Using Growing Degree Days, Agrometeorological Variables, Linear Regression, and Data Mining Methods to Help Improve Prediction of Sweetpotato Harvest Date in Louisiana

Arthur Villordon[1,4], Christopher Clark[2], Don Ferrin[2], and Don LaBonte[3]

Additional index words. *Ipomoea batatas*, heat units, phenology

Summary. **Predictive models of optimum sweetpotato (*Ipomoea batatas*) harvest in relation to growing degree days (GDD) will benefit producers and researchers by ensuring maximum yields and high quality. A GDD system has not been previously characterized for sweetpotato grown in Louisiana. We used a data set of 116 planting dates and used a combination of minimum CV, linear regression (LR), and several algorithms in a data mining (DM) mode to identify candidate methods of estimating relationships between GDD and harvest dates. These DM algorithms included neural networks, support vector machine, multivariate adaptive regression splines, regression trees, and generalized linear models. We then used candidate GDD methods along with agrometeorological variables to model US#1 yield using LR and DM methodology. A multivariable LR model with the best adjusted r^2 was based on GDD calculated using this method: maximum daily temperature (Tmax) – base temperature (B), where if Tmax > ceiling temperature [C (90 °F)], then Tmax = C, and where GDD = 0 if minimum daily temperature <60 °F. The following climate-related variables contributed to the improvement of adjusted r^2 of the LR model: mean relative humidity 20 days after transplanting (DAT), maximum air temperature 20 DAT, and maximum soil temperature 10 DAT (log 10 transformed). In the DM mode, this GDD method and the LR model also demonstrated high predictive accuracy as quantified using mean square error. Using this model, we propose to schedule test harvests at GDD = 2600. The harvest date can further be optimized by predicting US#1 yield using GDD in combination with climate-based predictor variables measured within 20 DAT.**

Optimum scheduling of the sweetpotato harvest date is essential in obtaining maximum yield of the economically important US#1 yield grade. Unlike other horticultural produce, sweetpotato storage roots will continue to gain size and weight if climatic conditions are favorable [U.S. Department of Agriculture (USDA), 2004]. Thus, roots with the highest grade of US#1 have the potential to enlarge beyond their optimal size and move into the lower grade of jumbo. On the other hand, harvesting the crop before the optimum US#1 yield is achieved generally results in a disproportionate amount of the lower grade small roots called canners relative to the premium US#1 yield grade. In Louisiana, growers generally schedule harvest dates for 'Beauregard' sweetpotato based on conducting test harvests starting at 90 d after transplanting (DAT). This is consistent with an earlier recommendation to "harvest several representative hills at regular intervals beginning when the plants are from 90 to 100 days old" (Edmond and Ammerman, 1971). However, it is not uncommon for growers to delay harvest of fields that have been planted at an earlier date due to a perceived delay in storage root sizing. Harvest scheduling is further complicated by the fact that the planting period can span 5 to 7 weeks, starting as early as 1 Apr. (southern and central Louisiana) or 1 May (north Louisiana), and ending around mid-July. These constraints are important considerations in sweetpotato harvest management practices. In other crops, the most common approach used for harvest scheduling is based on the relation of harvest date with accumulated degree days often in combination with other factors (Everaarts, 1999; Perry et al., 1997). Well-characterized degree day accumulation models and association with maturity or harvest scheduling are available for other crop species such as broccoli (*Brassica oleracea*), muskmelon (*Cucumis melo*), cucumber (*Cucumis sativus*), and cotton (*Gossypium hirsutum*) (Dufault, 1997; Jenni et al., 1998; Perry and Wehner, 1996; Viator et al., 2005). Heat unit summations or growing degree days (GDD) for vegetable production has been used for many years on crops with limited life span of quality in the field (Dufault, 1997). Heat units are also potentially useful for crops planted at different times

Portions of this paper were supported by the U.S. Department of Agriculture, by CSREES, and by RAMP Grant Award No. 370831201-02106 "Development of Grower Decision-making Tools to Reduce Risk and Enhance Sustainability of Southern Sweetpotato Pest Management Systems."

Approved for publication by the director of the Louisiana Agricultural Experiment Station as manuscript no. 2008–260–1699.

The mention of trademark, proprietary product or method, and vendor does not imply endorsement by the Louisiana State University AgCenter nor its approval to the exclusion of other suitable products or vendors.

We thank Dr. R. Viator for a thoughtful review of the manuscript. We also thank the anonymous reviewers who provided helpful critiques of the manuscript.

[1]Louisiana State University AgCenter, Sweet Potato Research Station, 130 Sweet Potato Road, Chase, LA 71324

[2]Louisiana State University AgCenter, Department of Plant Pathology and Crop Physiology, Chase, LA 71324

[3]Louisiana State University AgCenter, School for Plant, Environmental, and Soil Sciences, Chase LA 71324

[4]Corresponding author. E-mail: avillordon@agctr.lsu.edu.

Units

To convert U.S. to SI, multiply by	U.S. unit	SI unit	To convert SI to U.S., multiply by
2.54	inch(es)	cm	0.3937
0.4536	lb	kg	2.2046
1.1209	lb/acre	kg·ha^{-1}	0.8922
28.3495	oz	g	0.0353
(°F – 32) ÷ 1.8	°F	°C	(1.8 × °C) + 32

during the season (Wurr and Fellows, 1984) or under different microclimates (Wolfe et al., 1989).

Currently, we are unaware of any published GDD model for sweetpotato grown in Louisiana. Stoddard and Weir (2002) calculated GDD for California-grown sweetpotato with a base temperature (B) of 60 °F. Accumulated GDD has also been calculated for North Carolina growing conditions (Seem et al., 2003) with B = 70 °F. Neither model specified a maximum or ceiling temperature. The most suitable method and combinations of B and C for calculating GDD have been traditionally identified through two basic approaches: minimum CV (CV) (Dufault, 1997; Jenni et al., 1998) and linear regression (LR) (Stenzel et al., 2006; Viator et al., 2005). The CV method identifies candidate GDD accumulation methods through comparison of CV values from combinations of GDD methods, B, and C. The LR method identifies the candidate GDD method with the best linear fit with a character of interest like harvest date, yield, or phenological stage. Once a suitable GDD model has been identified, further regression modeling is typically performed to identify other predictor variables such as soil, climate, and other morphological variables (Jenni et al., 1996; Viator et al., 2005). Togari (1950) has previously documented that temperature in the first 20 DAT can exert significant influence on the final storage root yield.

Recently, Clapham and Fedders (2004) compared LR and neural networks (NN) in modeling vegetative development of berseem clover (*Trifolium alexandrinum*) as a function of GDD and concluded that NN were preferred when a priori knowledge of temperature thresholds was not available. NN, along with other adaptive and nonparametric methods are increasingly being used in agricultural research for predictive purposes. An important characteristic of these techniques is their adaptive nature with regard to learning by examples to solve problems (Park et al., 2005). These methods also include regression tree (RT) (Yang et al., 2003), support vector machine (SVM) (Maenhout et al., 2007), multivariate adaptive regression splines (MARS) (Turpin et al., 2005), and generalized linear models (GLM) (Benjamini and Leshno, 2005). Several of these methods are being used in data mining (DM) applications. DM involves the use of algorithms that explore data, develop models, and discover previously unknown patterns (Maimon and Rokach, 2005). DM approaches are increasingly being used in an agricultural context (Bui et al., 2006; Ekasingh, et al., 2005). DM is also considered the core of "knowledge discovery in databases" (KDD), an automatic, exploratory analysis and modeling of large data repositories (Maimon and Rokach, 2005). The DM modeling approach typically partitions a database into training (TRD) and testing (TED) data sets. Models are developed using the training partition and predictive accuracy is calculated using the testing data set. Depending upon the implementation of the DM software, a third partition is also created [i.e., validation data set (VAD)]. VAD is typically used to prevent overfitting during model development. Many of these "machine-driven" algorithms can reduce subjectivity and information loss due to data transformations to meet traditional parametric assumptions (Turpin et al., 2005). Previous work that used NN and RT to predict aflatoxin in peanut [*Arachis hypogaea* (Henderson et al., 2000)] and ending irrigation in cotton (Tronstad et al., 2003), respectively, provide examples of the use of DM methodology in developing GDD-based models.

