Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF
AGRICULTURAL STATISTICS 64(2) 2010 229-236

Using the Logistic pdf Model to Mitigate Autocorrelation in Growth Curve Analysis

James H. Matis[1*], Muhammed Jassem Al-Muhammed[2] and Wopke van der Werf[3]
[1]*Department of Statistics, Texas A&M University, USA*
[2]*Department of Mathematics, Damascus University, Syria*
[3]*Department of Plant Sciences, Wageningen University, The Netherlands*

Received 12 December 2009; Accepted 10 March 2010

SUMMARY

The well-known Verhulst-Pearl model in ecology,

$$y' = (\lambda - \delta \cdot y(t)) \cdot y(t)$$

where $y(t)$ denotes current population size, has a solution which may be written in the form of a logistic cumulative distribution function (cdf). This function is widely used to describe population growth curves. However population growth data are prone to serial correlation, which would complicate subsequent statistical inferences. The serial correlation in the data can be mitigated by fitting the first differences of the population data to a model for the rate of population growth. This function is the solution to the alternative mechanistic model,

$$y' = (\lambda - \delta \cdot Y(t)) \cdot y(t)$$

where $Y(t)$ is the integral of $y(s)$ from 0 to t, and it has the mathematical form of a logistic probability density function (pdf). A biologically meaningful parameterization of the logistic pdf model is provided to facilitate initial estimates for parameters in nonlinear curve fitting. We illustrate the procedure, demonstrating the problem of serial correlation in population data and the effectiveness of the suggested solution, by fitting two classic data sets.

Keywords: Verhulst-Pearl model, Cumulative size dependency, Aphid population size model.

1. INTRODUCTION

The logistic growth curve has a celebrated history, and is in widespread use today in the plant, animal and ecological sciences (see e.g. Renshaw 1991; Brown and Rothery 1993, Thornley and France 2007, Gotelli 2008). Google on the internet has hundreds of references. The logistic is usually the initial model of choice for describing S-shaped population growth curves, i.e. growth curves that plateau off after an initial exponential increase. However data fitted to this model often exhibit serial correlation, which complicates the subsequent statistical inferences (Franses 2002, Lindsey 2004). A common solution to this problem is to append some assumed error structure onto the model (see e.g. Berny 1989; Glasbey 1979). However doing so adds complexity to the analysis, and choosing a suitable model for the error structure is no trivial task. This paper proposes a potentially more efficient solution, namely taking the first differences of the data in order to analyze the incremental changes, or rate of change. If this solution is chosen, the first difference data should be fitted to a logistic model for the growth rate, which we call the 'logistic pdf' model.

Section 2 outlines the standard logistic curve and presents two examples of fitting the model directly. The problem of autocorrelation is illustrated. Section 3

[*] *Corresponding author* : J.H. Matis
E-mail address : matis@stat.tamu.edu

presents the logistic pdf model, including a parameterization that is stable and also facilitates the choice of initial estimates for iterative nonlinear curve fitting. Section 4 illustrates fitting the pdf model to the first differences of the data from the two previous examples, and demonstrates the advantages of this approach. Concluding remarks are given in Section 5.

2. THE STANDARD LOGISTIC GROWTH MODEL

2.1 Basic Model

The logistic growth curve was first suggested by Verhulst in 1838, and derived independently by Pearl and Reed in 1920 (Renshaw 1991). Let $y(t)$ denote population size at time t, and $y'(t)$ its derivative. The Verhulst-Pearl model is

$$y'(t) = \lambda y \cdot \left(1 - \frac{y}{K}\right) \tag{1}$$

with parameters $\lambda > 0$ and $K > y(0)$. The simple solution to (1) may be written as

$$y(t) = \frac{K}{1 + e^{(c - \lambda t)}} \tag{2}$$

with parameter c related to the initial value $y(0)$ as $c = \ln\left[\frac{K - y_0}{y_0}\right]$. Parameters λ, called the 'intrinsic growth rate' (units: time^{-1}), and K (same units as y), called the 'carrying capacity', are key descriptors of population dynamics in ecology. For subsequent convenience, we rewrite (2) as

$$y(t) = \frac{K}{1 + e^{-\lambda(t - t_{\max})}} \tag{3}$$

with new parameter $t_{\max} = c/\lambda$.

