MODELING SEASONAL WINE GRAPE DEVELOPMENT USING A MIXTURE TECHNIQUE

William J. Price

Bahman Shafii

Paul E. Blom

Julie M. Tarara

Nick Dokoozlian

See next page for additional authors

Author Information

William J. Price, Bahman Shafii, Paul E. Blom, Julie M. Tarara, Nick Dokoozlian, and Luis J. Sanchez

MODELING SEASONAL WINE GRAPE DEVELOPMENT USING A MIXTURE TECHNIQUE

William J. Price and Bahman Shafii
Statistical Programs
College of Agriculture and Life Sciences
University of Idaho

Paul E. Blom and Julie M. Tarara
USDA-ARS
Horticultural Crops Research Unit

Nick Dokoozlian and Luis J. Sanchez
E&J Gallo Winery
Modesto, CA

Biological growth data typically display an increasing sigmoidal pattern over time. Grape development is no exception and shows a similar general trend. A detailed examination of the growth process in grapes, however, reveals a few systematic deviations from this pattern. Specifically, grape development is often characterized by localized areas of growth plateaus leading to an overall growth pattern referred to as a double sigmoidal curve. Capturing and characterizing these local changes in growth is important as they represent important phases in grape development such as veraison. This paper utilizes a model adapted from the technique of mixture models to estimate the growth curve of grapes. The resulting model provides a more accurate description of the growth process and has parameter estimates directly related to the various phases of grape development. The model is demonstrated using data collected from an experimental trellis tension monitoring system in the Chardonnay grape varietie.

I. Introduction

In the United States, grapes are cultured on approximately 940, 000 acres (NASS, 2008), of which 91% are located in California. Virtually all US vineyards use a trellising system for directing vine growth and expediting vineyard management. An experimental system for monitoring wire tension within the trellis systems has been previously developed by Tarara, et al. (2004). In that system, trellis tension is continuously detected by load cells and recorded with automated data loggers. By collating the load cell output, a database of tension-based growth curves can be developed over several seasons, thereby allowing growers and processors to develop predictions of yield as well as potentially adjusting those predictions in real time as weather and vineyard conditions change.

Seasonal growth curves for grapes are typically described as having a "double sigmoidal" shape from which multiple phases of grape development can be inferred (e.g. Lewis, 1910; Coombe, 1960). These phases include berry cell division, an intermediate lag phase, and final berry expansion and ripening. Accurate identification of these stages is important for producers

and processors as they can affect the quality and management of the grape crop. The work presented here will utilize seasonal load cell data to estimate the multiple phases of growth based on a mixture model technique. The results for the predictive grape growth models will be demonstrated using Chardonnay wine grape data collected from Modesto, CA in 2007.

II. Methods

As with many biological growth models, seasonal grape growth can be generally described as a nonlinear process:

$$y_{ij} = g(x_i, \theta) + \epsilon_{ij} \quad (1)$$

where y_{ij} is the j^{th} replication of the cumulative growth response at time x_i, is a vector of unknown parameters to be estimated, and $_{ij}$ is an error term assumed to be IID N(0, ^{2}). The additive error structure assumed in this model is deemed adequate as the replicate observations within the trellis rows were separated by tens of meters and, hence, the resulting influence on the estimates and their variability can be expected to be negligible. In typical growth models, $g(x_i,)$ may take the form of a scaled cumulative normal (Eq 2a) or logistic (Eq 2b) distribution:

$$g(x_i,\theta) = C\int_x (1/\sqrt{2\pi\sigma})\ \exp(-(x_i - \mu)/2\sigma^2)dx\ ;\ \theta = (\mu,\sigma) \quad (2a)$$

$$g(x_i,\theta) = M(1/(1 + \exp(B(x_i - L))))\ ;\ \theta = (M,B,L) \quad (2b)$$

While such models describe a general sigmoidal pattern of increasing growth, they fail to account for the multi-phased growth stages observed in grapes. One means of capturing these multiple growth phases is the use of a mixture model form (e.g. Gilks et al., 1998, Flieshman, et al., 2003). This technique builds upon simple sigmoidal growth models, such as those given above, to describe a continuous multi-phased growth process. The general form of the mixture model may be given as:

$$G(x_i, \Theta, P) = C\sum_k p_k\ g_k(x_i, \theta_k)\ ;\ P = (p_1,p_2,p_3,\ldots,p_k)\ ;\ \Theta = (\theta_1,\theta_2,\theta_3,\ldots,\theta_k) \quad (3)$$

where the component functions $g_k(x_i, _k)$ are pdf forms entering the model with weights or proportions p_k under the condition of $p_k = 1$, and parameter vectors $_k$, k = 1, 2, 3, ... K. The individual component distributions often take on the form of a normal pdf, however, the mixture may also be composed of other distributions such as the logistic, exponential, gamma, etc. The exact mixture used, as well as the number of components, K, is largely controlled by the scientific situation at hand, the data, and the expected growth pattern.