Our study sought to identify the appropriate accumulated GDD model for scheduling the harvest of 'Beauregard' sweetpotato grown in Louisiana. We also sought to assess the feasibility of using accumulated GDD along with agrometeorological variables to improve harvest date prediction for sweetpotato based on a target yield using LR and DM approaches.

Materials and methods

DATA. Yield data from 118 planting dates spanning the years from 2002 to 2007 were compiled into a single database (GDDLA-Y). Storage roots were graded according to USDA standards (USDA, 2005): US #1 grade = 2 to 3-1/2 inches in diameter, 3 to 9 inches in length, maximum weight not more than 20 oz; canner = 1 to 2 inches in diameter, 2 to 7 inches in length; jumbo = larger versus the others, but marketable. These tests represented various planting times (May to July), cultural practices (irrigation regimes, weed control, and cropping patterns), management regimes (experimental station and on-farm replicated trials), and locations [northern and south central Louisiana (Table 1)]. The yield data were obtained from plots that were planted with 'Beauregard' and no significant year × US#1 yield interaction was detected. Yield in the US #1 grade ranged from 38 to 617 50-lb bushels per acre (mean = 306 bushels/acre). Days to harvest ranged from 83 to 166 DAT.

SAMPLING. Agrometeorological data were collected from the following Louisiana Agriclimatic Information System (LAIS) weather network stations (Louisiana Agriclimatic Information, 2008): Burden Center, Baton Rouge (BTR), R & D Research Farm, Port Barre (SC), Sweet Potato Research Station, Chase (CHS, NE10), and University of Louisiana at Monroe (NE). Values enclosed in parentheses represent location of test sites as defined in Table 1. Daily agrometeorological variables included maximum air temperature (MAXAIR), minimum air temperature (MINAIR), mean air temperature (MEANAIR), maximum soil temperature (MAXSOIL), minimum soil temperature (MINSOIL), mean soil temperature (MEANSOIL), radiation (RAD), maximum relative humidity (MAXRH), minimum relative humidity (MINRH), mean RH (MEANRH), and total rainfall (RAIN). Means (MAXAIR, MINAIR, MEANAIR, MAXSOIL, MINSOIL, MEANSOIL, RAD, MAXRH, MINRH, and MEANRH) or totals (RAIN) were calculated for the following periods: 5, 10, 20, and 30 DAT for each planting date. Normality analysis was performed on the combined yield and agrometeorological data set using SAS Analyst (version 9.2; SAS Institute, Cary, NC). Outliers were identified using the "filter outliers" node in SAS Enterprise Miner (version 4.5). Only the following climate-related variables met the Kolmogorov-Smirnov (K-S) test for normality: MAXSOIL 5 DAT (MAXSOIL5), RH 5 DAT (RH5, log 10 transformed), MINSOIL 30 DAT (MINSOIL30), MINRH30, MAXSOIL10 (log 10 transformed), MEANSOIL30, MAXAIR20, MINSOIL20, MINRH10, RH10, MINRH20, and

RH20. Following tests for normality, we performed correlation and variance inflation factor analyses in SPSS (version 15; SPSS Inc., Chicago) to test for variable independence. The following variables were identified to be independent and were used in all subsequent multiple variable LR experiments (forward stepwise selection, criterion for inclusion $P = 0.05$): RH20, MAXSOIL10, and MAXAIR20. Subsequent modeling experiments were performed on this reduced database, GDDLA-YMET (n = 116). RH20 ranged from 69% to 92% (mean = 79%, SD = 5.5), MAXSOIL10 ranged from 76 to 107 °F (mean = 89 °F, SD = 6.75), and MAXAIR20 ranged from 83 to 94 °F (mean = 89 °F, SD = 2.11).

Methods for calculating accumulated GDD. Eight methods for calculating GDD were used (Table 2). The following base (B) temperatures were used: 60, 65, and 70 °F. The following ceiling (C) temperatures were used: 80, 85, 90, 95, and 100 °F. For the purpose of this study, we used this notation (xx-XX) where xx = B and XX = C. GDD coefficients of variation were calculated using all combinations of methods (M), B, and C for each planting × harvest date (PH) combination.

LR modeling. Single-variable LR analysis (criterion for inclusion $P = 0.05$) was performed using US#1 as the dependent variable (DV) and GDD as the predictor variable. Subsequent multiple LR experiments (forward stepwise selection, criterion for inclusion $P = 0.05$) were performed using GDD, RH20, MAXSOIL10, and MAXAIR20 as predictor variables. SAS Analyst (version 9.2) was used to run the LR modeling experiments. Partial residual plots were generated using Statistica (version 8; Statsoft, Tulsa, OK).

DM methods. Insightful Data Miner (version 8; Insightul Corp., Seattle) was used to randomly generate five unique sets (five pseudoreplications) of training and testing data partitions from GDDLA-YMET. Each pseudoreplication was generated by specifying a unique number to "seed" the random sampling-based partitioning process. The proportion of training to testing data was 50%:50% (50% training:50% testing). Statistica Data Miner (version 8, Statsoft) was used to develop models from the training partition and measure predictive accuracy on testing partition. An overview of the DM-based model development ("DM mode") and testing is summarized in Fig. 1. The following algorithms were used: LR, SVM, NN, RT, MARS, and GLM. In most cases, the default software settings were used performing the DM algorithms. SVM: regression type 1, kernel = RBF, v-fold cross validation = 10, training = 1000 iterations. NN: training sample size = 80%; network = MLP, maximum hidden units = 13; minimum hidden units = 4; hidden neurons = identity, logistic, Tanh; output neurons = identity, logistic, Tanh. RT: stopping option for pruning = prune on variance, minimum n per node = 5, maximum number of nodes = 1000, v-fold cross validation = 10,

Table 1. Location of experimental test sites for 'Beauregard' sweetpotato in Louisiana.

Test site location in Louisiana	North latitude	West longitude	Soil taxonomic class
Baton Rouge (BTR)	30°24′26.9994″	91°8′44.9982″	Fine-silty, mixed, superactive, thermic Aeric Epiaqualfs
Chase (CHS)	32°5′43.08″	91°42′21.2394″	Fine-silty, mixed, active, thermic Typic Glossaqualfs
Northeast #5 location (NE5)	32°54′3.9594″	91°21′24.48″	Fine-silty, mixed, active, thermic Oxyaquic Fraglossudalfs-Fine-silty, mixed, active, thermic Typic Glossaqualfs
Northeast #7 location (NE7)	32°56′49.92″	91°18′14.4″	Fine-silty, mixed, active, thermic Typic Hapludalfs
Northeast #8 location (NE8)	32°52′36.1194″	91°20′26.5194″	Fine-silty, mixed, active, thermic Oxyaquic Fraglossudalfs-Fine-silty, mixed, active, thermic Typic Glossaqualfs
Northeast #9 location (NE9)	32°59′8.1594″	91°17′38.76″	Fine-silty, mixed, active, thermic Oxyaquic Fraglossudalfs-Fine-silty, mixed, active, thermic Typic Glossaqualfs
Northeast #10 location (NE10	32°0′43.5594″	91°38′45.2394″	Fine-silty, mixed, active, thermic Aquic Fraglossudalfs
Northeast #12 location (NE12)	32°49′59.9874″	91°39′35.9994″	Very-fine, smectitic, thermic Chromic Epiaquerts
South-central #1 location (SC1)	31°4′31.0794″	92°2′46.3194″	Fine-silty, mixed, active, thermic Typic Glossaqualfs
South-central #2 location (SC2)	31°3′32.76″	92°2′24 ″	Fine-silty, mixed, active, thermic Oxyaquic Fragiudalfs
South-central #3 location (SC3)	30°41′47.76″	92°9′18 ″	Fine-silty, mixed, superactive, thermic Aeric Epiaqualfs
South-central #4 location (SC4)	30°26′54.9594″	92°12′38.88″	Fine-silty, mixed, active, thermic Typic Hapludalfs
South-central #6 location (SC6)	31°11′18.96″	92°3′51.84″	Coarse-silty, mixed, superactive, nonacid, thermic Typic Udifluvents
South-central #11 location (SC11)	31°7′50.9982″	92°6′24.9978″	Fine-silty, mixed, superactive, thermic Aeric Epiaqualfs

Table 2. Methods used to calculate accumulated growing degree days (GDD) for 'Beauregard' sweetpotato grown in Louisiana.

