Combined Analysis to Characterize Yield Pattern of Greenhouse-grown Red Sweet Peppers

Wei-Chin Lin¹ and Dietmar Frey

Pacific Agri-Food Research Centre, Agriculture and Agri-Food Canada, P.O. Box 1000, 6947 #7 Highway, Agassiz, British Columbia, Canada V0M 1A0

Gordon D. Nigh and Cheng C. Ying

British Columbia Ministry of Forests and Range, Research Branch, P.O. Box 9519, Stn. Prov. Govt., Victoria, British Columbia, Canada V8W 9C2

Additional index words. modeling, neural networks, time series, regression, *Capsicum annuum*

Abstract. Understanding the irregular yield pattern of greenhouse-grown sweet peppers (*Capsicum annuum* L.) has been a challenge to researchers and greenhouse producers. Experimental data from 4 years, each consisting of 26 production weeks, were used in a time series analysis, neural network (NN) modeling, and regression analysis. Time series analysis revealed that weekly yield was influenced by yields from the preceding 2 weeks (Yd_1 and Yd_2), cumulative light 2 and 4 weeks prior (L_2 and L_4), and average 24-h air temperature 5 weeks prior (T_5). Cumulative light (L) data were transformed into kL by dividing by 1000 for subsequent NN modeling and regression analysis. These five inputs were used to establish a NN model, which illustrated the positive influence of Yd_1, kL_4, and kL_2 and negative influence of Yd_2 and T_5. Again, these five inputs were used in a regression analysis illustrating the positive influence of Yd_1 and the negative influence of Yd_2. Each input was further modified to include its squared value before entering the regression, which resulted in significant inputs of Yd_1, Yd_1 squared, and Yd_2 squared. Among these three analyses, the most consistent parameters were Yd_1 and Yd_2, confirming that the irregular yield pattern of greenhouse-grown peppers is of a biological nature. Environmental factors kL_2, kL_4, and T_5 did not show a consistent effect on yield in all three analyses, indicating yield pattern is less influenced by growing environment.

The irregular production patterns of greenhouse-grown sweet pepper (*Capsicum annuum* L.) have imposed a serious production challenge to greenhouse producers (Heuvelink and Korner, 2001). The nature of peaks and valleys in yield is not fully understood. It was found that in some cultivars such as Derby and Meteor, three preceding weekly yields gave a good estimate of current yield by using time series analysis (Verroens et al., 2006). They also found that the inclusion of environmental parameters could further improve the model up to 20%. In a recent attempt, Schepers et al. (2006) developed a model applicable to greenhouse peppers. The model was specifically developed for outputs of known biological variables and to simulate assimilate-dependent abortion and fruit set, yet the model was not fully

validated. Many attempts have been made to reduce such ups and downs in greenhouse pepper production, but only partial success has been achieved (Heuvelink et al., 2004). A better understanding of the irregular production pattern may lead to better solutions.

Recently, neural network (NN) models have been developed to predict weekly yields using commercial data (Lin and Hill, 2008). The authors recognized that the current and preceding yields influenced future yields more than known environmental factors. Similarly, the environmental effects were low and ignored in yield prediction (Verlinden et al., 2005). When the nature of the yield pattern can be understood and the influence of environment can be assessed, then a production strategy leading to even production may be possible. In the current study, experimental data were subjected to time series analysis to reveal the nature of the correlations among weekly yields and the possible correlation of yields with light and temperatures. The significant factors identified by time series analysis were further examined by NN modeling and regression analyses to elucidate the strength of their influence on pepper yield in greenhouse production. The objectives of the study were to identify the important factors and to describe their influences on pepper weekly yield.

