Abstract: Two empirical statistical models were developed to describe the spatial variation in defoliation by spruce budworm (*Choristoneura fumiferana* Clem.), as recorded by Ontario’s Forest Health Survey from 1967 to 1998. These models revealed a number of relationships between the spatial distributions of aerially detectable spruce budworm defoliation and bioclimatic conditions over the landscape. A classification tree model relates the northern and southern boundaries of defoliation to the relative abundance of different tree species that host spruce budworm. Between these boundaries, the classification tree uses the maximum winter temperature and the minimum temperature in May to describe where detectable defoliation occurred. A regression tree model uses a total of eight variables related to winter temperatures, forest composition, spring temperatures, summer temperatures, and precipitation to estimate the defoliation frequency in areas where defoliation was detected at least once from 1967 to 1998. High defoliation frequencies were associated with dry Junes (precipitation, <86 mm) and cool springs (mean minimum temperature < –2.7 °C). Conversely, low frequencies were associated with cold winters (mean minimum temperature < –23.3 °C; mean maximum temperature > –11.0 °C) in the north and a low abundance of host species (percentage of the basal area occupied by balsam fir, white spruce, and black spruce, <14.3%) in the south. Spatial autocorrelation in the bioclimatic variables had little effect on their relationships with the spatial distribution of the defoliation frequency.

Résumé : Deux modèles statistiques empiriques ont été développés pour décrire la variation spatiale dans la défoliation causée par la tordeuse des bourgeons de l’épinette (*Choristoneura fumiferana* Clem.) selon les données de l’inventaire de l’état de santé des forêts de l’Ontario de 1967 à 1998. Ces modèles mettent en évidence plusieurs relations entre la répartition spatiale de la défoliation causée par la tordeuse des bourgeons de l’épinette qui est décelable par voie aérienne et les conditions bioclimatiques dans l’ensemble du paysage. Un arbre de classification relie les limites septentrionale et méridionale de la défoliation et l’abondance relative des différentes espèces d’arbres hôtes de la tordeuse des bourgeois de l’épinette. À l’intérieur de ces limites, l’arbre de classification utilise la température hivernale maximum et la température minimum en mai pour déterminer où s’est produite une défoliation décelable. Un arbre de régression utilise un total de huit variables reliées aux températures hivernales, à la composition forestière, aux températures printanières et aux précipitations et températures estivales pour estimer la fréquence de la défoliation dans les régions où une défoliation a déjà été détectée au moins une fois entre 1967 et 1998. Des fréquences élevées de défoliation sont associées à un mois de juin sec (précipitation < 86 mm) et à un printemps frais (température minimum moyenne < –2,7 °C). À l’inverse, des fréquences faibles sont associées à des hivers froids (température minimum moyenne < –23,3 °C et température maximum moyenne > –11,0 °C) dans le nord et une faible abondance d’espèces hôtes (pourcentage de la surface terrière occupée par le sapin baumier, l’épinette blanche et l’épinette noire < 14,3 %) dans le sud. L’autocorrélation spatiale entre les variables bioclimatiques avait peu d’effet sur leurs relations avec la répartition spatiale de la fréquence de la défoliation.

Introduction

The role of climate in the dynamics of forest insect populations has long been a subject of debate (Uvarov 1931). Emerging issues, such as climate change, have given a new impetus to this question for both ecological and economic reasons. In the boreal forests of North America, insects are a major driver of ecosystem dynamics. During widespread,
periodic outbreaks, trees are attacked over vast areas, and tree mortality and growth reduction result in the loss of millions of cubic metres of wood. Severe outbreaks can also result in forest successional changes (Batzler and Popp 1985), related changes in the number and species of wildlife (Welsh 1983), increased risk of fire (Stocks 1987; McCullough et al. 1998; Fleming et al. 2002), and increased susceptibility to other insects and fungal pathogens (Shortle and Ostrofsky 1983; Houston 1984).

In Canada, annual wood volumes lost to insects were estimated to be 81 × 10⁶ to 107 × 10⁶ m³, with spruce budworm, Choristoneura fumiferana Clem. (Lepidoptera: Tortricidae), causing 41%–53% of the loss (Sterner and Davidson 1982; Power 1991; National Forestry Database 1995). Spruce budworm larvae feed on the flowers, cones, and youngest available foliage of its preferred hosts, balsam fir (Abies balsamea (L.) Mill.) and white spruce (Picea glauca (Moench) Voss). Occasionally, other species including black spruce (Picea mariana (Mill.) BSP) are attacked. Repeated removal of the current year’s foliage eventually causes tree mortality. Maps of cumulative defoliation indicate that in Ontario, 41 × 10⁶ ha has experienced moderate to severe defoliation for at least 1 year since 1941 (Candau et al. 1998). In 1981, at the peak of the last outbreak, the overall loss of wood volume to budworm was estimated at 16 × 10⁶ m³ in this province alone, whereas the total volume harvested was 20 × 10⁶ m³. During the same period in northeastern Ontario, tree mortality due to repeated defoliation averaged 10%–15% for white spruce, 17% for black spruce, and up to 85% for balsam fir (Meating et al. 1982).

Previously, Candau et al. (1998) showed that since 1941, aerially detected spruce budworm defoliation in Ontario has occurred in three zones of frequent defoliation separated by longitudinally oriented corridors of lower frequency. Each zone was characterized by “hot spots”, or areas where defoliation occurred in over one-third of all years observed. These hot spots were surrounded by approximately radial gradients of decreasing defoliation frequencies. It was suggested that environmental conditions (e.g., climate or host species) might explain aspects of the observed patterns.

Numerous authors have noted climatic influences on the spatial distribution of spruce budworm defoliation. Comparing the history of outbreaks in the lower St. Lawrence and Gaspé regions, Blais (1961) suggested that the prevailing climate of Gaspé (cooler, wetter summers and shorter frost-free periods) appears to be the main reason why conditions for spruce budworm outbreaks are marginal in this region, even though the preferred host species are more abundant here than in the lower St. Lawrence. In a general theory, Hardy et al. (1983) suggested that outbreaks follow a progression by bioclimatic zones: epicentres located in the zone of “normal abundance” serve as foci from which the infestation develops and spreads north and south to zones of “occasional abundance”, and under certain conditions, even farther north and south to zones of “possible abundance”. Each zone, originally solely defined by defoliation frequency, is also characterized by bioclimatic conditions. In the zone of normal abundance, climate is favourable, and host abundance and predators are responsible for fluctuations of insect populations. In the zone of occasional abundance, climatic conditions are generally less favourable, and they are almost always unfavourable in the zone of possible abundance.

