MULTIVARIABLE EMPIRICAL MODELING

OF ALS SYSTEMS USING POLYNOMIALS

David A. Vaccari, Ph.D.*
Stevens Institute of Technology
Hoboken, NJ 07030
Ph: 201/216-5570
Fax: 201/216-5352
Email: dvaccari@stevens-tech.edu
URL: http://attila.stevens-tech.edu/~dvaccari/
and
Julie Levri, M.S.
Stevens Institute of Technology
Hoboken, NJ 07030
Ph: 201/216-5337
Fax: 201/216-5352
Email: jlevri@stevens-tech.edu
*** Author to whom correspondence should be addressed**

***Life Support & Biosphere Science,* Vol. 6 pp. 265-271, 1999**

ABSTRACT

MULTIVARIABLE EMPIRICAL MODELING
OF ALS SYSTEMS USING POLYNOMIALS

David A. Vaccari and Julie Levri
Dept. of Civil, Environmental and Ocean Engineering
Stevens Institute of Technology
Hoboken, NJ 07030
Phone: 201/216-5570; Fax: 201/216-5352
email: dvaccari@stevens-tech.edu

Multivariable Polynomial Regression (MPR) was used to model plant motion time-series and nutrient recovery data for Advanced Life Support (ALS). MPR has capabilities similar to neural network models in terms of ability to fit multiple-input single-output nonlinear data. It has advantages over neural networks including: reduced overfitting; produces models that are more tractable for optimization, sensitivity analysis, and prediction of confidence intervals.

MPR was used to produce nonlinear polynomial time-series models predicting plant projected canopy area versus time and temperature. Temperature was found to not have a statistically significant effect. Models were developed to relate rate and extent of nutrient recovery to treatment parameters, including temperature and use of heat-pretreatment or nutrient supplementation.

These applications demonstrate MPR's capability to fill "gaps" in an integrated model of ALS. Fundamental models should be used whenever available. However, some components may require empirical modeling. Furthermore, even fundamental models often have empirical constituents. MPR models are proposed to satisfy these needs.

Keywords: Empirical, polynomial, modeling, nutrient recovery, plant motion.

Content Sentence

MULTIVARIABLE EMPIRICAL MODELING

OF ALS SYSTEMS USING POLYNOMIALS

David A. Vaccari, Julie Levri
Dept. of Civil, Environmental and Ocean Engineering
Stevens Institute of Technology
Hoboken, NJ 07030
Phone: 201/216-5570; Fax: 201/216-5352
email: dvaccari@stevens-tech.edu
URL: http://attila.stevens-tech.edu/~dvaccari/

Multivariable Polynomial Regression (MPR) was used to model plant motion time-series and nutrient recovery data for Advanced Life Support. Advantages over neural networks were described.

MULTIVARIABLE EMPIRICAL MODELING

OF ALS SYSTEMS USING POLYNOMIALS

David A. Vaccari and Julie Levri

***Life Support & Biosphere Science,* Vol. 6 pp. 265-271, 1999**

INTRODUCTION

An integrated system model consists of a number of individual subsystem models linked by their inputs and outputs. Integrated system modeling could lend critical support to the development of an advanced life support (ALS) system for long-term space exploration. Such a model could be validated through the construction of an experimental ALS system [1]. The system model could then be used to optimize design and operation of an ALS system [6]. It would be essential for troubleshooting operational problems during a mission.

Modules should be modeled using fundamental approaches whenever possible. However, many processes are too complex or not well enough understood for this approach. Biochemical processes often fall into this category. Even when fundamental process models exist, they often include empirical sub-modules. For example, fundamental plant growth models have been developed, but rely on empirical relations for photosynthetic efficiency.

However, it has been difficult to create multivariable empirical relations from experimental data. The photosynthesis model may account for CO_2 and temperature, but to add the simultaneous effects of temperature, nutrient level, and oxygen concentration would be difficult, even if the data existed.