Plant materials and data. The experiment was conducted in a Venlo-type greenhouse of a 35-m² experimental growing area located at Agassiz, BC. Red sweet peppers of commercially popular cultivars Edison, 444, and Forever were grown according to greenhouse vegetable production guidelines (British Columbia Ministry of Agriculture, Fisheries and Food, 1996). For the 1998 ('Edison') and 2005 ('Forever') crops, 7- to 8-week-old seedlings were obtained from commercial growers and transplanted at Agassiz on 7 Jan. 1998 and 20 Dec. 2004, respectively. For the 2003 ('444') and 2004 ('Forever') crops, seeds were sown in rockwool cubes at Agassiz in early Dec. 2002 and Jan. 2004, respectively, and transplanted 3 to 4 weeks later. Two plants were transplanted into a 34-L sawdust bag. Two stems of each plant were trained at the fifth node. There was 0.8 m between the north-south-oriented rows and 0.4 m between plants in each row. There were four rows of experimental plants with 20 plants per row. No artificial lighting or CO₂ enrichment was provided. The average 24-h temperature in the greenhouse compartment was targeted at 20 to 23 °C. The greenhouse was heated when the air temperature dropped below 19 °C and vented when the air temperature rose above 24 °C. Insect pests were controlled biologically and no fungicides were used. Fruit at full color stage were harvested two to three times each week according to commercial practice. The crop ended in mid-Nov. 1998, late Oct. 2003, mid-Oct. 2004, and mid-Sept. 2005. In these 4 years, production lasted 28 to 29 weeks, but only data from the first 26 production weeks were used for analysis. Half of the 80 plants were yellow cultivar; data for these cultivars were not included because they behave differently (Lin and Hill, 2007).

Greenhouse pepper plants are commercially planted in December and produce fruit continuously in the next year in British Columbia from early April to mid-November (British Columbia Ministry of Agriculture, Fisheries and Food, 1996). A set of data consisting of weekly harvests of four 1-year experiments was arranged in three different series specifically for time series analysis, NN modeling, and regression analyses. First, preparing for time series analysis, 4 years of data were combined to form the three time series of weekly yield (Yd), light (L), and temperature (T). These records are commonly kept by commercial operators; yield is a biological factor, whereas light and temperature are environmental. Weekly yield was calculated as kg·m⁻², L was measured as cumulative light (J·cm⁻²), and T was calculated by averaging measured daily 24-h air temperatures (°C). Production Week 1 occurred on calendar Week 14, 12, 12, and 10 in 1998, 2003, 2004, and 2005, respectively. In each year, there were 26 weekly records beginning with production Week 1 and arbitrarily ending on Week 26. The set of data was used for time series analysis using SAS/

Received for publication 9 Dec. 2008. Accepted for publication 27 Jan. 2009.

Agassiz Research Centre contribution no. 784. We thank G.S. Block, Therese Otis, and Darcy Schott for experimentation and skilful assistance in preparation of the manuscript. We thank two anonymous reviewers for their comments that improved our manuscript.

¹To whom reprint requests should be addressed; e-mail linw@agr.gc.ca.

ETS software (SAS Version 9.1.3; SAS Institute Inc., Cary, NC). Second, for the NN modeling, the same set of data was arranged in such a way that a case consisted of a Yd and its associated previous yields (Yd₁, Yd₂, and Yd₃), light (L₁, L₂, L₃, L₄, L₅, and L₆), and air temperatures (T₁, T₂, and T₃). The parameters in the immediately preceding week were designated as Yd₁, L₁, and T₁; 2 weeks prior as Yd₂, L₂, and T₂; and so on. The value (L) used in the time series analysis was transformed into kL by dividing L by 1000 for the NN modeling and regression analyses. The procedure for NN modeling (2002 release; BrainMaker Professional, Nevada City, CA) was previously described (Lin and Hill, 2008). Third, for the regression analyses (SAS Version 9.1.3), the NN data set was used as it was or with transformed values of each input parameter to include squared values such as Yd₁².

Time series analysis. The Yd data tended to be serially correlated; current measures were correlated with previous measurements, and time series analysis is specifically suitable for such data (Nemec, 1996). Three time series of yield (Yd), cumulative L, and 24-h average air T on a weekly basis were used in Proc ARIMA of SAS package (SAS, 2002–2003). The Autoregressive Integrated Moving Average (ARIMA) procedure was carried out according to examples illustrated by Nemec (1996).

