Journal of Agricultural and Resource Economics, 20(1):32–48

Importance of Melon Type, Size, Grade, Container, and Season in Determining Melon Prices

Russell Tronstad

Classification and Regression Trees (CART), a computer intensive nonparametric classification method, was used to model weekly Los Angeles wholesale prices (1990–93) for twelve different melon types. CART explained more of the variation in melon prices than did an ordinary least squares (OLS) regression with dummy variables. Explanatory variables ranked as the most-to-least important by CART are as follows: week, type of melon, year, size, grade, and shipping container. The most notable price change occurs when prices fall after 13 May.

Key words: binary split, CART, hedonic, relative importance, terminal node

Introduction

Weekly melon prices vary greatly across melon types (e.g., red-flesh watermelons versus seedless watermelons). Melons also receive a premium or discount depending on other "quality characteristics," such as size, grade, and shipping container. Consumers prefer some melons over others according to types and characteristics. For example, during the last week of January 1994, "one-label" grade honeydews received twice the price of good grade honeydews.

In his work with vegetables, Waugh was one of the first to consider the influence quality had on prices. Subsequent research has often used hedonic price analysis to determine the value of quality characteristics for agricultural goods (e.g., Bowman and Ethridge; Brorsen, Grant, and Rister; Goodwin et al.; Lenz, Mittelhammer, and Shi; Unnevehr and Bard; Veeman; and Wahl, Shi, and Mittelhammer). As noted by Epple, the two main goals of hedonic studies are to determine discounts and premiums and to estimate the demand and supply functions for attributes of the product. The focus here is on discounts and premiums.[1]

Collette and Wall identified the importance of timing, or seasonality, for the prices of cucumbers, eggplant, peppers, and tomatoes in Florida. Prices changed dramatically in just a few weeks, particularly in the spring. Tronstad, Huthoefer, and Monke argue for including seasonal factors in hedonic analysis since seasonal factors can influence the supply of quality characteristics. Ethridge and Davis and Wilson account for temporal price changes with a linear time trend and dummy variables for month or year.

Data availability is often a problem in hedonic analyses. For example, Goodwin et al. detected positive autocorrelation when analyzing factors that affect fresh potato prices. Four different varieties of potatoes were analyzed from 1982 to 1985 and the time periods of available data for each potato type differed by so much that it was not possible to follow appropriate autocorrelation adjustment techniques. Brorsen, Grant, and Rister were unable to find a "weekly farm price" for rice so they included a base mill price to account for changes

The author is an assistant specialist in the Department of Agricultural and Resource Economics at The University of Arizona.

[1] Quantity data are not available for melons with the quality characteristics considered. This precludes an analysis that estimates demand and supply functions.

in the rice bid/acceptance market. A nonparametric approach of Classification and Regression Trees (CART)[2] is used in this analysis to avoid problems associated with limited price quotes and restrictive model specifications.[3]

The objective of this article is to determine discounts and premiums due to various characteristics of wholesale melons. Characteristics considered are melon type, size, grade, shipping container, week, and year. CART results are described graphically and are probably easier for lay audiences to understand than parametric regression results. Using this information, grower and shipping practices may be altered to take advantage of price premiums or to avoid price discounts.

Data and Modeling Considerations

Weekly price data were obtained from the *Los Angeles Wholesale Fruit and Vegetable Report*, published by the U.S. Department of Agriculture, Federal State Market News Service. Data are from 3 January 1990 through 28 December 1993. Earlier price quotes for the Los Angeles market were unavailable. Monday price quotes were used unless a holiday preempted the weekly report, in which case, Tuesday was used for a price quote. The report often gives a price range for a particular grade, size, region, and type of melon. If a price range was given, the midpoint was used. Region of production is sometimes reported in specific terms (Imperial Valley) or sometimes in broad terms (California/Arizona). Due to the inconsistency of regional notation and the strong correlation between region and week of year, region was not considered as an explanatory variable.