A great variety of approaches exist for empirical process modeling. The method used in this work can be compared to several which are most widely known and applied: multilinear regression (MLR) [5], the related autoregressive moving average (ARMA) models [3], and artificial neural networks (ANN) [7]. Another approach is proposed: multivariate polynomial regression (MPR) has some advantages over the better-known methods. MPR models are multiple regression models with added terms for nonlinearity, linear and nonlinear interactions and lagged values of any variables. MPR models combine the MLR and ARMA model advantages of computational efficiency, parsimony and ability to show causative relationships, and the ANN advantages of ease of use and robust representation of complex nonlinear behaviors. MPR has several other advantages over ANNs. (1) Standard methods of statistical and other analysis such as hypothesis testing or sensitivity analysis can easily be used with it. (2) ANN models are difficult to communicate compactly. They consist of a large number of parameters and a carefully specified architecture. MPR models, on the other hand, are represented by a simple polynomial equation that is easy to communicate and to incorporate into applications.

This work applies MPR to the empirical modeling of experimental data generated by teams from the New Jersey NASA Specialized Center of Research and Training (NJ-NSCORT). The data used in this work were obtained from the Automated Plant Growth Monitoring Project and the Nutrient Recovery Project.

MULTIVARIABLE POLYNOMIAL REGRESSION

Multilinear regression models describe a dependent variable as a linear combination of two or more independent variables. For example, if *Y* is a linear function of both *Q* and *R*, the MLR model would be:

$$Y = a_0 + \mathrm{a}_1 Q + a_2 R \qquad [1]$$

The coefficients are found by a least squares procedure. If lagged values of the dependent or independent variables (values from previous intervals in a time series) can be included as independent variables, resulting in autoregressive models.

Multivariable polynomial regression (MPR) models are essentially multilinear regression models with added terms for nonlinearity, for both linear and nonlinear interactions, and for previous values of any variable. An interaction is a term involving the product of two or more variables. Addition of interaction terms to a model allows it to respond differently to a variable at different levels of another variable. An example is:

$$Y = a_0 + \mathrm{a}_1 Q + a_2 R + a_{12} Q \cdot R \qquad [2]$$

The last term on the right-hand-side is an interaction term. Because of it, in this model the slope of *Y* versus *Q* depends upon the value of *R*. Nonlinear and nonlinear interaction terms could also be added to the model. As an example, consider a dependent variable, *Y*, which depends upon three predictors, *Q*, *R* and *S*. In general:

$$Y = \Sigma\, a_i \cdot Q^{b_i} \cdot R^{c_i} \cdot S^{d_i} \qquad 0 < b_i, c_i, d_i < \mathrm{n} \qquad [3]$$

Typically, the exponents *a*, *b*, and *c* will be integers from zero to *n*, where n is the degree of the model. Other values may be used, such as negative integers or fractional values. The number of terms increases rapidly with *n* and the number of independent variables. However, the model can be made parsimonious (minimizes model complexity) by including only the terms which have coefficients which are significantly different from zero. In practice, this greatly reduces the number of terms in a model.

A stepwise algorithm has been developed to select terms for the model. To begin with, a fitting criterion, such as the sum of squares of the errors, is selected to compare models. Then, all models with only one term are tested. The term that produces the best value of the criterion is selected as the starting model. Then, the model is changed by one term at a time in successive steps by testing each possible one-term change to the model. That is, all models with one more term are tested, as well as each model with one term removed. The best of all of these becomes the new model. Ultimately, no further improvement is possible by a change in a single term. The result might not be the best model of all of the possibilities; however, in practical circumstances it is likely to be nearly so.