Because the component distributions of Eq. 3 are given in pdf form, the growth data must be decumulated (i.e. differenced between adjacent time points) prior to estimation. The peaks

and valleys present in these decumulated growth patterns typically have good biological relevance and interpretation. For example, in the grape growth data, these points represent peak rates of cell division, the onset and termination of lag phase, and the peak rate of berry enlargement during ripening.

Estimation of Eq. 3 may be carried out using maximum likelihood techniques, typically assuming a normal likelihood of the form:

$$\mathcal{L}(\theta) = f_{\theta}(x_1, x_2, x_3, \ldots, x_n) \quad (4)$$

where f (.) is the multivariate probability density associated with the observed data, assumed to be normal with mean defined by Eq 3 and variance ^{2}. Point estimation for relevant times in development, especially peak berry cell division or peak berry enlargement, may be given (closely approximated) by the individual component sample mean estimates. The intermediate lag phase, however, requires alternative methodologies such as analytical derivatives, numerical estimation, or derivations through the functional relationships between the component distributions. Inferential results on these estimates will be based on delta method approximations, or via computationally intense methods such as the bootstrap to be described in more detail in the following section.

All statistical computations and graphics were carried out using SAS 9.1.3 (2004). Relevant SAS codes are provided in an appendix.

III. Empirical Results

Trellis Data

In 2007, load cell data were collected at Modesto, CA on Chardonnay wine grapes. Six replicates (three positions along two trellis systems) were monitored beginning from berry set and continuing through harvest. Time increments were based on both Day of Year (DOY) and degree-day measurements, assuming a base temperature of 50^o F. Data were post processed (following collection, before analysis) to remove potential perturbations to the trellis wire systems due to temperature, precipitation, and mechanical disturbances. The cumulative, replicated growth data are shown in Figure 1. The time range of biological interest begin at DOY 130. Prior to this time the grape plants are producing vegetative growth and setting flower. DOY 130 marks the beginning of a berry cell division phase which continues to approximately 180 days followed by a berry enlargement phase running from DOY 190 to 240. An intermediate lag phase encompasses the approximate 10 day period between the two longer growth phases. The trellis tension response, reported as the raw signal from the load cells (mV) and shown in the figure, will eventually need to be calibrated to actual vine weights. For the work shown here, however, the modeling process will retain the mV units.

Estimation

To demonstrate the differences of simple from multi-phased sigmoidal growth, the simple logistic model from Eq. 2b was fitted to the Chardonnay data. The estimated parameters and associated values are given in Table 1. SAS codes are listed in Appendix 1a. The theoretical maximum, M, was estimated to be 624.5 mV, with a time to 50% of this maximum of L=160.7

days. These parameter estimates along with the rate estimate, K, were highly significant. While this seemed to be an adequate fit to the data, the predicted curve and residual plots given in Figures 2a and 2b indicated several problems with the model fit. Because of the multi-phased growth, the simple logistic model systematically over estimated the observed data during the initial berry division phase as well as the intermediate lag phase. Correspondingly, the residuals show a systematic, non-uniform pattern. Hence, a better model allowing for estimation of multiple growth phases was required.

As mentioned above, prior to applying the mixture model form to these data, the daily measurements must be differenced (decumulated). Figure 3 gives the decumulated data trends. In this form, the multiple phases of growth are quite evident showing two distinct peaks and valley in between. Because the data showed the two peaks, and the known growth pattern of grapes exhibits two growth phases (Coombe, 1960), a two- component mixture model form based on normal distributions was selected:

$$g(x_i, \theta_1, \theta_2) = C[p_1 N(\mu_1, \sigma_1^2) + (1 - p_1) N(\mu_2, \sigma_2^2)] \quad (5)$$

where C is a scaling constant, and $_1$, $_2$, $_1^2$, and $_2^2$ are the component normal distribution parameters. For the remaining demonstrations, the time increments, x_i, were taken to be in degree days (DD). While minimal differences were seen in these data between degree days and DOY, the degree day measurements were utilized in order to mitigate anticipated temporal effects expected when pooling the current data with future data sets.

The Eq.5 maximum likelihood estimates for Chardonnay are given in Table 2. Relevant SAS Codes are listed in Appendix 1b. All parameter estimates were significant ($p <= 0.0001$). The first growth phase had an estimated mean of $_1 = 565.01$ DD and a variance estimate of $_1^2 = 151.96$. It entered the model with a proportion of $p_1 = 0.639$. The mean in this case can be interpreted as the time of peak berry cell division. The second phase of growth, berry enlargement during ripening, can be expected to peak at approximately $_2 = 1143.36$ DD, or about 46 days later. The fitted curve and residuals are given in Figures 4a and 4b. These indicated a much improved fit over the standard logistic model with an expected curve that followed the data trends well and a residual pattern with little, if any, systematic deviations. In addition, the AIC model fit statistic is much lower than its simple logistic model counterpart, i.e. 2759.9 for the mixture model vs 6311.0 for the simple logistic model.