Twelve different melon types were considered: cantaloupes, honeydews, red-flesh watermelons, seedless watermelons, plus canary, casaba, crenshaw, mayan, orange-flesh, persian, santa claus, and sharlyn "specialty melons." A total of 5,186 price quotes were available. The number of observations for each melon type was: canary 138; cantaloupes 1,358; casaba 202; crenshaw 453; honeydews 1,217; mayan 31; orange-flesh 324; persian 84; santa claus 110; sharlyn 45; red-flesh watermelons 712; and red seedless watermelons 512.

A size is specified for almost all of the melon price quotes. The number given for size refers to the number of melons required to fill a standard carton. Thus, a size 18 for cantaloupes is a smaller melon than size 12. To standardize the size variable for all melons, size was standardized as

$$S_{mq}^{std} = \frac{(S_{mq} - \bar{S}_m)}{\sigma_m}, \tag{1}$$

where S_{mq}^{std} is the standardized size for the qth observation of melon type m, S_{mq} is the size number recorded, $\bar{S}_m$ is the mean size for melon type m, and σ_m is the sample standard deviation in size. Size was not given for regular and seedless watermelons in over half the price quotes. These watermelons were always sold in either bins or crates. Size was given

[2]CART is a computer package that is a trademark of California Statistical Software Inc., Lafayette, California, copyright 1984.

[3]In general, equilibrium prices cannot be linearly decomposed (Jones) and the demand functions for product characteristics cannot be consistently estimated (Epple) as modeled by OLS. The functional form of CART is less restrictive.

on 507 of 1,224 price quotes, or about 41%. CART provides procedures for handling missing data.

Grades of "one-label," "good quality," "fair quality," "fair condition," and "poor quality" are recorded. Over 90% of the price quotes (4,735) are "good quality," and 254 or almost 5% are of the highest grade, "one-label." Only 61, 131, and 5 price quotes are given for "fair quality," "fair condition," and "poor quality," respectively. Watermelon and honeydew price quotes are only reported for a grade of "good quality." Most of the grade variation occurs for cantaloupes and the specialty melons. The skin of cantaloupes and some specialty melons are much softer than watermelon and honeydews; thus, bruises or soft spots are much more likely to appear in the softer-skinned melons.

Watermelon prices are reported in cents per pound with about half of the price quotes for the shipping container of cartons and the rest for bins, crates, or mixed containers (cartons-crates or bins-crates). Bins, crates, and mixed containers are reported in 254, 276, and 103 price quotes, respectively, for both regular and seedless watermelons. All melon types except seedless and red-flesh watermelons are quoted in dollars per carton. To calculate the price per pound for other melons, a weight of 40 lbs. per one-half of a carton was used for cantaloupes and 30 lbs. per two-thirds of a carton of honeydew and specialty melons, as reported in the *Los Angeles Wholesale Fruit and Vegetable Report.*

To look at seasonality effects, week of year for the qth observation (w_q) was calculated as follows:

$$w_q = \frac{d_q}{7}, \tag{2}$$

where d_q is the day of the calendar year (1 to 365) for the qth observation or price quote. Seasonality should be strong in the data since melons are a perishable product with a short shelf life. The *1993 Produce Services Sourcebook* (The Packer) reports the post-harvest life for cantaloupes at 10 to 14 days, 14 to 21 days for watermelons and "specialty melons," and 21 to 28 days for honeydews that have been C_2H_4 treated. Melon production is very seasonal; most winter melons are shipped from Mexico and Central America, and some U.S. areas supply melons for only a one- to two-week period. Year was also included as an explanatory variable since weather conditions and resultant supply for a given week can vary dramatically from one year to the next.

The hedonic price model proposes that consumers derive satisfaction or utility from characteristics that goods possess, rather than just the goods themselves (Lancaster; Lucas; Rosen). Following Lucas, hedonic price functions are of the form

$$P_i = f(C_{i1}, \ldots, C_{ij}; \varepsilon_i), \tag{3}$$

where P_i is the observed price of commodity i, C_{ij} measures the "intrinsic value" of the jth quality characteristic for commodity i, and ε_i is a random error term. Epple argues that in general hedonic estimations should also include supply response functions. But given production and transportation lags, and the limited post-harvest life of melons, supply is assumed to be perfectly inelastic for a given week. Hence, there is no identification problem (Hanemann) and the hedonic-price estimates identify demand for quality characteristics.