The preceding weekly yields were found to be correlated with current yield (Yd). Both the Yd and L series were subject to prewhitening, a part of the ARIMA procedure that eliminates the autocorrelation and trends within each time series and allows for a valid crosscorrelation between the two time series to be performed. The identical crosscorrelation procedure was repeated for Yd and T. When there was a trend in time plot, differencing by period of 1 and moving average by 1 were executed.

Neural network models. The NN modeling procedure was described previously (Lin and Hill, 2008). The NN data set was modified from that used in the time series analysis to include current yield (Yd), light (kL), and average 24-h air T. Furthermore, the preceding yields (Yd₁, Yd₂, and Yd₃), light (kL₁, kL₂, kL₃, kL₄, kL₅, and kL₆), and 24-h average air temperatures (T₁, T₂, T₃, T₄, and T₅) were combined in a complete data record called a case. The parameters (Yd₁, Yd₂, kL₂, kL₄, and T₅) identified in time series analysis as being significant were used as inputs to predict current yield (Yd) in NN modeling (California Scientific Software, 1998). Sixty percent of available cases were randomly selected for training. The remaining 40% of cases were held back for the sole purpose of testing the model. NN modeling is an iterative process (Lin and Hill, 2008). A NN model was selected among many possible models based on the highest R² and lowest average root mean square (RMS) error obtained both for the training and testing

data. Percent error was calculated as (predicted – actual)/actual.

Regression analysis. The selected inputs from time series analysis (i.e., Yd₁, Yd₂, kL₂, kL₄, and T₅) were again used as inputs for regression analysis to predict Yd on the same data set as that used in NN modeling (Table 1). To explore the nonlinear nature of inputs, these inputs were transformed to include square terms (e.g., Yd₁²). Both data sets were analyzed with Proc REG of the SAS software package (SAS 9.1.3, 2002–2003 release).

Results

Time series analysis. The yields of greenhouse peppers from the preceding 2 weeks (Yd₁ and Yd₂) were significantly correlated with current yield, having correlation coefficients of 0.73 and –0.27, respectively. The light values from the second and fourth preceding weeks (L₂ and L₄) were also significantly correlated to Yd with correlation coefficients of 0.30 and –0.21, respectively, and temperature from 5 weeks preceding (T₅) was significantly correlated with Yd having a correlation coefficient of –0.19. The resulting forecasts (predicted values) are presented in Figure 1A. The residual errors (“predicted” minus “actual” yield, Yd) were randomly distributed around zero (Fig. 1B). The fitted model can be expressed as $Y_t = 0.192 + 0.731Y_{t-1} - 0.273Y_{t-2} + \text{error}$. In short, the ARIMA procedures identified that Yd was positively correlated with Yd₁ and L₂ and negatively correlated with Yd₂, L₄, and T₅.

Neural network models. The significant parameters (i.e., Yd₁, Yd₂, kL₂, kL₄, and T₅) identified from the ARIMA procedure were used as inputs to establish a NN model. The established NN model was selected with satisfactory training (R² = 0.59, error = 13.5%, RMS error = 15.6%, n = 62) and testing (R² = 0.52, error = 13.6%, RMS error = 16.9%, n = 42) statistics. A peak production week, represented by Case 14 (Week 12 of 1998), was used as an example to elucidate the relationship between individual inputs and yield (Fig. 2A–E) because peak production weeks are important to

commercial greenhouse operators. A graphic inspection of these inputs showed a nonlinear influence on Yd. For example, Yd₁ showed a positive influence in the range of 0.18 to 0.53 (increase of 0.35) (Fig. 2A) and Yd₂ showed a negative influence of 0.49 to 0.24 (decrease of 0.25) (Fig. 2B). Both previous light values showed a positive influence; kL₂ showed a positive influence in the range from 0.26 to 0.41 (increase of 0.15) (Fig. 2C), and kL₄ positively influenced Yd in the range from 0.30 to 0.39 (increase of 0.09) (Fig. 2D). The temperature 5 weeks prior (T₅) had a negative influence on Yd of 0.42 to 0.15 (decrease of 0.27) (Fig. 2E).