For the Chardonnay, the two-component mixture model form fitted the data well and provided useful estimates. Other distributional forms (e.g. gamma and beta distributions) which allow for skewed data patterns were also attempted (results not shown). However, in this particular case, they showed minimal improvement in overall fit and displayed high parameter correlations, indicative of model inadequacies. The alternative distributions should not be ruled out for future applications as climatic, biological or temporal effects on grape growth could require their utilization.

While the points of peak growth were easily obtained from the component mean estimates, the point of minimal growth, i.e. the lag phase, was more difficult to determine. One

obvious approach was to derive an estimate of the minimum through the analytic derivative of Eq. 4. This proves untenable, however, as a closed form for the derivative does not exist. An alternative might be to solve for the point at which the two component distributions intersect, i.e., the point where:

$$p_1 \exp(-(DD - \mu_1)^2/2\sigma_1^2) = (1 - p_1)\exp(-(DD - \mu_2)^2/2\sigma_2^2) \quad (6)$$

This solution, however, was unsatisfactory due to bias. For example, in the Chardonnay data, Figure 5 shows that the larger size of the proportion associated with the first component, p_1, forces the intersection point of the two growth phases to be greater than the actual minima. In fact, this "equivalency" solution would only be unbiased in the case where $p_1 = p_2$ which seems unlikely to occur in practice.

The final solution considered here was a numerical approximation technique. Specifically, the differential between adjacent predicted points, y_{diff}, was observed until a point was reached where the difference was sufficiently close to zero, i.e.:

$$y_{diff} = \min_i \left(\overline{y_{i+1}} - \overline{y_i} \right) \quad (7)$$

where the bar notation indicates a predicted value. Figure 6 demonstrates this process for the Chardonnay data. Here, the minimum was approximated at 889.9 DD. To be of inferential use, however, this estimate must be evaluated using statistical simulation. .

In this case, a bootstrap technique was applied to the data following Efron and Tibshirani (1993). In the bootstrap procedure, the residuals from the estimated model were randomly selected with replacement. Each sampled residual is then added back to a predicted value from the estimated model. The model is then re-estimated from the new bootstrapped growth series and the resulting model estimates recorded. This method assumes the model to be true and fixed. In addition, independence of vine rows and sampling positions within rows was assumed allowing for simple random sampling across the entire residual dataset. The sampling-estimation process was repeated B=1000 times resulting in 1000 estimates of the minimum (see appendix 2 for SAS codes). From these estimates, the sampling distribution of minima was formed and the 2.5 and 97.5 percentiles selected as lower and upper confidence bounds on the minimum. For the Chardonnay data, the minimum estimate of 889.9 was bounded by the values 869.8 DD to 904.7 DD. These values correspond to an actual calendar time of 3 - 4 days.

IV. Concluding Remarks

Grape growth exhibits a two-phased, "double sigmoidal" growth pattern. Models for predicting growth should, therefore, have the ability to accommodate and estimate these phases. In the examples shown here, the two-phased normal mixture form did well in achieving this

although other distributions may be considered for future work. Estimates obtained from the model naturally supply information on points of interest, such as the times of peak berry cell division and peak berry enlargement during ripening. Estimates and inference on the intermediate lag phase, however, require numerical and computationally intense methods. For the data at hand, lag time estimates indicate a short window where the lag phase occurs. Future work with this model should apply new data to provide estimation in the presence of annual temporal variation as well as providing a means for statistical validation of the models prior to their use in prediction.

V. References

Coombe, B. G. 1960. Relationship of growth and development to changes in sugars, auxins, and gibberellins in fruit of seeded and seedless varieties of Vitis vinifera. Plant Physiology, 35(2):241-50.

Efron, B. And R. J. Tibshirani. 1993. An Introduction to the Bootstrap (Monographs on Statistics and Applied Probability). Chapman Hall, 456pp.

Fleishman, S. J. and D. L. Burwen. 2003. Mixture models for the species apportionment of hydroacoustic data, with ech-envelope length as a discriminatory variable. ICES journal of Marine Science, 60: 592-598.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1998. Markov Chain Monte Carlo in Practice. Chapman and Hall/CRC, New York, New York. 486 pp.

Lewis. M. A. 1910. The development of the grape. Agricultural Journal of the Cape of Good Hope. 37: 528-551.

NASS. Noncitrus fruits and nuts, preliminary summary. January 2008. (Accessed 30 June, 2008. http://usda.mannlib.cornell.edu/usda/current/NoncFruiNu/NoncFruiNu-01-23-2008.pdf

SAS Institute Inc. 2004. SAS OnlineDoc® 9.1.3. Cary, NC: SAS Institute Inc.

Tarara, J. M., J. C. Ferguson, P. E. Blom, M. J. Pitts, F. J. Pierce. 2004. Estimation of grapevine crop mass and yield via automated measurements of trellis tension. Transactions of the ASAE, Vol. 47(2): 647-657.