Classification and Regression Trees

CART was used to analyze how the independent variables of melon type, size, grade, container, week, and year influence the dependent variable of melon price. To help understand the algorithm of CART, think of a jar full of marbles with each price quote representing one marble in the jar. Each marble or price quote has time, type, and quality characteristics stamped on it. The first question CART addresses is what variable and accompanying magnitude can be used to split the marbles into two jars (defined as a binary split) so that the prices in each jar are as close to one another as possible. The best split is weighted according to the number of marbles that are split left or right into separate jars. Then, subsequent binary splits occur using the same logic until all price quotes are placed into a terminal jar or "node." A predictor value is then assigned for each terminal node. Collectively, all binary splits and predictor values are referred to as the predictor tree.

For large data sets like this one, CART randomly picks two-thirds of the data as a "learning sample" and sets aside the other one-third (default and what was used for this analysis) of the data for testing the predictor tree constructed from the learning sample. Given a learning sample, three rules are needed to construct a predictor tree for a large data set: (*a*) a criterion for selecting a binary split at each node, (*b*) a rule for assigning a predictor value to every terminal node, and (*c*) a rule for determining when a node is terminal or optimal predictor tree size.

The classical measure of model accuracy is mean-squared error.[4] When using the criterion of minimizing mean-squared error, the first binary split of the learning sample is determined by iteratively searching all levels of independent variables as a possible split. The split that reduces the weighted[5] sum of squared errors the most is the split chosen by CART. For every node t, the sum of squared errors for each node is simply $\sum_{q \in t}(y_q - \bar{y}(t))^2$, where y_q is the qth observation in node t, and $\bar{y}(t)$ is the average of all melon prices selected for node t. The number c which minimizes $\sum_{q \in t}(y_q - c)^2$ is simply $c = \sum_{q \in t} y_q / Q$, where Q is the number of observations in node t.[6] Subsequent nodes are split following the same criterion until a very large tree is grown. This tree has no more than five observations (default number) in each terminal node or all values (e.g., melon prices) in a terminal node are equal.

CART selectively prunes branches off this large tree grown from the learning sample to select the right-sized tree. That is, the apparent error rate based on the learning sample will always appear small for the largest tree, since each observation could ultimately be classified in its own terminal node as a "perfect fit." But this tree would likely give spurious results when making predictions. Test sample data, selected at random in the beginning and not used in constructing the tree, are used to obtain a more reliable estimate for the "true error rate." The estimated true error rate of a predictor-tree, $R(d)$, is simply the mean-squared difference between the actual values of the test set and their predicted values from the constructed tree (based on the learning sample). More formally $R(d)$ is

[4] Alternative criteria that are less sensitive to outliers such as least absolute deviations are available in CART but yielded similar results to those presented.

[5] Weighted according to the number of observations going left and right, respectively.

[6] Splits could also be based on a linear combination of variables in CART. But this algorithm has a "limited search" and could get trapped on a local maxima so that there is no guarantee of a global optimum for the split. The simplicity of interpretation is also reduced with linear combinations and computation requirements are increased by at least a factor of five.

$$R(d) = \frac{1}{N^{ts}} \sum_{i=1}^{N^{ts}} (y_i^{ts} - d(y_i^{ls}))^2, \tag{4}$$

where N^{ts} is the number of observations in the test sample, y_i^{ts} is the ith observed value in the test sample, and $d(y_i^{ls})$ is the predicted terminal node value for the *i*th observation constructed from the learning sample. The estimate in (4) provides an unbiased estimate of $R(d)$ since observations for the learning and test sample are drawn independently from the same underlying probability distribution. $R(d)$ is generally large for very small trees, decreases as the tree size grows with a long flat valley, and then increases again as the tree becomes very large. The tree with the minimum true error rate is often referred to as $R(d)^*$, the "optimal tree." However, in order to select a more conservative tree, the Standard Error Rule (SER) was adopted. The standard error of the estimate for $R(d)$ is[7]

$$SE(R(d)) = \frac{1}{\sqrt{N^{ts}}} \left[\frac{1}{N^{ts}} \sum_{i=1}^{N^{ts}} (y_i^{ts} - d(y_i^{ls}))^4 - R(d)^2 \right]^{\frac{1}{2}}. \tag{5}$$