Method[z]	Description[y]
M1	[((Tmax + tmin)/2) – B] where if Tmin < 0, then GDD = 0.
M2	(Tmax – B) where if Tmin < 0, then GDD = 0.
M3	If Tmax > C, then Tmax = C and GDD = [((Tmax + Tmin/2) – B] where if Tmin < 0, then GDD = 0, or if Tmax ≤ C, then use M1.
M4	If Tmax > C, then Tmax = C and GDD = (Tmax – B) where if Tmin < 0, then GDD = 0 or if Tmax ≤ C, then use M2.
M5	If Tmax > C, then Tmax adj = C – (Tmax – C) and GDD = [((Tmax adj + Tmin)/2) – B] where if Tmin < 0 then GDD = 0 or If Tmax ≤ C, then use Equation M1.
M6	If Tmax > C, then Tmax adj = C – (Tmax – C) and GDD = (T max adj – B) where if Tmin < 0, then GDD = 0 or if Tmax ≤ C, then use M2.
TRI[x]	Calculations were performed using DEGDAY[w]
SINE[v]	Calculations were performed using DEGDAY

[z]Methods M1 to M6 were based on Dufault (1997), Jenni et al., (1996), and Perry et al., (1986).
[y]Tmax = maximum daily temperature, Tmin = minimum daily temperature, B = base temperature, C = ceiling temperature.
[x]TRI = single triangle method. This method used daily Tmin and Tmax to produce an equilateral triangle over a 24-h period. GDD were estimated by calculating the area between the two thresholds that is enclosed by the triangle (Zalom et al., 1983).
[w]DEGDAY is a spreadsheet used for calculating GDD (Snyder, 2005).
[v]SINE = single sine method. This technique used daily Tmin and Tmax to produce a sine curve over a 24-h period. GDD were estimated by calculating the area above the threshold and below the curve (Zalom et al., 1983).

standard error rule = 1. MARS: maximum number of basis functions = 21, degree of interactions = 1. GLM: model building method = forward stepwise, stepwise selection criterion = probability, criterion for best subset selection = r^2. Statistica Data Miner automatically generated and exported the "trained" model, applied it to the testing data set, and generated the measurements of predictive accuracy (Fig. 1). Several measurements of predictive accuracy were generated, but we only used mean square error (MSE) for comparing model performance where

$$\text{MSE} = \sum_{i=1}^{N} (E_i - O_i)^2/(N-1)$$

where E_i = predicted value of case i, O_i = observed value of case i, and N = number of observations. The exported model was in a predictive markup modeling language (PMML) format (Data Mining Group, 2008). To investigate the effect of the training sample size on model performance, new paired training and data sets were generated from GDDLAYMET (5 pseudoreplications) with the following proportion of training and testing data: 50%:50%, 70%:30%, and 90%:10%. DM experiments were performed with each set of training and testing data, along with evaluation of prediction accuracy.

Results

M1 [(Tmax + tmin)/2) – B] has been considered as the standard

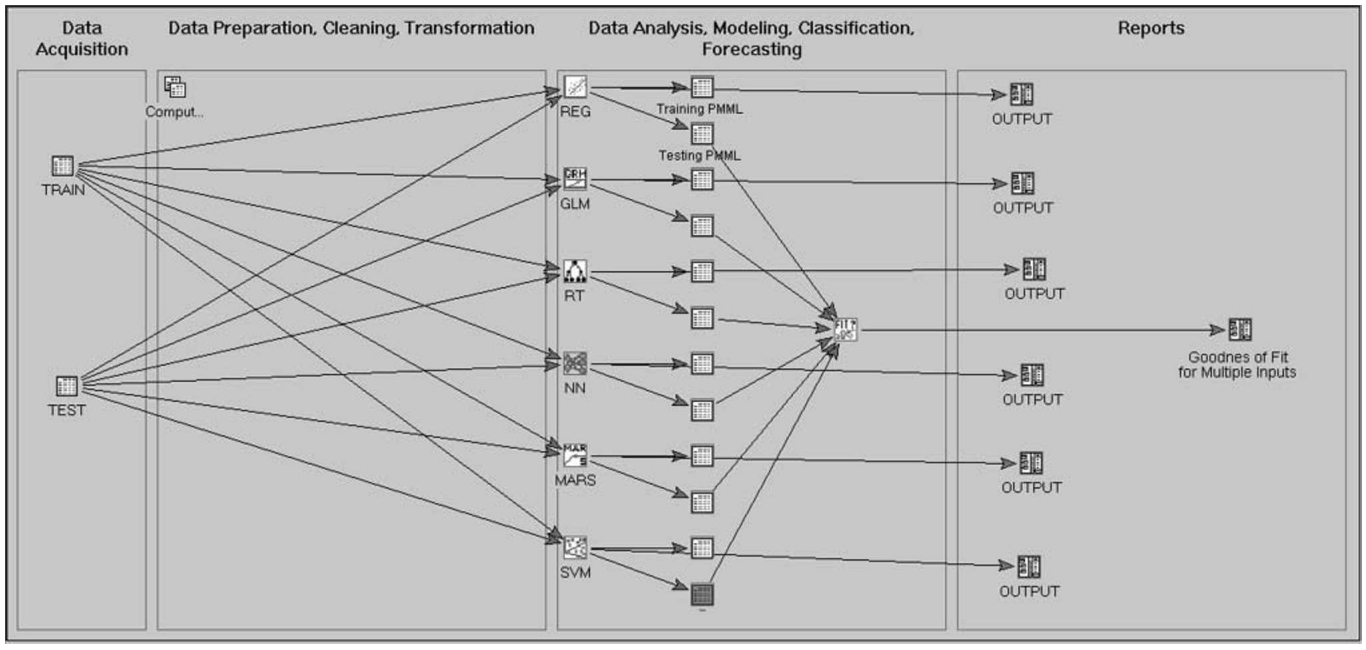

Fig. 1. Overview of the data mining process in Statistica Data Miner (version 8; Statsoft, Tulsa, OK). TRAIN = training data set, TEST = testing data set, REG = linear regression, GLM = generalized linear model, RT = regression tree, NN = neural networks, MARS = multivariate adaptive regression, SVM = support vector machine, FIT = calculation of model accuracy, OUTPUT nodes = output for each algorithm, Goodness of Fit for Multiple Inputs = calculation of multiple mean square error for each algorithm. These interconnected nodes represented one replication using the specific pair of TRAIN and TEST. These steps were repeated for n unique pairs of TRAIN and TEST.

Table 3. Coefficients of variation (normal font), adjusted r^2 (bold), and mean square errors (*italics*) for eight methods of calculating accumulated growing degree days (GDD) from transplanting to harvest in 'Beauregard' sweetpotato grown in Louisiana.

		GDD method[y]							
Base (°F)[z]	Ceiling (°F)	M1	M2	M3	M4	M5	M6	TRI	SINE
60	nc[x]	17.35[w] **0.04** *16298 S*[v]	13.85 **0.10** *15489 S*						
	80			18.41 **0.02** *17194 S*	15.67 **0.12**[u] *15518 S*	25.92 NS[t] *17665 G*	25.77 NS *17548 M*	12.51 **0.06** *16533 R*	12.51 **0.06** *16604 R*
	85			17.04 **0.03** *16540 S*	13.22[s] **0.13** *15224 G, R*	19.11 **0.02** *17390 R*	15.87 **0.04** *17194 R*	12.58 **0.06** *16499 S*	12.48 **0.06** *16411 S*
	90			16.85 **0.04** *17470 R*	12.99 **0.11** *16038*	16.99 **0.02** *17694 R*	13.42 **0.06** *16968 R*	12.8 **0.06** *17009 R*	12.78 **0.06** *16997 R*
	95			17.08 **0.04** *17028 R*	13.56 **0.09** *16048 S*	16.88 **0.03** *17125 R*	13.49 **0.07** *16399 R*	12.99 **0.06** *16913 R*	13.11 **0.06** *16866 R*
	100			17.3 **0.04** *17027 S*	13.8 **0.10** *17737 T*	17.26 **0.04** *17037 S*	13.77 **0.09** *17737 T*	13.04 **0.06** *17711 S*	13.11 **0.07** *17737 T*
65	nc	21.14 **0.02** *17313 R*	15.36 **0.07** *16448 R*						
	80			24.42 NS *17517 R*	17.63 **0.10** *15879 S*	49.47 NS *17737 G,R,T*	37.85 NS *17722 S*	12.63 **0.04** *16985 S*	12.61 **0.04** *16959 S*
	85			21.34 NS *15799 R*	13.98 **0.10** *16643 S*	26.39 NS *16017 R*	17.68 **0.01** *18843 R*	12.9 **0.04** *15298 R*	12.61 **0.04** *18025 S*
	90			20.94 NS *15794 R*	14.67 **0.07** *18065 R*	21.3 NS *15939 R*	14.49 **0.03** *18748 R*	13.34 **0.04** *15266 R*	13.14 **0.04** *18378 S*
	95			20.79 **0.01** *15750 R*	15.02 **0.06** *17675 S*	20.54 **ns** *15844 R*	14.89 **0.04** *18219 S*	13.7 **0.04** *15208 R*	13.74 **0.04** *16782 S*
	100			21.05 **0.01** *17345 R*	15.27 **0.06** *16542 R*	20.98 **0.01** *17377 R*	15.2 **0.06** *16638 R*	13.8 **0.04** *16962 R*	13.8 **0.05** *16972 S*