Various measures of the goodness-of-fit of the model may be computed and used as fitting criteria. All are based on the sum of squares of the error (*SSE*) based on the prediction errors using a dataset independent of the one used in the model identification and fitting. A measure that penalizes models for having too many terms is the mean square error (*MSE*), which is *SSE* divided by the degrees of freedom. The *MSE* is an unbiased estimate of the variance of the error. A more conservative criterion than the *MSE* is the model *F*-statistic:

$$F = \frac{(TSS - SSE) / n_p}{MSE} \qquad [4]$$

where *TSS* is the total sum of squares and n_p is the number of coefficients in the model (and represents the model degrees of freedom). Another statistic used to test individual model coefficients was the *t*-statistic for the hypothesis that the candidate term coefficient was different from zero: $t_p = a_p / \sigma_p$, where σ_p is the standard error of the coefficient a_p. (Note that t_p is related to the "*F*-to-add" criterion used in MLR, which in turn is different from the *F*-statistic given in equation 4.) In some cases the *t*-statistic was used to decide whether to add or remove a term from the model. In other cases a term was added only if two criteria were simultaneously true: (1) the *t*-statistic was significant at the 95% confidence level based on the fitting data; and (2) its addition produced an improvement in *MSE* or *F*-statistic in the test file.

When sufficient data are present, a cross-validation method can be applied. Ideally, the data are divided into three sets:

FITFILE -- Used to compute model coefficients and *t*-statistics;

TESTFILE -- Used to compute global goodness-of-fit statistics such as *MSE*, R^2, and *F*-statistic;

VALIDATION FILE -- Used as a final, independent check on how well the model predicts. This file is not used until all terms and coefficients are determined.

Use of a separate FIT and TEST files ensures that the model generalizes, that is that it does not fit only peculiarities in the fit file. This reduces overfitting and prevents “chance correlation”. When there are not sufficient data, then the same data are used for FIT and TEST. Sometimes validation may not be done. This is a “correlation” mode, used as a way of smoothing or identifying trends in the data, and not for making predictions.

MPR models are a polynomial form of the nonlinear autoregressive model with exogenous variables (NARX model) [2, 4, 8]. The MPR modeling technique was applied to plant motion and nutrient recovery data sets. The plant motion data sets were large, which allowed use of the cross-validation method. The nutrient recovery data sets had a small number of points; thus cross-validation could not be used.

PLANT MOTION MODELING METHODS

Data from the NJ-NSCORT Automated Plant Growth Monitoring Team [xxx] consisted of readings taken by machine vision of the total projected canopy area (TPCA) of tomato plants grown in a controlled environment. This project seeks to use machine vision to detect plant stress and to control automated culturing processes.

TPCA readings were taken by machine vision for five plants in each of the temperature treatments (described below) at thirty-minute intervals during the light period. In all cases, when the lights came on the plant was found to respond with a sharp, short-term increase in TPCA. Then TPCA would decrease again to a low point, followed by a slow increase. The Automated Plant Growth Monitoring Team of NJ-NSCORT desired to know if the TPCA could be predicted from environmental conditions and previous values, and to determine if treatment regime had a significant effect upon changes in TPCA. This study was undertaken to answer these questions.

Air temperature in the growth chambers was varied over three different treatments (or regimes). Five plants were grown under each regime, with TPCA measured every 30 minutes during the light period. Other physical conditions were controlled. Temperature regime 1 was the "Normal Temperature Regime", with air drybulb temperatures set at 21°C ±0.5°C during each photoperiod and 18°C ±0.5°C during each dark period. Two other regimes were used, although only regime 1 will be detailed here.

Two models were developed for predicting the TPCA of the plants in the normal temperature regime. The value of TPCA at time *t* is denoted as TPCA_t. Model A developed by regressing TPCA_t on time from the beginning of the light period, *t*, and the initial TPCA at the start of the photoperiod (TPCA_0). Model B is an autoregressive model, in which lagged variables for TPCA were added: TPCA_{t-1} and TPCA_{t-2}.

One type of model sought the effect of temperature: The data for all three temperature treatment regimes were combined into a single dataset. A coded dummy variable was defined which was a code for the temperature regime. Fitting was done using *t*, TPCA_0 and the temperature dummy variable as independent variables. For all models, data on three of the five plants in each regime were used as fitting data, a fourth plant was used as test data, and data on the fifth plant were used for validation.