Then, the tree with an expected risk or $R(d)$ closest to $R(d)^* + \gamma\, \mathrm{SE}(R(d)^*)$, where γ is the SER, is the smaller and more conservative tree selected. A SER of 5 was used here. The SER portrays a trade-off between tree complexity and accuracy. If the estimated true error rates from the sequence of pruned trees are relatively flat, a larger SER can be justified than when the true error rate rises steeply for smaller predictor trees. A "flat function" implies that little predictive accuracy needs to be given up for a much smaller tree. Regression trees constructed from a continuous variable are generally much "flatter" than classification trees constructed from a discrete dependent variable. This is because all values must be equal or below a small number before splitting ceases. Equality is much more likely with a limited number of discrete outcomes (classification) than a continuous variable (regression).

The missing observation algorithm of CART was used since some watermelons had no specified size. As above, the algorithm first determines the best split (s^*) of a node by testing splits for all variables. If a variable, say x_4 is missing some values, then the best split for x_4 is determined only from observations that contain a value for x_4. Suppose that the best split for x_4 is whether $x_4 \leq 2$ or $x_4 > 2$. Then CART searches through all possible splits on x_1 until it finds the split on x_1 that is most closely associated with $x_4 \leq 2$ or $x_4 > 2$. It repeats this procedure for all variables except x_4 until it ranks all splits that are most closely associated with the split of $x_4 \leq 2$ or $x_4 > 2$, defined as a surrogate split. Suppose the best surrogate split for $x_4 \leq 2$ or $x_4 > 2$ is whether $x_6 \leq 5$ or $x_6 > 5$. If the value of x_4 is missing then it sends a case to the left if $x_6 \leq 5$ and to the right otherwise. This procedure is analogous to replacing a missing value in a linear regression model by the nonmissing value with which it is most closely correlated. But Breiman et al. propose that CART is more robust than regression. When missing values are filled in by regressing on nonmissing values with linear regression, coefficients are computed by inversion of the covariance matrix. Thus, estimates are sensitive to these smaller eigenvalues, and a covariance matrix that is nearly singular is commonly the result if numerous observations are missing. An error associated with

[7]The proof of this derivation follows from the individual observations in (4) being independent of one another for a fixed learning sample. Thus, the variance is the sum of all the individual terms. Breiman et al. provides a proof.

assigning a case left or right with CART, due to a poor surrogate split, can be corrected in lower splits.

In determining which variables are most important for explaining melon prices, CART ranks variable importance by using surrogate splits. As alluded to above, the surrogate split, s_x, is the split for each variable x that most accurately predicts the action of the best linear split, s_x^*, at each node. The probability that s_x predicts s_x^* correctly is $p_{LL}(s_x^*, s_x) + p_{RR}(s_x^*, s_x)$, where $p_{LL}(\cdot)$ and $p_{RR}(\cdot)$ are the probability that both s_x and s_x^* send a case left and right, respectively. The measure of importance for variable x, $M(x)$, is

$$M(x) = \sum_{t \in T} \Delta I(s_x, t), \tag{6}$$

where T is the optimal tree selected from the test sample and SER, $\Delta I(\cdot)$ is the change in sum of squared errors (described as "node impurity" in CART terminology) from using the surrogate split (s_x) instead of the optimal split (s_x^*) for variable x at each node t, and other variables are as described above. Because the relative importance for ranking variables is more important than the actual level, values are normalized so that the most important variable has a ranking of 100. For further information on the CART method see Breiman et al., Efron and Tibshirani, Horowitz and Carson, and Tronstad and Gum.

Results

Figure 1 shows the predictor tree estimated by CART for describing how time variables and quality characteristics of melons sold in the Los Angeles wholesale market have influenced historical prices. Week of the year is the first variable to split price quotes: branch left if week is less than or equal to 18.9 and branch right if week is greater than 18.9. Panel A of figure 1 describes prices from 21 November (week 46.3) through 13 May (week 18.9), a period primarily supplied by imports from Mexico and the Caribbean and Central American countries. The remainder of the year is described in panel B of figure 1, a season when most melons are from the U.S. Between 1990 and 1993, shipments of U.S. grown melons have started as early as the first of May and ended as late as the last week of December.