(*Continued on next page*)

Table 3. (*Continued*) Coefficients of variation (normal font), adjusted r^2 (bold), and mean square errors (*italics*) for eight methods of calculating accumulated growing degree days (GDD) from transplanting to harvest in 'Beauregard' sweetpotato grown in Louisiana.

Base (°F)[z]	Ceiling (°F)	GDD method[y]							
		M1	M2	M3	M4	M5	M6	TRI	SINE
70	nc	35.15	23.49						
		NS	NS						
		17736 G	*17736 G*						
	80			42.95	22.35	30.54	54.45	13.04	12.97
				NS	**0.09**	NS	NS	**0.02**	**0.01**
				17736 G	*17744 M*	*17736 G*	*17524 M*	*17505 R*	*17524 M*
	85			32.1	15.2	52.71	20.58	13.68	12.92
				NS	**0.10**	NS	NS	**0.02**	**0.02**
				17736G	*15797R*	*17736 G*	*17736 G*	*17133 S*	*17233 S*
	90			30.65	18.65	32.8	16.88	14.57	13.95
				NS	**0.04**	**0.04**	NS	NS	**0.02**
				17737 G	*16910 S*	*17737 G*	*17317 S*	*16895 S*	*17363 S*
	95			33.5	21.56	33.2	20.81	15.25	15.14
				NS	NS	NS	NS	**0.03**	**0.02**
				17863 R	*17737 G*	*17602 T*	*17737 G*	*16790 S*	*17101 S*
	100			34.81	23.02	34.66	22.78	15.45	15.34
				NS	NS	NS	NS	**0.02**	**0.03**
				17737 G	*17737 G*	*17737 G*	*17737 G*	*16967 S*	*17393 R*

[z](°F – 32)/1.8 = °C.
[y]M1 to M6, TRI, and SINE as defined in Table 2.
[x]No ceiling temperature specified.
[w]Mean square error (MSE) method of calculation as described in Materials and methods; calculations were performed using Statistica Data Miner (version 8; Statsoft, Tulsa, OK). DM MSE values represent average of five experimental runs using five unique pairs of randomly sampled training (n = 58) and testing (n = 58) data partitions.
[v]R = linear regression, S = support vector machine, G = generalized linear model, M = multivariate adaptive regression; calculations were performed using Statistica Data Miner (version 8). Details of calculations are described in Materials and methods.
[u]Underlined values represent three best values for each method.
[t]None of the variables were entered into the model at $P = 0.05$.
[s]Similar MSE values were calculated for G and R.

method for calculating GDD and is frequently used for identifying alternative GDD models (Dufault, 1997). The CV values calculated for each M × C × B × PH ranged from 12.48 (SINE, 60–85) to 54.45 (M6, 70–80) (Table 3). The lowest CV for M1 was 17.35 (B = 60 °F). The adjusted r^2 values of single-variable LR models ranged from 0.01 (M5, 65–100) to 0.13 (M4, 60–85) (Table 3). The lowest overall MSE using the DM approach was 15208 calculated using REG (TRI, 65–95). The lowest observed MSE for M1 was 16298 (SINE). To investigate if further improvement in model accuracy (increased adjusted r^2, decreased MSE) was possible, we used SINE (60–80, 60–85), M4 (60–85, 60–80, 60–90), and TRI (60–80, 65–95) in multiple LR and DM experiments that included agrometeorological predictor variables (Table 4). These methods and combinations of B × C represented the three best-performing models in each approach (i.e., minimum CV, LR, and DM). Multiple variable LR models showed an improvement in adjusted r^2 values when agroclimatic variables (MAXSOIL10 log 10 transformed, RH20, and MAXAIR20) were used along with candidate GDD methods (Table 4). The M4 (60–90)-based multivariable model ranked first in terms of adjusted r^2 value at 0.42, had the lowest MSE (DM mode), and represents the best candidate for predicting US#1 yield for the current data set. M4 (60–85) was ranked second in LR and DM.

Partial residual plots summarized the effect of each predictor variable after factoring out the effects of other covariables (Fig. 2). In DM mode, SVM (M4, 60–90), REG (M4, 60–90), and REG (SINE, 60–85) were the best-performing models for predicting US#1 yield using GDD and agrometeorological variables (Table 5). RT and NN models were ranked near the bottom in terms of predictive accuracy. Increasing the size of the training data set (with a consequent decrease in the size of the testing data set) did not result in significant increase of predictive accuracy for any model except for NN (Fig. 3). While the other models showed a slight and gradual decrease in MSE, the NN MSE showed instability over the various training sample sizes [i.e.,

Table 4. Linear regression equations, adjusted r^2 values, and mean square errors (MSE) of data mining models describing US#1 yield based on various growing degree day (GDD) methods and agrometeorological factors for 'Beauregard' sweetpotato grown in Louisiana.

Base (°F)[z]	Ceiling (°F)	GDD[y] method	Adjusted r^2	MSE[x]	Linear regression equations
60	90	M4	0.42	10085	Y = 548.71 + 0.14 M4 – 1333.34 MAXSOIL10 + 14.50 MAXAIR20 + 8.10 RH20
60	85	M4	0.40	10476	Y = 594.95 + 0.14 M4 – 1304.10 MAXSOIL10 + 14.80 MAXAIR20 + 6.89 RH20
60	85	SINE	0.38	10557	Y = 786.35 + 0.19 SINE – 1434.81 MAXSOIL10 + 14.00 MAXAIR20 + 8.42 RH20
60	80	TRI	0.37	10569	Y = 721.29 + 0.21 TRI – 1425.47 MAXSOIL10 + 14.57 MAXAIR20 + 8.37 RH20
60	80	SINE	0.37	10574	Y = 766.81 + 0.21 SINE – 1439.97 MAXSOIL10 + 14.46 MAXAIR20 + 8.28 RH20
65	95	TRI	0.37	10754	Y = 1121.24 + 0.19 TRI – 1436.49 MAXSOIL10 + 8.78 RH20 + 10.61 MAXAIR20
60	80	M4	0.36	11198	Y = 936.18 + 0.12 M4 – 1317.44 MAXSOIL10 + 13.15 MAXAIR20 + 6.19 RH20

[z](°F – 32)/1.8 = °C.
[y]M1 to M6, TRI, and SINE as defined in Table 2.
[x]MSE of linear regression (LR) model in data mining (DM) mode. MSE method of calculation as described in Materials and methods; calculations were performed using Statistica Data Miner (version 8; Statsoft, Tulsa, OK). In DM mode, LR MSE values represent average of five experimental runs using five unique pairs of randomly sampled training (n = 58) and testing (n = 58) data partitions.
[w]Variable selection used: forward stepwise linear regression performed with $P = 0.05$ as criterion for inclusion. Calculations performed in SAS Analyst (version 9.2; SAS Institute, Cary, NC). Modeling database = GDDLA – YMET (n = 116), as described in Materials and methods. MAXSOIL10 = mean maximum soil temperature 10 d after transplanting (DAT); RH20 = mean relative humidity 20 DAT; MAXAIR20 = mean maximum air temperature 20 DAT.