Parameter	Estimate	Std Err	Lower	Upper
M	624.500	3.906	616.800	632.100
K	0.049	0.001	0.047	0.051
L	160.700	0.411	159.900	161.500

Table 1. Estimated parameters for the simple sigmoidal logistic model and 95% confidence bounds.

Parameter	Estimate	Std Err	p-value
p1	0.639	0.010	<.0001
$\mu 1$	565.010	3.439	<.0001
$\mu 2$	1143.360	6.105	<.0001
$\sigma 1$	151.960	3.149	<.0001
$\sigma 2$	127.640	4.155	<.0001
C	6199.970	103.700	<.0001

Table 2. Estimated parameters for the two component normal mixture model.

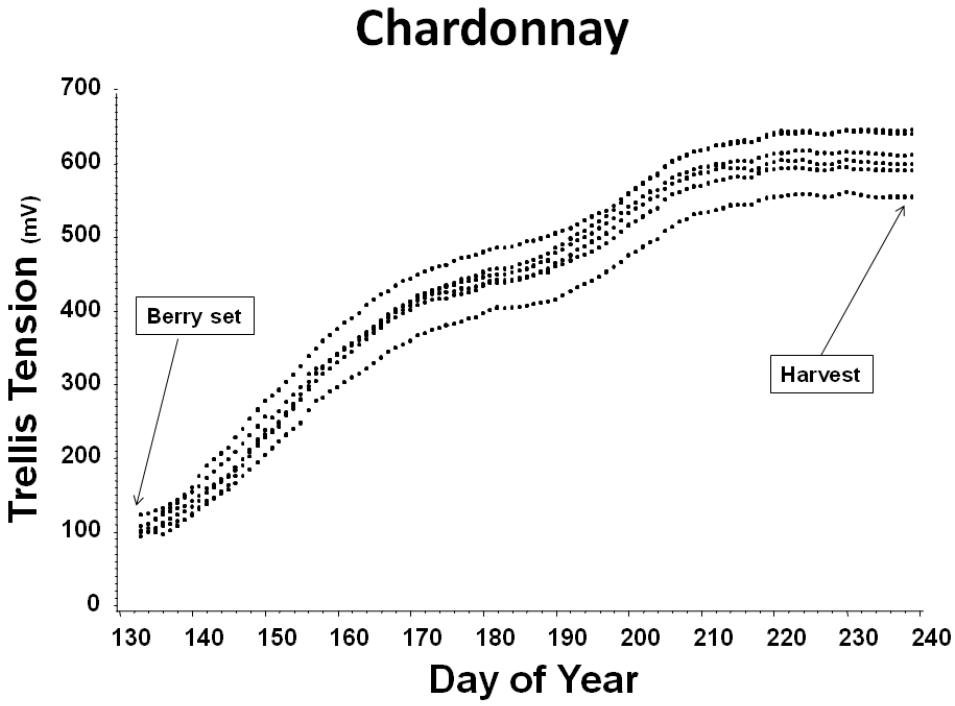
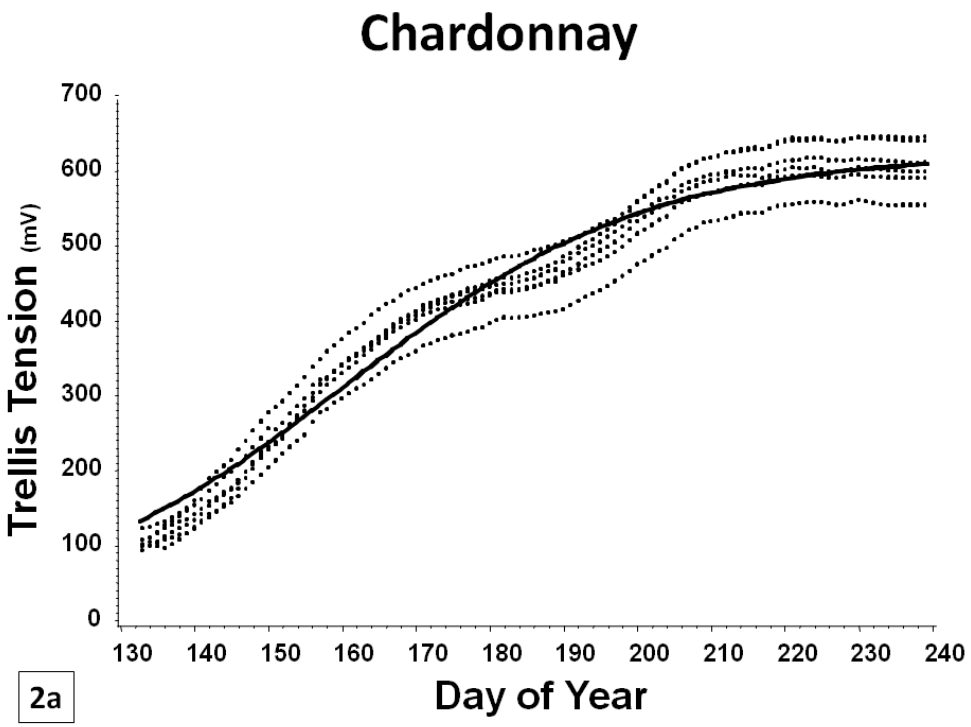


Figure 1. Cumulative growth data for Chardonnay grape varieties.