A price estimate is obtained by branching left or right to fit the conditions specified until a terminal node is reached. For example, in figure 1, say production is targeted for a "one-label" cantaloupe with week 14 delivery to the Los Angeles market. In following the bolded line path, week 14 is less than 18.9 so branch left to the next criterion of grade. A grade of "one-label" requires a branch right to the next criterion of melon type. Number 2 defines the melon type of a cantaloupe and the next level falls into terminal node 10 with a price prediction of \$0.60/lb. Table 1 gives the learning sample and test sample values for all the observations or cases that fall into each terminal node in figure 1. For example, node 10 has 40 cases in the learning sample with an average value of \$0.598 and a learning sample standard deviation of \$0.150. The test sample has similar values with an average of \$0.573 and standard deviation of \$0.130.

Terminal node 12 has only three observations in the learning sample and zero observations in the test sample. This node represents price quotes for three "one-label" canary melons sold in the spring of 1990 (no orange-flesh melons were in this node). The split between terminal nodes 11 and 12 indicates that the three observations in 12 with an average price of \$0.55/lb. were "outliers" from the other 22 prices in the learning sample with an

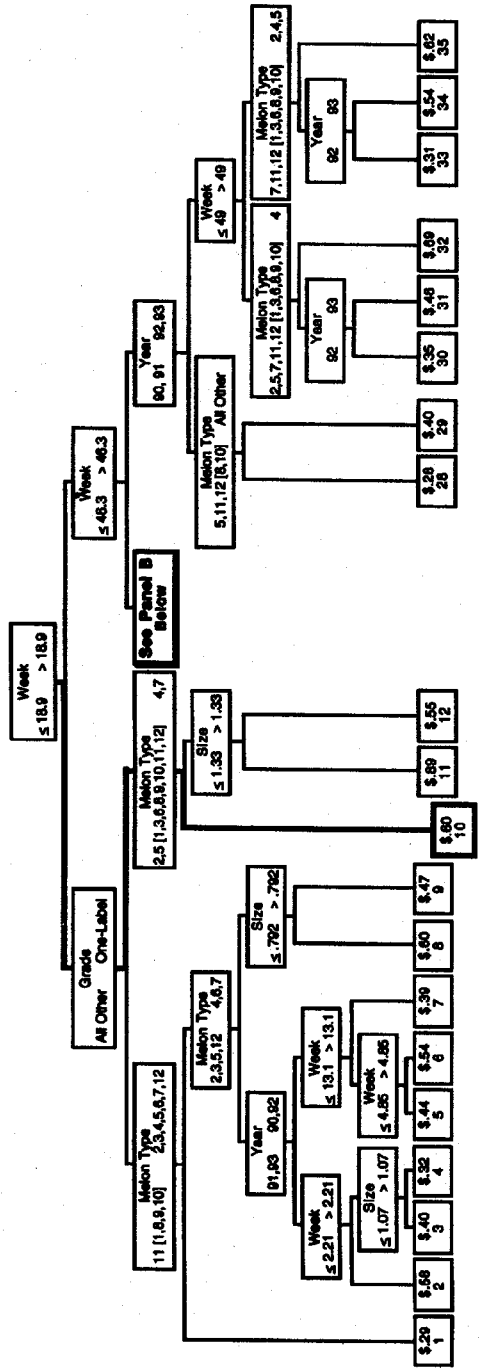
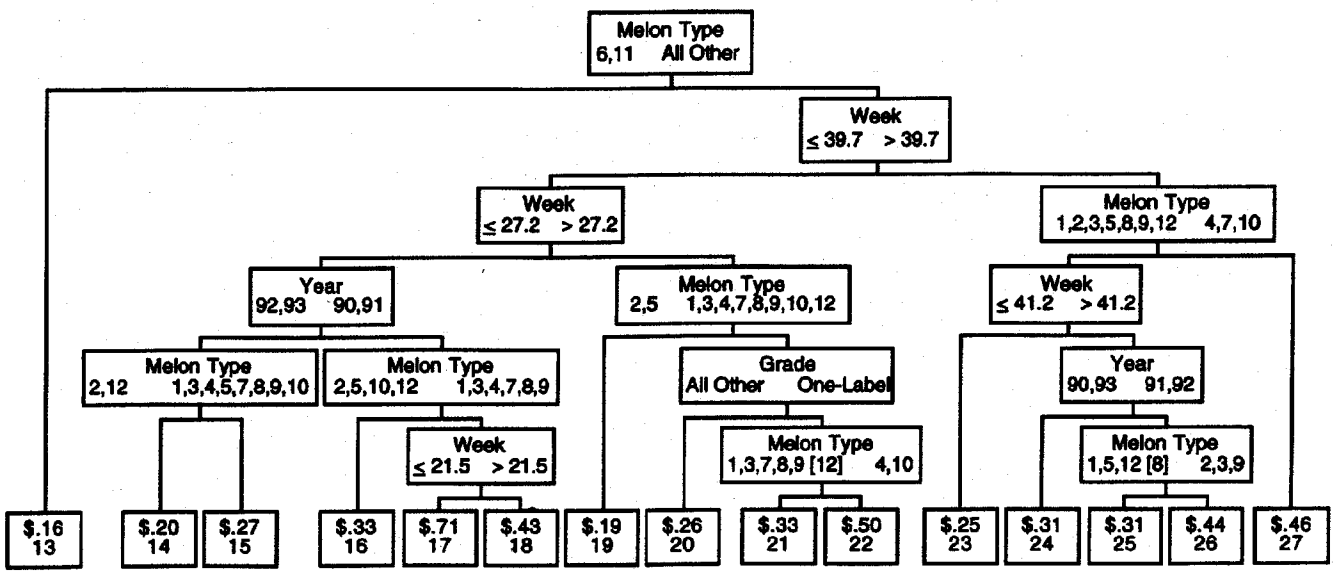