The process began by fitting a simple multilinear model. Coefficients were determined by a least-squares method. Terms were added by the stepwise procedure based on *t*-statistic for the FITFILE and MSE in the TESTFILE. Terms with higher exponents and interaction terms were then successively examined for model improvement. In models A and B, exponents were restricted to the range from -4 to +4. The high polynomial degree made it possible to capture rapid changes in curvature. Keep in mind that the MPR model can be considered a truncated Taylor polynomial, which with an infinite number of terms can describe any functional relationship. The *t*-statistic criterion was also used to test whether to remove terms after each addition cycle. The *t*-statistic is less conservative than the *F*-statistic; that is, models using the former will tend to have more terms. After the best model was created using the *t*-statistic, the procedure was continued with criterion changed to “maximize F-statistic”. This much more conservative criterion usually had the effect of pruning the models, making them more parsimonious. In some cases, only linear terms remained in the model.

SOLIDS REDUCTION MODELING METHODS

A similar procedure was used to analyze data that was generated by the NJ-NSCORT Nutrient Recovery Team. Here, the purpose is correlation or fitting, and not prediction. Development of a predictive capability would require larger amounts of data. In the nutrient recovery study, fungal treatment of a slurry consisting of tomato plant leaves and stems was carried out in shaker flasks to examine the extent of biodegradation of inedible plant material. Experiments were carried out at three temperatures, and half of the samples were heat-treated before inoculation with the fungus *Phanaerochaete chrysosporium*. Four of the experiments were supplemented with nutrients. Percent total solids reduction (PTSR) was measured after 16 and 32 days.

MPR modeling was done with PTSR as dependent variable, and the independent variables were temperature (T), a dummy variable for nutrient addition (N) and a dummy variable for heat-pretreatment (H). Because of the small amount of data, no cross-validation was attempted. All the available data were used in the fitting procedure.

PLANT GROWTH MODELING RESULTS

Figures 1 and 2 show models A and B, respectively, with the data on tomato plant TPCA. Each figure shows a plot of the validation dataset, the predictions for the validation data, and the resulting errors.

Model A is:

$$\begin{aligned} TPCA_t = \quad & 1.196\ TPCA_0 - 0.232\ TPCA_0 \cdot t + 1.989 \cdot 10^4\ t^3 - 1.771 \cdot 10^4\ t^2 \\ & + 2.725 \cdot 10^3\ t + 5.893 \cdot 10^{-6}\ TPCA_0^{\ 2} \cdot t^3 \end{aligned} \qquad [5]$$

The *t*-statistics of the coefficients: 244.0, -12.0, 9.18, -8.04, 5.08, and 4.23, respectively

$R^2 = 0.989$; MSE = $1.164 \cdot 10^6$; F-Statistic = $8.07 \cdot 10^3$

Model B, the autoregressive model, is:

$$TPCA_t = +2.4235\ TPCA_{t-2} - 1.4226\ TPCA_{t-2}^{\ 2} \cdot TPCA_{t-1}^{\ -1} \qquad [6]$$

t-statistics of the coefficients: 122.6, -72.0, respectively

$R^2 = 0.999$; Mean square error = $1.573 \cdot 10^5$; F-Statistic = $1.72 \cdot 10^5$

In spite of having an R^2 which would commonly be thought to be very good, model A does not adequately predict the data. It is not capable of predicting the magnitude of the initial peak. What did significantly improve the fit was to include the lagged TPCA (model B). The MSE was reduced by about an order of magnitude when compared to A. This was not a surprise, since predictions are being made using more recent data.

The final modeling attempt on the plant motion data was for the purpose of determining if temperature variation affected TPCA. The stepwise procedure did not find a statistically significant term containing a temperature dummy variable. The models were compared using another *F*-statistic. This was computed as the ratio of their MSE values. In this case the *F*-statistic is 1.0073, with degrees of freedom 1602 for the numerator, and 1601 for the denominator. The probability of a larger *F* (the probability that the two MSE's are different by chance) is 44%. Thus we can reject the hypothesis that the two MSE's are different, and therefore we conclude that temperature does not significantly affect TPCA. This is an example of the use of standard statistical methods with MPR models.