Equations (2) and (3) with $K = 1$ have the form of a cumulative distribution function, or 'cdf', of the logistic probability distribution in statistics. This is curious, as the previous derivation has nothing to do with a random variable, but is instead the solution to a differential equation with no randomness. The fact that (2) and (3) have the form of a logistic cdf is very helpful, as many properties of the logistic cdf are well-known in statistics. Due to this correspondence, we call the previous logistic growth model in (2) and (3) the *logistic cdf* model.

The Verhulst-Pearl model in (1) is often reparameterized as

$$y'(t) = (\lambda - \delta y(t)) \cdot y(t) \tag{4}$$

where $\delta = \lambda/K$. In a mechanistic interpretation of model (4), the intrinsic growth rate λ is interpreted as the *per capita* birth rate of the population. The corresponding *per capita* death rate is δy, where δ is a death rate coefficient. The model is called 'density dependent', as the death rate is a function of current size, y. The positive root in (4), $K = \lambda/\delta$, gives the carrying capacity.

2.2 Two Classic Examples

Pearl fitted the logistic curve in 1927 to some yeast data given in Carlson (1913, the data are also given in Renshaw 1991). We illustrate first fitting the model to this classic data set, using standard nonlinear least squares (Neter *et al*. (1996)), as implemented in SPSS (2007). Letting $C(t)$ denote the observed population size at time t, we assume regression model

$$C(t) = y(t) + \varepsilon \tag{5}$$

where ε denotes an *independent* random error term with constant variance. The Carlson data and the resulting fitted logistic curve are illustrated in Fig 1. The parameter estimates, with standard errors in parentheses, are K = 663.0 (1.7) for the carrying capacity, λ = 0.547 (0.006) /hr for the intrinsic growth rate, and $t_{\max}$ = 7.81 (0.022) hr. The estimated equation, which fits the data very well (Fig. 1), is

$$C(t) = \frac{663.0}{1 + e^{-0.547(t - 7.81)}} \tag{6}$$

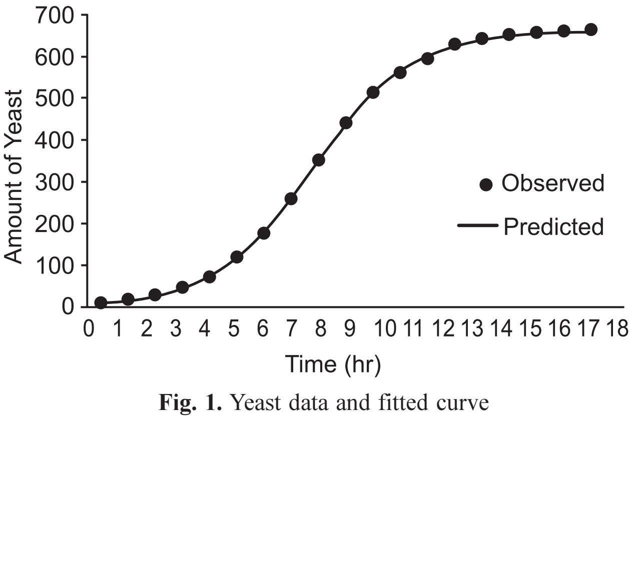

Fig. 1. Yeast data and fitted curve

The parameter estimate for the death rate coefficient is $\delta = 0.547/663 = 8.25 \times 10^{-4}$ (1.12×10^{-5}), where the standard error of δ is calculated from the standard errors of λ and K by the linear approximation:

$$SE(\delta) = \frac{SE(\lambda)}{K} + \frac{\lambda\, SE(K)}{K^2} \text{ (Taylor 1997).}$$

As an illustration of a modern example, consider US population size data for 1790-1990, given by Lipken and Smith (2009). The data in 10-yr intervals and the fitted curve are illustrated in Fig. 2.