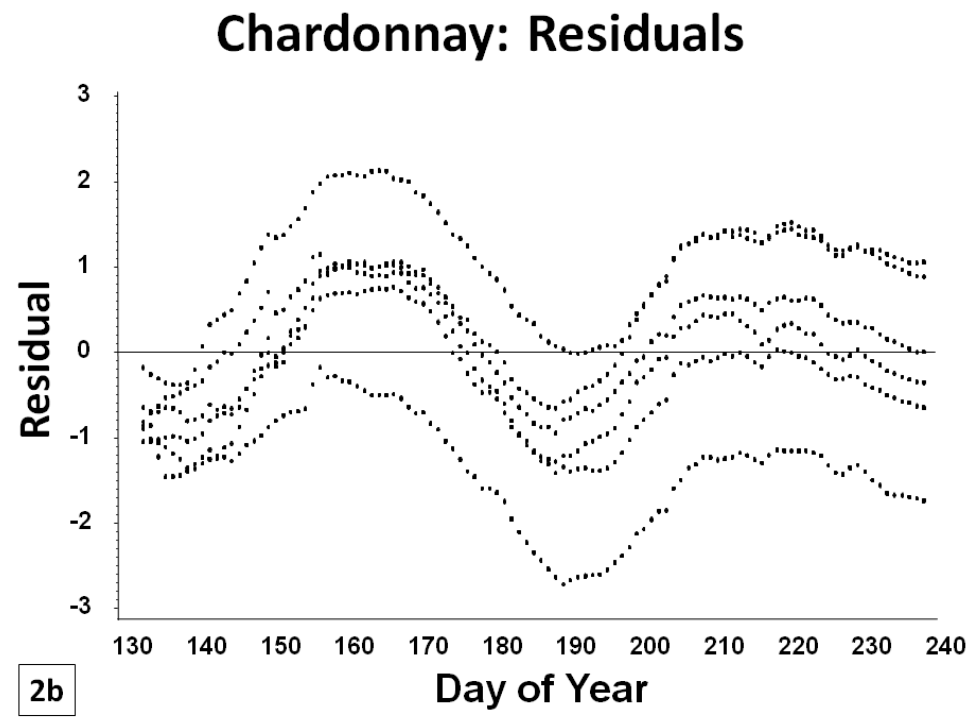


Figure 2. Estimated logistic model (a) and corresponding standardized residuals (b) for the Chardonnay data.