SOLIDS REDUCTION FITTING RESULTS

Figures 3 and 4 illustrate the original data on percent total solids reduction with the fitted curves. In the models below, *T* represents temperature regime, and *H* is the dummy variable for heat-pretreatment. The variable for nutrient addition was not found to be significant for either dataset. That is, nutrient addition had no significant effect on solids reduction by fungus.

Figure 3 shows the results at 16 days of biodegradation. The resulting model is linear with respect to temperature and heat-pretreatment, as well as an interaction between both variables. Thus the sensitivity (derivative) of the model with respect to temperature depends upon heat-pretreatment. The corresponding model is:

$$\text{PTSR} = +1.657\ T + 25.261\ H - 0.7733\ T \cdot H \qquad [7]$$

H = 1 for With Heat Treatment and H = 2 for Without Heat Treatment.

t-statistics of the coefficients: 24.6, 11.9, -11.6, respectively

R^2 = 0.852, Mean square error = 9.063, F-Statistic = 21.1

Figure 4 shows the results after 32 days of biodegradation. The model shows a moderate curvilinear interaction effect. The model for PTSR after 32 days is:

$$\text{PTSR} = +80.517 - 8.8715\ H \cdot T - 310.89\ T^{-1} \cdot H^{-1} \qquad [8]$$

t-statistics of the coefficients: 13.4, -3.66, -2.48, respectively

R^2 = 0.5532; Mean square error = 8.610; F-Statistic = 4.54

The fitted curves in Figures 3 and 4 both show that heat pre-treatment significantly enhances degree of solids destruction, with the exception of low temperatures at 16 days. The effect of heat treatment is a classic interaction effect that cannot be described with linear models.

CONCLUSIONS

1. Multivariate polynomial regression is a method that, like artificial neural networks, can describe nonlinear behaviors empirically, but produces simpler and easier-to-manipulate models.

2. Modeling results show that temperature does not significantly affect plant motion under the experimental conditions.

3. MPR modeling enables the use of statistical methodology developed for linear models to be applied to nonlinear models with interaction terms.

4. Nutrient additions do not significantly improve biodegradation of inedible plant solids.

5. MPR modeling of solids reduction at 16 and 32 days shows an interaction effect that could not be described with linear models.

6. The models also suggest that heat pretreatment enhances the solids reduction for fungal treatment of inedible tomato plant biomass.

ACKNOWLEDGEMENTS

This work was funded by NASA through the New Jersey - NASA Specialized Center of Research and Training (NJ-NSCORT). The assistance of the following people in providing data and other assistance is gratefully appreciated: Peter Ling, Christos Christodoulatos, Kimberley Davies, Syamalendu Baumik, and Tsan-Liang Su.

BIOGRAPHIES

Dr. David A. Vaccari is an Associate Professor in the Department of Civil, Environmental and Ocean Engineering at Stevens Institute of Technology. He teaches and conducts research in the areas of biological and physicochemical processes for the treatment of industrial and hazardous wastes. His research specialties include mass transfer processes, and process control and modeling of the activated sludge process, especially in nonlinear time-series analysis of processes. Dr. Vaccari holds B.S., M.S. and Ph.D. degrees in environmental science, and a M.S. in chemical engineering, all from Rutgers University. He is a registered Professional Engineer in the State of New Jersey, and a member of WEF and IAWQ.

Ms. Julie Levri is a Research Assistant and Ph.D. candidate in the Department of Civil, Environmental and Ocean Engineering at Stevens Institute of Technology. Her research is in modeling of integrated systems for advanced life support systems. She has a B.E. in civil engineering from Vanderbilt University and an M.S. in civil engineering from the University of New Mexico.

REFERENCES

1. Averner, M.M.; MacElroy, R. D., The CELSS program: an overview of its structure and use of computer modelling. Controlled Ecological Life Support System - First Principal Investigators Meeting, NASA-CP-2247 Dec. 1982.

2. Bard, Y.; Lapidus, L., Nonlinear system identification. Ind. Eng. Chem. Fundam., v9, n4, p628-633; 1970.

3. Box, G.E.P.; Jenkins, G.M., Time Series Analysis, Forecasting and Control. Holden-Day, 1976.

4. Chen, S.; Billings, S.A., Representations of nonlinear systems: the NARMAX model. Int. J. Control, v49, n3, 1013-1032; 1989.

