

Modeling the mountain pine beetle epidemic using a probabilistic clustering approach

The University of British Columbia iGEM 2011 Team

Model Creators: Jacob Toth, Joe Ho, Samuel Wu

Model Advisor: Shing H. Zhan

Coder: Samuel Wu

Objective

To predict the impact of the mountain pine beetle (MPB) epidemic on the lodgepole pine trees in British Columbia (BC), we developed a mathematical model that describes the expansion of the MPB population in BC. Using the model, we simulated the spread of MPB population from 2011 to 2020. The model predicted key locations in BC where we can strategically install the iSynthase “TrapBox” product to effectively control the MPB epidemic. Lastly, we propose a formula to estimate the budget for implementing a control strategy.

Model Description

A MPB population of size N is described as a collection of n subpopulations. At year t , subpopulation i is centered at $c_t^i = (x_t^i, y_t^i)$ and is normally distributed along the West-East axis (x) with the standard deviation $\sigma_x^{i,t}$ and along the North-South axis (y) with $\sigma_y^{i,t}$. The *subpopulation expansion rate* is defined as the rate of change in the mean subpopulation standard deviation. Two subpopulation expansion rates are defined to reflect two-dimensional expansion: one for the West-East axis (s_x) and the other for the North-South axis (s_y). Suppose that subpopulations expand linearly with time, we can estimate the subpopulation expansion rate along each axis using standard linear regression. The slope is interpreted as the subpopulation expansion rate. Thus, we can conveniently extrapolate the standard deviation of subpopulation i at time $t+k$ along each axis:

$$\sigma_x^{i,t+k} = \sigma_x^{i,t} + s_x k \quad (1)$$

$$\sigma_y^{i,t+k} = \sigma_y^{i,t} + s_y k \quad (2)$$

The growth of the MPB population is assumed to follow a linear model. Then, the growth rate (v) is simply the slope of the regression equation that predicts MPB population size over time.

The above parameters can be estimated from available Geographical Information System (GIS) data that indicate the distribution of MPB across BC. Below, we describe the statistical methods used to estimate those parameters.

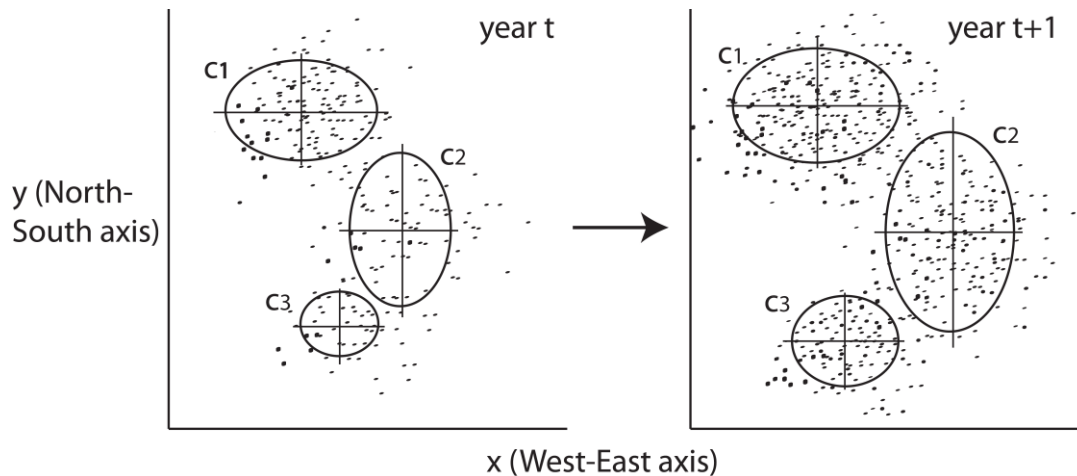


Figure 1: The mountain pine beetle population is distributed along the North-South axis and the West-East axis. The subpopulations c_1 , c_2 , and c_3 are shown to expand along both dimensions from year t to year $t+1$.

Materials and Methods

Converting GIS data to Cartesian coordinates

We obtained GIS data showing the distribution of the MPB population in BC between 1999 and 2010 from the BC Ministry of Forests, Lands and Natural Resource Operations website (1). We converted the GIS data into raster images at a resolution of 60 dpi (jpeg format) using the commercial ArcGIS software (2). The raster images were converted into Cartesian coordinates using the utilities in the Python Image Library (3).

Probabilistic clustering and parameter estimation

Using a model-based clustering approach, we identified clusters of coordinates (interpreted as subpopulations) that fit two-dimensional normal distributions whose standard deviations are allowed to vary. Each data set contains an optimal number of such clusters, which comprise a normal mixture model. To find the optimal number of clusters, we fitted the data set of year 1999 to a range of possibilities (from 1 to 50). The optimal cluster number was selected according to the Bayesian Information Criterion (4). The number of subpopulations is remained fixed in the current model. The mean and covariance matrix of each cluster in the mixture model having the best-fitting cluster number were optimized. Using the parameter values from the optimized mixture model, we estimated the subpopulation expansion rates and the population growth rate using linear regression.

Simulation

We simulated the expansion of the MPB population from year 2011 to 2020 using the estimates obtained from the clustering analysis. Taking year 2010 as the initial point, the standard deviation along the West-East axis and that along the North-South axis were extrapolated for each subpopulation using Equation 1 and Equation 2, respectively. The subpopulations inferred in the year 2010 data set are assumed to be stationary. A pattern of MPB population distribution was predicted for each year from 2011 to 2020, and then overlaid onto

the pattern observed in year 2010. We superimposed the predicted MPB population distribution on a 1-to-20,000 scale map of BC (available at <http://gis2.forestry.ubc.ca/BCDD.html>).

