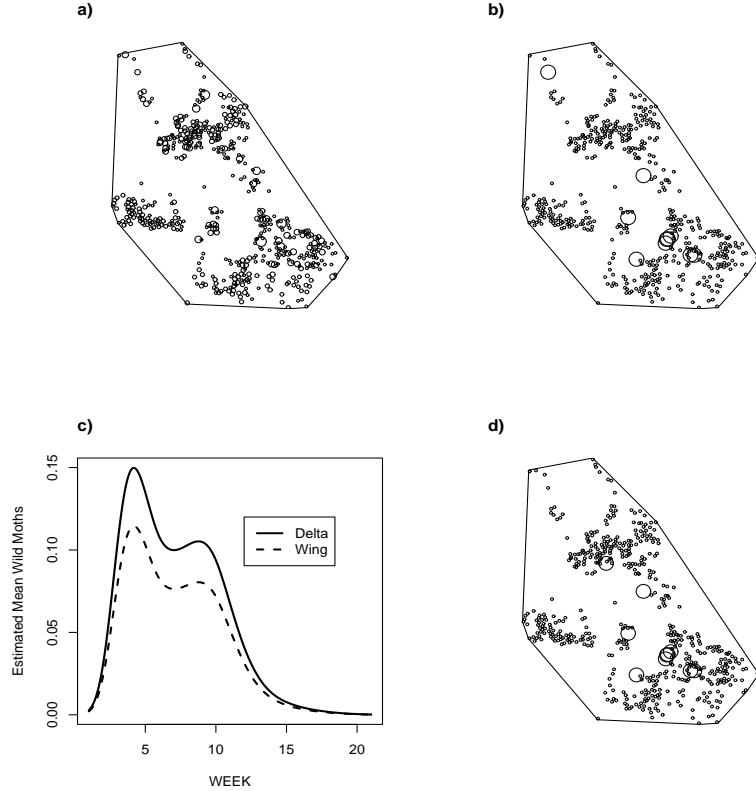


Figure 2–4: Zero-heavy analyses of counts of wild moths. (a) Posterior probability of being in the zero component - zero-heavy model. (b) Largest residual relative risks identified through the zero-heavy model analysis of the aggregated counts. (c) Estimated mean wild moth count over time from the Hidden Markov model. (d) Largest residual relative risks identified through the Hidden Markov model analysis. In (b) and (d) the positions associated with the ten largest estimates are identified.

at site  $i$  in week  $t$ ,  $i = 1, \dots, 506, t = 1, \dots, m_i$ , and, as previously let  $\mathbf{x}_i$  denote the covariates associated with site  $i$ . Conditional on latent variables  $Z_i(t)$ , and random effects  $b_i$ , the counts  $Y_i(t)$  are independent with distributions

$$[Y_i(t)|\{Z_i(t)\}, \{b_i\}] \equiv \begin{cases} \sim \mathcal{P}(\mu_i(t)), & \text{if } Z_i(t) = 1; \\ 0, \text{ with probability 1,} & \text{if } Z_i(t) = 0; \end{cases}$$

and, for each  $i = 1, \dots, 506$ , the sequence  $\{Z_i(1), \dots, Z_i(m_i)\}$  is assumed to follow a Markov chain.

We make several assumptions which permit an exploratory analysis of the data in a timely manner. In particular, we assume a two-state homogeneous Markov chain with transition probabilities  $p_{01}$  and  $p_{10}$  representing the probabilities of moving between states;  $p_{01}$  is the probability of moving out of the zero state and into the non-zero state, for example. The probability of initially being in the non-zero state is defined as  $p_{\text{init}}$ . Less restrictive assumptions would allow these probabilities to depend on time, covariates and random effects; dependence on time would specifically be quite interesting and important to explore. These stationary probabilities then provide an estimate of the overall transition probabilities.

The logarithm of the conditional mean of the non-zero state is modeled as a function of the covariates, time and random effects. We use only the spatially correlated random effect as earlier analyses identified this as the predominant variance component:

$$\log\{\mu_i(t)\} = \mathbf{x}_i^\top \beta + g(t) + \alpha_i,$$

where  $g(t)$  is of the same form as in Section 3: a cubic B-spline with inner knots at 6, 10 and 14 weeks. The model specification is completed by assigning weakly informative prior distributions to the model parameters:  $p_{\text{init}} \sim \mathcal{B}(1, 1)$ ,  $p_{01} \sim \mathcal{B}(1, 1)$ ,  $p_{10} \sim \mathcal{B}(1, 1)$ ,  $\beta_0 \sim \mathcal{U}(-\infty, \infty)$ ,  $\beta_j \sim \mathcal{N}(0, 10^3)$ , and the inverse of the variance component  $\sigma_\alpha^2$  is distributed as  $\mathcal{G}(0.01, 0.01)$ . The  $\beta$ s refer to the coefficients of the covariates and the spline basis functions ( $\beta_0$  being the intercept). Markov chain Monte Carlo sampling is based on two chains of length 100,000 run in parallel. Diagnostics indicate a burn-in length of 30,000 for each chain to be sufficient.

Table 2–5: Estimates from the Fit of the Hidden Markov Model

	Posterior mean (SD)	95% Credible interval
Tree Area	0.0028 (0.00081)	(0.0012, 0.0044)
Elevation	-0.0050 (0.0037)	(-0.012, 0.0023)
Trap Type	0.27 (0.26)	(-0.25, 0.78)
North	-0.11 (0.23)	(-0.58, 0.33)
$\sigma_\alpha^2$	16.15 (2.12)	(12.4, 20.68)
$p_{\text{init}}$	0.30 (0.16)	(0.06, 0.67)
$p_{01}$	0.34 (0.046)	(0.26, 0.44)
$p_{10}$	0.13 (0.031)	(0.07, 0.19)

Table 2–5 provides estimates of the parameters and 95% credible intervals. Although we include Elevation, Trap Type and North, it seems that when focussing specifically on non-zero counts in the manner of this analysis, these are no longer significant. Though this was not explored, based on previous results, it seems likely that elevation plays a role in transitions into the zero-state, with a greater probability of such transitions for higher elevations. The estimated probability of moving from the zero-state into the non-zero state from one week to the next is 0.34. There is a reasonably high probability, namely 0.66, of staying in the zero state from one week to another, and similarly a high probability of staying in the non-zero state, 0.87, from one week to the next. This suggests a tendency for certain sites to consistently attract wild moths, and provides support for the aggregated analysis of the previous section. For the non-zero state, the estimated mean wild moth count as a function of time for the average Tree Area and Elevation, and for a site on a non-north-facing slope are plotted in Figure 2–4c. Figure 2–4d displays the posterior mean estimate of the site-specific random effects. Though the mean counts per trap are quite low, residual relative risks can be quite high. For example, counts attain over twenty times the mean at some sites. The first and third quartiles of the residual relative risks are 0.44 and 2.13. The large residual relative risks identified in Figure 2–4d are again clustered in the south-central region.

## 5. DISCUSSION

The variables Tree Area, Trap Type, Elevation and Aspect explain variation in the counts of wild moths over the sites. Larger counts are associated with larger values of Tree Area, lower Elevation and Delta traps; sites facing north tend to have fewer counts of wild moths. There are many areas where no or very few moths are observed. Indeed, this is a dominant element of the data. However, we consistently observe relatively high wild moth abundance in specific areas. Apart from Aspect, the same variables are associated with the relative proportions of wild moths out of

the total moths caught. Also, there are several regions where the ratio of trapped sterile to wild moths was greater than 40:1, with high probability. All analyses identified the same set of sites as having the largest nine residual relative risk values. In fact, this agreement is not surprising as several of the corresponding counts are relatively large. The two zero-heavy analyses in Section 4 identified the same sites as having the largest five residual relative risks, these being somewhat different than the top five based on the raw data.

Several directions for future work have been discussed. Note that pest control and management practices may vary among orchards. Thus, it may be appropriate to include such orchard-specific effects in the model in order to provide better identification of orchards with a high risk of wild moth abundance. It would also be quite interesting to contrast the results of similar analyses using data for the years 2000 and 2001. Finally, simpler models may handle this data well; for example, the negative binomial distribution might be an interesting alternative to consider. Statistical aspects of extensions to the zero-heavy analyses to allow more general structures need development. In particular, it would be helpful to allow the probabilities associated with the states to include spatially correlated random effects and to permit correlations between random effects in the two components of the models. In this way we may permit sites with low probability of belonging to the zero state to have high probability of a large count, for example. More general correlation structures for random effects might also be considered, for example temporally correlated or spatio-temporally correlated. Space-varying coefficient models could be investigated where we allow covariate effects to vary over space, a special case being random spline models where the coefficients of the spline basis are spatially varying from trap to trap. This could help to identify whether peak periods of moth abundance vary spatially, for example.

#### ACKNOWLEDGEMENTS

P. Gill and C. B. Dean were supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank the referees and Case Studies Editor for valuable comments which led to a significant improvement of the article.

---

Farouk NATHOO: [nathoo@stat.sfu.ca](mailto:nathoo@stat.sfu.ca)  
Laurie AINSWORTH: [lmainsw@sfu.ca](mailto:lmainsw@sfu.ca)  
*Department of Statistics and Actuarial Science  
Simon Fraser University, Burnaby  
British Columbia, Canada V5A 1S6*

Paramjit GILL: [pgill@ouc.bc.ca](mailto:pgill@ouc.bc.ca)  
*Department of Mathematics and Statistics  
Okanagan University College, Kelowna  
British Columbia, Canada V1V 1V7*

Charmaine B. DEAN: [dean@stat.sfu.ca](mailto:dean@stat.sfu.ca)  
*Department of Statistics and Actuarial Science  
Simon Fraser University, Burnaby  
British Columbia, Canada V5A 1S6*

## Analyse/Analysis 2:

# Étude de l'influence de la topographie et des caractéristiques des vergers sur l'efficacité du programme d'épandage d'insectes stériles pour le carpocapse de la pomme (*Laspeyresia pomonella*)

Gaétan DAIGLE, Thierry DUCHESNE, Emmanuelle RENY-NOLIN et  
Louis-Paul RIVEST

### 1. TRAITEMENT PRÉLIMINAIRE DES DONNÉES

Les caractéristiques des trappes dans les vergers sont les variables explicatives les plus importantes pour modéliser l'abondance des carpocapses et l'étendue des dommages. Cette section décrit les transformations appliquées à ces variables pour faciliter leur interprétation.