Figure 1. Predictor tree of weekly L.A. wholesale melon prices ($/lb.), 1990–93

[a]Size ≤ (>) x.xx implies a size number that is less than or equal to (greater than) x.xx standard deviations away from the average size number for each melon type. Larger numbers indicate a smaller melon. Items in brackets indicate that no price quotas were in the learning or test samples for the node characteristics specified.

Panel B. Week greater than 18.9 and less than or equal to 46.3



Melon Type	Average Size	SD in Size
1. Canary	6.043	1.323
2. Cantaloupes	16.064	5.155
3. Casaba	5.691	1.460
4. Crenshaw	5.504	1.390
5. Honeydew	6.879	1.796
6. Mayan	5.532	0.863
7. Orange-Flesh	7.543	1.679
8. Persian	5.476	1.174
9. Santa Claus	5.736	1.438
10. Sharlyn	5.989	1.495
11. Red-Flesh Watermelons	4.738	0.972
12. Seedless Watermelons	5.160	0.866

Grades: One-Label, Good Quality, Fair Quality, Fair Condition, Poor Quality

Figure 1. Continued

Table 1. Number of Cases, Average Values, and Sample Standard Deviations by Terminal Node for Learning and Test Sample Values Given in Figure 1

	Learning Sample			Test Sample		
Node	No. Cases	Avg. ($/lb.)	SD ($/lb.)	No. Cases	Avg. ($/lb.)	SD ($/lb.)
1	160	0.293	0.060	61	0.308	0.093
2	40	0.580	0.150	20	0.561	0.150
3	308	0.403	0.077	161	0.394	0.085
4	80	0.322	0.084	40	0.328	0.087
5	60	0.437	0.120	44	0.402	0.160
6	137	0.545	0.150	65	0.523	0.140
7	82	0.394	0.086	48	0.401	0.081
8	109	0.602	0.150	60	0.596	0.160
9	41	0.472	0.160	21	0.473	0.150
10	40	0.598	0.150	26	0.573	0.130
11	22	0.892	0.150	16	0.825	0.170
12	3	0.550	0.024	0		
13	298	0.156	0.052	144	0.157	0.050
14	125	0.204	0.073	67	0.179	0.070
15	188	0.267	0.080	100	0.282	0.097
16	160	0.325	0.100	76	0.358	0.120
17	13	0.714	0.170	7	0.568	0.160
18	61	0.425	0.100	29	0.429	0.087
19	332	0.186	0.060	137	0.189	0.050
20	484	0.257	0.067	208	0.261	0.071
21	52	0.327	0.050	22	0.333	0.051
22	17	0.496	0.048	12	0.461	0.110
23	59	0.253	0.054	22	0.310	0.100
24	133	0.313	0.073	54	0.325	0.066
25	48	0.310	0.052	25	0.330	0.054
26	60	0.441	0.073	45	0.421	0.068
27	39	0.460	0.110	22	0.441	0.086
28	58	0.279	0.110	21	0.320	0.120
29	77	0.396	0.130	34	0.378	0.120
30	51	0.345	0.097	30	0.331	0.088
31	43	0.484	0.100	21	0.463	0.085
32	12	0.689	0.072	7	0.576	0.039
33	11	0.311	0.090	9	0.346	0.081
34	11	0.543	0.044	5	0.588	0.016
35	69	0.617	0.120	44	0.626	0.180
Total	3483			1703		

average price of $0.892/lb. Given the small number of cases, it is not surprising that no cases were found in the test sample. A procedure such as least absolute deviations rather than least-squares regression removes these "outliers" from being in their own node. But these outliers are salient features of the data by regression criteria since they were sorted into their own terminal node, even though their relative numbers are small.

