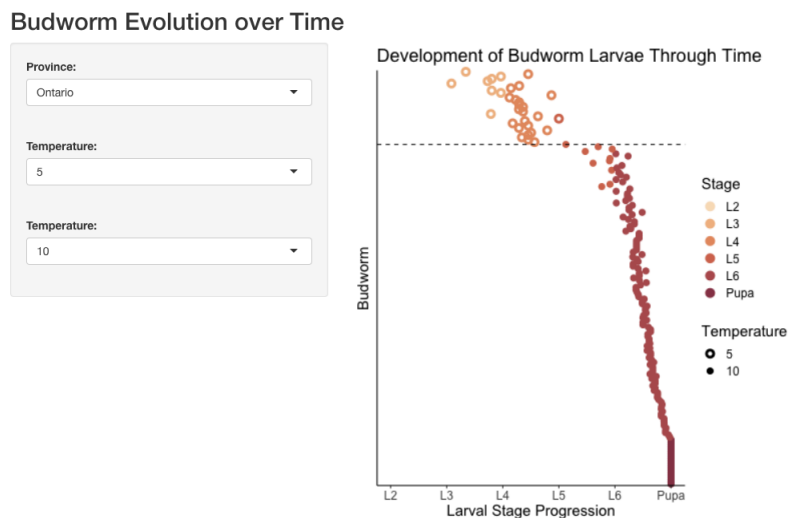# Modelling the Life Cycle of the Spruce Budworm

Martin Stelmach and Kala Studens

## 1 The Story

1.1 Plot 1: Previewing the Data in Shiny

The first graphic that we created was a Shiny app for the user to do some initial data exploration. Since the data consists of five different colonies and seven temperature treatments, it is useful for the user to have the option of subsetting the data and only examining specific pieces at a time. This graphic highlights some of the most important aspects of the data: the variation of individuals within one population, and the overall differences between different temperature treatments. The "Province" toggle allows the user to examine these relationships for the colony of their choosing. The animation dimension allows the user to see how the individual insects progress through their respective larval stages over time. We would like the user to understand the distribution of development rates of the insects, and the overall differences in development rates across different colonies. Upon using the toggles in this graphic and comparing temperature treatments, it becomes apparent that temperature has a strong effect on larval development. It is also clear that some individuals develop more quickly than others, and that this needs to be handled in a spruce budworm life cycle model.

Figure 1: Budworm Evolution Over Time

1.2   Plot 2: Fitting the Model

Regniere et al. [1] introduced a nonlinear model describing the relationship between temperature and development for a given stage in the spruce budworm's life cycle. They also included a parametric model describing the variation of development rates around the median, irrespective of temperature. We fit both models to the budworm lab data using nonlinear least squares; the median model was weighted by the number of observations at each temperature treatment. The variation component of the model was fit by dividing the development rate of each individual in the population by the median rate for their respective temperature treatments, and then pooling these values to fit a two-parameter distribution of variation "multipliers". More model details are included in Appendix A. The second graphic shows the results of these fits, with the observed points overlaid. Figure 2 shows this plot for the L3 development stage.