#### 1.1 La densité du verger.

Pour caractériser la densité des arbres dans le verger, nous avons utilisé deux variables associées à la disposition des arbres dans le verger, Within Row Spacing, WRSP, Between Row Spacing, BRSP, et deux variables associées à la taille des arbres eux-mêmes, Tree height, TRHE et Basal Diameter, BADIA. Pour les combiner en une seule, nous avons effectué une analyse en composantes principales de la matrice de variances-covariances sur l'échelle logarithmique. Les corrélations étaient toutes positives, ce qui indique la présence d'un effet taille. En effet, les grands arbres ont un diamètre élevé et sont plantés dans des vergers où les arbres sont relativement espacés. La première composante principale explique 86% de la variabilité. L'exponentielle de cette première composante,  $TR\_AR = WRSP^{0.59} BRSP^{0.15} TRHE^{0.17} BADIA^{0.60}$  a été utilisée dans les analyses qui suivent. Une grande valeur de  $TR\_AR$  correspond à une faible densité d'arbres autour d'une trappe.

#### 1.2 La latitude et la longitude.

Les coordonnées  $X$  et  $Y$  donnent la longitude et la latitude des trappes. Dans ce système d'axes, la rive du lac Okanagan peut être représentée par une droite de pente  $-1.37$ . La direction de cette droite est un vecteur unitaire de  $\text{arctan}(-1.37) = -54$  degrés. Une rotation de  $-36$  degrés est requise pour faire coïncider la rive du lac avec l'axe  $y$ . Le changement d'axes s'effectue sur les données centrées,  $X_c$  et  $Y_c$ . Les nouvelles coordonnées  $X_N$  et  $Y_N$  sont alors données par

$$\begin{pmatrix} X_N \\ Y_N \end{pmatrix} = \frac{1}{1000} \begin{pmatrix} \cos(-36) & -\sin(-36) \\ \sin(-36) & \cos(-36) \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \end{pmatrix}.$$

La division par 1000 fait en sorte que  $X_N$  et  $Y_N$  sont en kilomètres. Ils varient dans l'intervalle  $(-4.4, 6)$ . La variable  $X_N$  est associée à la distance au lac;  $X_N > 2.5$  donne un point dans le lac. La variable  $Y_N$  détermine le déplacement le long de la rive du lac. Les nouvelles coordonnées sont à peu près non corrélées, avec  $r(X_N, Y_N) = 0.03$ , où  $r$  représente la corrélation de Pearson. Avec les données originales  $r(X, Y) = -0.47$ . Il faut noter que la rive du lac Okanagan est pentue et que l'altitude GPSH est corrélée avec  $X_N$  et  $Y_N$ ,  $r(X_N, GPS\_HGT) = -0.75$  et  $r(Y_N, GPS\_HGT) = 0.41$ . L'interprétation de cette première corrélation est claire : lorsque  $X_N$  augmente, on s'approche de la rive du lac Okanagan et l'altitude diminue.

### 1.3. La variable aspect.

ASPECT donne la direction, en degrés, de la pente du terrain où se trouve une trappe. Pour une pente orientée franc nord ASPECT = 0 degrés, ou 360 degrés, alors que pour une orientation franc est ASPECT = 90 degrés. La raideur de la pente est évaluée par la variable SLOPE, qui donne essentiellement la tangente de l'angle de l'inclinaison de la pente par rapport à l'horizontale. Ainsi si SLOPE = 0, ASPECT n'est pas défini. L'importance de la variable ASPECT est donc liée à SLOPE. Ainsi dans les analyses les variables SL\_SIN\_ASP = SLOPE  $\times$  sin(ASPECT) et SL\_COS\_ASP = SLOPE  $\times$  cos(ASPECT) sont utilisées. Ces variables prennent la valeur 0 si SLOPE = 0.

Nous avons jusqu'ici traité de 7 variables explicatives, TR\_AR,  $X_N$ ,  $Y_N$ , GPS\_HGT, SLOPE, SL\_SIN\_ASP, SL\_COS\_ASP. Les autres variables décrivant les trappes sont STYLE, une variable dichotomique, GRD\_CULTIV et IRRIGATION, deux variables qualitatives décrivant le mode de culture du verger, et trois variables associées à la distance de l'agglomération urbaine la plus proche, URB200, URB100 et DISTANCE. Dans les analyses, des variables centrées sont utilisées et  $Z$  dénote le vecteur de ces variables pour une trappe.

## 2. MODÉLISATION TEMPORELLE ET SPATIALE DE L'ABONDANCE DES CARPOCAPSES

Pour décrire l'arrivée des carpocapses dans les trappes, on utilise un processus de Poisson non stationnaire, avec le temps  $t$  dans l'intervalle (0, 1). Le temps  $t = 0$  correspond à une semaine avant le début de la période de collecte et  $t = 1$  est le temps de la dernière visite. Le tableau 3-1 donne les dates correspondant à ces deux événements pour les trois années de l'étude. Si  $Z$  dénote le vecteur des variables géographiques et physiques mesurées pour une trappe, le modèle utilisé postule que l'intensité du processus de Poisson est donnée par

$$\lambda(t|Z, \alpha, \psi) = \lambda_0(t) \exp(Z^\top \beta + \alpha + \psi), \quad t \in (0, 1), \quad (1)$$

où  $\beta$  un vecteur de paramètres inconnus associés aux variables explicatives,  $\alpha$  représente un effet de trappe résiduel et  $\psi$  est associé au verger.

La notation suivante est utilisée :

- l'indice  $i$ ,  $i = 1, \dots, I$  représente le verger, en 2001 le nombre total de vergers est de  $I = 236$ ;  $j = 1, \dots, N_i$  est associé aux trappes. Le nombre de trappes  $N_i$  par verger varie entre 1 et 14. Le total des trappes utilisées est  $\sum N_i$ . Le vecteur des variables explicatives centrées de la trappe  $j$  du verger  $i$ , ou plus simplement de la trappe  $(i, j)$ , est appelé  $Z_{ij}$ .
- $k = 1, \dots, N_{ij}$  représente les visites effectuées à la trappe  $(i, j)$ . Nous appelons  $t_{ijk} \in (0, 1)$  le temps de la  $k$ ième visite à la trappe  $(i, j)$ ; de plus par définition  $t_{ij0} = 0$  ce qui correspond à la date d'apparition des carpocapses. En 2001, le 22 avril est le temps 0.
- $m_{ijk} = (t_{ijk} + t_{ij(k-1)})/2$  est le point milieu de l'intervalle de temps entre la  $(k - 1)$ ième et la  $k$ ième visite à la trappe  $(i, j)$ .
- $n_{ijk}$  représente, pour  $i = 1, \dots, I$ ,  $j = 1, \dots, N_i$ , et  $k = 1, \dots, N_{ij}$  le décompte des carpocapses fait au temps  $t_{ijk}$ ; ce nombre représente l'effectif des insectes capturés entre les temps  $t_{ij(k-1)}$  et  $t_{ijk}$ .
- $r_{ijk}$  est une variable indicatrice prenant la valeur 1 si, lors de la  $k$ ième visite à la trappe  $(i, j)$ , la règle d'au moins 40 stériles pour 1 sauvage n'est pas respectée.

Nous sommes ainsi en présence d'un processus de dénombrement dans chaque trappe. Le processus augmente de un lorsqu'un carpocapse est capturé. Les temps exacts d'arrivée des insectes sont inconnus. En fait ces temps sont censurés par intervalle, car à chaque insecte capturé est

Tableau 3–1 : Quelques statistiques descriptives concernant la capture des insectes aux trois années de l'étude. On y retrouve le nombre total de trappes ( $\sum N_i$ ), le nombre total d'insectes sauvages et stériles capturés, les dates correspondant à  $t = 0$  et à  $t = 1$ , le nombre maximal de jours durant lesquels une trappe a été en opération, et la proportion moyenne, par rapport au nombre maximal, des jours d'utilisation d'une trappe (% visite). Les dates sont données dans le format jj/mm.

Année	$\sum N_i$	Sauvages	Stériles	$t = 0$	$t = 1$	nb jours	% visite
1999	506	1069	282144	01/05	25/09	148	90%
2000	500	290	151740	24/04	18/09	148	77%
2001	454	263	113401	22/04	02/09	134	65%

associé un intervalle de temps  $(t_{ij(k-1)}, t_{ijk})$ . L'arrivée de carpocapses sauvages étant un événement relativement rare, plusieurs techniques d'analyse peuvent être envisagées.

### 2.1. Les modèles à risques proportionnels.

Les modèles présentés dans le chapitre 8 de Therneau & Grambsch (2000) auraient pu être utilisés en imputant les temps exacts d'arrivée par les points milieux  $m_{ijk}$  des intervalles  $(t_{ij(k-1)}, t_{ijk})$ . Nous nous sommes intéressés à modéliser le temps de la première capture d'un carpocapse sauvage. La variable dépendante pour une trappe est donc donnée par l'intervalle  $(t_{ij(k-1)}, t_{ijk})$  de première capture. Cette variable est censurée à droite par le temps de la dernière visite lorsqu'aucun insecte sauvage n'est capturé. La forme (1) du modèle correspond alors au modèle avec risques proportionnels de Cox. Plusieurs approches ont été présentées dans la littérature pour estimer les paramètres de ce modèle à partir de données censurées par intervalle ; voir par exemple Alioum & Commenges (1996). Pour simplifier les analyses, nous avons cependant imputé les temps d'arrivée par les milieux  $m_{ijk}$  des intervalles de première observation. Les résultats pour le modèle de Cox présentés au tableau 3–2 ont été obtenus à l'aide de la procédure PHREG de SAS.