Figures 2, 3, and 4 plot estimated CART terminal nodes versus average cantaloupe, honeydew, and watermelon prices by week of calendar year.[8] These figures illustrate the importance of seasonality for melon prices. Both the average price and CART terminal nodes plotted demonstrate how prices can plummet after 13 May. This sharp drop reflects the first large influx of new crop shipments from temperate areas of California, Florida, and Texas, plus continued imports from Mexico and Central America. It is not surprising that growers seek out warmer microclimates and use mulches or foams in an effort to achieve a harvest one or two weeks ahead of the spring price drop.

Red-flesh watermelon and mayan melon prices bottom out for all years after week 18.9 at around $0.16/lb. and remain there until late fall. CART gives these two melon types the same price for all years between weeks 18.9 and 46.3, the longest period for all melons. Prices continue to drop for other melon types until week 27.2, 11 July, when they hit bottom for the year. Prices are most stable from year to year during the midsummer season, between weeks 27.2 and 39.7. During this period, terminal price nodes for all melons (i.e., 13, 19, 20, 21, and 22) are the same for all years. Yearly price uncertainty is lower for these weeks than any other period, but prices are least favorable then too. In considering all melon types, when week is less than 18.9 or greater than 46.3 (i.e., panel A), several terminal nodes yield a price-predictor that is greater than $0.50/lb. But for the period dominated by U.S. production (i.e., between 13 May and 21 November, which is presented in panel B), only one terminal node is above $0.50/lb., node 17. This node only includes "specialty melons" (canary, casaba, crenshaw, orange-flesh, persian, and santa claus) and is for a very short window, between 13 May and 1 June, in 1990 and 1991.

Melon prices start to increase in the fall around 6 October (week 39.7), albeit more gradually than the price drop in the spring. Terminal nodes for shorter time periods in the fall than spring reflect the more gradual price increase. U.S. melon supplies increase in the spring more sharply with the new crop than production drops off in the fall. Melons are planted only after minimum soil temperatures have been reached. Few heat units accumulate in early spring so that most melons are ready for harvest at about the same time, even if plantings are a few weeks apart. Also, a light freeze in the spring would kill most seedlings, but many vines may keep producing after a light freeze in the fall. Post-harvest life also appears to affect how quickly prices increase in the fall. Cantaloupes have the shortest post-harvest life and they exhibit the sharpest price increase in the fall.