Forest characteristics can also influence spruce budworm defoliation through various mechanisms. In general, older stands of balsam fir, as well as stands with higher densities and a greater proportional content of balsam fir, are more prone to defoliation (Balch 1946; Mott 1963; Su et al. 1996). This may be due to effects of forest characteristics on insect dispersal, reproductive success, and (or) survival (Morris 1963; Kemp and Simmons 1979).

The previous work on how climate and forest composition influence where spruce budworm defoliation occurs is generally of two types: either statistically rigorous analyses of data from a very limited spatial extent, perhaps even just three or four neighboring sites; or subjective assessments of the geographical patterns in time-series of landscape-scale maps. One exception is a study of the potential impact of climate change on spruce budworm defoliation in the northeastern and north-central United States by Williams and Liebhold (1997). These authors developed a model of the presence or absence of defoliation in relation to climate and forest conditions. However, their data imposed limitations on their work. First, because their defoliation data (i.e., 1954–1980) spanned less than a complete outbreak cycle (Candau et al. 1998; Royama 1984) and outbreak cycles can be 5–6 years out of phase in neighboring zones (Candau et al. 1998), spatio-temporal bias can result. Second, as Williams and Liebhold acknowledge, their work depends critically on the assumption that spruce budworm defoliation occurs only within the spruce–fir forest type. Forest type data are collected by remote sensing, and “a forest type must cover a significant amount of land on the AVHRR map in order to be conspicuous” (Zhu and Evans 1992, p. 28). Thus, Williams and Liebhold (1997) implicitly excluded as possible predictor variables the proportion of host species present and any interactions of this variable with climatic factors. They also implicitly excluded the possibility of spruce budworm defoliation in other forest types in which spruce and fir are minor components.

We tried to minimize these difficulties in our models by using data from all years in which defoliation was recorded in Ontario during the last complete outbreak cycle (1967–1998) and by including the host species effects quantitatively (proportional representation), rather than merely qualitatively (presence or absence). We statistically evaluated the influences of climate and forest characteristics on the frequency of spruce budworm defoliation using classification and regression tree models (Breiman et al. 1984). Our objectives were to determine (1) the extent to which the large-scale spatial patterns in defoliation can be explained by factors reflecting climate and forest characteristics; and (2) which of these factors are most important for defoliation.

**Materials and methods**

Digital maps of defoliation and bioclimatic data were saved in a set of raster grids using 1 km × 1 km pixel size. These maps were analyzed in a search for spatial relationships between bioclimatic factors and (1) the presence or absence of defoliation and (2) defoliation frequency. For each analysis, a random sample of 600 pixels was drawn from the grids. To
begin the first analysis, the map of defoliation frequency was reclassified into areas that were defoliated at least once before the sample was drawn and areas never defoliated. The presence or absence of defoliation was studied separately to better identify the bioclimatic factors related to the northern and southern boundaries of defoliation (the eastern and western boundaries of spruce budworm defoliation occur beyond the range of our data). In the second analysis, the sample was drawn from the area of defoliation only. In each analysis, we fitted a multivariate model relating defoliation to bioclimatic factors. The models were then applied to the rest of the data to produce maps of estimated defoliation frequencies. These maps were later compared with corresponding maps of observed defoliation. We purposely limited our study to Ontario: although the spruce budworm range extends far beyond this province, important differences in defoliation survey methods and spraying histories can introduce bias when data from neighboring jurisdictions are combined (Candau et al. 1998).

We used the software R (Ihaka and Gentleman 1996) for all statistical analyses. Regression tree analyses were performed with the R library RPART (Therneau and Atkinson 1997).