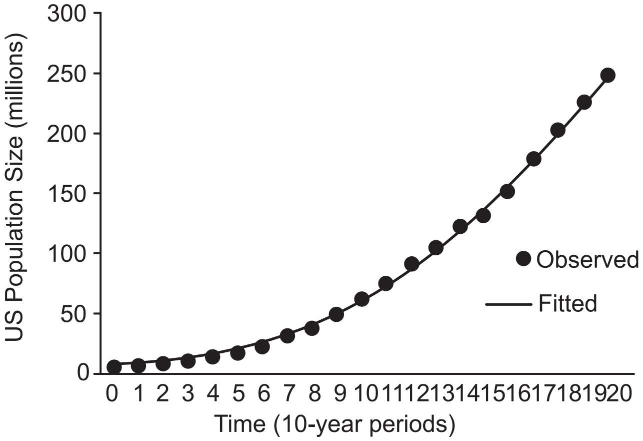

Fig. 2. US population data and fitted logistic curve

The fitted logistic model is

$$C(t) = \frac{387.7}{1 + e^{-0.227(t-17.6)}} \tag{7}$$

which describes the data remarkably well over this 200-year period. An ecologist would note that the estimate, with again the standard error in parentheses, of carrying capacity is $K = 387.7$ (30.5) million, and of the intrinsic growth rate is $\lambda = 0.227$ (0.011)/10 yr. period.

2.3 A Statistical Caution

An observed population growth curve is a time series, and due to its cumulative nature, its observations are likely to be serially correlated. Serial correlation, also called autocorrelation, violates the assumption of independent observations in regression model (5). A consequence of such violation is that the estimated standard errors of the parameter estimates from standard nonlinear least squares would be too small, indicating greater apparent precision than warranted (Neter *et al.* 1996). Serial correlation is readily assessed by the Durbin-Watson d statistic. The expected value of d under the null hypothesis of no serial correlation is $d = 2$ (Neter *et al.* 1996).

Serial correlation is indeed evident in the plot of residuals in Fig. 3 for the US population data. Such correlation is apparent visually as there are only four 'runs' (i.e. changes of sign) for the 21 data points. The Durbin-Watson test statistic for the US population data is $d = 0.56$, well below the lower critical value $d_{L,\,0.05} = 1.22$, hence there is strong statistical evidence of positive serial correlation in the residuals. Consequently the assumed regression model in (5) is not appropriate for these data.

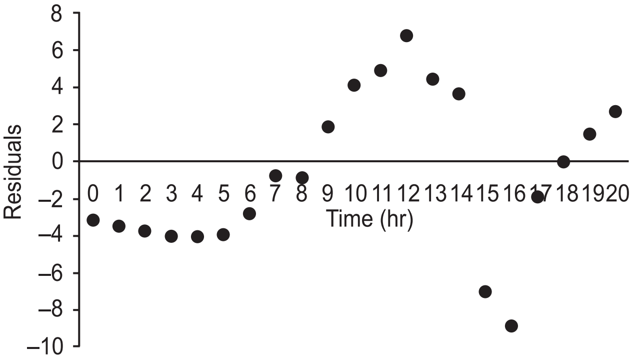

Fig. 3. Residual plot for US population data

We *do not* claim that every (cumulative) growth curve has significant serial correlation, indeed the yeast data has no doubt become a celebrated data set in part because it does not have significant serial correlation (Its $d = 1.96$). The lack of serial correlation in this data set is not surprising due to its carefully controlled lab conditions. However we *do* suggest that many population growth data sets have sizeable error terms, with naturally occurring perturbations accumulated over time. For such growth curves with substantial serial correlation, we propose using the recently developed *logistic pdf* growth model outlined below.