### 2.2. Les modèles de régression paramétrique pour des durées de vie.

Les modèles de régression paramétrique sont des alternatives au modèle de Cox qui permettent d'incorporer explicitement la censure par intervalle. En utilisant, dans (1), l'expression paramétrique  $\lambda_0(t) = ct^\alpha$ , on peut utiliser la régression Weibull, avec données censurées par intervalle, pour estimer les paramètres. Pour transformer les paramètres estimés  $\hat{\gamma}$  de la régression Weibull en paramètres log-linéaires, on utilise la transformation  $\hat{\beta} = -\hat{\alpha}\hat{\gamma}$  (voir Klein & Moeschberger 1997, p. 378). L'erreur-type de  $\hat{\beta}$  est ensuite calculée par linéarisation. La procédure LIFEREG de SAS a été utilisée pour ajuster une loi Weibull.

### 2.3. La régression de Poisson.

Cette méthode s'applique tant aux insectes stériles qu'aux insectes sauvages. La plupart des analyses sont basées sur cette technique. Pour modéliser l'effet temporel, une série de Fourier d'ordre  $m$  a été utilisée,

$$\log \lambda_0(t) = a_0 + \sum_{j=1}^m \{a_j \sin(2\pi jt) + b_j \cos(2\pi jt)\}. \quad (2)$$

Des séries jusqu'à l'ordre 5 ont été ajustées. Toutes les intensités estimées présentées dans ce travail sont basées sur des séries d'ordre 4 qui offrent un bon compromis biais-variance pour la modélisation des carpocapses tant sauvages que stériles.

En vertu du modèle postulé, la distribution du nombre de carpocapses observés dans la trappe  $(i, j)$  au temps  $t_{ijk}$ ,  $n_{ijk}$ , est une Poisson de paramètre

$$\int_{t_{ij(k-1)}}^{t_{ijk}} \lambda(t|Z_{ij}, \alpha_{ij}, \psi_i) dt = \exp(Z_{ij}^\top \beta + \alpha_{ij} + \psi_i) \int_{t_{ij(k-1)}}^{t_{ijk}} \lambda_0(t) dt.$$

L'intégrale de  $\lambda_0(t)$  est une expression complexe qui dépend des paramètres  $a$  et  $b$  de la série de Fourier. Lors de l'ajustement des modèles statistiques, nous approximons cette intégrale par la méthode du rectangle,

$$\int_{t_{ij(k-1)}}^{t_{ijk}} \lambda_0(t) dt \approx (t_{ijk} - t_{ij(k-1)}) \lambda_0(m_{ijk}). \quad (3)$$

Étant donné que la majorité des visites sont hebdomadaires ou bimensuelles, les différences  $t_{ijk} - t_{ij(k-1)}$  sont en grande majorité inférieures à 0.1 et ces approximations devraient être précises. Une étude par simulation présentée à la section 3.4 confirme cette hypothèse.

Les paramètres d'intérêt sont les coefficients  $a$  et  $b$  des séries de Fourier et le vecteur  $\beta$ . La caractérisation des composantes intra-trappe et intra-verger des erreurs n'est pas un objectif de l'étude. Ainsi, lors des analyses nous avons utilisé la régression de Poisson standard, sans effet aléatoire pour les trappes et les vergers. La régression de Poisson s'exprime sous la forme du modèle log-linéaire suivant,

$$\begin{aligned} \log E(n_{ijk}) = \\ \log(t_{ijk} - t_{ij(k-1)}) + a_0 + \sum_{\ell=1}^m \{a_\ell \sin(2\pi\ell m_{ijk}) + b_\ell \cos(2\pi\ell m_{ijk})\} + Z_{ij}^\top \beta, \end{aligned} \quad (4)$$

où  $\log(t_{ijk} - t_{ij(k-1)})$  est une variable offset. Les variances ont été estimées à l'aide d'estimateurs sandwich qui mesurent la variabilité inter-verger. L'estimateur de variances-covariances,  $\hat{\Sigma}$ , de  $(\hat{a}, \hat{b}, \hat{\beta})$  s'écrit en fonction de  $Z_{ijk}^*$ , le vecteur de toutes les variables explicatives, y compris celles associées à la modélisation du temps, et de  $\hat{\mu}_{ijk}$ , la valeur prédictive pour  $n_{ijk}$  de la façon suivante

$$\begin{aligned} \hat{\Sigma} = & \left\{ \sum_{ijk} Z_{ijk}^* Z_{ijk}^{*\top} \hat{\mu}_{ijk} \right\}^{-1} \\ & \left[ \sum_i \left\{ \sum_{jk} Z_{ijk}^* (n_{ijk} - \hat{\mu}_{ijk}) \right\} \left\{ \sum_{jk} Z_{ijk}^* (n_{ijk} - \hat{\mu}_{ijk}) \right\}^\top \right] \left\{ \sum_{ijk} Z_{ijk}^* Z_{ijk}^{*\top} \hat{\mu}_{ijk} \right\}^{-1} \end{aligned} \quad (5)$$

(voir Cameron & Trivedi 1998, p. 320). Pour étudier la présence d'une interaction entre le temps et les variables associées aux trappes, des interactions  $\sin(2\pi m_{ijk}) \times Z_{ij}$  et  $\cos(2\pi m_{ijk}) \times Z_{ij}$  ont été incluses dans le modèle. Puisque  $\beta_c Z_{ij} \cos(2\pi m_{ijk}) + \beta_s Z_{ij} \sin(2\pi m_{ijk})$  peut s'écrire  $r_{max} Z_{ij} \cos\{2\pi(m_{ijk} - t_{max})\}$ , où  $r_{max}$  est la valeur maximale de l'interaction et  $t_{max}$ , le temps où cette interaction est maximale, les interactions temporelles sont rapportées comme des couples  $(r_{max}, t_{max})$  au Tableau 3-3. La majorité des calculs ont été faits à l'aide de la procédure GENMOD de SAS.

### 3. DÉTERMINATION DES FACTEURS ASSOCIÉS À L'ABONDANCE DES CARPOCAPSES

L'année cible pour les analyses est 2001. La sélection de modèle a été faite à partir des données de 2001 avec la régression de Poisson pour l'abondance des carpocapses, sauvages et stériles, et avec la régression logistique pour la conformité à la règle du 40 pour 1. Toutes les variables explicatives

Tableau 3–2 : Estimations des paramètres du modèle final pour la capture des carpocapses sauvages et de leurs erreurs-types (e.t.).

Année	1999		2000			2001		
	Modèle	Poisson	Variable	Poisson	Weibull	Cox		
		$\hat{\beta}$ (e.t.)		$\hat{\beta}$ (e.t.)	$\hat{\beta}$ (e.t.)	$\hat{\beta}$ (e.t.)		
$X_N$	0.17 (0.10)		0.17 (0.15)	0.58 (0.20)	0.37 (0.14)	0.41 (0.14)		
$X_N^2$	-0.28 (0.09)		-0.18 (0.07)	-0.21 (0.08)	-0.24 (0.10)	-0.23 (0.10)		
TR_AR	0.36 (0.11)		0.66 (0.17)	0.60 (0.24)	0.17 (0.16)	0.17 (0.16)		

disponibles ont été considérées dans la procédure de choix des modèles. Le modèle sélectionné, pour chaque type d'analyse, a ensuite été ajusté aux données de 1999 et 2000. De plus, les paramètres de 2001 pour l'abondance des carpocapses sauvages ont été réestimés à partir des temps de première capture en utilisant la régression de Weibull et le modèle de Cox.

### 3.1. *Carpocapses sauvages.*

L'analyse des données de 2001 a débuté par une sélection pas à pas parmi les 13 variables explicatives disponibles. Puisque PROC GENMOD de la version 8 de SAS n'a pas de procédure automatique pour faire une telle sélection, nous avons utilisé PROC LOGISTIC en prenant comme variable dépendante une variable dichotomique indiquant la présence d'au moins un insecte sauvage. En fixant respectivement à .30 et .10 les seuils d'entrée et de sortie de la procédure de sélection, il est alors possible d'identifier toutes les variables potentiellement associées à la présence des insectes.

La régression de Poisson, avec estimateur de variance Sandwich (5), a ensuite été utilisée pour faire la sélection définitive des variables explicatives au seuil de 5%. L'ajustement du modèle a été vérifié en incluant des termes quadratiques, des interactions entre variables explicatives et avec le temps. Cette étape a permis l'ajout de  $X_N^2$ . Finalement des trappes extrêmes, ayant de grands résidus pour la régression de Poisson, ont été identifiées et retirées de l'analyse. Ceci a permis de vérifier la stabilité des estimations présentées au tableau 3–2.

L'interprétation des paramètres pour 2001 révèle que pour des trappes près du lac (avec  $X_N \approx 1.5$ ) l'intensité du processus de Poisson augmente de 50% par rapport aux trappes avec  $X_N = 0$ . Elle diminue de plus de 50% loin du lac (avec  $X_N < -1$ ). De plus une augmentation d'un écart-type ( $\sigma = 0.77$ ) de la variable TR\_AR augmente  $\lambda(t)$  de 58 %. Les insectes sauvages sont donc plus abondants dans les vergers moins denses.

La régression de Weibull et le modèle de Cox ont été utilisés, tel que décrit à la section 2, pour valider les estimations faites en 2001. La différence la plus importante est le paramètre pour TR\_AR qui est beaucoup plus faible pour les analyses de temps de première apparition des carpocapses sauvages. Une analyse, non présentée ici, sur les données de 1999 explique ce phénomène. En 1999, une interaction significative  $TR\_AR \times t$ , montre que l'importance de la variable TR\_AR dépend du temps. Pour  $t$  voisin de 0 son paramètre est essentiellement nul ; il atteint son maximum, 0.65, à  $t = 0.4$ . Une faible interaction  $TR\_AR \times t$  est donc une explication possible pour les écarts entre les analyses basées sur les données complètes et celles qui utilisent seulement les temps de première apparition.