Table 1. Minimum, mean, maximum, and standard deviation of the defoliation frequency by spruce budworm and biogeoclimatic variables in the 600-cell random sample used in analyzing the spatial distribution of the defoliation frequency in relation to bioclimatic conditions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Name</th>
<th>Min.</th>
<th>Mean</th>
<th>Max.</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defoliation frequency (no. years of defoliation)</td>
<td>Freq</td>
<td>1.00</td>
<td>7.19</td>
<td>19.00</td>
<td>3.21</td>
</tr>
<tr>
<td>Minimum temperature, winter (°C)</td>
<td>TminWinter</td>
<td>−26.04</td>
<td>−20.71</td>
<td>−10.69</td>
<td>3.07</td>
</tr>
<tr>
<td>Maximum temperature, winter (°C)</td>
<td>TmaxWinter</td>
<td>−12.92</td>
<td>−8.55</td>
<td>−2.42</td>
<td>2.36</td>
</tr>
<tr>
<td>Minimum temperature, spring (°C)</td>
<td>TminSpring</td>
<td>−9.31</td>
<td>−5.48</td>
<td>−0.06</td>
<td>1.98</td>
</tr>
<tr>
<td>Minimum temperature, summer (°C)</td>
<td>TminSummer</td>
<td>5.40</td>
<td>7.69</td>
<td>11.11</td>
<td>1.31</td>
</tr>
<tr>
<td>Maximum temperature, summer (°C)</td>
<td>TmaxSummer</td>
<td>17.15</td>
<td>20.73</td>
<td>23.42</td>
<td>1.07</td>
</tr>
<tr>
<td>Precipitation, summer (mm)</td>
<td>PrepSummer</td>
<td>63.59</td>
<td>83.01</td>
<td>99.55</td>
<td>6.41</td>
</tr>
<tr>
<td>Minimum temperature, May (°C)</td>
<td>MINT5</td>
<td>−0.04</td>
<td>2.46</td>
<td>6.34</td>
<td>1.43</td>
</tr>
<tr>
<td>Minimum temperature, June (°C)</td>
<td>MINT6</td>
<td>5.30</td>
<td>7.62</td>
<td>11.27</td>
<td>1.35</td>
</tr>
<tr>
<td>Minimum temperature, July (°C)</td>
<td>MINT7</td>
<td>8.87</td>
<td>10.95</td>
<td>14.07</td>
<td>1.30</td>
</tr>
<tr>
<td>Minimum temperature, August (°C)</td>
<td>MINT8</td>
<td>7.30</td>
<td>9.75</td>
<td>13.40</td>
<td>1.27</td>
</tr>
<tr>
<td>Maximum temperature, May (°C)</td>
<td>MAXT5</td>
<td>12.80</td>
<td>15.98</td>
<td>18.77</td>
<td>1.26</td>
</tr>
<tr>
<td>Maximum temperature, June (°C)</td>
<td>MAXT6</td>
<td>16.84</td>
<td>20.93</td>
<td>23.56</td>
<td>1.07</td>
</tr>
<tr>
<td>Maximum temperature, July (°C)</td>
<td>MAXT7</td>
<td>19.90</td>
<td>23.96</td>
<td>26.56</td>
<td>1.06</td>
</tr>
<tr>
<td>Maximum temperature, August (°C)</td>
<td>MAXT8</td>
<td>19.06</td>
<td>22.04</td>
<td>24.95</td>
<td>1.01</td>
</tr>
<tr>
<td>Precipitation, May (mm)</td>
<td>PCP5</td>
<td>44.17</td>
<td>69.55</td>
<td>86.22</td>
<td>8.34</td>
</tr>
<tr>
<td>Precipitation, June (mm)</td>
<td>PCP6</td>
<td>54.76</td>
<td>87.11</td>
<td>114.92</td>
<td>9.50</td>
</tr>
<tr>
<td>Precipitation, July (mm)</td>
<td>PCP7</td>
<td>55.37</td>
<td>87.44</td>
<td>104.74</td>
<td>9.77</td>
</tr>
<tr>
<td>Precipitation, August (mm)</td>
<td>PCP8</td>
<td>63.59</td>
<td>87.94</td>
<td>114.94</td>
<td>7.59</td>
</tr>
<tr>
<td>Elevation (m above sea level)</td>
<td>ELEV</td>
<td>30.01</td>
<td>314.32</td>
<td>532.14</td>
<td>434.54</td>
</tr>
<tr>
<td>% basal area occupied by balsam fir and white spruce</td>
<td>FbSw</td>
<td>0.00</td>
<td>10.10</td>
<td>47.05</td>
<td>7.43</td>
</tr>
<tr>
<td>% basal area occupied by balsam fir, white spruce, and black spruce</td>
<td>FbSwSb</td>
<td>0.00</td>
<td>43.38</td>
<td>99.60</td>
<td>25.20</td>
</tr>
<tr>
<td>% basal area occupied by age-class 0–20 years</td>
<td>AGE10</td>
<td>0.00</td>
<td>11.56</td>
<td>97.35</td>
<td>14.64</td>
</tr>
<tr>
<td>% basal area occupied by age-class 20–40 years</td>
<td>AGE30</td>
<td>0.00</td>
<td>6.12</td>
<td>66.38</td>
<td>10.55</td>
</tr>
<tr>
<td>% basal area occupied by age-class 40–60 years</td>
<td>AGE50</td>
<td>0.00</td>
<td>14.49</td>
<td>84.21</td>
<td>15.86</td>
</tr>
<tr>
<td>% basal area occupied by age-class 60–80 years</td>
<td>AGE70</td>
<td>0.00</td>
<td>23.66</td>
<td>81.36</td>
<td>18.68</td>
</tr>
<tr>
<td>% basal area occupied by age-class 80–100 years</td>
<td>AGE90</td>
<td>0.00</td>
<td>18.09</td>
<td>95.00</td>
<td>17.13</td>
</tr>
<tr>
<td>% basal area occupied by age-class 100–120 years</td>
<td>AGE110</td>
<td>0.00</td>
<td>10.15</td>
<td>74.34</td>
<td>11.67</td>
</tr>
<tr>
<td>% basal area occupied by age-class 120–140 years</td>
<td>AGE130</td>
<td>0.00</td>
<td>7.78</td>
<td>74.62</td>
<td>11.02</td>
</tr>
<tr>
<td>% basal area occupied by age-class 140–160 years</td>
<td>AGE150</td>
<td>0.00</td>
<td>6.06</td>
<td>86.69</td>
<td>11.61</td>
</tr>
<tr>
<td>% basal area occupied by age-class 160–180 years</td>
<td>AGE170</td>
<td>0.00</td>
<td>1.32</td>
<td>69.67</td>
<td>4.81</td>
</tr>
<tr>
<td>% basal area occupied by age-class 180–200 years</td>
<td>AGE181</td>
<td>0.00</td>
<td>0.26</td>
<td>22.38</td>
<td>1.56</td>
</tr>
</tbody>
</table>

Note: Winter months, Dec.–Feb.; spring months, Mar.–May; summer months, May–Aug.

Defoliation data

In Ontario, the Forest Insect and Disease Survey (FIDS) of the Canadian Forest Service (CFS) has conducted extensive aerial surveys of large-scale spruce budworm defoliation since 1941. Aerial surveys map discolouration of foliage and other changes in tree appearance caused by the feeding of budworm larvae (Sippell 1983; Howse 1995). Three levels of defoliation are defined on the basis of percentage of foliage damaged: light (0%–25%), moderate (26%–75%), and severe (76%–100%). Mapping is done from an aircraft and involves sketching areas defoliated on 1 : 125 000 or 1 : 250 000 maps. Later these maps are compiled and transferred to smaller scale maps (e.g., 1 : 600 000).

Candau et al. (1998) analyzed the annual maps of moderate to severe spruce budworm defoliation in Ontario between 1941 and 1996. (Records of light defoliation were ignored because these records are widely considered to be relatively unreliable (Sippell 1983).) In 1941, when surveys were operationally initiated in their present large-scale format in Ontario, an outbreak was already underway. This outbreak lasted until 1963. Our analyses focus on the only completely surveyed outbreak, which started in 1967, reached a maximum...

**Bioclimatic data**

Before starting these analyses, we created composite variables as bioclimatic predictors for our models. These variables were created on the basis of spruce budworm biology and phenology (Table 1) and focused on spatial variation in climate, vegetation, and elevation. The elevation data were derived from a digital elevation model (Mackey et al. 1994). The historical climate data were spatial interpolations of monthly minimums and maximums for temperature (°C) and precipitation (mm) from 471 meteorological stations across Ontario, eastern Manitoba, and western Quebec over the period 1961–1990 (Price et al. 2000). These data have a spatial resolution of 1 km.