3. THE LOGISTIC pdf GROWTH MODEL

3.1 Review of Derivation of New Logistic Growth Model

We review first the derivation of this different logistic model. Its mechanistic formulation was first proposed in mathematical terms by Kindlmann (1985) to describe aphid populations, and is given as:

$$y'(t) = (\lambda - \delta Y(t)) \cdot y(t) \tag{8}$$

where

$$Y(t) = \int_0^t y(s)ds \tag{9}$$

$Y(t)$ denotes the time-integrated past population size, or 'cumulative density', since the initial time (t = 0). In simple terms, model (8) changes the *per capita* death rate from $\delta\, y(t)$ in (4) to $\delta\, Y(t)$. Prajneshu (1998) derived the first analytical solution to the differential equation in (8), in the form:

$$y(t) = \frac{ae^{-bt}}{(1+de^{-bt})^2} \tag{10}$$

where *a, b*, and *d* are positive parameters. Matis *et al.* (2007) shows that the solution to (8) can also be written using stable parameters, which simplify fitting the model to data, as:

$$y(t) = \frac{4y_{\max} \cdot e^{-b(t-t_{\max})}}{(1+e^{-b(t-t_{\max})})^2} \tag{11}$$

The parameters in this model are naturally interpretable, as parameter $y_{\max}$ denotes the maximum of $y(t)$ in (11) and $t_{\max}$ the time of this maximum. Parameter *b* is a relative rate defined subsequently. The biological meaning of the parameters in equation (11) facilitates the choice of initial parameters in iterative nonlinear estimation procedures. Ross *et al*. (2010) gives a slightly different parameterization of the model.

Once the observed data are fitted to model (11), the estimates of $y_{\max}$, $t_{\max}$, and *b* may be used to estimate the parameters of underlying model (8). One can show that

$$\lambda = b \cdot \frac{d-1}{d+1} \tag{12}$$

where

$$d = e^{bt_{\max}} \tag{13}$$

and corresponding relationships for parameters δ and $y(0)$ are given in Matis *et al*. (2007).

Parameter *d* is typically very large, hence it follows from (12) that parameter *b* is an accurate approximation for the *per capita* birth rate λ.

Solution (11) could alternatively be written as

$$y(t) = K \cdot p(t) \tag{14}$$

where $p(t)$ is the logistic probability density function, or 'pdf', defined as

$$p(t) = \frac{b \cdot e^{-b\,(t-t_{\max})}}{(1+e^{-b(t-t_{\max})})^2} \tag{15}$$

and *K* is the constant (Matis *et al*. (2009a))

$$K = \frac{4y_{\max}}{b} \tag{16}$$

The logistic pdf (also called the sech-squared pdf) is also well-known in statistics. It is symmetric with heavier tails than a Normal pdf with the same mean and variance (Johnson and Kotz 1970).

We call new model (11) the logistic pdf model, due to (14). It seems curious again that mechanistic model (8) also, which is devoid of any randomness, would have as its solution a scaled form of a well-known probability distribution in statistics. We note, however, that one can formulate a stochastic analog to model (8), in which the current count *y(t)* is a random variable. As a contrast to 'deterministic' model (11), which is the solution to a single differential equation, Matis *et al.* (2005) presents an example of an 'exact' solution to a stochastic analog of (8) based on the solution of a system of over 180,000 (Kolmogorov) differential equations.