Les paramètres estimés pour 1999 et 2000, présentés au tableau 3–2, sont relativement semblables à ceux de 2001. L'estimation pour  $X_N$  est cependant plus petite en 1999 et en 2000 qu'en 2001. Ainsi, en 2001, la proximité au lac est associée à une augmentation des carpocapses sauvages pour la première fois.

La figure 3–1 compare les estimations de  $\lambda_0(t)$ , l'intensité temporelle de l'abondance des carpocapses sauvages, pour les trois années de l'étude. Même si les périodes d'observation pour les trois années ne sont pas rigoureusement identiques, il est intéressant de mettre les courbes sur le

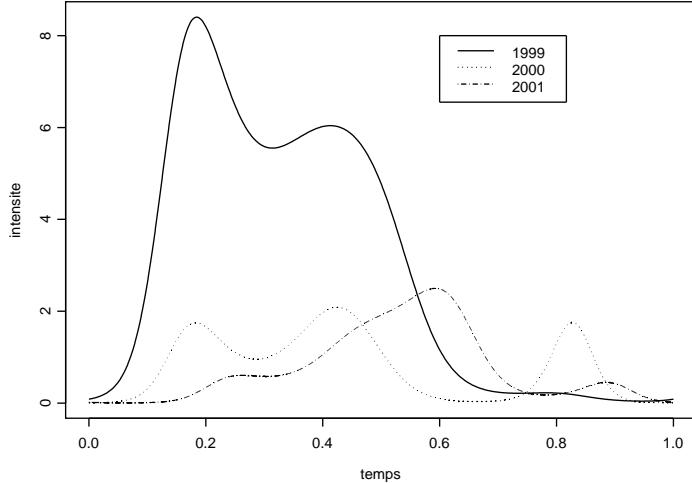


Figure 3–1: Estimation de l’intensité  $\lambda_0(t)$  d’arrivée des carpocapses sauvages pour trois années.

même graphique. Les carpocapses sauvages étaient plus abondants en 1999. En 1999 et 2000, on voit deux maxima locaux à  $t = 0.2$  et  $t = 0.42$  qui correspondent au 19 mai et au 17 juin environ. En 2001, l’arrivée des carpocapses sauvages est plus tardive avec un maximum à  $t = 0.61$ , ce qui correspond au 12 juillet.

La figure 3–2 compare l’estimation de l’intensité  $\lambda_0(t)$  obtenu avec les données complètes et celle qui utilise seulement les temps de première apparition. Cette dernière utilise la méthode de Ramlau-Hansen (voir Klein & Moschberger, 1997, section 6.2) avec le noyau de Epanechnikov et une fenêtre de  $b = 0.3$  pour lisser l’estimation du risque. Les deux estimations ont la même forme. Cependant, le maximum aux environs de  $t = 0.6$  est beaucoup plus accentué pour la régression de Poisson que pour le modèle de Cox. La présence en (1) d’effets aléatoires  $\alpha$  et  $\psi$  pour la variabilité inter-trappes et inter-vergers est sans doute responsable de ces différences. La fonction de risque estimée par la régression de Poisson est en effet égale à  $\lambda_0(t) \exp(Z^\top \beta) E\{\exp(\alpha + \psi)\}$ , alors que le risque cumulé pour le modèle de Cox s’écrit

$$-\log E \left[ \exp \left\{ - \int_0^t \lambda_0(s) ds \exp(Z^\top \beta + \alpha + \psi) \right\} \right] \leq \int_0^t \lambda_0(s) ds \exp(Z^\top \beta) E\{\exp(\alpha + \psi)\}.$$

L’inégalité précédente est une conséquence de celle de Jensen: si  $X$  est une variable aléatoire et  $a$  est une constante,  $E(e^{-aX}) \geq e^{-aE(X)}$ . Il y a donc une explication théorique aux écarts observés à la Figure 3–2: en présence d’effets aléatoires inter-verger et inter-trappe, l’ajustement d’un modèle avec risques proportionnels sous-estime l’intensité  $\lambda_0(t)$ .

### 3.2. Carpocapses stériles.

Une régression de Poisson telle que présentée à la section 2 a été utilisée pour les analyses. Vu l’abondance de données, seules les composantes du modèle qui sont significatives au seuil de 1% ont été retenues. Le STYLE de trappe est le facteur le plus important ; l’emploi des trappes WING, qui représentent environ 90% des trappes disponibles, multiplie l’intensité de capture par 4. En fait la performance des trappes DELTA dépend fortement de l’altitude GPS\_HGT. Une faible altitude

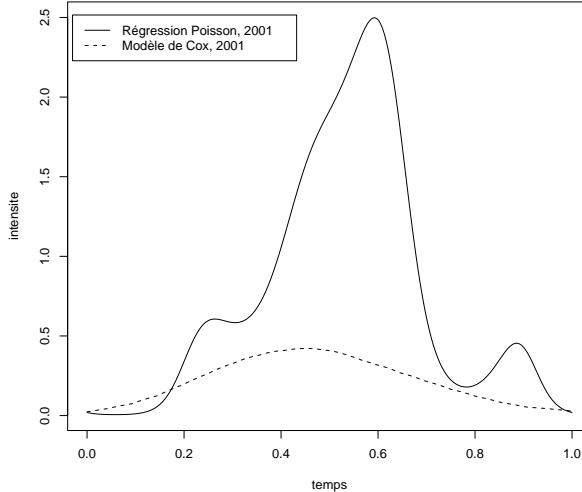


Figure 3–2: Estimation de l'intensité  $\lambda_0(t)$  obtenue à l'aide de la régression Poisson et du modèle de Cox.

est associée à des probabilités de capture plus importantes. La deuxième variable en importance pour 2001 est  $X_N$ . La relation entre la distance au lac est, en moyenne, l'inverse de celle pour les sauvages. Il y a plus de stériles loin du lac (lorsque  $X_N < 0$ ) ; à  $X_N = -1.5$ , l'intensité augmente de 50%. Les deux interactions significatives  $X_N \times t$  et  $X_N \times \text{GPS\_HGT}$  soulignent que cette relation varie dans le temps et avec l'altitude. Cette association négative entre  $X_N$  et l'intensité des stériles est unique à 2001. Ainsi en 2000, la proximité de la rive du lac est liée à une augmentation des stériles.

La figure 3–3 présente les intensités des processus de Poisson pour les stériles de 1999 à 2001. Il y a avait plus d'insectes stériles en 1999 ; les courbes pour 2000 et 2001 sont très semblables.

### 3.3. Respect de la règle du 40 stériles pour 1 sauvage.

La variable dépendante pour cette analyse est  $r_{ijk}$  qui vaut 1 si la règle n'est pas respectée lors de la visite de la trappe  $i$  au temps  $t_{ijk}$  et 0 sinon, pour  $i = 1, \dots, I$ ,  $j = 1, \dots, N_i$  et  $k = 1, \dots, N_{ij}$ . La probabilité de succès de cette variable de Bernoulli dépend à la fois du temps d'observation et des caractéristiques des trappes. Cela suggère de considérer le modèle suivant

$$\text{logitE}(r_{ijk}) = \lambda_0(t_{ijk}) + Z_{ij}^\top \beta,$$

où  $\lambda_0(t)$  représente une fonction lisse du temps écrite à l'aide d'une série de Fourier comme dans les sections précédentes. Cette formulation du modèle n'inclut pas un effet de trappe résiduel. Pour éviter qu'un tel effet ne biaise les estimations d'erreurs-types à la baisse, ces dernières sont estimées à l'aide d'un estimateur de variance de type sandwich associé à la variabilité inter-verger. Les variables explicatives importantes sont les mêmes que celles pour l'abondance des stériles présentée au tableau 3–2 et les estimations des paramètres des tableaux 3–2 et 3–4 sont très voisines. Finalement soulignons que l'approche présentée dans Maul (1999) aurait pu être envisagée pour réaliser cette partie de l'analyse.

Pour étudier les fluctuations temporelles de la probabilité de violer la règle du 40 pour 1, on utilise la probabilité de non-conformité évaluée pour des variables explicatives nulles,  $\exp \hat{\lambda}_0(t) / \{1 +$

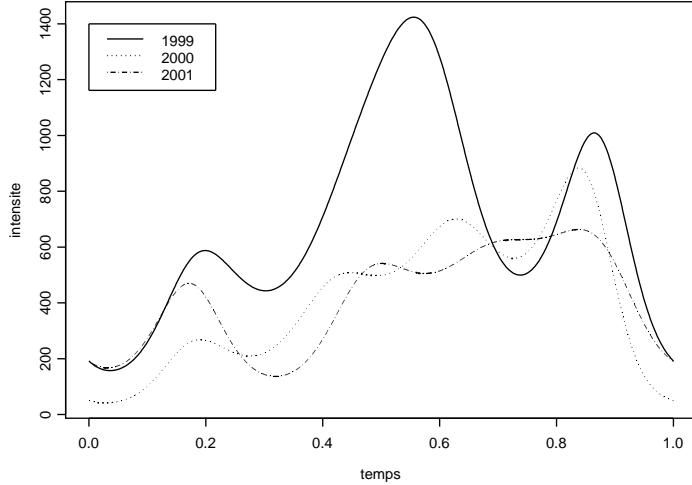


Figure 3–3: Intensité estimée  $\hat{\lambda}_0(t)$  pour les stériles.

$\exp \hat{\lambda}_0(t)\}$ . La figure 3–4 présente le graphique de ces probabilités estimées pour les trois années. L'allure générale des courbes est similaire à celles de la figure 3–1. L'année 2001 se caractérise par un maximum relativement tardif à  $t = 0.62$ , ce qui correspond au milieu de juillet. Les probabilités de la figure 3–3 étant relativement faibles, les paramètres associés aux variables explicatives peuvent s'interpréter sur le mode multiplicatif comme à la section 3.1. Augmenter d'un écart-type la variable TR\_AR ou se rapprocher du lac en passant de  $X_N = 0$  à  $X_N = 1.5$  augmente la probabilité de violer la règle du 40 pour 1 de 50%.