The vegetation data were based on the forest resource inventory conducted by the Ontario Ministry of Natural Resources (OMNR 1996a). For this inventory, forest characteristics were determined at the stand level with a combination of airphoto interpretation and ground surveys, and then they were summarized over a raster grid varying in resolution between 5 km × 5 km and 20 km × 20 km (OMNR 1996b). For each pixel of the grid, the data included the percentage of the total basal area for (1) dominant trees in each 20-year age-class (denoted as AGE10 to AGE181); (2) balsam fir and white spruce (FbSw); and (3) balsam fir, white spruce, and black spruce (FbSwSb). FbSw accounts for the tree species on which spruce budworm primarily feeds, whereas FbSwSb covers all major hosts in Ontario. Stand age was assessed on the basis of dominant trees only. The OMNR vegetation data cover the entire area of the federal government FIDS defoliation surveys, as well as areas not reported by FIDS that are north and south of the area of defoliation. OMNR forest resource inventory data are not available in the northernmost part of the province (Fig. 1), but aerially visible defoliation is quite rare there (G. Howse, CFS, Sault Ste. Marie, Ontario, personal communication).

**Data set construction**

We partitioned our climate data into predictor variables reflecting different stages, periods, and events in the spruce budworm’s life cycle: mean minimum and mean maximum
temperatures during the winter (December to February); mean minimum and mean maximum temperatures and mean precipitation during the summer (May–August); and monthly minimum and maximum temperatures and mean precipitation during the feeding and adult periods (1 May to 31 August). Because of the importance of late spring frosts in limiting outbreak progress (Cerezke and Volney 1985), we also included spring (March to May) mean minimum temperature as a potential predictor variable. Because even a few predictor variables can interact in many ways and lead to spurious results in statistical tests, we restricted our investigations to these 18 climate variables. By contrast, Williams and Liebhold (1997) examined 24 climatic predictors, the average temperature and precipitation for each month (1948–1987).

To build our data set, we saved spatially referenced data on defoliation frequency and the bioclimatic predictor variables in a set of grids at 1-km resolution, representing 594,840 pixels. The choice of a 1-km resolution in the final data set corresponds to the finest resolution in the original data (i.e., the climate data). Although this necessarily introduced some redundancy in the coarser variables (e.g., forest composition), random sampling limited this redundancy and kept the samples used in fitting the models to a manageable size.

We then drew two samples of 600 pixels randomly, without replacement, but stratified according to whether (1) defoliation was present or absent for the first analysis; and (2) different defoliation frequency classes occurred for the second analysis. This sampling procedure represents a compromise between minimizing the loss of information, accounting for different resolutions in the original data, and keeping the final data set at a manageable size.

Classification and regression tree modelling

We used classification and regression trees (CARTs) (De’ath and Fabricius 2000), which are used for categorical and continuous response variables, respectively, to model the defoliation occurrence and frequency, respectively. Compared with classical methods for predictive modelling (e.g., generalized linear model and logistic multiple regression), CART models do not require the restrictive assumptions of (1) Gaussian relationships between response and predictor variables; (2) uniform effects of predictors and their interactions on the response over their range of values; and (3) constant interactions among predictors over their range of values (Muñoz and Felicísimo 2004). Regression trees present several advantages over linear discriminant analysis and multiple regression. They can capture nonlinear and non-additive behaviour, as well as general interactions among predictors (Breiman et al. 1984), such as relationships between a response variable and certain predictors that are conditional on the values of other predictors. Regression trees can also accommodate both continuous and categorical predictor variables, without transformation.

CART models are built in two stages: a forward recursive algorithm for “growing” the tree, followed by a second stage in which the tree is “pruned back”. In the first stage, the data are progressively divided into ever more homogeneous subsets, until all the variation is explained and it is impossible to continue. A univariate, binary partitioning algorithm (Breiman et al. 1984) is used to “grow” the CART tree. At each step, the algorithm finds a potential predictor variable and a particular value of this predictor to split the total sample (or subsets of the total sample resulting from partitions at previous steps) into two groups. The predictor and its “split point” value are chosen to maximize the resulting reduction in misclassification error for classification trees and in variance for regression trees. Partitioning continues until a preset minimum number of observations or a predefined and acceptably low level of overall deviance (e.g., heterogeneity in the values of the response variable) has been reached within the node at the end of each branch of the tree. Such nodes are known as leaves. In each leaf, the classification tree estimates the probability of class membership, and the regression tree estimates the mean value of the response variable.

As a CART model grows, observations are partitioned into ever smaller subgroups of increasing homogeneity. When these subgroups become very small, the tree tends to describe peculiarities of the sample data, rather than generally valid relationships. To limit such over-fitting, a cross-validation procedure is used to prune and eliminate unreliable partitions (Breiman et al. 1984). For each branch, the cross-validation algorithm randomly separates the original data set into 10 mutually exclusive subsets and then uses each subset once to independently test subtrees grown on the 9 remaining subsets. The mean square error of prediction, averaged over the 10 runs, is used to assess the quality of each tested subtree. The subtree structure that minimizes the averaged mean square error of prediction is retained in the final CART model. Finally, the residuals are analyzed to check that they satisfy the criteria (Draper and Smith 1981) for ordinary least squares regression. This is likely unnecessarily strict, because CART models are typically more robust in dealing with poorly behaved data (from a statistical standpoint) than is standard regres-
Dealing with multicollinearity

In linear or nonlinear regressions, correlated variables included in the model often produce misleading coefficients (Draper and Smith 1981). CART analysis, on the other hand, allows only one of any set of correlated variables to enter the model at any given split. Although this prevents the model from including correlated variables at the same split, multicollinearity renders the model unstable by creating opportunities for alternative splits that perform nearly as well. Consequently, the variable involved in a split might change with even a small change in the data. The predictive ability of the model, however, is not affected by multicollinearity among predictors.

One common approach to dealing with multicollinearity is to drop collinear variables from the analysis (Legendre and Legendre 1998), but in CART this approach reduces efficiency in finding the best predictor at each split. Moreover, as CART splits the data into progressively smaller groups in fitting the tree, the collinearity among predictor variables may change, and variables that were highly correlated over all the data may be less so in subsets of the data.