3.2 Mechanistic Application of the New Logistic pdf Model

Both Kindlmann, the Czech ecologist who first formulated (8), and Prajneshu, the Indian mathematician who obtained solution (10), sought specifically to describe aphid population growth. Theoretical reasons why the aphid family satisfies the mechanistic assumptions in (8) are reviewed in Matis *et al.* (2007). This model formulation has also been verified empirically by fitting model (11) successfully to a number of aphid species, including the mustard aphid in Prajneshu (1998), the cotton (*aka melon*) aphid in Ross *et al*. (2010), and the pecan and the soybean aphids (in Matis *et al*. (2006) and (2009), respectively). These past studies demonstrate the versatility of the model to describe observed aphid abundance curves, whether the peak count is high or low, the spread narrow or wide, or the time of maximum short or long. Research is also in progress to use this pdf model to describe gypsy moth population curves (Matis *et al.* (2010)).

4. USE OF THE LOGISTIC pdf MODEL WITH FIRST DIFFERENCES

4.1 First Differences Procedure

We consider now using the logistic pdf model in (11) for another purpose, i.e. as a potential solution for the problem of serial correlation in cumulative data, as described in Section 2. A simple and well-known procedure for analyzing time series data in economics is to transform the data by taking "first differences" (Neter *et al.* 1996). For this procedure, one would obtain the differences between consecutive values of the independent variable and also the differences between consecutive values of the dependent variable. Under certain conditions, most notably under a first-order autoregressive error assumption, the residuals from a regression model relating these first difference variables would no longer be serially correlated (Neter *et al.* 1996).

We propose an adaptation of this standard procedure. With $C(t)$ denoting the cumulative count, let $D(t)$ denote the difference between consecutive $C(t)$ values, ie. $D(t) = C(t) - C(t-1)$. These first differences of the cumulative counts for *equally spaced data* are proportional to the estimated derivatives of the function. Hence assuming that the cumulative $C(t)$ counts follow the logistic cdf model, it follows that the $D(t)$ incremental changes would follow the logistic pdf model.

We are not aware of the logistic pdf model being used previously for this purpose, and we propose this as a general method for growth curve analysis. It is plausible to assume a regression model in which the $D(t)$ incremental changes are independent. Under that assumption the $C(t)$ counts could not be independent as assumed in (5) as each $C(t)$ is the sum of the past $D(t)$. Clearly, one cannot *prove* that the residuals of $D(t)$ are uncorrelated. However one can easily investigate *empirically* for any given data set whether the serial correlation has been substantially reduced under this transformation. The important point is that *when there is serial correlation in the C(t) counts, the transformed D(t) data for the new pdf model, which retains equivalent parameters, are likely to have far less serial correlation*. We investigate this first differences approach for the two data sets in Section 2.

4.2 Reanalysis of Carlson's Yeast Data

Consider now fitting logistic pdf model (11) to the first differences, illustrated in Fig. 4, of the yeast data. The fitted curve is

$$D(t) = \frac{(4)\times(91.56)\times e^{-0.56(t-8.29)}}{(1+e^{-0.56(t-8.29)})^2} \tag{17}$$

which fits the data well, as apparent in Fig. 4.