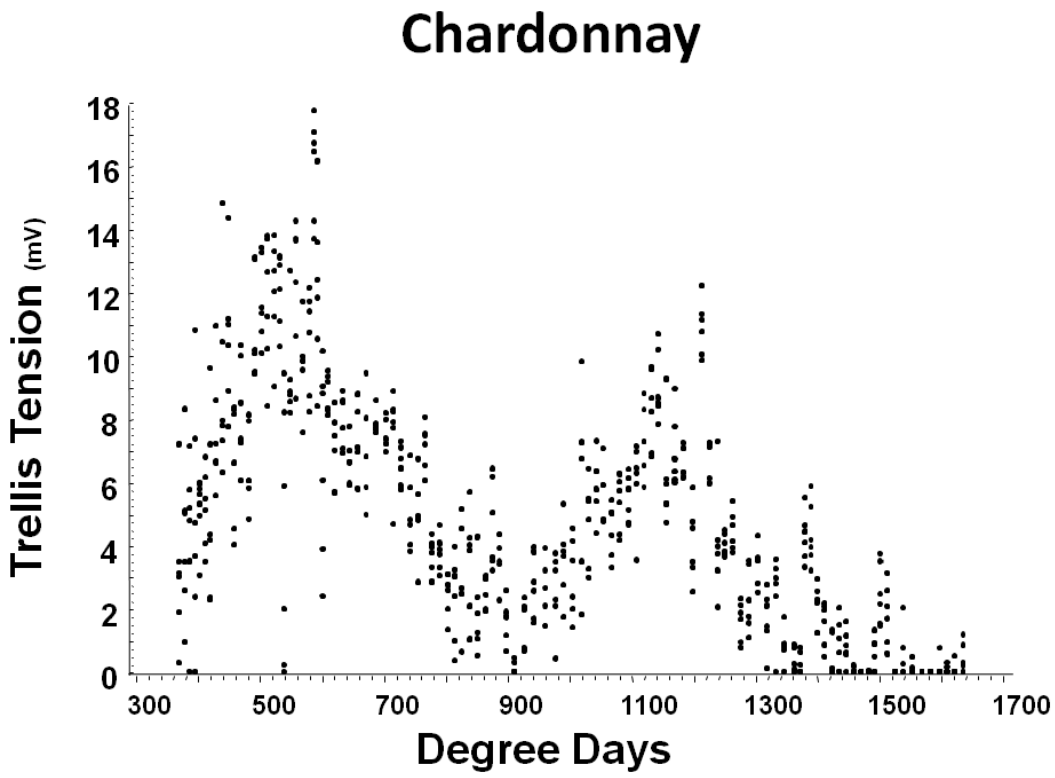
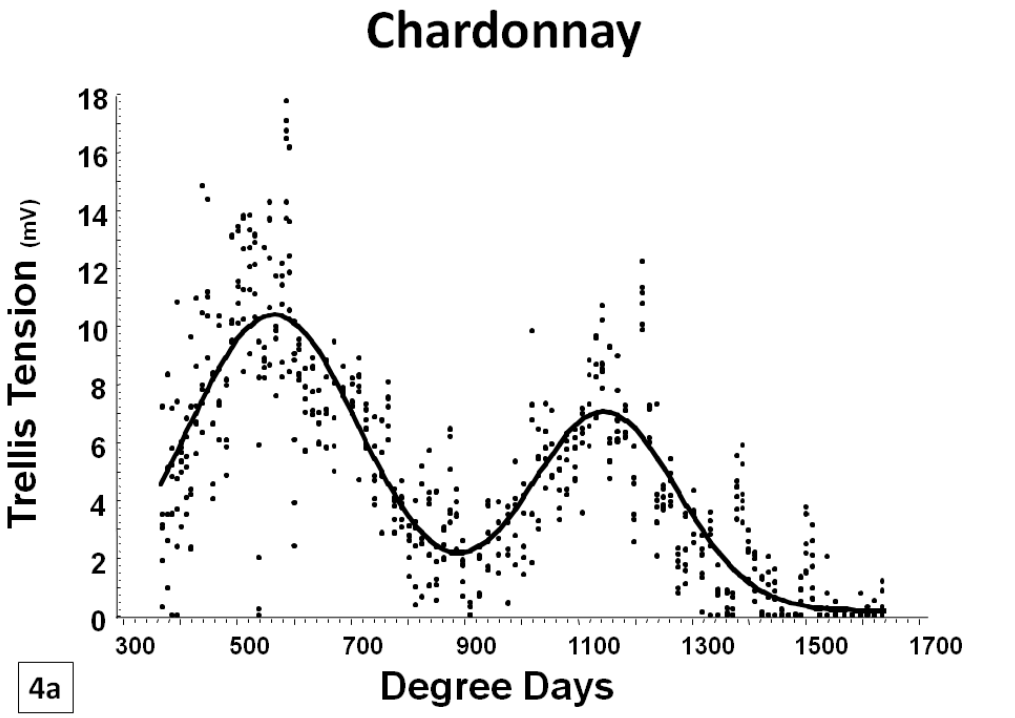


Figure 3. Decumulated data for Chardonnay growth data.



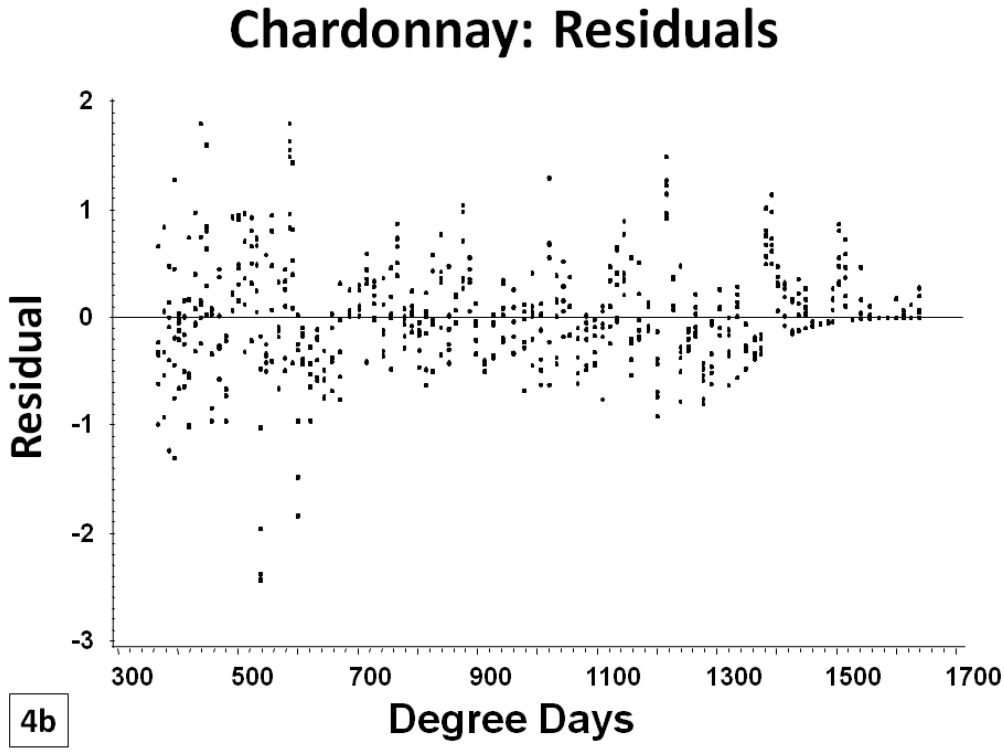


Figure 4. Estimated two-component mixture model form (a) and associated standardized residuals (b) for the Chardonnay data.