Regression analysis. The selected inputs from ARIMA (i.e., Yd₁, Yd₂, kL₂, kL₄, and T₅) were again used as inputs for the regression analysis. When input values were used without any transformation, the resulting R² was 0.42 (Table 1). The significant inputs were Yd₁ (estimate = 0.679, P < 0.0001) and Yd₂ (estimate = –0.264, P = 0.0117). When the input values were transformed to include the squared term of each input, the regression model had a slightly higher R² value of 0.45 with significant parameters Yd₁ (estimate = 1.31, P < 0.0001), Yd₁² (estimate = –0.723, P = 0.0062), and Yd₂² (estimate = –0.313, P = 0.0034), but kL₂, kL₄, and T₅ were not significant.

Discussion

Time series analysis. The time series analysis revealed that Yd was positively correlated with Yd₁ and L₂ but negatively correlated with Yd₂, L₄, and T₅. This study confirmed that preceding weekly yields either positively (Yd₁) or negatively (Yd₂) affected current yield (Yd), which is similar to our previous observation that Yd and Yd₁ positively or negatively affected yield 1 week into the future (Lin and Hill, 2008). The positive influence of L₂ was similar to a previously established commercial NN model (Lin and Hill, 2008) and to that of Verlinden et al. (2005). It was unexpected, however, that L₄ would have a negative influence on Yd, which is opposite to previous observations that L₃ positively affected yield 1 week into the future (Lin and Hill, 2008).

Table 1. Linear regression model without (A) and with (B) transformed inputs.

Regression inputs	Parameter	df	Estimate	SE	t value	Pr > t ²
Original (A) ^a	Intercept	1	0.317	0.296	1.07	0.2863
	Yd ₁	1	0.679	0.0101	6.76	<0.0001
	Yd ₂	1	–0.264	0.103	–2.57	0.0117
	kL ₂	1	0.00856	0.00596	1.44	0.1539
	kL ₄	1	0.00275	0.00591	0.46	0.6430
	T ₅	1	–0.0105	0.0130	–0.81	0.4217
Transformed (B) ^a	Intercept	1	0.134	0.293	0.46	0.6474
	Yd ₁	1	1.31	0.256	5.12	<0.0001
	Yd ₁ ²	1	–0.723	0.258	–2.80	0.0062
	Yd ₂ ²	1	–0.313	0.104	–3.00	0.0034
	kL ₂	1	0.00713	0.00581	1.23	0.2222
	kL ₄	1	0.000640	0.00564	0.11	0.9099
	T ₅	1	–0.00551	0.0127	–0.43	0.6653

^aTest was done at a significance level of $\alpha = 0.05$.

^bOriginal inputs (A), R² = 0.42 (n = 104).

^cInputs with transformed terms (B), R² = 0.45 (n = 104).

Fig. 1. Autocorrelation of red pepper weekly yields (kg·m⁻²) based on $Y_t = 0.192 + 0.731Y_{t-1} - 0.273Y_{t-2} + \text{error}$: observed versus forecast in time plot (A) and residual plot (B). X-axis represents 104 weeks consisting of 26 production weeks of each of 4 years (1998, 2003, 2004, and 2005) in a time series. Y-axis is weekly yield.

The influence of temperature T₅ on yield pattern observed in this study appears to be negligible, and it is similar to the conclusion that temperature is not pertinent in a simulation model for fruit abortion and yield in sweet pepper (Wubs et al., 2006).