Instead of dropping collinear variables, we tested the importance of each variable with the random Forest procedure (Breiman 2001; Liaw and Wiener 2002). In this procedure, CART models were constructed by using a different random subsample of the originally selected 600 pixels used to build the models; and by finding the best split among the randomly chosen subset of the predictors, at each node. We constructed 200 trees for each of the two analyses, with 10 predictors randomly chosen at each split. The importance of each predictor was estimated as the mean decrease in accuracy in the test sample (the 10% of the data held back for testing the tree) when data for only that predictor are per-
The importance of each variable in predicting the presence or absence of defoliation. Importance is measured by the mean percentage increase in misclassification rate (or mean decrease in accuracy) when the values of each variable are permuted. If a variable is important for the model, the permutation of its values will decrease the accuracy of the model considerably.

Effects of spatial correlation

When data are spatially autocorrelated, the assumption of independent observations is violated, and this leads to potential bias when the analyst is statistically testing for explicit causal factors (Legendre 1993). Although CART models are non-parametric and robust (Diniz-Filho et al. 2003), we assessed both the raw data and the residuals from the regression tree for spatial autocorrelation. Our goal in testing for spatial autocorrelation was not to remove it, as would be necessary in standard regression, but rather to determine the scale of spatial autocorrelation and hence possibly gain insight into its cause. Moran’s I coefficient (Cliff and Ord 1981) was used to calculate correlograms of the spatial autocorrelation for lag distances up to 100 km, both for the residuals of the regression tree and for the observed defoliation frequency. The statistical significance of the coefficient of spatial autocorrelation was calculated by using a permutation test (Lichstein et al. 2002) with 999 permutations.

Results

As expected, defoliation frequencies observed in the period 1967–1998 exhibited a spatial pattern (Fig. 1) broadly similar to that reported by Candau et al. (1998) for 1941–1996. The total area defoliated at least once (392 310 km² from the period 1967 to 1998; 413 000 km² from the period 1941 to 1996) covers a continuous east–west “defoliation belt” consisting of three zones running from east to west. In each zone, areas of high defoliation frequencies (hot spots) are surrounded by gradients of decreasing frequencies.

Presence or absence of defoliation

The classification tree for the presence or absence of defoliation (Fig. 2) has four leaves and an overall misclassification error rate of 15%. Of 31 potential predictors, only 3 were retained: the percentage of FbSw, the maximum winter temperature (TmaxWinter), and the minimum May temperature (MINT5). The analysis showed that defoliation tended to occur where species FbSw occurred over >4.6% of the basal area. Defoliation was rare where these host species accounted for <4.6% of the basal area, unless the climate was fairly mild (i.e., TmaxWinter > –11.3 °C and MINT5 < 4.4 °C).

A map of host species distribution and climate isolines (Fig. 3) reveals that the area where primary host species (i.e., FbSw) account for >4.6% of the basal area is coincident with much of the range of content (see Fig. 1). Climate isotherms (TmaxWinter = –11.3 °C and MINT5 = 4.4 °C) refine the northern and southern boundaries of defoliation, respectively.

Figure 4 confirms the importance of both TmaxWinter and the prevalence of the spruce budworm’s primary host species, FbSw. Minimum winter temperature (TminWinter) and the minimum spring temperature (TminSpring) also appear to be important. However, MINT5, which was selected in the particular classification tree shown in Fig. 2, does not appear to be generally important.

Defoliation frequency

The final regression tree for describing spatial variation in defoliation frequency has 11 leaves (Fig. 5). Out of 31 potential predictors, only 8 were retained in the construction of the tree. Four partitions are based on values of each winter temperature variable (TminWinter and TmaxWinter), and one is based on values of each forest composition variable (FbSw and FbSwSb). The remaining partitions involve the minimum August temperature (MINT8), maximum May temperature (MAXT5), precipitation in June (PCP6), and TminSpring. The lowest mean defoliation frequency (i.e., 2.1 years) was associated with the lowest winter temperatures (TminWinter < –23.3 °C and TmaxWinter < –11.0 °C). Mean defoliation frequencies of 8.2, 8.8, and 11.0 years were observed where moderate winter conditions (TminWinter > –23.3 °C and TmaxWinter > –11.6 °C) prevailed and where spruce budworm host trees FbSwSb accounted for >14.3% of the basal area.

The leftmost branch of the regression tree, for which TminWinter < –23.3 °C, corresponds to the area of low defoliation frequencies located north of isotherm TminWinter = –23.3 °C (Fig. 6) and up to the northern limit of defoliation (see Fig. 1). The other main branch, for which TminWinter > –23.3 °C, corresponds to the region of defoliation (Fig. 1) lying south of the isotherm for TminWinter = –23.3 °C (Fig. 6).

Progressing from left to right through the right-hand main branch of the regression tree (see Fig. 5), we come to the leaf for which FbSwSb < 14.3%, which relates to the southernmost area of low defoliation frequencies. (The estimated mean frequency for this area is 5.9 years.) The leaf for which TmaxWinter < –11.6 °C is associated with an area.
of low frequencies (mean = 3.3 years) in the northern part of the western zone (Figs. 1, 6). The subtree defined by TminWinter < –22.2 °C corresponds to an east–west band of low to intermediate defoliation frequencies, located south of the area of low frequencies that was defined by the first (TminWinter < –23.2 °C) subtree. The leaf for which MAXT5 > 17.9 °C relates to an area of lower frequencies, located in the southwestern corner of the western zone. Finally, the rest of the tree (i.e., MAXT5 < 17.9 °C) corresponds to areas of intermediate to high defoliation frequencies, located in the centre of the defoliation belt. For instance, areas of high defoliation frequencies (mean = 11.0 years) in the eastern zone (e.g., point A in Fig. 6) are defined by PCP6 < 86.0 mm and TminSpring < –2.7 °C. The complexity of the PCP6 = 86.0 mm isoline is due to the interaction of continentality, strong lake effects, and topography on precipitation patterns, particularly in the central and southern parts of the province (Baldwin et al. 2000).

The cross-validated estimate of the proportion of variance explained was 0.57. This last measure is roughly equivalent to the $R^2$ of a standard regression model when it is tested on data totally independent of the data used to build the model. (The $R^2$ typically reported in standard regression measures how well a model fits the same data that were used to build the model and, hence, is a much less severe and realistic test of predictive ability.)