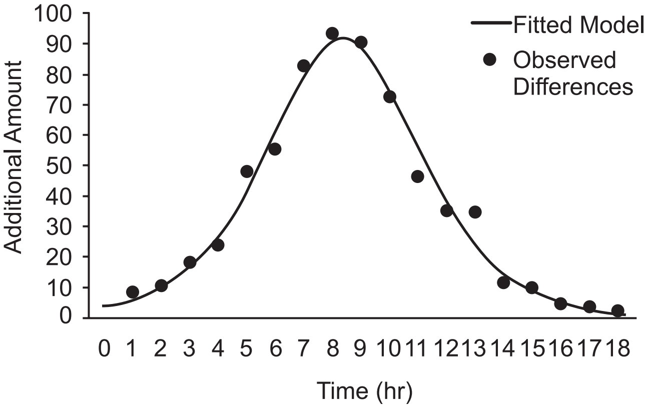

Fig. 4. Yeast data differences and fitted curve

The parameter estimates for model (11) are $y_{\max} = 91.56$ (2.69), $t_{\max} = 8.29$ (0.092) hr and $b = 0.559$ (0.020) / hr. The estimate of $y_{\max} = 91.56$, using (16), corresponds to $K = 654$ (42.7). These point estimates are close to the estimates of $K = 663.0$ (1.7), $t_{\max} = 7.81$ (0.022) hr, and $b = 0.547$ (0.006) /hr from model (6). Note, however, that the standard errors for the estimates from model (11) are at least three times larger than those from model (3). This might be expected, in part as a perturbation from the expected amount of yeast produced in a given hour stands out more clearly when expressed directly as in Fig. 4 than when accumulated with all past hourly amounts of yeast as in Fig. 1.

The residuals in Fig. 5 show no apparent pattern of serial correlation (as Durbin-Watson $d = 1.47$, above the critical value $d_{U,\,0.05} = 1.39$). This conclusion is expected, because as noted previously the residuals from the cumulative data in Section 2.2 do not indicate any serial correlation either (with $d = 1.96$). The residual at $t = 13$ hr is the largest in magnitude for the logistic pdf. If it were possible, it would be instructive to examine the circumstances of this original yeast experiment to determine whether anything unusual occurred at around hour 13.

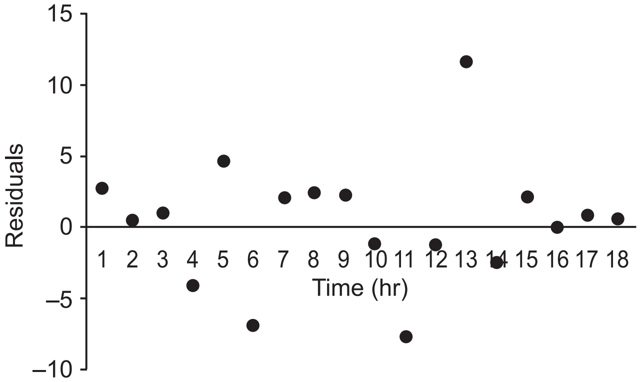

Fig. 5. Residual plot for yeast data differences

4.3 Reanalysis of US Population Data, 1790-1990

Consider now the US population data in Section 2. The first differences and the fitted model are illustrated in Fig. 6. One feature that stands out immediately in this figure is the lack-of-fit for periods 15 and 17 in the graph. These periods correspond to the US depression years 1930-1940 and to the post-war boom years of 1950-1960, respectively. The lack-of-fit for these two periods is even more dramatic in the subsequent residual plot. On the other hand, these two effects are barely noticeable in the cumulative plot in Fig. 2. This illustrates an advantage of the analysis based on first differences, namely better diagnostics for model lack-of-fit.

The fitted curve in Fig. 6 is

$$D(t) = \frac{(4)\times(23.89)\times e^{-0.181(t-20.67)}}{(1+e^{-0.181(t-20.67)})^2} \tag{18}$$

which fits adequately. It appears, however, that the first differences are just approaching the peak in Fig. 6, and have not yet started to decrease. Thus though the $C(t)$ counts seem to fit the logistic cdf model adequately in Fig. 2, it is apparent from the $D(t)$ incremental counts in Fig. 6 that the model should be regarded with caution until it is validated by a decreasing trend in future observations, as an extrapolation of the model would predict.

The parameter estimates are $y_{\max}$ = 23.89 (3.03) million, $t_{\max}$ = 20.67 (2.95) periods, and b = 0.181 (0.038)/period. $y_{\max}$ corresponds, using (16), to K = 528.2 (177.8). These estimates differ considerably from the estimates of K = 387.7 (30.5), $t_{\max}$ = 17.57 (0.72) and b = 0.227 (0.011) for the cumulative data fitted to (3). These differences indicate there is still considerable instability in model definition due to the lack of crucial data to estimate accurately the size and time of the peak in model (11).