5. Draper, N.R.; Smith H., Applied Regression Analysis. John Wiley; 1966.

6. Drysdale, A.; Thomas, M.; Fresa, M.; Wheeler, R., OCAM, A CELSS modeling tool description and results", 22nd Int'l Conf. On Environmental Systems, Seattle, WA, July 1992.

7. Tang, Z.; de Almeida, C.; Fishwick ,P.A., Time series forecasting using neural networks vs. Box-Jenkins methodology. Simulation 57:5, 303-310; 1991.

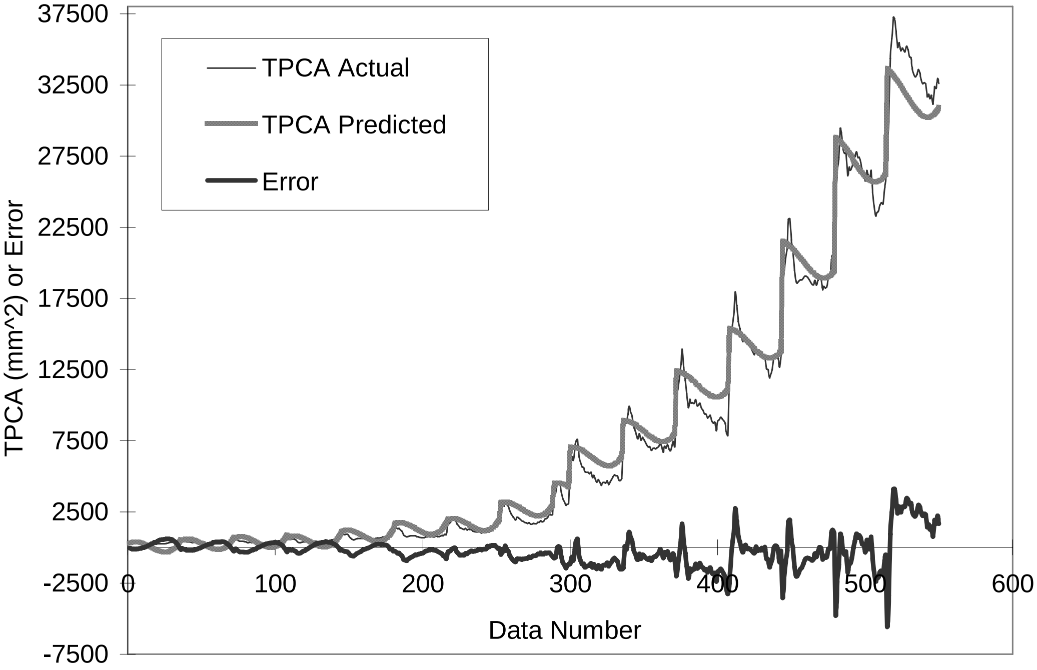
8. Wu, X.; Çinar, A., An automated knowledge-base system for nonlinear system identification. Gensym User Society Meeting, May 26-28, Cambridge, MA; 1993.

Christodoulatos, C., D.A. Vaccari, G.P. Korfiatis, S. Baumik, K. Davies, and T.-L. Su, “Nutrient Recovery and Biodegradation of Inedible Tomato plant Residues by Activated Sludge Cultures and Phanerochaete Chrysosporium,” Life Support and Biosphere Science, vol. 5 pp. 53-61 (1998).

Ling, Peter, Rutgers University Dept. of Bioresource Engineering, personal communication, (1996).

List of Figures

Figure 1. Model A, Non-autoregressive model of normal temperature regime.