MSE increased from 50%:50% to 70%:30%, and then decreased at 90%:10% (Fig. 3)]. The relationship between GDD, harvest dates, climatic variables, and actual and predicted US#1 yield for 2007 field trials are presented in Table 6. The accumulated GDD at harvest ranged from 2774 (214.97 bushels/acre) to 3389 (477.05 bushels/acre). Based on yield and GDD estimates of GDDLA-YMET, GDD = 2600 appears to be the suitable GDD-based scheduling of test harvest. In GDDLA-YMET, nine of 10 harvest dates before GDD = 2600 resulted in US#1 yield below 300 bushels/acre (data not shown). This yield level has been suggested by growers in Louisiana as the "break even point" (K. Thornhill, personal communication).

Discussion

Our results indicate that alternative methods of calculating accumulated GDD lowered CVs, increased LR-adjusted r^2 values, and reduced MSE relative to M1. Several candidate models were identified using multiple approaches (i.e., minimum CV, LR, and DM). Inclusion of agrometeorological predictor variables in single-variable GDD models increased model goodness-of-fit (as measured by adjusted r^2) and predictive ability (as measured by MSE) and helped to identify the best candidate for calculating GDD. The use of DM methods helped to quantify predictive accuracy of the candidate models. MSE is generally considered the most robust measure of overall predictive model performance (Schwartz et al., 1997). Our results are consistent with previous reports that heat units alone cannot explain the entire development of a crop (Arnold, 1960; Perry and Wehner, 1996). Our data also suggested a ceiling temperature of 90 °F for computing accumulated GDD for 'Beauregard' sweetpotato grown in Louisiana. Using a different method for calculating accumulated GDD may introduce more variation, thereby reducing prediction precision (Dufault, 1997). Regardless of the calculation method, Higley et al. (1986) emphasized that degree days are never more than estimates of developmental time. Although Arnold (1959) has suggested that a base temperature for a linear heat unit system may not be identical with all of the physiological requirements of the plant, the empirically derived base and ceiling values were consistent with known optimal temperature ranges in sweetpotato growth and yield. For example, in a review of the effect of atmospheric and soil factors that influenced sweetpotato growth and yield, Ravi and Indira (1999) noted that night air temperature below 59 °F suppressed storage root formation while promoting shoot growth. At air temperatures greater than 86 °F, an increase in indole acetic acid oxidase activity caused reduction in storage root formation and growth (Ravi and Indira, 1999). Taking all of these into consideration, using a different method for calculating GDD for 'Beauregard' sweetpotato in Louisiana might lead to less precise scheduling of time-critical management activities (e.g., test harvest and harvest). For example, prevailing temperatures are generally lower in "early" planting dates in May [5-year mean Tmax = 84.1 °F, mean minimum temperature (Tmin) = 62.4 °F] compared with "late" planting dates in July (5-year mean Tmax = 88.4 °F, mean Tmin = 70.5 °F) in two northeast Louisiana locations (CHS, NE10). If M1 or M2 (both methods do not require a ceiling temperature) were used to calculate GDD, an earlier harvest date for "late" planted sweetpotato would have been predicted due to rapid accumulation of GDD associated with higher Tmax compared with the "early" planting date. When the negative effect of high temperature (greater than 86 °F) on storage root initiation is considered (Ravi and Indira, 1999), it is apparent that rapid accumulation of

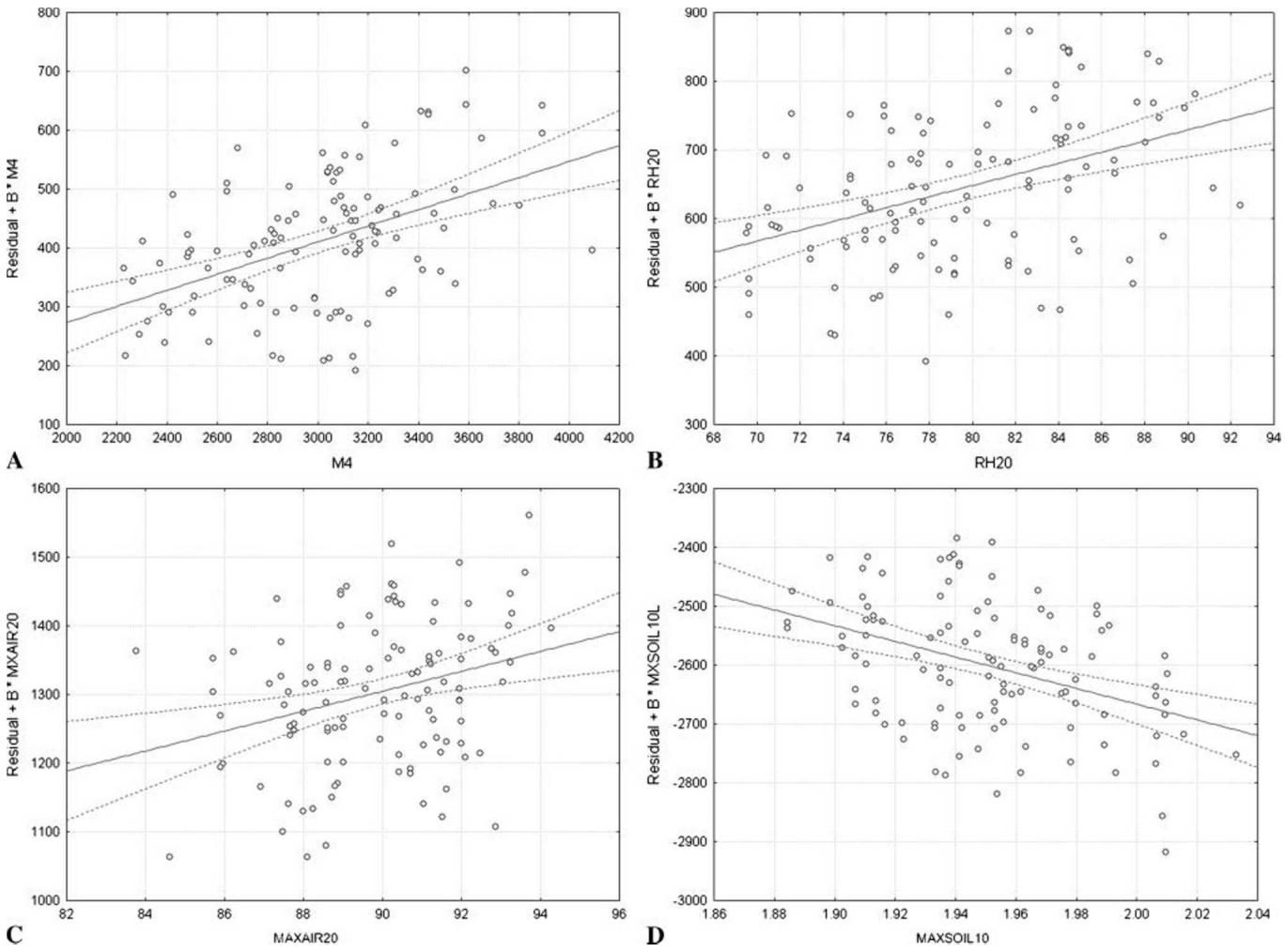

Fig. 2. Partial residual plots of the effect of a predictor variable after adjusting for the effects of other covariables on 'Beauregard' sweetpotato US#1 storage root yield in Louisiana. Effect of accumulated growing degree days (M4) on US#1 storage root yield after adjusting for RH20, MAXAIR20, and MAXSOIL10 (log 10 transformed) (A); effect of RH20 on US#1 storage root yield after adjusting for M4, MAXAIR20, and MAXSOIL10 (log 10 transformed) (B); effect of MAXAIR20 on US#1 storage root yield after adjusting for M4, RH20, and MAXSOIL10 (log 10 transformed) (C); effect of MAXSOIL10 (log 10 transformed) on US#1 storage root yield after adjusting for M4, RH20, and MAXAIR20 (D). M4 = maximum daily temperature (Tmax) – base temperature (B), where if Tmax > ceiling temperature {C [90 °F (32.2 °C)]}, then Tmax = C, and where GDD = 0 if Tmin < 60 °F. MAXSOIL10 = mean maximum soil temperature 10 d after transplanting (DAT); RH20 = mean relative humidity 20 DAT; MAXAIR20 = mean maximum air temperature 20 DAT. Plots were generated using Statistica (version 8; Statsoft, Tulsa, OK). Dashed lines represent 95% confidence interval for regression line (solid line) (Statsoft, 2008).