#### 3.4. Précision de l'approximation (3).

Pour évaluer la précision de (3), la méthode du rectangle, pour calculer des intégrales de l'intensité nous avons effectué une petite étude par simulation à partir des paramètres estimés pour l'abondance des carpocapses sauvages en 2001. Les vraies valeurs des paramètres  $a$  et  $b$  de (1) et de  $\beta$  sont fixées aux estimations de 2001. Pour chaque  $n_{ijk}$ , l'intégrale

$$\int_{t_{ij(k-1)}}^{t_{ijk}} \lambda_0(s) ds$$

est évaluée à l'aide d'un méthode d'intégration numérique. Ceci permet de calculer le paramètre  $\lambda_{ijk}$  de la loi de Poisson pour  $n_{ijk}$ . Un échantillon Monte-Carlo est ensuite simulé en générant les  $n_{ijk}$  selon des lois de Poisson de paramètres  $\lambda_{ijk}$  et la méthode d'estimation décrite à la section 2, qui utilise (3), permet de calculer des estimations Monte-Carlo des paramètres. Cette procédure est répétée 500 fois et la moyenne des 500 estimations calculées donne l'espérance de l'estimateur des paramètres. Le biais de  $(\hat{a}, \hat{b}, \hat{\beta})$  est donné par la différence entre les vecteurs des espérances et celui des vraies valeurs. Pour  $\hat{\beta}$  le biais le plus important est celui pour  $X_N^2$  qui est inférieur à 5%. La figure 3–5 montre de plus que l'intensité estimée calculée à l'aide des moyennes de estimations Monte Carlo de  $a$  et  $b$  est très voisine de la vraie intensité. Ainsi (3) est assez précise et son utilisation ne compromet en rien la qualité des estimateurs de la régression de Poisson donnée par (4).

Tableau 3–3 : Estimations des paramètres du modèle final pour la capture des carpocapses stériles et de leurs erreurs-types (e.t.). Les interactions avec le temps sont rapportées comme une paire  $(r_{\max}, t_{\max})$  où  $r_{\max}$  donne la valeur maximale de l’interaction et  $t_{\max}$  donne le temps où cette interaction est maximale. L’interaction est minimale au temps  $t_{\max} + 1/2$ ; elle est alors donnée par  $-r_{\max}$ . Un paramètre suivi d’un astérisque signifie que ce dernier est significativement non nul au seuil de 1%.

Année	1999	2000	2001
Var.	$\hat{\beta}$	$\hat{\beta}$	$\hat{\beta}$
$X_N$	-0.03	0.09*	-0.13*
$X_N^2$	0.01	0.08*	0.12*
GPS_HGT	-0.0004	0.002*	-0.002*
STYLE DELTA	-1.32*	-1.42*	-1.40*
STYLE WING	0 (0)	0 (0)	0 (0)
$X_N \times \text{GPS\_HGT}$	0.0008*	0.001*	0.0013*
$\text{GPS\_HGT} \times \text{STYLE DELTA}$	-0.002	-0.004*	-0.008*
Inter.	$(r_{\max}, t_{\max})$	$(r_{\max}, t_{\max})$	$(r_{\max}, t_{\max})$
$X_N \times t$	(-0.03, 0.13)	(0.05, 0.42)	(0.12, 0.33)*
$\text{GPS\_HGT} \times t$	(0.002, 0.15)*	(0.003, 0.48)*	(0.002, 0.08)*

Tableau 3–4 : Estimations des paramètres du modèle pour le respect de la règle du 40 pour 1 et de leurs erreurs-types (e.t.).

Année	1999	2000	2001
Var.	$\hat{\beta}$ (e.t.)	$\hat{\beta}$ (e.t.)	$\hat{\beta}$ (e.t.)
$X_N$	0.22 (0.07)	0.19 (0.12)	0.47 (0.14)
$X_N^2$	-0.22 (0.05)	-0.19 (0.06)	-0.24 (0.07)
TR_AR	0.34 (0.10)	0.57 (0.15)	0.54 (0.17)

#### 4. ASSOCIATION ENTRE LES MESURES DE DOMMAGES DES POMMES ET LES CARACTÉRISTIQUES DU VERGER

En 1999, 2000, et 2001, des dommages ont été constatés dans respectivement 10, 5 et 28 vergers. Pour modéliser l’ampleur des dommages en fonction des caractéristiques des vergers nous utilisons un modèle Tobit. Ce modèle est un modèle normal censuré que l’on peut définir à l’aide d’une variable latente  $Y_i* = \mu + Z_i^\top \beta + \epsilon_i$ , où  $\beta$  est un vecteur de paramètres inconnus et  $\epsilon_i$  est une variable aléatoire de distribution  $N(0, \sigma^2)$ . On observe  $Y_i = Y_i*$  si  $Y_i* > 0$  et  $Y_i = 0$  sinon. Les données égales à 0 sont donc considérées comme étant censurées à gauche. On a modélisé, pour les trois années de l’étude, les dommages en fonction des deux variables explicatives retenues aux sections 3.1 et 3.3, TR\_AR et  $X_N$ . Les résultats apparaissent au tableau 3–5.

En 2000 et 2001, la variable TR\_AR est la seule variable significative dans les analyses des dommages. L’estimation du paramètre est cependant beaucoup plus élevée en 2001. De plus, le seuil observé en 2001 pour la distance au lac  $X_N$  est de 10%. Cette variable semble donc être associée aux dommages.

#### 5. CONCLUSION

En 2000 et 2001, les captures de carpocapses sauvages et stériles sont semblables. Toutefois, il y a eu beaucoup plus de dommages dans les vergers en 2001. Peut-on expliquer ce phénomène ? Les

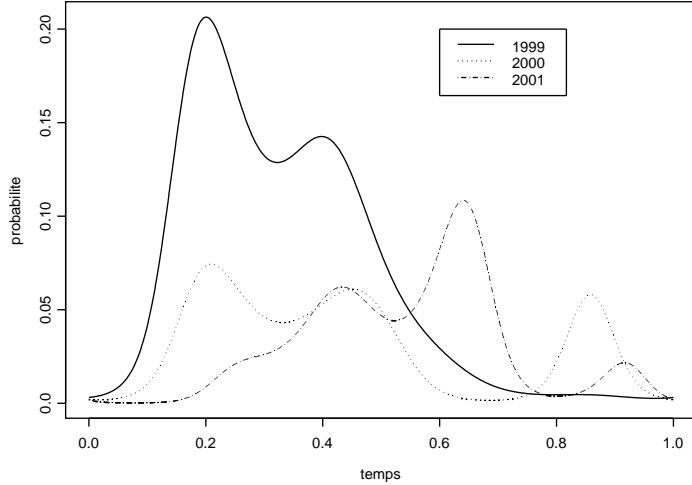


Figure 3–4: Probabilité estimée de non-conformité à la règle du 40 stériles pour 1 sauvage.

Tableau 3–5 : Estimation des paramètres de la régression tobit pour les dommages et de leurs erreurs-types (e.t.).

Année	1999	2000	2001
Var.	$\hat{\beta}$ (e.t.)	$\hat{\beta}$ (e.t.)	$\hat{\beta}$ (e.t.)
$\mu$	-0.0074 (0.0013)	-0.0077 (0.0022)	-0.0218 (0.0041)
$X_N$	-0.0008 (0.0005)	0.0011 (0.0007)	0.0022 (0.0014)
TR_AR	0.0007 (0.0009)	0.0025 (0.0010)	0.0086 (0.0028)
$\sigma^2$	0.0062 (0.0009)	0.0046 (0.0012)	0.016 (0.0024)

analyses présentées ici permettent de formuler certaines hypothèses. En 2001, l'arrivée tardive des carpocapses sauvages fait en sorte que la règle du 40 stériles pour 1 sauvage n'était pas respectée en juillet dans un nombre important de vergers. En juillet, les pommes ont atteint une grosseur appréciable et elles sont peut-être plus sensibles aux carpocapses. Une deuxième caractéristique de 2001 est une diminution des insectes stériles près du lac. Cette diminution est associée à une augmentation des insectes sauvages et des dommages aux vergers près du lac. Soulignons également qu'en 2000 et 2001 les vergers moins denses, correspondant à de grandes valeurs de TR\_AR, sont favorables à la présence de carpocapses sauvages et sont ainsi associés à des dommages plus importants.

Notre dernière hypothèse concerne le monstre Ogopogo un résident du lac Okanagan qui aurait, paraît-il, favorisé la prolifération des carpocapses sauvages près du lac en 2001. Peut-être faudrait-il prendre des moyens musclés pour éloigner Ogopogo de la rive du lac Okanagan au cours des prochaines années.

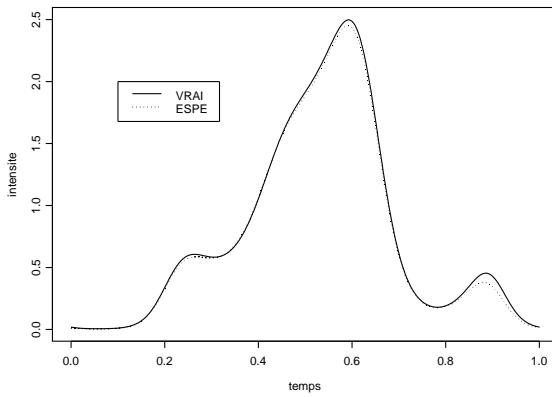


Figure 3-5: Comparaison de la vraie intensité (VRAI) et de l'espérance de l'intensité estimée (ESPE) par la régression Poisson ( 4).