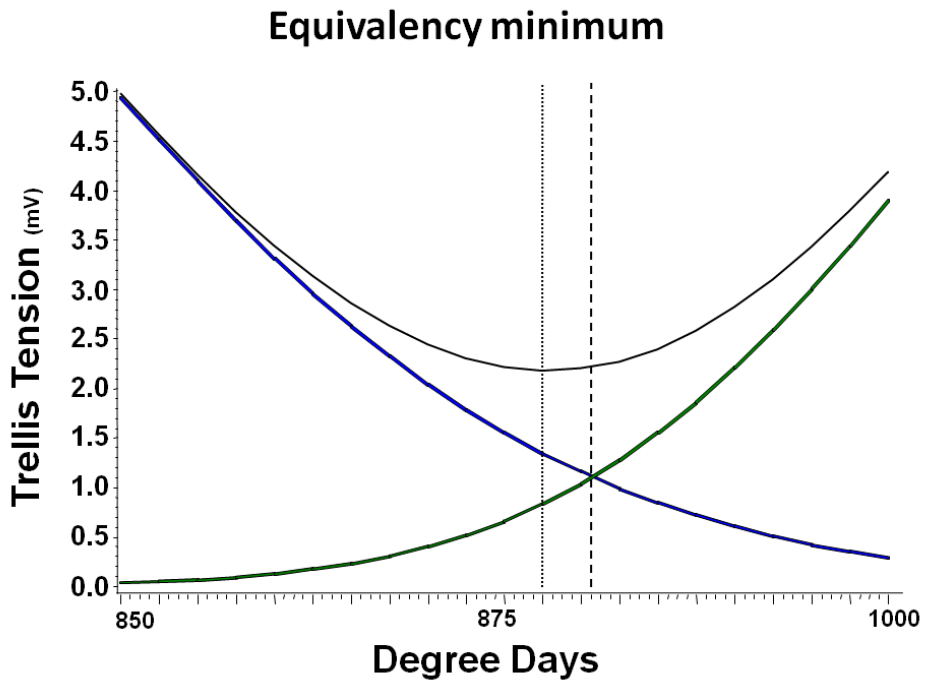


Figure 5. Point of component equivalency (large dash line) and actual model minimum (small dotted line) for the Chardonnay data.

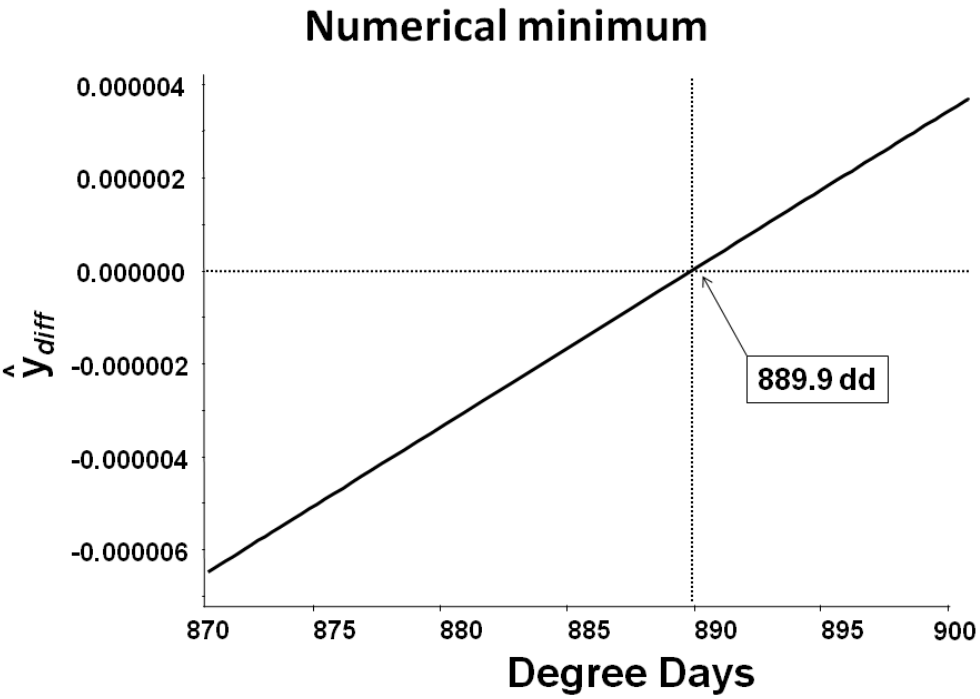


Figure 6. Point of numerical minimum for the Chardonnay data.