GDD due to high Tmax does not translate to an early harvest or higher potential yield. This is in part demonstrated in Table 6 where the 28 June planting date achieved GDD = 2600 at 89 d (compared with 91 d for the 22 May planting date). Even when harvested at a comparatively later date (106 DAT), the US#1 storage root yield of the "late" planting was similar to the 22 May planting (92 DAT).

The application of DM to agricultural field data has previously been described (Frank et al., 2004; Witten and Frank, 2005), and several examples exist in the scientific literature (Bui et al., 2006; Ekasingh et al., 2005). As scientific instruments continue to generate massive data sets, the KDD approach and, in particular DM methods, play important and enabling roles (Fayyad et al., 1996). We used DM methodology to provide estimates of predictive accuracy wherein "trained models" were applied on "test (nontraining) data." In the DM mode, the best LR model (M4, 60–90) was ranked second in terms of predictive ability (Table 5). The best-performing DM model used SVM (M4, 60–90) (Table 5). SVMs are a set of unsupervised learning methods originally developed to solve classification problems, but were later extended to the domain of regression problems (Vapnik et al., 1997). SVMs have been used in agricultural research to predict soil moisture (Gill et al., 2006) and maize hybrid performance (Maenhout et al., 2007). Several studies have demonstrated that DM-based techniques, especially NN and RT, matched or exceeded the predictive accuracy of LR-derived models (Clapham and Fedders, 2004; Park et al., 2005). In our work, the

Table 5. Mean square error (MSE) of data mining models describing the relationship between US#1 yield and various growing degree day (GDD) methods and agrometeorological factors for 'Beauregard' sweetpotato grown in Louisiana.

Base (°F)[z]	Ceiling (°F)	GDD method[y]	Model[x]	MSE[w]
60	90	M4	SVM	11,343.90
60	90	M4	REG	11,709.40
60	85	SINE	REG	12,194.70
60	85	SINE	SVM	12,284.70
60	85	M4	REG	12,613.90
60	85	M4	SVM	12,636.60
60	80	M4	SVM	12,668.70
60	80	SINE	SVM	12,776.10
60	80	TRI	SVM	12,815.50
65	95	TRI	REG	12,845.10
60	80	TRI	REG	12,917.10
60	80	SINE	REG	12,941.00
60	80	M4	REG	13,384.90
60	80	M4	GLM	13,562.80
65	95	TRI	SVM	13,583.00
60	90	M4	GLM	13,586.00
60	85	M4	GLM	13,655.00
60	90	M4	MARS	13,744.00
60	80	TRI	MARS	13,791.80
60	80	SINE	MARS	13,910.10
65	95	TRI	GLM	14,056.80
60	80	TRI	GLM	14,123.40
60	80	SINE	GLM	14,150.40
65	95	TRI	MARS	14,306.20
60	85	SINE	GLM	14,405.20
60	85	M4	MARS	14,410.30
60	85	SINE	MARS	14,430.40
60	80	TRI	RT	14,739.30
60	80	M4	MARS	14,845.40
65	95	TRI	RT	15,466.70
60	80	M4	RT	15,474.20
60	85	M4	RT	15,593.60
60	80	SINE	RT	16,241.20
60	85	SINE	RT	16,521.20
60	90	M4	RT	17,071.40
60	85	SINE	NN	47,837.80
60	80	TRI	NN	93,841.80
60	80	SINE	NN	446,374.00
60	80	M4	NN	1,211,640.00
60	85	M4	NN	9,786,120.00
65	95	TRI	NN	1.77E + 10
60	90	M4	NN	1.55E + 17

[z](°F – 32)/1.8 = °C.
[y]M1 to M6, TRI, and SINE as defined in Table 2.
[x]SVM = support vector machine, MARS = multivariate adaptive regression, NN = neural networks, REG = linear regression, RT = regression tree, GLM = generalized linear model. Calculations were performed using Statistica Data Miner (version 8; Statsoft, Tulsa, OK). In most cases, default software settings were used. Details of calculations are described in Materials and methods.
[w]Values represent average of five experimental runs using five unique pairs of randomly sampled training (n = 58) and testing (n = 58) data partitions. Method of calculating MSE is defined in Materials and methods; calculations were performed using Statistica Data Miner (version 8).

predictive accuracy of NN and RT models were ranked lowest (high MSE), suggesting that the trained models likely overfit the training data, leading to high MSE estimates. Overfitting pertains to cases where the model gives good results when applied to the training data, but yields poor results when applied to a new set of observations [i.e., testing data (Levin and Zahavi, 2005)]. RT- and NN-based models have been documented as prone to produce models that overfit data (Khoshgoftaar and Allen, 2001; Zhang, 2005). The relatively small sample size of the modeling database likely contributed to the instability of the NN-derived models. While a data set of 100 to 10,000 records is considered "large" in traditional statistics, in DM, 10^4 may be considered a small sample size (Benjamini and Leshno, 2005). During model development, NN further splits the training data into the actual training sample and a validation sample. Thus, the actual sample size used in NN model building is smaller than the initial sample size (Zhang, 2005). A sample size of 40 and 74 is considered sufficient for classification and time series problems, respectively (Zhang, 2005). However, Kim (2008) reported that LR techniques were superior to DM regardless of the number of variables and sample size when continuous independent variables were used, and that NN were better when categorical variables were involved. This information can be used for future studies where categorical variables are involved. Our results underscore the necessity of comparing the results derived from DM-oriented methodology with traditional LR approaches to verify the suitability or advantages of using one method or the other (Zhang, 2005). The DM approach can be used in conducting preliminary assessment for the presence of nonlinear relationships, especially in large data sets. Current DM software applications have built-in functions that automate this procedure. DM software applications also include tools that are potentially useful in crop growth modeling work. For example, various DM software feature "deployment" modules that automatically generate computer programming code (e.g., C and C++) so that "trained" models can be incorporated into compiled programs that underlie deterministic growth models or decision support systems.

The main premise of using nonlinear methods in modeling plant development is that growth response to temperature is often nonlinear (Clapham and Fedders, 2004; Park et al., 2005). Our experimental results indicate that the variables used in this study failed to show nonlinearity, and various DM algorithms that fit nonlinear functions or interactions did not contribute to increased predictive accuracy. The partial residual

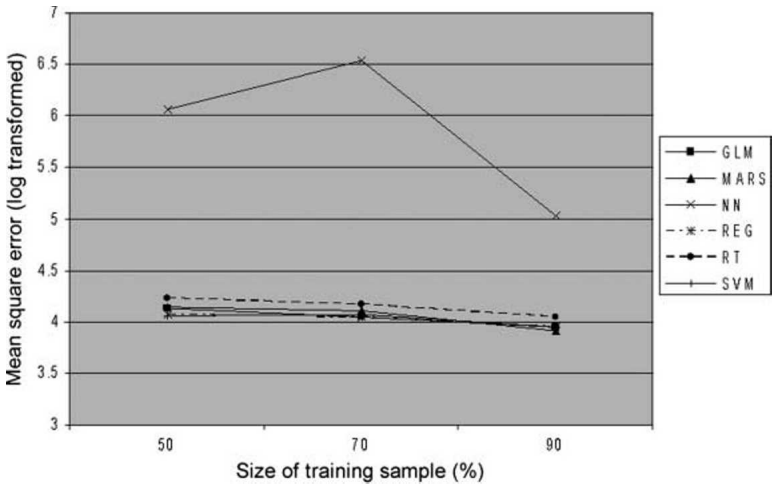

Fig. 3. Average mean square error (log transformed) of various data mining models generated with different sizes of training samples. REG = linear regression, GLM = generalized linear model, RT = regression tree, NN = neural networks, MARS = multivariate adaptive regression, SVM = support vector machine. Calculations performed in Statistica Data Miner (version 8; Statsoft, Tulsa, OK). Modeling database = GDDLA-YMET (n = 116).

plots (Fig. 2) help to show the linear relationship of these predictor variables. It is possible that certain growth stages of the sweetpotato may respond nonlinearly to temperature or that certain other variables may demonstrate nonlinear relationship to covariables and yield. Kays (1985) cited studies that documented that early in the growing season, crop growth rate (CGR) was initially slow; highest CGR was achieved 70 to 98 DAT, after which it decreased. Kays (1985) also noted that for a planting date of 16 June, net photosynthesis tended to increase until the middle of September (about 87 DAT), and then gradually decreased as harvest approached. This decline was attributed to a decline in gross photosynthesis rather than an increase in respiration. It appears that sweetpotato respond linearly to temperature for most of the growing season, but this response becomes less linear as the harvest approaches. At this time, we are unable to account for the existence of this nonlinear relationship due to the experimental limitations imposed by our modeling data set.