As a final test of the regression tree (see Fig. 5), it was run on the entire grid from which the 600 sample pixels were drawn, to estimate the defoliation frequency for each pixel. Comparing the spatial pattern of the resulting map of predicted defoliation (Fig. 7) with that of the corresponding map of observed defoliation (see Fig. 1) shows that the two maps appear broadly similar: areas of higher defoliation frequency are centrally located and surrounded by areas of lower frequency.

The main differences (Table 2) occur (1) in areas where 10–12 years of defoliation were observed, especially in the western zone; and (2) in the failure to predict any of the areas where defoliation was most frequent (13–19 years). These differences are a direct consequence of the structure of the regression tree: the highest value the tree can predict is 11.0 years (see Fig. 5, leaf at bottom right), so no value from the highest class can be predicted. This leaf is also the only one predicting a value for the 10- to 12-year defoliation frequency class.
and it occurred in only $N = 49$ (i.e., 8.2%) of the 600 sample pixels used to build this tree. This percentage is consistent with the total predicted area of this defoliation class (9.8%), but it underestimates the observed area (24.1%) (Table 2).

High defoliation frequencies were observed in pockets scattered throughout the centre of the region of defoliation. Because these pockets had little latitudinal gradient (see Fig. 1), most bioclimatic variables were of little use in locating them. These relatively small pockets (3% of the total area of the defoliation belt) were not well explained by our forest composition and age data either.

The map of residuals (i.e., observed – predicted defoliation frequencies) calculated for each pixel (Fig. 8) confirms that the model underestimates defoliation frequency in frequently defoliated areas. The map also reveals that defoliation frequency is overestimated in (1) the corridor between the western and central zones; (2) the southern part of the defoliation belt; and (3) some areas in the northwest. Overall, the distribution of differences between observed and predicted values shows that in 80% of the pixels, the error in estimating defoliation frequencies falls between −3.1 and 2.8 years. This amounts to an error rate ((observed – predicted frequencies) / (observed frequency)) of <10% in these pixels during the 32-year period 1967–1998. The distribution of residuals (not shown), plotted against the quartiles of the standard normal distribution, revealed a slight over-dispersion, but no outliers.

Figure 9 confirms the importance of winter temperatures ($\text{TminWinter}$ and $\text{TmaxWinter}$), spring temperatures ($\text{TminSpring}$), and primary host species content ($\text{FbSw}$). Other variables used in the regression tree, such as $\text{MAXT5}$, $\text{MINT8}$, and $\text{PCP6}$, have lower importance, indicating that they could be over-fitting some peculiarities of the sample set. All variables excluded by the regression tree have low importance.

Defoliation frequency exhibits a spatial autocorrelation that is strong and statistically significant ($\alpha = 0.01$), up to the maximum distance for which it can be estimated (Table 3). This is expected from the spatial patterns of defoliation (Fig. 1) (Candau et al. 1998). Residuals from the regression tree...
show statistically significant spatial autocorrelation for only relatively short (<30 km) distances.

**Discussion**

This paper focuses on the cumulative, large-scale patterns of defoliation from the last spruce budworm outbreak in Ontario. We do not deal with the temporal dynamics inherent in this outbreak. In addition, because defoliation is a consequence of the physiological interaction between host trees and insect population density, the level of defoliation may not always accurately reflect population density. In fact, the relationship between defoliation and spruce budworm population density is nonlinear (Fleming and van Frankenhuyzen 1992) and varies with host species content (MacKinnon and MacLean 2003; Nealis and Régnière 2004). Thus, precise inference of population density from defoliation is difficult. At best, defoliation data can be considered a qualitative proxy for spruce budworm abundance, especially at high densities. On the other hand, it is defoliation, not insect population density, that impacts the forest and therefore is of immediate concern to forest managers.

One possible concern with our data is the degree to which large-scale insect control efforts may have distorted relationships between defoliation and bioclimatic variables. There is reason to suspect that, as Hardy et al. (1986) imply, spraying generally has had little effect on the large-scale patterns of spruce budworm defoliation in Ontario. The largest spray program occurred in 1986 (Howse et al. 1995), but this covered only 1.7% of the 8.75 million ha of moderately and severely defoliated forest that year (Howse 1995). Furthermore, Lysyk’s (1990, p. 261) analysis of Ontario survey data from 1968 to 1988 indicated that “sprays … overall, did not confer a very high level of foliage protection.” Similar results have been reported for some other jurisdictions (e.g., Fleming et al. 1984).

Certain general characteristics of CART models proved useful in interpreting our results. For instance, when CART models are applied to spatial data, the predictor variables that operate at large spatial scales are generally used early in the model, while predictor variables that operate at smaller scales are typically used near the leaves (Moore et al. 1991). Accordingly, our results suggest that among the predictor variables we considered, it was (1) the contribution of primary

---

**Fig. 7.** Defoliation frequency predicted by the regression tree (Fig. 5). This map was obtained by predicting the frequency for each pixel of a grid at 1-km² resolution. To facilitate the comparison, classes are identical to those in Fig. 1.
host species to forest species composition and (2) the minimum winter temperature that acted over the largest spatial scales in influencing (1) where spruce budworm defoliation occurred and (2) where it occurred at different frequencies, respectively.

**Classification tree of the occurrence of defoliation**

Our classification tree (Fig. 2) suggests that the primary determinant (first split) of where aerially detectable defoliation occurred in Ontario during the 1967–1998 outbreak was host species composition. Specifically, defoliation occurred in 94% of 316 sampled sites where the percentage of FbSw basal area exceeded 4.6%. Although the presence of host species is an obvious requirement for an outbreak to develop, it is not necessarily the primary limiting factor. For example, it has been reported by Blais (1968) for the Atlantic region and Volney and Fleming (2000) for the north that within the spruce budworm range, harsh climate can prevent the development of outbreaks, even where host species are plentiful. In northern Ontario, the limit of defoliation is located >200 km south of the limit of FbSw distribution (Frank 1990; Nienstaedt and Zasada 1990).

Where FbSw constitute <4.6% of the forest basal area, the situation is more complex. In the north, where the average TmaxWinter is below −11.3 °C, defoliation was detected in only 8% of 128 sampled sites. In the south, where the average MINT5 is above 4.4 °C, defoliation was detected in only 1% of 74 sample sites. Between these climatic boundaries, however, defoliation was detected in 96% of 82 sampled sites (even though FbSw constituted <4.6% of the forest basal area). This last result leads one to question what Williams and Liebhold (1997) acknowledged as the critical assumption in their work: that spruce budworm defoliation occurs only within the spruce–fir forest type. Our results suggest that spruce budworm defoliation is aerially detectable even in areas where spruce and fir are not the major components of the forest.