Gaétan DAIGLE: [Gaetan.Daigle@mat.ulaval.ca](mailto:Gaetan.Daigle@mat.ulaval.ca)

Thierry DUCHESNE: [duchesne@mat.ulaval.ca](mailto:duchesne@mat.ulaval.ca)

Emmanuelle RENY-NOLIN: [Reny-Nolin@mat.ulaval.ca](mailto:Reny-Nolin@mat.ulaval.ca)

Louis-Paul RIVEST: [lpr@mat.ulaval.ca](mailto:lpr@mat.ulaval.ca)

*Département de mathématiques et de statistique, Université Laval  
Québec (Québec), Canada G1K 7P4*

## Discussion of the analysis of codling moth data from the Okanagan Sterile Insect Release Program

Sylvia R. ESTERBY, Howard THISTLEWOOD, Bob VERNON and Scott SMITH

### 1. INTRODUCTION

The objectives of the case study (Vernon, Thistlewood & Smith 2005) concern primarily moth populations, and secondarily the resulting fruit damage. Moth populations change in size rapidly, are highly mobile spatially, and have offspring which cause fruit damage each generation. An important index for fruit-growers is the count of the moths within a standard area of orchard and a standard time interval. Adherence to consistent field protocols allows changes in the number of moths counted to be interpreted as changes in population size. In the current dataset, two populations are measured, sterile moths and native (wild) moths, whose interactions are critical to the success of the control program. The models and analyses reported by the two teams are primarily concerned with assessing how season, trap attributes, and topographical variables affect the counts or some function of the counts, of one or both populations.

A feature of the SIR data set, typical in biology, is the different spatial and temporal scales of observation. Variables provided for the case study were referenced to two spatial levels: trap and orchard, and to three temporal levels: weekly trap counts, annual orchard damage percentages, and one-time trap attributes recorded when the trap was first deployed. Also, the spatial distribution of traps and orchards is highly irregular, typical of biological data in mountain areas, due to topography or land use (Figure 1–1, Vernon, Thistlewood & Smith 2005)

The breadth of the questions posed and extent of the data permit various approaches, which, for this case study, differ in the data included, use of derived variables, models and methods of estimation. Nonetheless, several physical attributes of orchards or trap sites were consistently related to the response variable. The two analyses provide a clear picture of which effects are separable using these data and demonstrate the intimate connection between the model, method of estimation and results. In the present discussion, our emphasis is on how specific models were formulated to answer general questions, and on common approaches, novel features, or effectiveness of the analyses, in answering the questions.

In their analysis, Nathoo, Ainsworth, Gill and Dean, referred to as NAGD hereafter, used only the 1999 moth counts to address trap-attribute effects. They developed a succession of models using the total wild moth count at a trap over the monitoring period, the proportion of wild moths relative to the total number of wild and sterile moths at a trap each week, and the count of wild moths at a trap each week. As for Daigle, Duchesne, Reny-Nolin and Rivest, referred to as DDDR, addressed all questions posed in the study objectives and used the data for wild moths, sterile moths and damage over the three years. They also used several different models, but their emphasis was on answering each of the questions in turn.

## 2. DERIVED TRAP-ATTRIBUTE AND LOCATION VARIABLES

Derived variables were constructed by one or both of the teams to account for: (i) orchards of different tree density ranging from the traditional orchard with large, well-spaced trees to the new trellis-style orchards with small, high-density trees; (ii) different exposures due to the mountainous region; and (iii) geographic location.

The importance of orchard type was emphasized in the introductory paper, where a derived variable, Tree Area, a product of within-row spacing and between-row spacing, was related to basal diameter. Tree Area, representing the crown of the tree, was the variable chosen by NAGD for orchard type. Each team, with differences in the details, calculated derived variables as principle components based on four variables: within-row spacing, between-row spacing, basal diameter and tree height. The first principle component calculated by DDDR explained 89% of the variation, and when transformed back to the scale of the original variables (given the name TR\_AR), consisted of a product of the four variables where within-row spacing and basal diameter were more important than between-row spacing and tree height. Biologically, the basal diameter is a better predictor of tree age and potential suitability as moth habitat than is tree height and, similarly, within-row spacing may better reflect moth habitat than between-row spacing. Comparability of the variables, tree area and TR\_AR, is relevant to the discussion of the results.

The variables slope and aspect contain useful information about conditions in mountainous terrain. NAGD took north-facing exposure to be most important and used a variable that corresponded to either north-facing slope or not a north-facing slope. DDDR defined a more complex variable that included values for no slope and for a combination of the magnitude of non-zero slope and aspect.

DDRR considered the most important features that could be obtained from the geographical co-ordinates to be the distance from the shore of Lake Okanagan,  $X_N$ , and the displacement along the shore,  $Y_N$ . This lead to two variables that were practically uncorrelated but each was correlated with elevation of the trap.

### 3. COMPARISON OF MODELS AND CONCLUSIONS, 1999 WILD MOTHS

The assessment of the effect of the trap attributes on wild moth counts was based on an assumption of counts being independent Poisson and a log-linear relationship between the Poisson mean and covariates. In the NAGD analysis, this is conditional on random trap effects, where the latter included the spatial relationship between traps through the definition of neighbourhoods. For the DDDR analysis, the log-linear model included the spatial effect through the variables  $X_N$  and  $Y_N$ , as fixed effects.

#### 3.1. Wild moth counts.

For both teams, only two effects were consistently identified as being associated with either level or variability of wild moth counts: 1) the effect of orchard type as the variable TR\_AR of DDDR, and Tree Area variable of NAGD, and 2) spatial effect as a fixed effect through variables  $X_N$  and  $Y_N$  of DDDR and the random effects through neighbourhoods of NAGD. Numerically, it is possible to compare a coefficient of variation for the estimates of the parameters for TR\_AR and Tree Area for the 1999 counts. However, the analyses of NAGD use total counts at a trap over the monitoring period in the log-linear model for the conditional mean and this model with zero-heavy component, and use count at each monitoring time only when the analysis is based on the zero-heavy hidden Markov model. For the orchard-type variable, the values of the estimate or posterior mean and of the standard error or standard deviation, depending upon the method of analysis, are very consistent over the three models of NAGD. The results in order of presentation lead to coefficients of variation of 36%, 32% and 29% for Tree Area, using the corresponding measure of variability divided by estimate or mean. This is comparable to the coefficient of variation of 31% for TR\_AR based on the Poisson regression for 1999 of DDDR. Thus orchard effect is clearly dominant, wild moth counts are higher in traditional orchards with larger trees, and the uncertainty in interval estimates for this effect is similar for these four different analyses. The tree-size effect is well accepted biologically, so the analyses here point out, that: (i) having taken into account other variables, the effect of tree size is dominant; and (ii) the standard deviation associated with the tree-size effect is about one third the size of the effect.

Comparison of the results for the spatial component shows agreement on the existence of such a component; how best to model it is less clear. NAGD models include random effects for heterogeneity and spatial correlation. NAGDs Figures 2-2, 2-3 and 2-4 show the location of large residual relative risks or large random effects, interpretable as traps where the count or proportion is greater than can be accounted for by the covariates. These higher values are in the south-central region, with a couple of exceptions. Thus, the mapping of effects after the analysis provides the information about the location of large and small effects.

In contrast, the DDDR variables,  $X_N$ , distance from the lake, and  $Y_N$ , distance along the shore, provide a description of the general tendencies of change in moth counts. It is informative to plot the polynomial (coefficients from Table 3-2 of DDDR), which gives the adjustment due to the variable  $X_N$  (Figure 4-1). The 1999 curve shows that the reduction in the number of moths is smallest for values of  $X_N$  near 0, approximately central to the range -1 to 1.5. Since  $X_N$  is distance to the shore, and the boundary on the right side of the polygon enclosing the traps in the NAGD figures corresponds approximately to the shore of the lake, qualitatively there is agreement between the two analyses of a tendency for higher counts in 1999 to be located centrally in the region (Figure 2-2 NAGD).

No other covariates were significant for the DDDR model and the 1999 data, whereas NAGD found, on the basis of total counts, that higher numbers of wild moths were associated with lower elevation and, to a lesser extent, lower numbers on north-facing slopes and higher numbers in Delta traps. Such conclusions are biologically of interest, but under both the zero-heavy model and the zero-heavy hidden Markov model, the variables "North-facing slope" and "trap type" were no

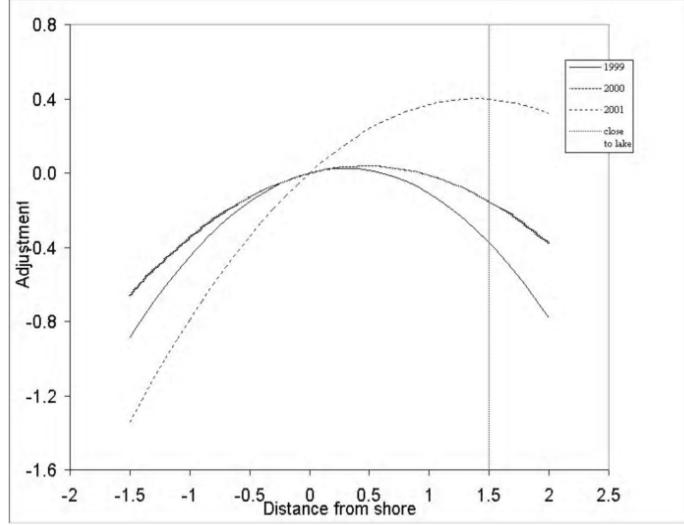


Figure 4–1: Effect of distance from the shore of Lake Okanagan in the Poisson regression models fitted for wild moth counts by DDDR. The vertical line is at  $X_N = 1.5$ .

longer even marginally significant. NAGD discuss some limitations of the zero-heavy models and the attribution of effect to alternative terms, each being plausible. For example, either elevation plays a role in the assignment of a trap to the zero-component or elevation is an effect that reduces the moth count. Similarly,  $X_N$  and  $Y_N$  may explain variability, which is attributed to terrain effects, under the different model for spatial effects of NAGD.

### 3.2. Seasonality.

Plots of the expected number of wild moths versus time, when based on the zero-heavy hidden Markov model (NAGD Figure 2–4c) and of the Poisson intensity versus time (DDRR Figure 3–1), show a very similar pattern, consistent with: an initial large peak at about week 4, from overwintered larvae, and a smaller peak at about week 8, due to the second generation of moths. Also noteworthy is adequate flexibility of both the Fourier series of order 4 (DDRR) and the B-splines (NAGD) for modelling the component for change over time.