The empirically derived models in this study were specific to the range of environments where the yield trials were conducted. To some extent, such models can be calibrated for use in other locations outside of Louisiana. Empirical models can be used as explanatory tools for identifying the hidden structure of crop growth

Table 6. Agrometeorological variables, accumulated growing degree days (GDD) to test harvest and harvest, and actual and predicted US#1 storage root yields of 'Beauregard' sweetpotato grown in Louisiana.[z]

Location	Planting date	Harvest date	Agrometeorological variables[y] MAXSOIL10 (°F)	MAXAIR20 (°F)	RH20 (%)	Accumulated GDD[x] To test harvest	To harvest	US#1 yield (no. 50-lb bushels/acre)[w] Actual	Predicted[v]
CHS[u]	22 May	22 Aug.	76.64	87.76	82.64	2,609 (91)[t]	2,609 (92)[s]	324.06	343.33
	22 May	29 Aug.	76.64	87.76	82.64		2,849 (99)	343.03	376.93
	30 May	4 Sept.	79.18	89.67	83.90	2,618 (90)	2,618 (97)	420.63	363.51
	30 May	11 Sept.	79.18	89.67	83.90		3,038 (104)	526.37	422.31
	7 June	12 Sept.	81.82	91.19	84.12	2,614 (89)	2,854 (97)	419.07	401.42
	7 June	21 Sept.	81.82	91.19	84.12		3,114 (106)	460.75	437.82
	14 June	18 Sept.	81.45	90.14	85.07	2,614 (89)	2,814 (96)	427.63	390.92
	14 June	28 Sept.	81.45	90.14	85.07		3,108 (106)	553.58	432.08
	22 June	5 Oct.	82.36	88.95	88.69	2,615 (89)	3,089 (105)	534.92	435.05
	22 June	19 Oct.	82.36	88.95	88.69		3,389 (119)	494.22	477.05
	28 June	12 Oct.	82.00	87.62	92.45	2,616 (89)	3,091 (106)	308.67	449.03
BTR	15 May	13 Aug.	81.36	85.90	69.62	2,601 (96)	2,502 (90)	101.23	183.53
	15 May	7 Sept.	81.36	85.90	69.62		3,252 (115)	279.48	217.13
	6 June	7 Sept.	89.73	92.00	69.62	2,629 (82)	2,774 (93)	148.63	214.97
	6 June	1 Oct.	89.73	92.00	69.62		3,419 (117)	205.60	273.77
	26 June	1 Oct.	87.36	91.62	73.64	2,625 (69)	2,833 (97)	174.94	290.54
	26 June	19 Oct.	87.36	91.62	73.64		3,199 (115)	155.77	325.68

[z]Data obtained from field experiments conducted in 2007.
[y]MAXSOIL10 = mean maximum soil temperature 10 d after transplanting (DAT), log 10 transformed values used for calculation; RH20 = mean relative humidity 20 DAT; MAXAIR20 = mean maximum air temperature 20 DAT; (°F – 32)/1.8 = °C.
[x]Accumulated GDD calculated using method M4 as defined in Table 2.
[w]Storage roots were graded according to USDA standards (USDA, 2005): US #1 grade = 2 to 3–1/2 inches (5.1–8.9 cm) diameter, 3 to 9 inches (7.6–22.9 cm) length, maximum weight not more than 20 oz (567.0 g); canner = 1 to 2 inches (2.5–5.1 cm) diameter, 2 to 7 inches (5.1–7.8 cm) length; jumbo = larger vs. others, but marketable; 1 50-lb bushel/acre = 56.0426 kg·ha^{-1}.
[v]Predicted US#1 yield calculated using M4, base temperature = 60 °F, ceiling temperature = 90 °F, and agrometeorological variables as defined in Table 4.
[u]CHS = Chase, LA; BTR = Baton Rouge, LA. Details of locations described in Table 1.
[t]Values enclosed in parentheses = days after transplanting.
[s]Values enclosed in parentheses = days to harvest.

processes (Park et al., 2005). Results from this work have practical implications for the sweetpotato harvest scheduling in Louisiana. One direct application of our work would be to schedule test harvests using GDD = 2600 instead of using calendar days. Based on 2007 planting dates, the number of days required to reach GDD = 2600 ranged from 69 to 91 d. Scheduling a test harvest and the ability to forecast yield can help commercial growers and crop consultants to further fine tune decisions concerning harvest dates based on a target yield level. For example, the harvest can be scheduled based on yield potential of fields (planting dates) if a threat of extended severe weather emerges during this period. Seem et al. (2003) also proposed using GDD as a method for 'Beauregard' to better compete with weeds if it is planted during periods where maximum GDD is accumulated rapidly. The GDD model can be incorporated into predictive models of sweetpotato crop phenology. To our knowledge, a well-defined phenological model does not exist for the species, and past reviews of sweetpotato yield physiology (Kays, 1985; Ravi and Indira, 1999) do not mention the development or existence of such models. Such models can open areas for further investigation, including the role of accumulated GDD in scheduling herbicide application, fertilizer application, irrigation, and pest management.

Currently, some commercial weather monitoring stations include software that calculates GDD using specific preprogrammed methods (i.e., triangle, sine, and modifications of these basic methods). Our results indicate that the SINE method (60–85) can be used with a very slight reduction in model accuracy. This will allow growers and crop consultants to use such software without further modification. This research represents a preliminary step toward helping to account for field-level yield variability in sweetpotato. Future studies should be able to improve the predictive performance of current models through the addition of other soil- and plant-related predictor variables. Rainfall-related variables were excluded from the current modeling experiments due to violation of normality assumptions. Even when all agrometeorological variables were included in DM mode, none of the rainfall-related measurements were included as predictor variables in the best-performing models (data not shown). Soil moisture and nutrient measurements were available for some locations for the current study, but were excluded in the analysis because this would have reduced the size of the modeling database.

Conclusions

This research indicates the potential for using GDD-based models to help predict sweetpotato harvest dates in Louisiana. Several methods and combinations of B and C showed better goodness-of-fit and predictive accuracy when compared with the standard method of calculating GDD. In addition to the conventional methods of identifying candidate GDD methods, we also considered using adaptive algorithms associated with DM methodology. LR- and DM-based regression approaches identified similar candidate models. Using accumulated GDD, our results suggest that test harvests can be done at about GDD = 2600 and harvesting can start shortly thereafter. Further calibration is necessary to improve the predictive ability of the current model. Future studies will likely investigate the potential modulating effect of moisture stress and other management variables.

Literature cited

Arnold, C.Y. 1959. The determination and significance of the base temperature in a linear heat unit system. Proc. Amer. Soc. Hort. Sci. 74:430–445.

Arnold, C.Y. 1960. Maximum-minimum temperatures as a basic for computing heat units. Proc. Amer. Soc. Hort. Sci. 76: 683–692.

Benjamini, Y. and M. Leshno. 2005. Statistical methods for data mining, p. 565–588. In: O. Maimon and L. Rokach (eds.). Data mining and knowledge discovery handbook. Springer, New York.

Bui, E.N., B.L. Henderson, and K. Viergever. 2006. Knowledge discovery from models of soil properties developed through data mining. Ecol. Modell. 191:431–446.

Clapham, W.M. and J.M. Fedders. 2004. Modeling vegetative development of berseem clover (*Trifolium alexandrinum* L.) as a function of growing degree days using linear regression and neural networks. Can. J. Plant Sci. 84:511–517.

Data Mining Group. 2008. Data Mining Group. 1 Mar. 2008. <http://www.dmg.org/>.

Dufault, R.J. 1997. Determining heat unit requirements for broccoli harvest in coastal South Carolina. J. Amer. Soc. Hort. Sci. 122:169–174.