Using a different approach and spruce budworm defoliation data between 1938 and 1980 for Ontario, Quebec, and the Maritime Provinces, Hardy et al. (1986) also delimited a central area of defoliation, with isotherms (zone B, Fig. 3); and the northern and southern limits of defoliation, with vegetation data (zones A and C, Fig. 3). These authors expected
most outbreaks to occur in their central area of defoliation. However, the northern limit of this central area (defined by an isotherm) is clearly located south of the northern limit of the isotherm (TmaxWinter = –11.3 °C) found by our classification model. The southern limit of zone B matches the southern limit defined by the classification tree.

Regression tree of defoliation frequency

In the regression tree of defoliation frequency (Fig. 5), the first four levels of partition are based on the percentage of host tree species and on winter temperatures. Partitions related to low winter temperatures correspond to areas of low defoliation frequencies located in the northern part of the defoliation belt. Hence, low winter temperatures may seem to be an important limiting factor on second instar survival (and consequently on defoliation) in the north (where these lows are most extreme). This appears to contradict earlier reports that cold temperatures do not kill large numbers of overwintering larvae, since in the field they can survive prolonged periods of temperatures of –40 °C or below (Blais 1958). In fact, range maps of spruce budworm outbreaks show that survival is possible as far north as Norman Wells (65°18′N, 125°30′W) in the Northwest Territories (Shepherd et al. 1995), which is far higher in latitude (and under a harsher winter climate) than northern Ontario. However, temperature requirements for the insect to complete its development may vary along its range (Volney and Cerezke 1992).

Winter temperatures may also interact with host species composition through the synchrony of the insect with its hosts (Régnière 1982). Low defoliation frequencies in the northern part of the defoliation belt may also reflect a distinct geographic cline in fecundity explained by differences in egg weight, with larger eggs (lower fecundity) in the north (Harvey 1983). This cline may itself reflect the harshness of winter conditions in the northern latitudes, as it has been shown, under laboratory conditions, that survival of overwintering larvae from larger eggs is higher than survival of larvae from small eggs (Harvey 1985).

Most of the larval feeding stages occur in June, so weather conditions during this month could be critical. Laboratory and field investigations in Ontario have shown that under wet conditions larvae do not feed readily and remain relatively inactive (Wellington 1949a, 1949b, 1950). Consequently, development rates decrease with increased PCP6 (Greenbank 1956). A temporal association between drier conditions in June and the development of spruce budworm outbreaks has been suggested in the past (Wellington et al. 1950; Greenbank 1956; Pilon and Blais 1961). Our results show that the same association seems to hold spatially: high defoliation frequencies in the northern part of the defoliation belt may also reflect a dispersal pattern that cold temperatures do not kill large numbers of overwintering larvae, since in the field they can survive prolonged periods of temperatures of –40 °C or below (Blais 1958). In fact, range maps of spruce budworm outbreaks show that survival is possible as far north as Norman Wells (65°18′N, 125°30′W) in the Northwest Territories (Shepherd et al. 1995), which is far higher in latitude (and under a harsher winter climate) than northern Ontario. However, temperature requirements for the insect to complete its development may vary along its range (Volney and Cerezke 1992).

Winter temperatures may also interact with host species composition through the synchrony of the insect with its hosts (Régnière 1982). Low defoliation frequencies in the northern part of the defoliation belt may also reflect a distinct geographic cline in fecundity explained by differences in egg weight, with larger eggs (lower fecundity) in the north (Harvey 1983). This cline may itself reflect the harshness of winter conditions in the northern latitudes, as it has been shown, under laboratory conditions, that survival of overwintering larvae from larger eggs is higher than survival of larvae from small eggs (Harvey 1985).

Most of the larval feeding stages occur in June, so weather conditions during this month could be critical. Laboratory and field investigations in Ontario have shown that under wet conditions larvae do not feed readily and remain relatively inactive (Wellington 1949a, 1949b, 1950). Consequently, development rates decrease with increased PCP6 (Greenbank 1956). A temporal association between drier conditions in June and the development of spruce budworm outbreaks has been suggested in the past (Wellington et al. 1950; Greenbank 1956; Pilon and Blais 1961). Our results show that the same association seems to hold spatially: high defoliation frequencies are associated with low precipitation in June (PCP6 < 86 mm).

High defoliation frequencies also appear to be associated with a cool spring (TminSpring < –2.7 °C) followed by dry conditions in the summer (PCP6 < 86 mm). Conversely, low defoliation frequencies are found where MAXT5 > 17.9 °C. Spring climate has potential effects on larval emergence, survival, feeding activity, and dispersal (Miller 1958). The relation between emergence dates and the temperature regime in late April and early May has been established (Rose and Blais 1954). After emergence, however, the effects of

Table 2. Percentage of total area affected (392 310 km²) during the last spruce budworm outbreak in Ontario that experienced various frequencies of aerially detected defoliation, according to observation (Fig. 1) and regression tree prediction (Fig. 7).

<table>
<thead>
<tr>
<th>Defoliation frequency (years)</th>
<th>Observed (%)</th>
<th>Predicted (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–3</td>
<td>16.8</td>
<td>11.8</td>
</tr>
<tr>
<td>4–6</td>
<td>23.9</td>
<td>28.7</td>
</tr>
<tr>
<td>7–9</td>
<td>32.2</td>
<td>49.7</td>
</tr>
<tr>
<td>10–12</td>
<td>24.1</td>
<td>9.8</td>
</tr>
<tr>
<td>13–19</td>
<td>3.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Table 3. Moran’s I coefficient of spatial autocorrelation for the defoliation frequency observed between 1967 and 1998 and for the residuals of the regression tree model for 10-km distance classes, calculated on a random sample of 600 points.