Time series analysis served two purposes: to provide a useful means to illustrate the correlation among yields in sweet peppers and to identify potential influencing environmental factors for further analysis in NN modeling and regression analyses. Time series analysis is a powerful tool to help researchers select significant input parameters (e.g., Yd₁, L₂, and T₅) for NN modeling and regression analysis from potential factors (i.e., yield, light and temperature) identified from prior knowledge or theory. There are two issues concerning the use of time series analysis. First, time series analysis is much more difficult to interpret than NN modeling and regression analyses. Second, time series analysis requires continuous data series that meet minimum requirements that are difficult to obtain. In this study, we were forced to combine 4 years of data in a

chronological series, because it was not effective to run a time series analysis only based on 1-year data of 26 weeks. Therefore, we limited ourselves to rely on time series analyses for the identification of significant input parameters that could be used for subsequent exploration by NN modeling and regression analysis.

Neural network models. The NN models accommodate the nonlinear relationship between inputs and output (Hoshi et al., 2000). It appears that the nonlinear nature of these inputs has been elucidated by using NN modeling, which had R^2 values of 0.59 and 0.52 in training and testing, respectively. The NN models are useful in revealing how each input influences the predicted Yd (Fig. 2A–E). The relationship between input (e.g., Yd₁) and output (Yd) in NN models was case-specific; the shape of the curve can change from one production week (case) to the next. By using experimental data, this study provided additional evidence to support our previous conclusion that NN modeling can be a practical approach for pepper yield prediction in commercial production

(Lin and Hill, 2008). The R^2 obtained in this experimental NN model is lower than the R^2 values obtained in previous commercial models (Lin and Hill, 2008). This is likely because we have restricted the model to the five significant inputs identified by time series analysis compared with the 13 inputs used in the previous commercial model. The experimental NN model may be considered to be more effective considering the fact that those inputs have been preselected by time series analysis. When NN modeling is to be used for commercial purposes, our NN model can serve as a starting point and be updated with site-specific, multiyear data but will require additional training and testing using a commercially available NN software package (e.g., BrainMaker Professional, NeuralWorks Predict[®]) as discussed by Lin and Hill (2008).

Regression analysis. One of the objectives of this study was to explore the nature of the irregular weekly yield pattern and to find how each factor may lead to such irregularity. After dissecting the series of yield, light, and temperature in time series analysis, five significant factors (i.e., Yd₁, Yd₂, kL₂, kL₄, and T₅) were identified for subsequent regression analysis involving two types of inputs from a data set. One consisted of the original values of five factors, which resulted in a regression model with an R^2 of 0.42 (Table 1). The other, which involved original and squared values of five factors, resulted in a regression model with an R^2 of 0.45 (Table 1). This analysis revealed the nonlinear nature of the inputs. For example, the changes in Yd are better described by a quadratic function of Yd₁ than by Yd₁ itself (Table 1).

Combined method. The current study illustrates the necessity and usefulness of using more than one approach to understand the nature of irregular yield pattern in greenhouse-grown sweet peppers. First, the time series analysis provided some significant inputs that other methods may have failed to detect. Second, NN models can confirm the practicality of time series analysis and be used to elucidate the nonlinear nature of the contributing inputs (Hoshi et al., 2000). Third, the nature of the inputs can be explored by fitting a regression model with and without transformed input parameters. The former inputs (e.g., Yd₁) revealed linear responses, whereas the latter (e.g., Yd₁²) showed a nonlinear response. Based on our experimental data, pepper yield (Yd) was primarily influenced by Yd₁ and Yd₂.

Summary. The current study illustrates, for the first time, how inputs can be identified by time series analysis and how each input can be examined by NN modeling and regression analysis to explain that the weekly yields are more under a biological influence (Yd₁ and Yd₂) than any environmental factor such as L₂, L₄, or T₅. This conclusion is similar to the conclusion drawn by Verlinden et al. (2005) and by our previous NN modeling (Lin and Hill, 2008). The significant influence of Yd₁ and Yd₂ is persistent through all three methods of analysis. On the other hand,

Fig. 2. A neural network (NN) model of greenhouse pepper illustrated the relation between weekly yield (Y_d) and five selected inputs: (A) yield 1 week prior, Y_{d_1} ; (B) yield 2 weeks prior, Y_{d_2} ; (C) cumulative light 2 weeks prior, kL_2 ; (D) cumulative light 4 weeks prior, kL_4 ; and (E) average air temperature 5 weeks prior, T_5 . The data of a sample week, production Week 12, 14 to 20 June 1998, was used for this simulation (see “Results”).