Figure 2: L3 Development Rate Curves
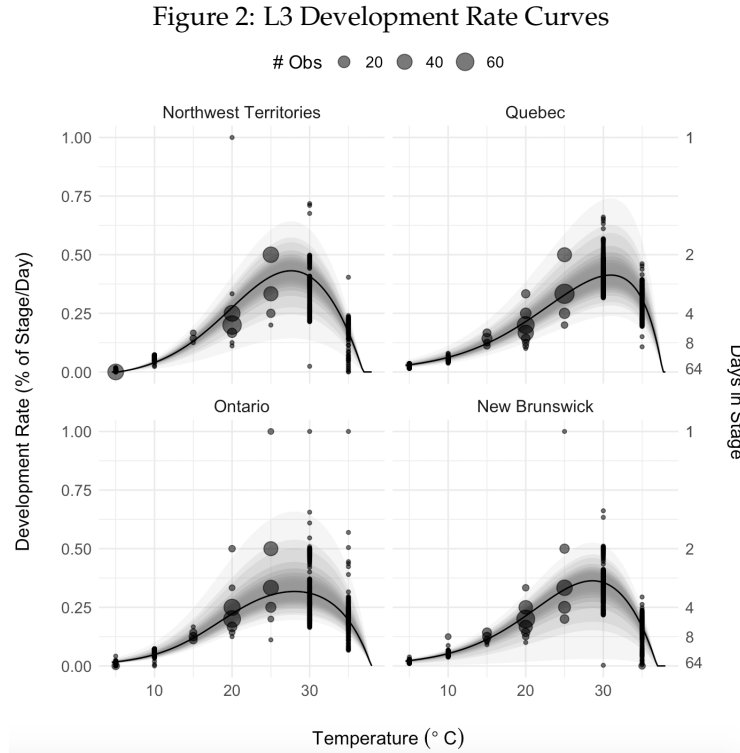
The black line is the development rate curve resulting from the point estimates for the median model. The gray ribbons with decreasing alpha values surrounding this development curve represent different quantiles of the distribution obtained using the point estimates for the variation model. Since the curve represents the 50th percentile, each level l of the ribbon represents the area between the $50 - 4l$th percentile and the $50 + 4l$th percentile of the distribution of the variability multiplier until the 2nd-98th perfcentiles. Also included are the 1st-99th percentile ribbons and the 0.1st and 99.9th percentile ribbons to show potential extreme values. On the whole, the visualization represents the information that the model retains from the data. The observed points are overlaid on top of these model estimates to give a visual representation

of the model fit, and points lying outside of the ribbons indicate some discrepancy between observed and modelled values. Part of the utility of this model is that once it is parameterized it lends itself well to simulations. These plots show the values that will be used for the simulated development, with darker alpha signifying values that are more likely.

1.3 Plot 3: Stacked Density Model Output

The final graphic is a stacked density plot that represents the output of the parameterized models. Once the parameters are obtained, the model can be used to simulate the behaviour of a population of spruce budworm from L2 to Pupa, based on a given temperature regime. The graphic shows the proportion of the population of insects in each stage at a given time. The temperature inputs for the model were taken from weather stations near each of the locations where the insects were collected. Thus, for each province facet, the graphic shows how we would have expected insects in those regions to have developed in a given year. The implementation of shiny allows the user to select a year of interest to assess the progression of each colony and to compare the colonies to each other. If the user is curious about the composition of the budworm populations on a given date, they can hover over the plot they are interested in, and a vertical line spanning all of the facets will appear in line with their cursor. A tooltip also appears, showing the province, the date and the proportion of insects in each stage on that date in the simulation for that province.

Figure 3: Stacked Density Output of Model
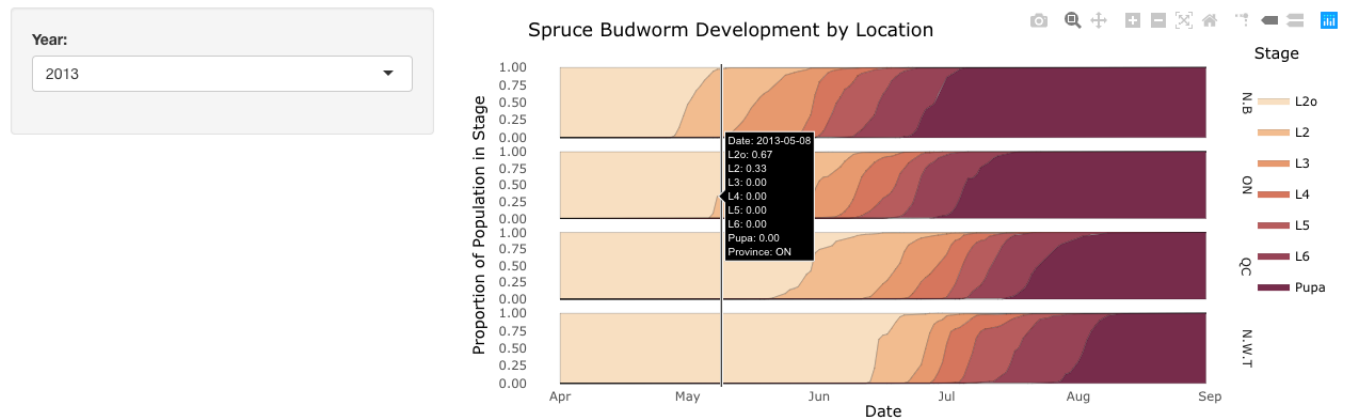
2 Techniques and considerations