<table>
<thead>
<tr>
<th>Distance class</th>
<th>Moran’s I</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–10</td>
<td>0.63**</td>
</tr>
<tr>
<td>10–20</td>
<td>0.73**</td>
</tr>
<tr>
<td>20–30</td>
<td>0.55**</td>
</tr>
<tr>
<td>30–40</td>
<td>0.44**</td>
</tr>
<tr>
<td>40–50</td>
<td>0.36**</td>
</tr>
<tr>
<td>50–60</td>
<td>0.32**</td>
</tr>
<tr>
<td>60–70</td>
<td>0.31**</td>
</tr>
<tr>
<td>70–80</td>
<td>0.23**</td>
</tr>
<tr>
<td>80–90</td>
<td>0.20**</td>
</tr>
<tr>
<td>90–100</td>
<td>0.17**</td>
</tr>
</tbody>
</table>

Note: **, statistical significance at α = 0.01.

© 2005 NRC Canada
temperature and precipitation on survival are difficult to assess because of population losses and movement associated with the dispersal of second-instar larvae on air currents. Low defoliation frequencies in areas of higher spring temperatures could be explained if the spruce budworm populations, which emerged early (likely as a result of warm weather), later experienced heavy mortality because of (1) the vulnerability of the larvae at this stage to late frosts (Cerezke and Volney 1995) or (2) faster insect development at a time when trees are held back by frozen ground.

Forest age did not appear as an important factor in explaining the patterns of defoliation frequency, although some (e.g., Blais 1961, 1968; Mott 1963) literature suggests that older stands are more susceptible to spruce budworm. Spatial patterns of forest age-classes may explain in part why age was not retained in the regression tree: forest age is influenced by disturbances operating at different spatial scales (e.g., clearcuts, fires, insects), creating complex spatial patterns that could not explain the broad patterns of spruce budworm defoliation. Moreover, we used 20-year age-classes and 32 years of defoliation data, so stands moved between age-classes over the course of data collection.

We analyzed the frequency distribution of the residuals (not shown) to validate the results of the regression tree. The dispersion and the variance of the residuals were minimal for the lowest and the highest predicted values. A likely explanation can be found in the spatial distribution of the defoliation frequencies in comparison with the spatial distribution of the climate (particularly temperature) variables. Because the thresholds related to temperature have a latitudinal pattern (Fig. 6), defoliation frequencies that have a similar pattern will likely be better captured by the regression tree and, therefore, exhibit less variation in their residuals. This is the case for the lowest defoliation frequencies, and to a lesser extent, for frequencies in the range of 10–12 years of defoliation. In contrast, the spatial patterns of intermediate defoliation frequencies (from 4 to 9 years) are complex and do not coincide as well with the patterns of our climate variables.

That the regression tree variables (i.e., the bioclimatic variables and forest composition) account for the long-distance spatial autocorrelation observed in defoliation frequency is indicated by the absence of significant long-distance (>30 km) spatial autocorrelation in the residuals (Table 3). However, the regression tree variables cannot explain the short-distance (<30 km) spatial autocorrelation. This threshold is comparable to the 36-km mean distance within which spatial autocorrelation occurred in variogram analyses of population data in New Brunswick (Fleming et al. 1999). Possible explanations for the short-distance spatial autocorrelation include (1) local population processes (e.g., density dependence); and (2) site characteristics (e.g., stand age). No data on spruce budworm local population fluctuations were available at the scale of our study.

Conclusion

A strength of our analysis is that bioclimatic variables were limited to those with clear biological relevance. We think that this approach allows stronger inferences than studies that screen for correlations between observed distributions and numerous bioclimatic variables with unspecified biological effects. We also chose a modelling technique (i.e., CART) known for its power and robustness (Breiman et al. 1984). This technique revealed and accounted for a number of nonlinear relationships and complex, conditional, interactions among variables, which would have been very difficult to handle with the usual linear approaches. We conducted two independent assessments of the reliability of the model (i.e., cross-validation and comparison of model output with observed values of data independent from those used to build the model).

Finally, our study relies on a static view (cumulative defoliation frequency) of a dynamic process (spatiotemporal oscillations of the spruce budworm ecosystem). The CART models describe relations between the occurrence and the frequency of moderate to severe defoliation averaged over 32 years (1967–1998), climatic conditions averaged over 30 years (1961–1990), and average forest composition. Therefore, thresholds identified by the models indicate average environmental conditions that affect spruce budworm populations directly through survival and reproduction or indirectly through a component of the budworm ecosystem: forest composition, natural enemies, or competitors (Fleming 1996). However, when average bioclimatic conditions are already close to the limit of population maintenance, the frequency and severity of catastrophic deviations can be more important than the average conditions in explaining the defoliation frequency (Pilon and Blais 1961). The development of dynamic, spatially explicit models to simulate the dynamics of spruce budworm disturbance regimes will allow us to test this hypothesis.

Factors reflecting climate and forest characteristics were able to describe many, but not all, aspects of the large-scale spatial patterns in defoliation. For instance, at one extreme, the lowest defoliation frequencies have broad latitudinal patterns (Fig. 1). These patterns are well described by climatic factors, such as those represented by isotherms (Figs. 3, 6) that are also oriented latitudinally. We have suggested biological reasons why climate may largely determine where these low frequencies occur. At the other extreme, the complex, spotty patterns of the highest defoliation frequencies (Fig. 1) coincide with patterns generated by neither our forest nor our climatic variables. Consequently, the highest defoliation frequencies that we observed are not predicted by the regression tree (Fig. 5). These results may explain why the regression tree was able to account for the spatial autocorrelation observed in defoliation frequency at all but the shortest (<30 km) distances. At these relatively short distances, site factors (e.g., stand age) and population processes (e.g., dispersal) may interact in ways that cannot be described by factors reflecting climate and forest characteristics alone.

In general, for describing the large-scale spatial patterns in defoliation frequency, the most important forest characteristics were reflected in species composition. Key climate variables were reflected in winter and spring temperatures and precipitation in June. We found forest age was not important, largely because it varied over much finer spatial scales.

Acknowledgments

We thank A. Hopkin, R. Fournier, G. Howse, and B. Sajan (all at CFS) for their help in providing, verifying, and inter-
preparing insect records. We also thank D. McKenney (CFS) for providing historical climate data; and L. Watkins (Ontario Ministry of Natural Resources) for the vegetation data. Tim Burns and Lorne Keatley provided expertise with the geographic information system and assistance with the research.

References

Balch, R.E. 1946. The spruce budworm and forest management in the Maritime Provinces. Canada Department of Agriculture, Entomology Division, Ottawa, Ont. Publ. 60.


© 2005 NRC Canada