Edmond, J.B. and G.R. Ammerman. 1971. Sweet potatoes: Production, processing, marketing. AVI Publishing, Westport, CT.

Ekasingh, B., K. Ngamsomsuke, R.A. Letcher, and J. Spate. 2005. A data mining approach to simulating farmers' crop choices for integrated water resources management. J. Environ. Mgt. 77:315–325.

Everaarts, A.P. 1999. Harvest date prediction for field vegetables. A review. Gartenbauwissenschaft. 64:20–25.

Fayyad, U., D. Haussler, and P. Stoloroz. 1996. KDD for science data analysis: Issues and examples, p.50–56. In: E. Simoudis, J. Han, and U. Fayyad (eds.). p. 50–56. Proc. Second Knowledge Discovery and Data Mining Conf. 2–7 Aug. 1996, Menlo Park, CA.

Frank, E., M. Hall, L. Trigg, G. Holmes, and I.H. Witten. 2004. Data mining in bioinformatics using Weka. Bioinformatics 20:2479–2481.

Gill, M.K., T. Asefa, M. Kemblowski, and M. McKee. 2006. Soil moisture prediction using support vector machines. J. Amer. Water Resour. Assoc. 42:1033–1046.

Henderson, C.E., W.D. Potter, R.W. McClendon, and G. Hogenboom. 2000. Predicting aflatoxin contamination in peanuts: A genetic algorithm/neural network approach. Appl. Intell. 12:183–192.

Higley, L.G., L.P. Pedigo, and K.R. Ostlie. 1986. DEGDAY: A program for calculating degree-days, and assumptions behind the degree-day approach. Environ. Entomol. 15:999–1016.

Jenni, S., D. Cloutier, G. Bourgeois, and K.A. Stewart. 1996. A heat unit model to predict growth and development of muskmelon to anthesis of perfect flowers. J. Amer. Soc. Hort. Sci. 2:274–280.

Jenni, S., K.A. Stewart, G. Bourgeois, and D. Cloutier. 1998. Predicting yield and time to maturity of muskmelons from weather and crop observations. J. Amer. Soc. Hort. Sci. 123:195–201.

Kays, S.J. 1985. The physiology of yield in the sweet potato, p. 79–132. In: J. Bouwkamp (ed.). Sweetpotato products: A natural resource for the tropics, pages. CRC Press, Boca Raton, FL.

Khoshgoftaar, T. and E.B. Allen. 2001. Controlling overfitting in classification-tree models of software quality. Empir. Softw. Eng. 6:59–79.

Kim, Y.S. 2008. Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. Expert Syst. Appl. 34:1227–1234.

Levin, N. and J. Zahavi. 2005. Data mining for target marketing, p. 1261–1304. In: O. Maimon and L. Rokach (eds.). Data mining and knowledge discovery handbook. Springer, New York.

Louisiana Agriclimatic Information. 2008. Louisiana agriclimatic information. 1 Mar. 2008. <http://www.lsuagcenter.com/weather/index.asp>.

Maenhout, S., B. De Baets, G. Haesaert, and E. Van Bockstaele. 2007. Support vector machine regression for the prediction of maize hybrid performance. Theor. Appl. Genet. 115:1003–1013.

Maimon, O. and L. Rokach. 2005. Introduction to knowledge discovery in databases, p. 1–20. In: O. Maimon and L. Rokach (eds.). Data mining and knowledge discovery handbook. Springer, New York.

Park, S.J., C.S. Hwang, and P.L.G. Vlek. 2005. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. Agr. Systems 85:59–81.

Perry, K.B. and T.C. Wehner. 1996. A heat unit accumulation method for predicting cucumber harvest date. HortTechnology 6:27–30.

Perry, K.B., T.C. Wehner, and G.L. Johnson. 1986. Comparison of 14 methods to determine heat unit requirements for cucumber harvest. HortScience 21:419–423.

Perry, K.B., Y. Wu, D. Sanders, J.T. Garrett, D. Decoteau, R. Nagata, R. Dufault, K.D. Batal, D. Granberry, and W. Mclaurin. 1997. Heat units to predict tomato harvest in the southeast USA. Agr. For. Meteorol. 84:249–254.

Ravi, V. and P. Indira. 1999. Crop physiology of sweet potato. Hort. Rev. (Amer. Soc. Hort. Sci.) 23:277–338.

Schwartz, M.D., G.J. Carbone, G.L. Reighard, and W.R. Okie. 1997. A model to predict peach phenology and maturity using meteorological variables. HortScience 32:213–216.

Seem, J.E., N. Creamer, and D.W. Monks. 2003. Critical weed-free period for Beauregard sweetpotato (*Ipomoea batatas*). Weed Technol. 17:686–695.

Snyder, R.L. 2005. Degree days. 18 Aug. 2008. <http://atm.ucdavis.edu/~biomet/DegreeDays/DegDay.htm>.

Statsoft. 2008. Statsoft electronic textbook. 18 Aug. 2008. <http://www.statsoft.com/textbook/stathome.html>.

Stenzel, N.M.C., C.S.V.J. Neves, C.J. Marur, M.B.S. Scholz, and J.C. Gomes. 2006. Maturation curves and degree-days accumulation for fruits of 'Folha murcha' orange trees. Scientia Agricola 63:219–225.

Stoddard, S.C. and B. Weir. 2002. Sweetpotato research trials 2002 research progress report. Univ. California Coop. Ext. 15 Aug. 2008. <http://cemerced.ucdavis.edu/files/19101.pdf>.

Togari, Y. 1950. A study of tuberous root formation in sweet potato. Bul. Natl. Agr. Expt. Sta. Tokyo 68:1–96.

Tronstad, R., J.C. Silvertooth, and S. Husman. 2003. Irrigation termination of cotton: An economic analysis of yield, quality, and market factors. J. Cotton Sci. 7:86–94.

Turpin, K.M., D.R. Lapen, E.G. Gregorich, G.C. Topp, M. Edwards, N.B. McLaughlin, W.E. Curnoe, and M.J.L. Robin. 2005. Using multivariate adaptive regression splines (MARS) to identify relationships between soil and corn (*Zea mays* L.) production properties. Can. J. Soil Sci. 85:625–636.

U.S. Department of Agriculture. 2004. The commercial storage of fruits, vegetables, and florist and nursery stocks. 18 Aug. 2008. <http://www.ba.ars.usda.gov/hb66/contents.html>.

U.S. Department of Agriculture. 2005. United States standards for grades of sweetpotatoes. 18 Aug. 2008. <http://www.ams.usda.gov/AMSv1.0/getfile?dDocName=STELPRDC5050330>.

Vapnik, V., S. Golowich, and A. Smola. 1997. Support vector method for function approximation, regression estimation, and signal processing, p. 281–287. In: M. Mozer, M. Jordan, and T. Petsche (eds.). Advances in neural information processing systems 9. 2–5 Dec. 1996. Denver, CO.

Viator, R.P., R.C. Nuti, K.L. Edmisten, and R. Wells. 2005. Predicting cotton boll maturation period during degree days and other climatic factors. Agron. J. 97:494–499.

Witten, I.H. and E. Frank. 2005. Data mining: Practical machine learning and techniques. Morgan Kaufman, San Francisco.

Wolfe, D.W., L.D. Albright, and J. Wyland. 1989. Modeling row cover effects on microclimate and yield: I. Growth response of tomato and cucumber. J. Amer. Soc. Hort. Sci. 114:562–568.

Wurr, D.C.E. and J.R. Fellows. 1984. The growth of three crisp lettuce varieties from different sowing dates. J. Agr. Sci. Cambridge 102:733–745.

Yang, C.C., S.O. Prasher, P. Enright, C. Madramootoo, M. Burgess, P.K. Goel, and I. Callum. 2003. Application of decision tree technology for image classification using remote sensing data. Agr. Systems 76:1101–1117.

Zalom, F.G., P.B. Goodell, L.T. Wilson, W.W. Barnett, and W.J. Bentley. 1983. Degree-days: The calculation and use of heat units in pest management. Univ. California Div. Agr. Natural Resources Lflt. 21373.

Zhang, P.G. 2005. Neural networks, p. 487–516. In: O. Maimon and L. Rokach (eds.). Data mining and knowledge discovery handbook. Springer, New York.