2.1 Plot 1

The larval stages are differentiated using a sequential, HCL-based colour scheme. The colours begin light and become progressively darker as the larvae develop. The temperatures are differentiated by point shape,

along with a horizontal line separating individuals in different temperature treatments. The points in the graphic travel through the larval stages in such that the individuals are sorted by development speed; the individuals who complete their development first appear at the top. This allows the user to get a better sense of when the first and last points arrive to give a sense of the width of the distribution of development times within one treatment level.

The main problem with the graphic is the render time. Originally the graphic was made using `ggplotly` and the animation functioned fairly nicely with a slider showing the amount of days. The problem with this is that it became very difficult to create the type of graphic we would have liked with the interactive properties we wanted and have the animation look as nice. The problem started becoming the division of the points and the aesthetic incorporation of a more complex graphic. The points started almost running away from the rest and with a large number of data points, the whole graphic looked clusted and confusing. That being said the `ggplotly` was originally selected due to the shorter render time that `gganimate`. The latter was selected due to the nicer animations qualities and adjustments to the visualizations from a technical perspective. An effective method that decreased time significantly was, in the shiny app, rather than call the segment of data the user selected in the animation, the data was divided before hand upon user selection into a new data frame which was then used in the `gganimate` animation. Although it seems obvious, this ended up making a significant difference in render time and was one of the only ways that the animation could be sped up. Another problem with the plot was implementing a horizontal line. Because the line was created by taking the top ID of the lower set of temperature data selected, many bugs occurred if a user selected '10' and '5' instead of '5' and '10' because there would be a swap in the selection. This was fixed with an if statement that checked which data set was above and which was below, however this had to be done inside the image render instead of the global reactive environment.

2.2   Plot 2

When looking at the results of the model, it is most interesting to see the shape and amplitude of the development curve that results for each population in each stage, along with the modeled variation of individuals around the median. We are also more interested in comparing the curves for different provinces to each other than comparing different larval stages for the same province, so province was selected as the variable to facet over. We attempted to facet by both province and stage, but the image became much too condensed and difficult to read. Another important aspect of the modelling that we wanted to highlight was the variability in development of individuals within a population. This was most effectively shown using fading ribbons around the median development curve. It could have been interesting to compare different provinces on the same plot but this would have also cluttered the plotting area and clouded the more important message of variability around the curve. The alpha values associated with the ribbons decreases to a value close to zero as the quantiles move further from the median. This means that low alpha values are associated with simulation values that are less likely. The left side of the plots show the linear

development per day scale that is used to fit the model, while the right side shows the reciprocal value of days until completion of the stage. The maximum value on the plots represents a development rate of 1; that is it only takes one day for complete development. The plots are anchored at zero because zero development is a possibility, but negative development is not. The facets are roughly positioned geographically; there are not necessarily pairs of provinces that are better grouped together, so the "best" ordering of the facets is not immediately apparent. Finally, the points representing the observed values were sized by the number of observations occurring at those values. This was necessary since the data are only observed on a daily basis, leading to discrete groups of points at rates corresponding to integer days. These points were made semi-transparent to allow the viewer to see how they overlapped with one another. The legend for point size was included at the top of the graphic to create space on the right side of the plot and to allow the secondary axis title to stand out more clearly.