there was no one environmental factor consistently significant in all three analyses. For example, the positive influence of L_2 was observed in time series analysis and in NN modeling but became nonsignificant in multiple linear regression. Similarly, the negative influence of T_5 was observed in time series analysis and NN models but was not significant in regression analysis. L_4 was negative in time series but positive in NN models. A combination of conventional and nonconventional approaches appears to be necessary in elucidating the complex nature of weekly yield pattern of greenhouse-grown peppers and in modeling weekly yield of greenhouse-grown peppers. In viewing the advantages and

limitations of these three methods, NN modeling provides a practical solution to yield prediction of greenhouse-grown peppers. The proposed NN modeling can be strengthened by time series analysis in selecting proper input factors and be assisted by regression analysis in revealing the strength of each influencing factor.

Literature Cited

British Columbia Ministry of Agriculture, Fisheries and Food. 1996. Greenhouse vegetable production guide for commercial growers. 1996/97 Edition. Extension Systems Branch, BC Ministry of Agriculture, Fisheries and Food, Victoria, BC, Canada.

- California Scientific Software. 1998. BrainMaker Professional. California Scientific Software, Nevada City, CA.
- Heuvelink, E. and O. Korner. 2001. Parthenocarpic fruit growth reduces yield fluctuation and blossom-end rot in sweet pepper. *Ann. Bot. (Lond.)* 88:69–74.
- Heuvelink, E., O. Korner, and L.F.M. Marcelis. 2004. How to reduce yield fluctuations in sweet pepper? *Acta Hort.* 633:349–355.
- Hoshi, T., T. Sasaki, H. Tsutsui, T. Watanabe, and F. Tagawa. 2000. A daily harvest prediction model of cherry tomatoes by mining from past averaging data and using topological case-based modeling. *Comput. Electron. Agr.* 29:149–160.
- Lin, W.C. and B.D. Hill. 2007. Neural network modelling of fruit colour and crop variables to predict harvest dates of greenhouse-grown sweet peppers. *Can. J. Plant Sci.* 87:137–143.
- Lin, W.C. and B.D. Hill. 2008. Neural network modelling to predict weekly yields of sweet peppers in a commercial greenhouse. *Can. J. Plant Sci.* 88:531–536.
- Nemec, A.F.L. 1996. Analysis of repeated measures and time series analysis: An introduction with forestry examples. *Biom. Inf. Handb.* 6 Res. Br., B.C. Min. For., Victoria, BC. Work. Pap. 15/1996. 29 May 2008. <<http://www.for.gov.bc.ca/hfd/pubs/Docs/Wp/Wp15.pdf>>.
- SAS. 2002–2003. SAS version 9.1.3. SAS Institute Inc., Cary, NC.
- Schepers, H., W. Kromdijk, and L. van Kooten. 2006. The conveyor belt model for fruit bearing vegetables: Application to sweet pepper yield oscillations. *Acta Hort.* 718:43–50.
- Verlinden, B.E., B.M. Nicolai, C. Sauviller, and W. Baets. 2005. Bell pepper production prediction based on color development distribution, solar radiation and glass house temperature data. *Acta Hort.* 674:375–380.
- Verroens, P., B.E. Verlinden, J. Lammertyn, B. De Ketelaere, B.M. Nicolai, and C. Sauviller. 2006. Time series analysis of *Capsicum annuum* fruit production cultivated in greenhouse. *Acta Hort.* 718:97–103.
- Wubs, A.M., M.J. Bakker, E. Heuvelink, L. Hemerik, and L.F.M. Marcelis. 2006. Stochastic simulation of fruit set in sweet pepper, p. 40–50. *Proc. Plant Growth Modeling and Applications, PMA'06. Second International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications*; 13–17 Nov. 2006; Beijing, China.