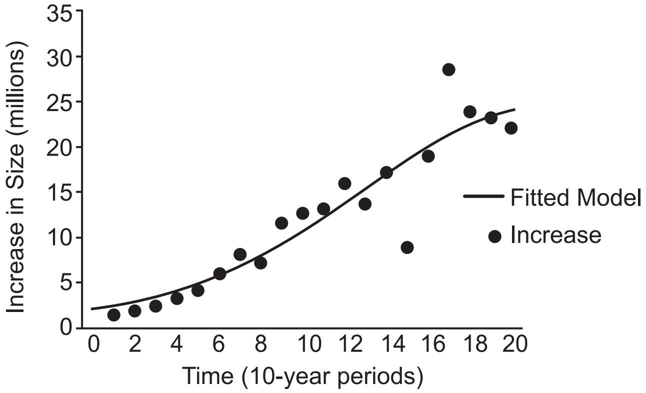

Fig. 6. US population differences with fitted curve

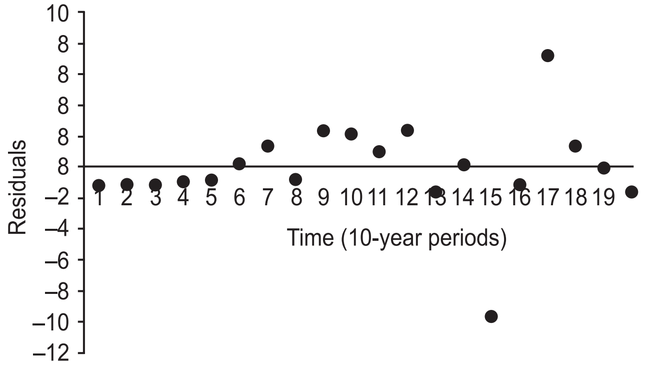

Fig. 7. Residual plot of US population differences

Fig. 7 shows the plot of residuals. There is no obvious pattern to these residuals, as opposed to the striking serial correlation apparent in Fig. 3 for the cumulative data. The Durbin-Watson test statistic is d = 1.77 (with critical value $d_{U,\,0.05}$ = 1.41). Therefore the hypothesis of independent observations is not rejected, as it was for the cumulative data, which indicates clearly that the objective of reducing serial correlation has been achieved. Hence a regression model which assumes independent errors for the first differences is appropriate for analyzing these data.

The standard errors of the estimates from (18) are again at least three times than those from (6). In this case, the ones based on first differences are obviously more appropriate statistically due to their little or no serial correlation.

5. DISCUSSION AND CONCLUSIONS

This paper considers the common case where the $C(t)$ observations are equally spaced over time. In cases where that is not so, one could use a slight variation in which the dependent variable is the rate of change, say $R(t) = D(t)/\Delta t$. Model (11) is the direct derivative of the cdf model, hence the $R(t)$ could be fitted directly to (11).

If experimenters were aware of both the cdf and pdf model approaches, they might prefer the cdf model approach because its graphs obviously tend to be smoother. However, the error estimates for the parameters, as shown in the two examples, may be based on incorrect error assumptions, which would invalidate statistical inferences on those parameter values. We thus make a case for using the pdf model as an alternative to the cdf model.

1. When there is an indication of serial correlation in the residuals from the cdf model, the residuals from the pdf model may have far less serial correlation. Statistical tests should be used to verify this assertion, and if it is correct, a data analysis based on the pdf model should be considered.
2. With little or no serial correlation in the residuals, the estimated standard errors of the parameter estimates from the pdf model are larger, and arguably more credible than those in the cdf model.
3. Model "lack-of-fit" is more apparent in the pdf than in the cdf model, making the former a better diagnostic tool for finding outlying or unusual observations which may have heuristic value and also suggest areas for future model refinements.

ACKNOWLEDGEMENT

We acknowledge helpful suggestions from Mark van Wijk.