2.3 Plot 3

Since the output of the model exists as a matrix containing the percentage of the population in each stage of development at each day, we needed a way to convey proportions over time. It was decided that a stacked density plot with time on the x-axis would be the most effective way to convey this. The time window was trimmed based on the months when development was expected to occur. The colours used to represent stage were selected to be the same as the first graphic to maintain continuity. Unlike the previous shiny applet, this one does not have a long render time since there is no animation. Instead the graphic is made with `ggplot` and a `plotly` attribute is added to give the user precise quantitative information about what is occurring in the plot on a given date. A vertical line was added to the tooltip to provide a reference to the other colonies on the same date. The facets were ordered by development speed; it was estimated that the New Brunswick colony would complete their development first, followed by Ontario, Quebec and then the Northwest Territories. The shiny app was implemented to allow the user to toggle through years of interest.

We discovered that the `ggplotly` commands did not work well with the pre-defined titles in the `ggplot` object. Commands that grow the margins and set padding levels between the axis titles and the plot do exist, but they only work for new titles that are specified within the `ggplotly` command. This would have been good to do, but ggplotly only seemed to recognize the first facet within the plot which made the y-axis title shift to the very top of the plotting area, cutting most of the words off. It was decided that we provide the shiny app with shortened province titles and a shortened legend title to prevent the labels from running into one another. Figure 4 shows one of the original `ggplot` objects inputted into `ggplotly`, demonstrating the intended consideration for label positioning and spacing.
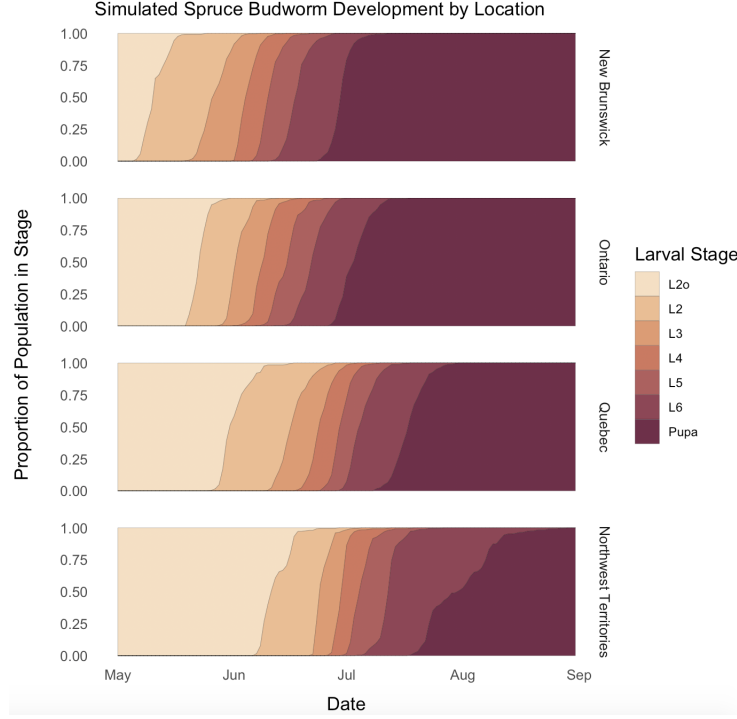
Figure 4: Static 2014 Stacked Density Plot

References

[1] Jacques Régnière, Rémi St-Amant, and Pierre Duval. Predicting insect distributions under climate change from physiological responses: spruce budworm as an example. *Biological Invasions*, 14(8):1571–1586, 2012.

A Model Details

The age a of individual i in stage j is written as $a_{ij}(t) = \delta_{ij} \int_t r_j(T_t)dt$, where $r_j(T_t)$ represents the median development rate at temperature T_t, and δ_{ij} represents an individual i's development relative to the median in stage j. The equation for the population's median development rate r_j is

$$r_j(T_t) = \begin{cases} \beta_1 \left[\frac{1}{1+e^{\beta_2 - \beta_3 \tau}} - e^{(\tau-1)/\beta_4} \right] & \text{if } T_b \leq T \leq T_m, \text{ where } \tau = \frac{T - T_b}{T_m - T_b} \\ 0 & \text{otherwise} \end{cases}$$

where T_b is the minimum temperature threshold, and T_m is the maximum. The equation for the variation component of the model is

$$\delta = 1 - \frac{1}{k} \log \left[\frac{\epsilon^{-q} - 1}{0.5^{-q} - 1} \right]$$