Appendix 1a. Estimation of simple logistic model.

```
/*********************************************************/
/**  NLMIXED used to get ML estimates of simple                    **/
/**  logistic model.  Lc is cumulated growth data, DOY is day of year.   **/
/*********************************************************/

proc nlmixed data=gallo1;
        parms m=700 k=1 l=200 sigma=20;
        yhat = m/(1 + exp(-k*(doy-l)));
        model lc ~normal(yhat, sigma);
        predict  yhat out=pred;
        predict lc - yhat out=resid;
run;
```

Appendix 1b. Estimation of two component mixture model.

```
/*********************************************************/
/**  NLMIXED used to get ML estimates of two- component mixture  **/
/**  model based on Normal PDFs. Diff is differenced cumulative       **/
/**   growth data and dd are degree day measurements.                 **/
/*********************************************************/

proc nlmixed data=gallo1;
        parms p=.75 mu1=565 mu2=1143 sigma1=151 sigma2=127 C=6200 sigma=4;
        yhat=C*(p*PDF("NORMAL",dd,mu1,sigma1)+(1-p)*PDF("NORMAL",dd,mu2,sigma2));
        model diff ~normal(yhat, sigma);
        predict yhat out=pred;
        predict diff-yhat out=resid;
run;
```

Appendix 2. SAS Bootstrap program for identifying percentile intervals on two component minimum.

```
/*************************************************************/
/****  Define bootstrap macro.                                ***/
/*************************************************************/
%MACRO BOOT(ITER);

        /*************************************************************/
        /****  Run bootstrap loop ITER times                          ***/
        /*************************************************************/

        %DO I = 1 %TO &ITER ;
        /*********************************************************************/
        /****  Use SURVETSELECT to grab random sample with replacement.       ***/
        /*********************************************************************/

                proc surveyselect method=urs n=642 outhits data=resid out=sample;
                        id pred;

        /*************************************************************/
        /****  Name change and add residuals to model                 ***/
        /*************************************************************/
                data sample (keep = resid);
                        set sample;
                        resid = pred;
                data sample (keep = boot dd yhat resid diff);
                        merge pred sample;
                        boot = yhat + resid;

        /*************************************************************/
        /****  Fit model to bootstrap sample and save results.        ***/
        /*************************************************************/
                ods output ParameterEstimates=ests;
                proc nlmixed data=sample;
                        parms p=.75 mu1=565 mu2=1143 sigma1=151 sigma2=127 C=6200 sigma=4;
                        mu=C*(p*PDF("NORMAL",dd,mu1,sigma1)+
                    (1-p)*PDF("NORMAL",dd,mu2,sigma2));
                        model boot ~normal(mu, sigma);

        /*************************************************************/
        /****  Numerically find minimum.  We know where to look.      ***/
        /*************************************************************/
                proc transpose data=ests out=ests;
                        var estimate;
                        id parameter;
                data ests;
                        set ests;
                        do dd=850 to 910 by .001;
                                mu=C*(p*PDF("NORMAL",dd,mu1,sigma1)+
                    (1-p)*PDF("NORMAL",dd,mu2,sigma2));
                                output;
                        end;
                data ests;
                        set ests;
                        f1 = mu-lag1(mu);
                        if f1=. then delete;
                        test=abs(f1);
```

```
                proc sort data=ests;
                        by test;
                data ests;
                        set ests;
                        if _n_=1;

        /*************************************************************/
        /****  Add the current minima to theset of all bootstrap results          ***/
        /*************************************************************/
        data boots (keep=iter pred dd);
                        set boots ests;
                        iter=&ITER;
        %END;
%MEND;

/*************************************************************/
/****  Fit Original data using two component distribution          ***/
/*************************************************************/
proc nlmixed data=gallo1;
        parms p=.75 mu1=565 mu2=1143 sigma1=151 sigma2=127 C=6200 sigma=4;
        mu=C*(p*PDF("NORMAL",dd,mu1,sigma1)+ (1-p)*PDF("NORMAL",dd,mu2,sigma2));
        model diff ~normal(mu, sigma);
        predict mu out=pred;
        predict diff-mu out=resid;

/*************************************************************/
/****  Quick and dirty variable name change to avoid conflicts.          ***/
/*************************************************************/

data pred (drop = pred);
        set pred;
        yhat = pred;

/*************************************************************/
/****  Start a dataset for all bootstrap results.                  ***/
/*************************************************************/
data boots;
        iter=0;

/*************************************************************/
/****  Turn off listing and start bootstrap macro.          ***/
/*************************************************************/
ods listing close;
%BOOT(1000);
run;

/*************************************************************/
/****  Turn listing back on.  Use UNIVARIATE, SUMMARY or          ***/
/****  INSIGHT on dataset BOOTS to find percentiles.          ***/
/*************************************************************/
ods listing;
```