### 3.3. Overflooding (Sterile:Wild) Ratio.

The success of the SIR program depends upon the number of sterile male moths being sufficient to inhibit successful mating of wild male moths, taken as a minimum of a 40:1 (overflooding) ratio of sterile/wild males (Vernon, Thistlewood & Smith 2005). NAGD considered the number of wild moths at a trap at time  $t$ , conditional on the total number of wild and sterile moths at time  $t$ , to be a binomial random variable, whereas DDDR took the response variable to be Bernoulli with value 1 if the 40:1 ratio is not achieved on the basis of the observed counts and 0 if it is achieved.

In assessing covariate effects on the basis of these response variables, the parameter estimates obtained by DDDR were very similar to those for wild moth counts, and tree area and elevation were significant for the NAGD analysis, as they had been for the analysis on total wild moth counts. In contrast, trap type is highly significant in the NAGD analysis of proportions (NAGD

Table 2–3) but not of counts, with the estimated odds of catching a wild moth being 5.5 times greater with delta traps than with wing traps (95% credible interval 3.1 to 9.5).

With the variables  $X_N$ ,  $X_N^2$  and TR.AR in the model, at the time of highest wild moth count,  $t = 0.2$  or week 4, the probability of nonconformance (ratio < 40:1) was about 0.20 (DDRR Figure 3–4). On the basis of the binomial random variable of NAGD, nonconformance corresponds to a binomial proportion of greater than  $1/41=0.024$ . NAGD identify sites with high probability of conformance with the 40:1 overflooding ratio. For individual traps, 28 of the 506 sites have a probability of 0.95 of not attaining the 40:1 ratio (Figure 2–3d).

### *3.4. Trap efficiencies or biases.*

The model for total wild moth counts of NAGD gives a point estimate of 1.6 times as many wild moths caught in delta traps as in wing traps (NAGD Table 2–2) and the analysis of sterile moths by DDRR gives an estimate of 3.7 times as many sterile moths caught in wing traps as in delta traps (DDRR Table 3–3). Assuming a low bias for wild moths in wing traps and for sterile moths in less frequently deployed delta traps, and using intensities of moths at  $t = 0.2$  from DDRR models as realistic values for the number of moths, sterile:wild ratios of 75, 20.3 and 120 were calculated for true (no trap bias), delta trap and wing trap (Table 4–1). Thus, nonconformity (below 40:1 ratio) might appear to be the case in delta traps when it is not, due to the effect of the trap bias. The DDRR analysis did not find trap significant whereas the NAGD analysis did, suggesting another case of the different models ascribing effects to different components, i.e., nonconformity versus trap effect.

The same values for numbers of moths and estimates of biases may be used to calculate the odds of catching a wild moth in a delta trap relative to a wing trap (Table 4–1). The value obtained, 5.69, agrees extremely well with the estimate of the odds from the binomial analysis of NAGD.

### *3.5. Hidden Markov model.*

The ability to estimate the probability of wild moths arriving at a trap given that there were no wild moths at the trap on the previous sampling occasion and the probability of there being no wild moths given that there were wild moths at the last sampling, is of considerable practical importance in determining factors that could contribute to re-infestation and resistance to the control measure in use. For example, the estimates 0.66 and 0.87, obtained by NAGD, of the probability of staying respectively in the non-zero or zero state over a season, relative to the 0.34 probability of moving from zero to non-zero state could suggest local infestation rather than immigration or importation. Scenarios that correspond to these probabilities are reinfestation in areas where moth populations are poorly controlled and importation of larvae in bins. The suggestions by NAGD for further refinements to allow the probabilities to depend on time and covariates seems particularly appropriate, but may not yield useful results for the sparse counts of this data set.

## 4. OTHER ANALYSES

### *4.1. Wild and sterile moth counts, 1999 to 2001.*

The analyses by DDRR of the entire 3-year data set provide a number of additional conclusions about covariate effects and seasonal patterns. There are higher numbers of wild moths in 1999. The differences in the patterns for Poisson intensity over years (Figure 3–1 DDRR) may be interpreted in terms of moth density and timing of the wild moth broods over the season of activity. The timing of the first and second peaks, possibly the over-wintered and second brood, for 1999 and 2000 is essentially the same, but there is a third peak in 2000 of almost equal amplitude to the

first two. This, contrasted with the lateness of the largest peak in 2001, leads to the question as to whether the diapausing larvae of the late third brood in 2000 failed to overwinter successfully or, as adults, they emerged late in 2001. Also, the late major peak in 2001 may have resulted in more moths being active at a critical time in 2001 resulting in more damage, as detected in the analysis of DDDR.

Table 4–1: Calculation of overflooding ratio and odds of catching a wild moth

	Number of moths		Ratio	Proportion	Odds
	Wild	Sterile	Sterile/Wild	Wild/Total	$P_{delta}/P_{wing}$
True <sup>a</sup>	8	600	75	0.013	
Delta trap	8	600/3.7	20.3	0.047	5.69
Wing trap	8/1.6	600	120	0.008	

<sup>a</sup>1999, t=0.2, number of wild moths from Figure 3–1 and sterile moths from Figure 3–3, DDDR.

The most important variable in the analysis of sterile moth counts was the trap style, with the count in Wing traps being about 4 times that in Delta traps (i.e., 3.7 used in Table 4–1), in agreement with what was previously known about trap differences. Distance from the lake and elevation were significant in 2000 and 2001 (DDRR Table 3–3), and in 2001 the effect of distance from the lake was approximately the reverse of that for wild moths (DDRR Table 3–2), i.e. sterile moths decreased while wild moths increased near the lake. Significant interactions showed that the effect of distance from the lake and the probability of capture in a Delta trap vary with elevation, and the effect of distance from lake and of elevation change over time. Thus, for the much more abundant sterile moths, the model is able to capture complexities, possibly related to sterile moth vigour and the effect of the change in temperature, which in turn is related to elevation and time. Also to be noted are the generally higher sterile moth intensity in 1999 and the presence of at least 3 peaks in the plots of intensity versus time (DDRR Figure 3–3). These observations raise questions as to whether the first was by design, i.e. larger numbers of sterile moths released because of the higher wild moth counts in 1999, and the second was because of variability in the vigour of the sterile moths released at different times.

#### 4.2. Time of first appearance of wild moths in 2001.

The key idea behind the analysis by DDDR is to consider time of first arrival of wild moths at a trap as analogous to failure in the context of survival distributions, thus making the well-developed methodology and software for this latter topic available. The analyses conducted for 2001 using both Cox's proportional hazard model and Weibull regression model for survival time (DDRR Table 3–2), show a much smaller effect for TR\_AR than in models based on counts, attributed to a small TR\_AR × t interaction. The plot of intensity versus time lacks the 3-peak feature of the plot based on counts (DDRR Figure 3–2). The latter form of plot is consistent with the interpretation of the multiple peaks for intensity of counts being due to successive broods. The first occurrence of wild moths at a trap would happen only for the over-wintered brood, given that there is such a brood at that location.

#### 4.3. Orchard damage.

The analysis by DDDR is based on the tobit model and left censoring at 0 %, using the variables  $X_N$  and TR\_AR. Only TR\_AR was significant, and only for 2000 and 2001, with a much larger

effect in 2001. Use of the Tobit model can be justified since censoring acknowledges the possibility of a lower percentage of damage than could be estimated from the size of samples taken. A zero percentage damaged is a reasonable indication that there is censoring for this sample, although the censoring point might be taken as a value other than 0.

## 5. FURTHER CONSIDERATIONS

The methods used in the analysis papers (NAGD and DDDR) have met a requisite feature of a good analysis, that of matching the method of analysis to the design of the data collection program. Thus, for example, an estimate of tree-size effect could be given, having taken into account other factors that affect moth abundance, including temporal pattern and spatial location. It is reassuring that the dominant effects detected, using this data set from an operational program, agree with results that were previously known, qualitatively or quantitatively, to field scientists. Also important are the concurrence on dominant effects from the two sets of quite different analyses, including the associated level of variability for the orchard type/tree size effect, and the further insight into effects, including interactions, which may only be quantified by models where the factors are included simultaneously.

The ranges of variable values and the levels of different variables that occur together cannot be controlled in observational studies such as this, and a further problem is posed by count data with low values or frequent zeros. The data set was chosen to provide a large range of values for the attributes, but nonetheless for wild moths, with sparse counts, only a couple of effects could be consistently detected. The real potential of the methods used in the case study is better shown by the analysis of the sterile moth data, where the more abundant sterile moths, because they are occurring at all locations, give more information about the effect of explanatory variables, singly and in combination, than obtained from 0 or low counts of the wild moths. However, conclusions drawn are still based only on the particular combinations of variables that happened to occur in the data set, and thus subject to limitations about applicability to other regions or times. A useful additional analysis for such data sets would be a characterization of the combinations of variables in the data set.

The nature of the region from which the data were obtained, with its highly irregular boundaries and considerable vertical relief (Figure 1–1, Vernon, Thistlewood & Smith 2005), poses challenges for spatial analysis. The neighbourhoods (NAGD) based on isotropic distance do not seem ideal, but may have worked because of a good choice of distance. Similarly, it would be interesting to know if there remains some, presumably weak, small scale spatial structure in the residuals from the analysis of DDDR, that corresponds to the location of large relative risks in the NAGD analysis (Figure 2–2), since, at least for the 2001 analysis, mention was made by DDDR of confirming the stability of the model by removing large residuals. To some extent, the two ways of modelling the spatial component are complementary, not only in terms of local versus general tendency characterizations, but also in demonstrating the interaction between which effects are found to be significant and the way that they are modeled.