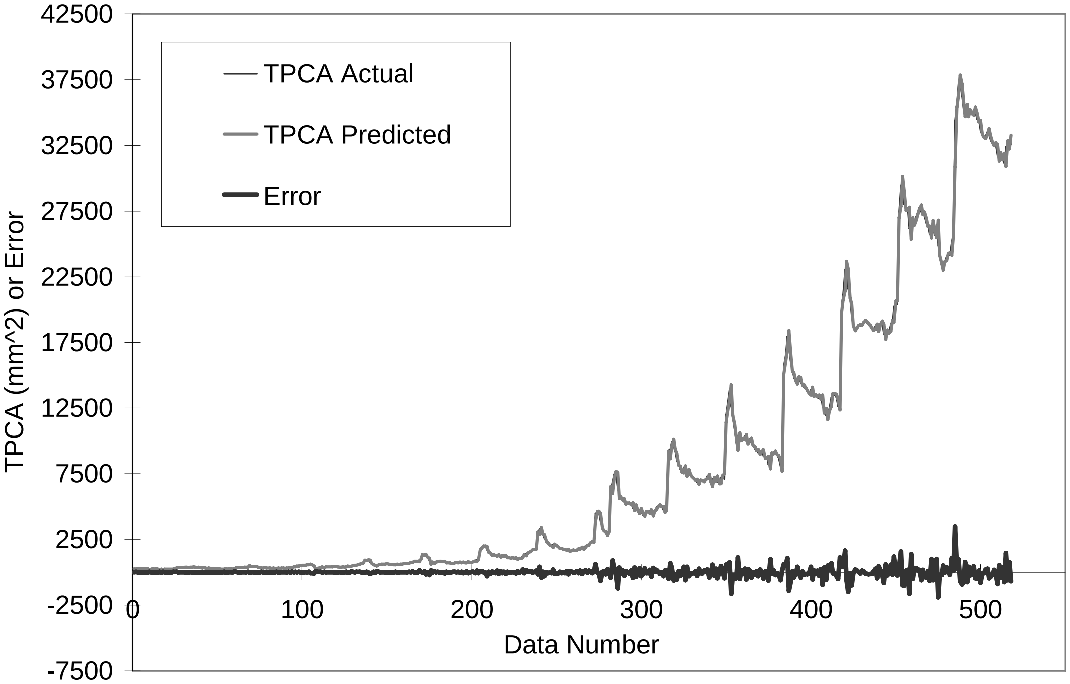


Figure 2. Model B, Autoregressive model of normal temperature regime.

Figure 3. Percent Total Solids Reduction (PTSR) in Inedible Biomass after 16 days of Fungal Treatment.