REFERENCES

Berny, J. (1989). New concepts in deterministic growth curve forecasting. *J. Appl. Statist.,* **16**, 95-120.

Brown, D. and Rothery, P. (1993). *Models in Biology: Mathematics, Statistics and Computing*. Wiley, New York.

Carlson, T. (1913). Uber Geschwindigkeit und Grosse der Hefevermehrung in Wurse. *Biochemishe Zeitschrift,* **57**, 313-334.

Franses, P.H. (2002). Testing for residual autocorrelation in growth curve models. *Tech. Forecasting Soc. Change*, **69**, 195-204.

Glasbey, C.A. (1979). Correlated residuals in non-linear regression applied to growth data. *J. Appl. Statist.*, **28**, 251-259.

Gotelli, N.J. (2008). *A Primer of Ecology*, 4th Edition. Sinauer Associates, Sunderland, MA.

Johnson, N.L. and Kotz, S. (1970). *Continuous Univariate Models – 2*. Wiley, New York.

Kindlmann, P. (1985). A model of aphid population with age structure. In : *Mathematics in Biology and Medicine*, V. Capasso, E. Grosso and S.L. Paveri-Fontana, (eds.) Proceedings, Bari, 1983: Lecture Notes in Biomathematics, Springer, Berlin, pp. 72-77.

Krebs, C.J. (2008). *Ecology – The Experimental Analysis of Distribution and Abundance*, 6th Edition. Pearson Education Press, New York.

Lindsey, J.K. (2004). *Statistical Analysis of Stochastic Processes*. Cambridge University Press, New York.

Lipken, L. and Smith, D. (2009). Logistic growth model. *J. of Online Mathematics and Its Applications* | Logistic Growth Model. (mathdl.maa.org/mathDL).

Matis, J.H., Kiffe, T.R., Matis, T.I. and Stevenson, D.E. (2005). Nonlinear stochastic modeling of aphid population growth. *Maths. Biosci.,* **198**, 148-168.

Matis, J.H., Kiffe, T.R., Matis, T.I. and Stevenson, D.E. (2006). Application of population growth models based on cumulative size to pecan aphids. *J. Ag. Biol. Environ. Stat.,* **11**, 425-445.

Matis, J.H., Kiffe, T.R., Matis, T.I., Jackman, J.A. and Singh H. (2007). Population size models based on cumulative size, with application to aphids. *Ecol. Modelling,* **205**, 81-92.

Matis, J.H., Kiffe, T.R., van der Werf, W., Costamagna, A. C., Matis T.I. and Grant W.E. (2009). Population dynamics models based on cumulative density dependent feedback: A link to the logistic growth curve and a test for symmetry using aphid data. *Ecol. Modelling,* **220**, 1745-1751.

Matis, J.H., Al alouni U. and Matis, T.I. (2010). Use of the logistic pdf model as a general model for population growth - A case study with gypsy moth data. Manuscript.

Neter, J., Kuttner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied Linear Statistical Models*, 4th Edition. Irwin, Chicago, Il.

Prajneshu. (1998). A nonlinear statistical model for aphid population growth. *J. Ind. Soc. Agric. Res.,* **51**, 73-78.

Renshaw, E. (1991). *Modeling Biological Populations in Space and Time*. Cambridge University Press, New York.

Ross, G.J.S., Prajneshu, and Sarada, C. (2010). Reparameterization of nonlinear statistical models: A case study. *J. Appl. Statist.* (to appear).

SPSS. (2007). *SPSS 16.0 for Windows*. SPSS Inc., Chicago, IL.

Taylor, J.R. (1997). *An Introduction to Error Analysis: The Study of Uncertainty in Physical Measurements*. 2nd Edition. University Science Books, Sausalito, CA.

Thornley, J.H.M. and France, J. (2007). *Mathematical Models in Agriculture: Quantitative Methods for the Plant, Animal and Ecological Sciences*. 2nd Edition, CABI Publishing.