During preparation of the case study data set, the question arose as to whether trap style should be included as a variable or whether to simply exclude the Delta-trap data. The former route was taken and some results about the effect of trap style emerged. One concern would be that, by including the Delta-trap data, we are not able to detect the effects of interest. The consistent effect of the orchard-type variable and the identification of spatial regions of higher abundance can be stated confidently. Interactions, including those with Delta-trap type, detectable for the higher sterile moth counts (Table 3–3 DDDR), suggests that the issue about efficiency for both wild and sterile moths in Wing and Delta traps needs further consideration. If a program uses only wing traps, then an analysis might be done including only Wing-trap data, but this would result in a loss of regional coverage for this data set. However, since typical operational data have included

both wing and delta traps, the present analysis is useful in considering how to prepare annual summaries based on such data.

There are two well-developed areas of statistical application that are related to the case-study data set. The first area is the development of entomological sampling plans used for the control of agricultural pests, and the second, the role of the spatial distribution of a species in inferring the existence of underlying processes in ecology. Probability distributions are used to develop sampling plans (e.g., Binns & Nyrop 1992; Park & Cho 2004) and, for data consisting of counts, the starting position might be a Poisson distribution under homogeneity leading to, e.g., a negative binomial distribution under heterogeneity. In the second case, the Poisson distribution is considered to represent the case of a random distribution of organisms in space, and data not fitted by a Poisson distribution would be considered to arise from a non-random spatial distribution, typically under-dispersed or over-dispersed. An alternative approach is given by methods that characterize the nature of the non-randomness, such as that implemented in the SADIE (Spatial Analysis by Distance IndicEs) software which provides a means to measure overall spatial pattern for a set of spatially referenced data (Perry & Dixon 2002). The primary relationship to the methods used in the case study is the idea that counts, if homogeneous, may be characterized by a Poisson distribution, but if not, we need to account for the heterogeneity.

More closely related to the situation of monitoring for moths using standardized traps and consistent protocol is the point-count monitoring of bird populations in programs such as the North American Breeding Bird Survey (Peterjohn, Sauer & Robbins 1995). The point-counts are taken under a standardized protocol at fixed locations and are used as an index to population size, just as in the case of moth counts in the pheromone traps. There are similar considerations for covariates, time, and space. The data have been analyzed by a progression of statistical methods starting with route-regression, linear regression on transformed counts at a site, and more recently including estimating equations and hierarchical models (Link & Sauer 1994, 2002).

#### ACKNOWLEDGEMENTS

In addition to the acknowledgements of the introductory paper, the Case Studies Editor would like to thank Paramjit Gill for suggesting the case study, and PARC co-authors and SIR colleagues for their patience during the very extensive consultations involved in preparing the data set and documentation for the analyst teams, and in writing the introductory and discussion papers.

---

Sylvia R. ESTERBY: [sylvia.esterby@ubc.ca](mailto:sylvia.esterby@ubc.ca)  
Mathematics, Statistics and Physics, University of British Columbia Okanagan  
Kelowna, British Columbia, Canada V1V 1V7

Howard THISTLEWOOD: [thistlewoodh@agr.gc.ca](mailto:thistlewoodh@agr.gc.ca)  
Pacific Agri-Food Research Centre, Agriculture and Agri-Food Canada  
Summerland, British Columbia, Canada V0H 1Z0

Bob VERNON: [vernonbs@agr.gc.ca](mailto:vernonbs@agr.gc.ca)  
Pacific Agri-Food Research Centre, Agriculture and Agri-Food Canada  
Agassiz, British Columbia, Canada V0M 1A0

Scott SMITH: [smithcas@agr.gc.ca](mailto:smithcas@agr.gc.ca)  
Pacific Agri-Food Research Centre, Agriculture and Agri-Food Canada  
Summerland, British Columbia, Canada V0H 1Z0

## COMBINED REFERENCES/BIBLIOGRAPHIE COLLIGÉE

- A. Alioum & D. Commenges (1996). A proportional hazard model for arbitrarily censored and truncated data. *Biometrics*, 52, 512–524.
- R. Balshaw & C. B. Dean (2002). A semi-parametric model for recurrent event panel data. *Biometrics*, 58, 324–331.
- L. Bernardinelli, D. G. Clayton & C. Montomoli (1995). Bayesian estimates of disease maps – How important are priors? *Statistics in Medicine*, 14, 2411–2431.
- J. Besag, J. York & A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1–21.
- M. R. Binns & J. P. Nyrop (1992). Sampling insect populations for the purpose of IPM decision making. *Annual Review of Entomology*, 37, 427–453.
- N. E. Breslow & D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9–25.
- N. E. Breslow & X. Lin (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82, 81–91.
- A. C. Cameron & P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge University Press, New York.
- D. Dendy (1999). The worm in the apple: contesting the codling moth in British Columbia. In *Beyond the City Limits: Rural History in British Columbia*. (R. W. Sandwell, ed.). University of British Columbia Press, pp. 142–156.
- M. J. Dobbie & A. H. Welch (2001). Modelling correlated zero-inflated count data. *Australian and New Zealand Journal of Statistics*, 43, 431–444.
- S. Dorn, P. Schumacher, C. Abivardi, R. Meyhofer & M. W. Brown. (1999). Global and regional pest insects and their antagonists in orchards: spatial dynamics. *Agriculture, Ecosystems and Environment*, 73, 111–118.
- V. A. Dyck & M. G. Gardiner (1992). Sterile-insect release program to control the codling moth *Cydia pomonella* (L.) (Lepidoptera: Olethreutidae) in British Columbia, Canada. *Acta Phytopathologica et Entomologica Hungarica*, 27, 219–222.
- D. L. Johnson (1989). Spatial autocorrelation, spatial modeling, and improvements in grasshopper survey methodology. *Canadian Entomologist*, 121, 579–588.
- D. L. Johnson, O. O. Olfert, M. G. Dolinski & L. Harris (1996). GIS-based forecasts for management of grasshopper populations in Western Canada. In *Proceedings of the FAO Symposium on Agricultural Pest Forecasting and Monitoring*, Québec (Québec) Canada, pp. 109–112.
- G. J. R. Judd, M. G. T. Gardiner & D. R. Thomson (1996). Commercial trials of pheromone-mediated mating disruption with Isomate-C R to control codling moth in British Columbia apple and pear orchards. *Journal of the Entomological Society of British Columbia*, 93, 23–34.
- J. P. Klein & M. L. Moeschberger (1997). *Survival Analysis. Techniques for Censored and Truncated Data*. Springer, New York.
- D. Lambert (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34, 1–14.
- J. F. Lawless & C. Nadeau (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37, 158–168.
- W. A. Link & J. R. Sauer (1994). Estimating equations estimates of trend. *Bird Populations*, 2, 23–32.
- W. A. Link & J. R. Sauer (2002). A hierarchical analysis of population change with application to Cerulean Warblers. *Ecology*, 83, 2832–2840.
- Y. MacNab & C. B. Dean (2001). Autoregressive spatial smoothing and temporal spline smoothing for mapping rates, *Biometrics*, 57, 949–956.

- H. F. Madsen & P. J. Procter (1982). *Insects and Mites of Tree Fruits in British Columbia*. British Columbia Ministry of Agriculture and Food. Queen's Printer, Victoria, British Columbia, Canada.
- H. F. Madsen & T. M. Vakenti (1973). Codling moth: Use of codlemone baited traps and visual detection of entries to determine need for sprays. *Environmental Entomology*, 2, 677–679.
- C. Marston, K. Barg & J. Cranston (2002). A GIS Analysis of Spatial Factors Affecting the Resistance of Codling Moth to Sterile Insect Release Treatments. Unpublished GIS Student Project, British Columbia Institute of Technology, Burnaby, British Columbia, Canada.
- A. Maul (1999). Sequential monitoring in large comparative studies with low event rates, an application to environmental epidemiology. *Environmetrics*, 10, 439–455.
- J. Mullahy (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341–365.
- H. Park & K. Cho (2004). Use of covariates in Taylor's Power Law for sequential sampling in pest management. *Journal of Agricultural, Biological and Environmental Statistics*, 9, 462–478.
- J. N. Perry & P. M. Dixon (2002). A new method to measure spatial association for ecological count data. *Ecoscience*, 9, 133–141.
- B. G. Peterjohn, J. R. Sauer & C. S. Robbins (1995). Population trends for the North American Breeding Bird Survey. In *Ecology and Management of Neotropical Migratory Birds* (T. E. Martin & D. M. Finch, eds.). Oxford University Press, pp. 3–39.
- M. D. Proverbs & J. R. Newton (1975). Codling moth control by sterile insect release: importation of fruit and fruit containers as a source of reinfestation. *Journal of the Entomological Society of British Columbia*, 72, 6–9.
- M. D. Proverbs, J. R. Newton & C. J. Campbell (1982). Codling moth: A pilot program of control by sterile insect release in British Columbia. *Canadian Entomologist*, 114, 363–376.
- J. Sun & L. J. Wei (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B*, 62, 293–302.
- B. C. Sutradhar & Z. Qu (1998). On approximate likelihood inference in a Poisson mixed model. *The Canadian Journal of Statistics*, 26, 169–186.
- T. M. Therneau & P. M. Grambsch (2000). *Modeling Survival Data. Extending the Cox Model*. Springer, New York.
- H. Thistlewood & G. Judd (2002). Movement and dispersal of codling moth, *Cydia pomonella*, in rural and orchard areas. *IOBC/WPRS Bulletin*, 26, 111–116.
- H. M. A. Thistlewood, G. J. R. Judd & M. E. O. Clodius (2004). Sustainable management and monitoring of codling moth, *Cydia pomonella*, in an area-wide program employing SIT. (IAEA eds.), pp. 79–87. *Proceedings of the Second Research Coordination Meeting, Improvement of Codling Moth SIT to Facilitate Expansion of Field Application, Stellenbosch, South Africa*, March 8–12, 2004. Working Paper IAEA-314-D4-RC876. IAEA, Vienna, Austria.
- L. A. Waller, B. P. Carlin, H. Xia & A. E. Gelfand (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92, 607–617.
- Y. Yasui & S. Lele (1997). A regression method for spatial disease rates: an estimating function approach. *Journal of the American Statistical Association*, 92, 21–32.