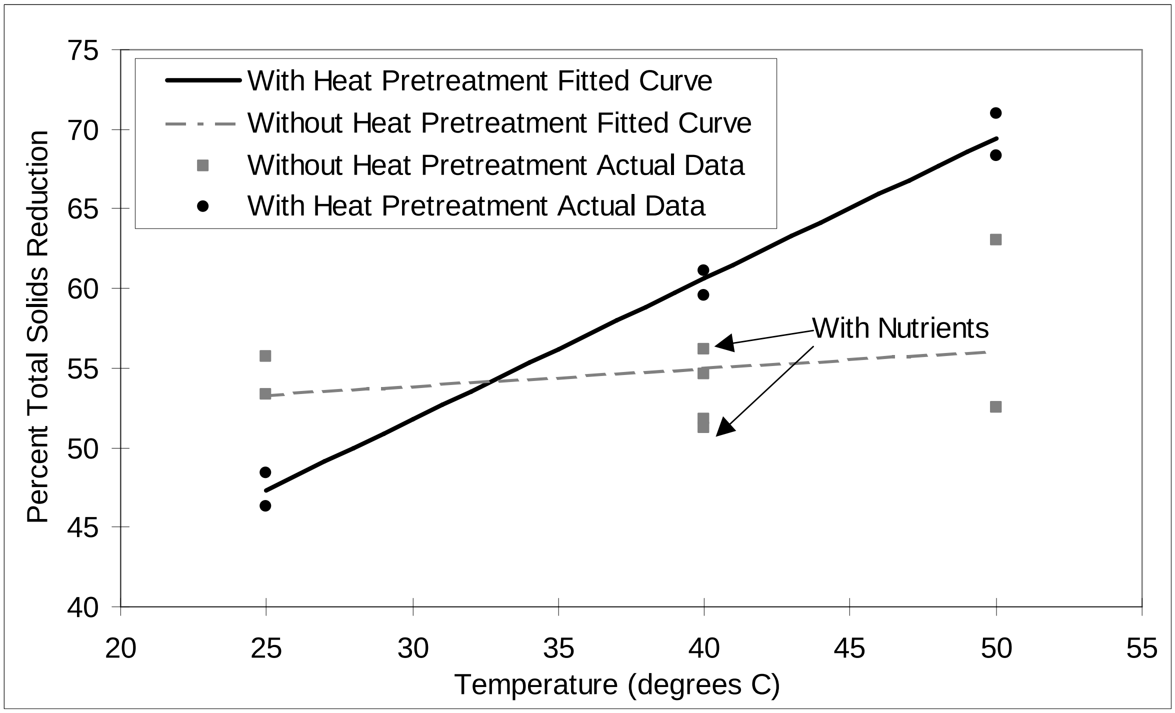
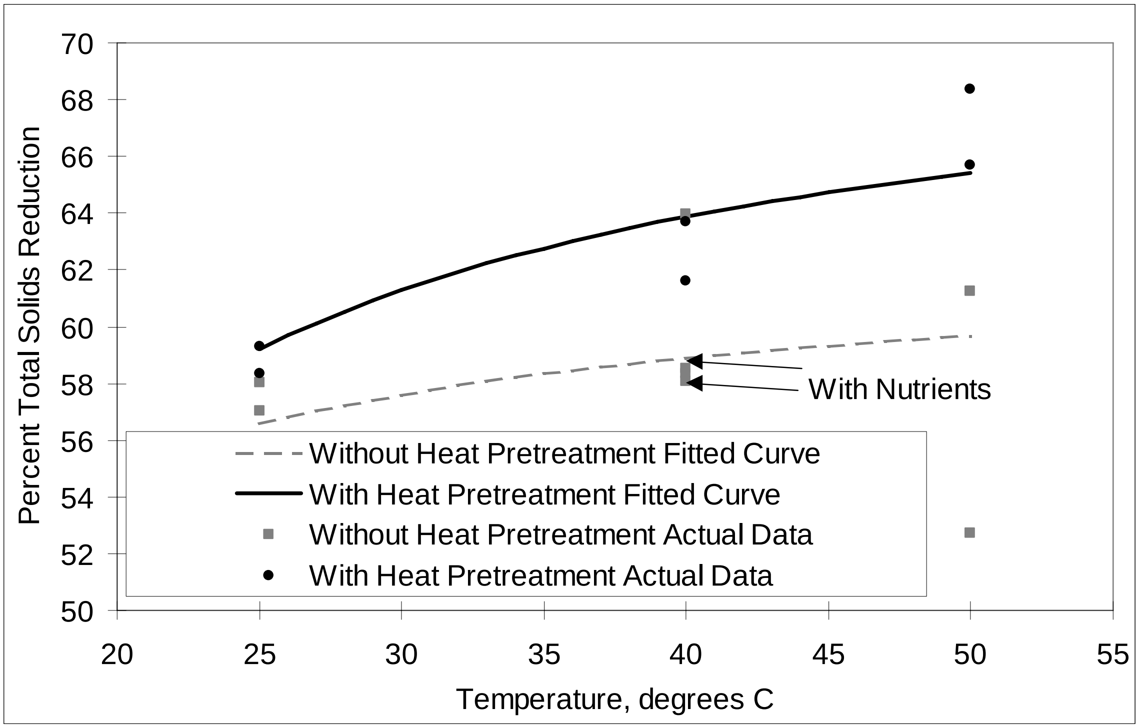


Figure 4. Percent Total Solids Reduction of Inedible Biomass after 32 days of Fungal Treatment.


Percent Total Solids Reduction
Temperature, degrees C
With Nutrients
Without Heat Pretreatment Fitted Curve
With Heat Pretreatment Fitted Curve
Without Heat Pretreatment Actual Data
With Heat Pretreatment Actual Data