Model implementation

All statistical analysis was conducted in the R computing environment (5). The normal mixture modeling and simulations were done using the package mclust (6). The clusters were modeled as ellipsoidal structures that are normally distributed along its major and minor axes (using the “VEV” covariance option in mclust).

Derivation of the cost formula

Supposed that, at year t , we design a strategy to control the MPB epidemic by year t' . The premise of the strategy is to surround each of the constituent subpopulations of the MPB population with a product that hampers MPB expansion. Here, we assume that the product (which is the iSynthase “TrapBox” in the context of the UBC iGEM 2011 project) is completely effective against MPB spread. First, using the current model, we predict the MPB epidemic situation in year t' . Then, utilizing the theoretical knowledge gained from the prediction, we estimate the cost required to implement the strategy with a cost model proposed below.

The cost of the strategy is determined by the duration of product installation and by the amount of the product needed to contain the MPB population. For subpopulation i , the product is installed along the circumference of the ellipsoid structure that encompasses 99.7% of the subpopulation. Thus, the major radius, a_i , and the minor radius, b_i , of the ellipsoid are related to the standard deviations of the distribution:

$$a_i = \max(3\sigma_x^{i,t'}, 3\sigma_y^{i,t'}) \quad (3)$$

$$b_i = \min(3\sigma_x^{i,t'}, 3\sigma_y^{i,t'}) \quad (4)$$

The circumference of the ellipsoid structure (S) is a function of the two radii:

$$S(a_i, b_i) = 4a_i E\left(\sqrt{1 - \left(\frac{b_i}{a_i}\right)^2}\right) \quad (5)$$

where $E(p)$ is the complete elliptic integral of the second kind:

$$E(p) = \int_0^1 \frac{\sqrt{1 - p^2 z^2}}{\sqrt{1 - z^2}} dz \quad (6)$$

The duration of product installation (F_i) for subpopulation i is defined as:

$$F_i = \max\left(\frac{\sigma_x^{i,t'} - \sigma_x^{i,t}}{s_x}, \frac{\sigma_y^{i,t'} - \sigma_y^{i,t}}{s_y}\right) + \varphi \quad (7)$$

This estimates the number of years it takes for the subpopulation to reach the product and cease expanding. φ denotes the additional number of years to continually apply the product to ensure that the MPB subpopulation is controlled. Finally, the cost (C) is a function of the duration of product installation and the amount of product used:

$$C = R \sum_{i=1}^{n_t} (F_i \times S(a_i, b_i)) \quad (8)$$

where R is the cost of the product per unit of circumference per year. The cost estimated using this method can be incorporated into the budget for implementing a control strategy.

Model extensions

Emergence of new MPB subpopulations

The current model assumes that the number of subpopulations remains fixed. In reality, however, new MPB subpopulations can emerge. To model the origin of new subpopulations, we define a *subpopulation birth rate*. This parameter can be estimated by:

$$\beta = \frac{1}{T} \sum_{t=1}^T (n_t - n_{t-1}) \quad (9)$$

where T is the total number of years for which MPB distribution data is available (assuming that the years are uniformly spaced). A new subpopulation is assumed to arise from an existing subpopulation. Thus, the initial center of a new subpopulation is expected to be located closely to an existing one. Using this assumption, the initial center of a new subpopulation can be simulated in two steps: 1) randomly select an established subpopulation and 2) then randomly draw a point from a bimodal distribution composed of two evenly massed normal distributions distanced from the center of the parent subpopulation by twice the standard deviation of the parent subpopulation.

Dynamic interactions among MPB and lodgepole pine trees

The parasite-host relationship between MPB and lodgepole pine trees can be modeled using fundamental ecological principles. First, we can describe the population distribution of the lodgepole pine trees in BC using the current model for MPB. The current model should be able to capture the contraction of lodgepole pine tree subpopulations as a direct consequence of the expansion of MPB subpopulations. Second, we can model the probabilistic interaction between the lodgepole pine tree and MPB subpopulations. Overlapping lodgepole pine tree and MPB subpopulations represent the interactions between the two organisms. The Lotka-Volterra model can be integrated into the model to enable the probabilistic subtraction of the lodgepole pine tree subpopulations that overlap with MPB subpopulations.

Effectiveness of the iSynthase “TrapBox” product

While we cannot determine the impact of the iSynthase “TrapBox” product on the MPB population in the wild, we can make explicit assumptions about the effectiveness of the product.

We can parameterize the “porousness” of a iSynthase “TrapBox” barrier, modeled by ellipsoid structures. This directly reflects the effectiveness of the product in preventing the growth of MPB subpopulations surrounded by the product.

Acknowledgement

We are grateful to Dr. Jerry Maedel from the UBC Department of Forestry and Adrian Walton from the BC Ministry of Forests, Lands and Natural Resource Operations for directing us to the sources of the MPB distribution data.

References

1. British Columbia Ministry of Forests, Lands and Natural Resource Operations. <http://www.for.gov.bc.ca/hre/bcmapb/Year7.htm>
2. ArcGIS. Version 10. <http://www.esri.com/software/arcgis/index.html>
3. Python Image Library. Version 1.1.7. <http://www.pythonware.com/products/pil/>
4. Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics*. 6:461-4.
5. R Development Core Team. 2011. R: a language and environment for statistical computing. <http://www.R-project.org>
6. Fraley C, Raftery A. MCLUST version 3 for R: normal mixture modeling and model-based clustering. <http://cran.r-project.org/web/packages/mclust/index.html>