A price premium was most consistently exhibited for crenshaw, orange-flesh, and sharlyn melons. Red-flesh watermelons and mayan melons were at the lowest price level ($0.156/lb. for node 13) for most of the U.S. harvest season. Seedless watermelons receive a premium relative to red-flesh watermelons, except after week 46.3. For other weeks, the premium ranged anywhere from a high of $0.25/lb. (weeks between 4.85 and 13.1 for 1990 and 1992) to a low of $0.05/lb. (weeks between 18.9 and 27.2 for 1992 and 1993).

---

[8]Cantaloupe prices plotted are for a size number less than 21.58 [i.e., 16.064 + (1.07)(5.155)] and all grades except "one-label." Honeydew and watermelon prices reflect a good quality grade with a size number less than 8.80 for honeydews and all sizes for watermelons. Average prices were calculated by first taking the average of all price quotes for each week of a year. Then the average of each week for all years was taken so that each year would have equal weight.

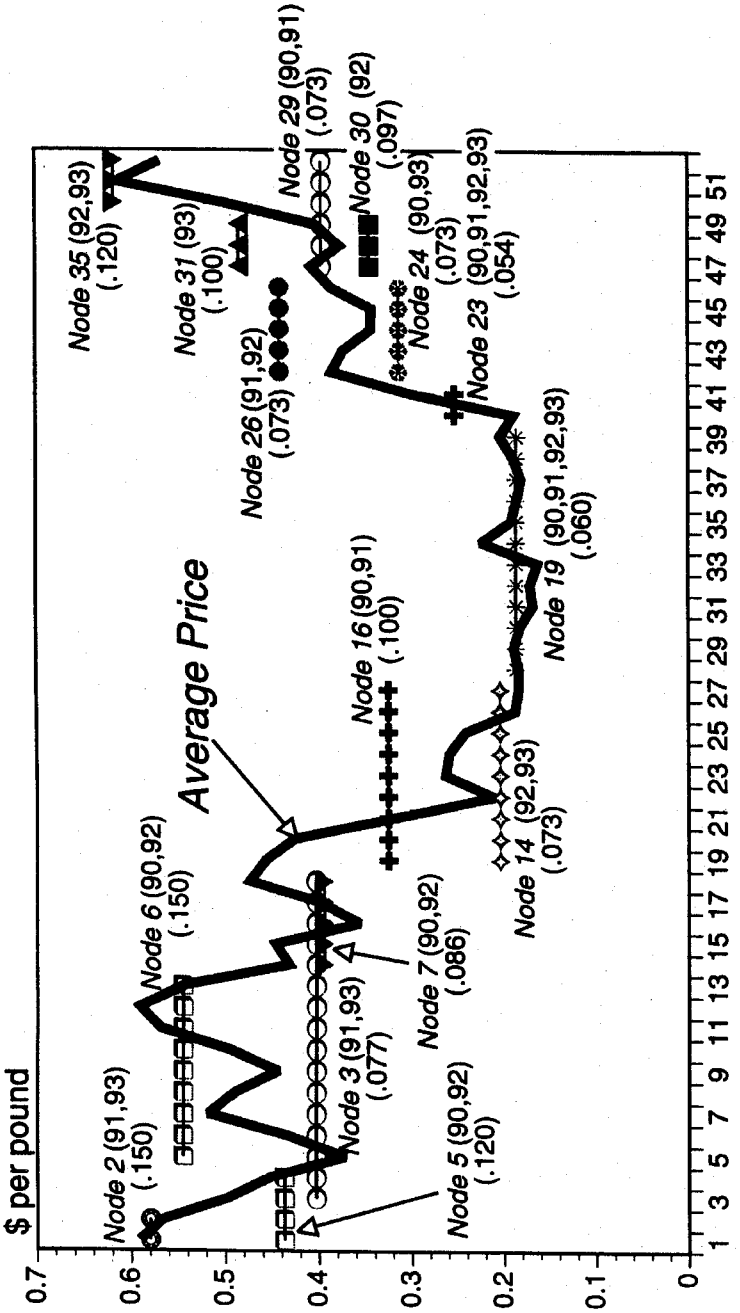


Figure 2. CART terminal nodes versus average cantaloupe prices, 1990–93

Note: Size number less than 21.58 and all grades except "one-label." Sample standard deviations of the "learning sample" are given in parenthesis below each terminal node number. Applicable year(s) are given to the right of each node number.

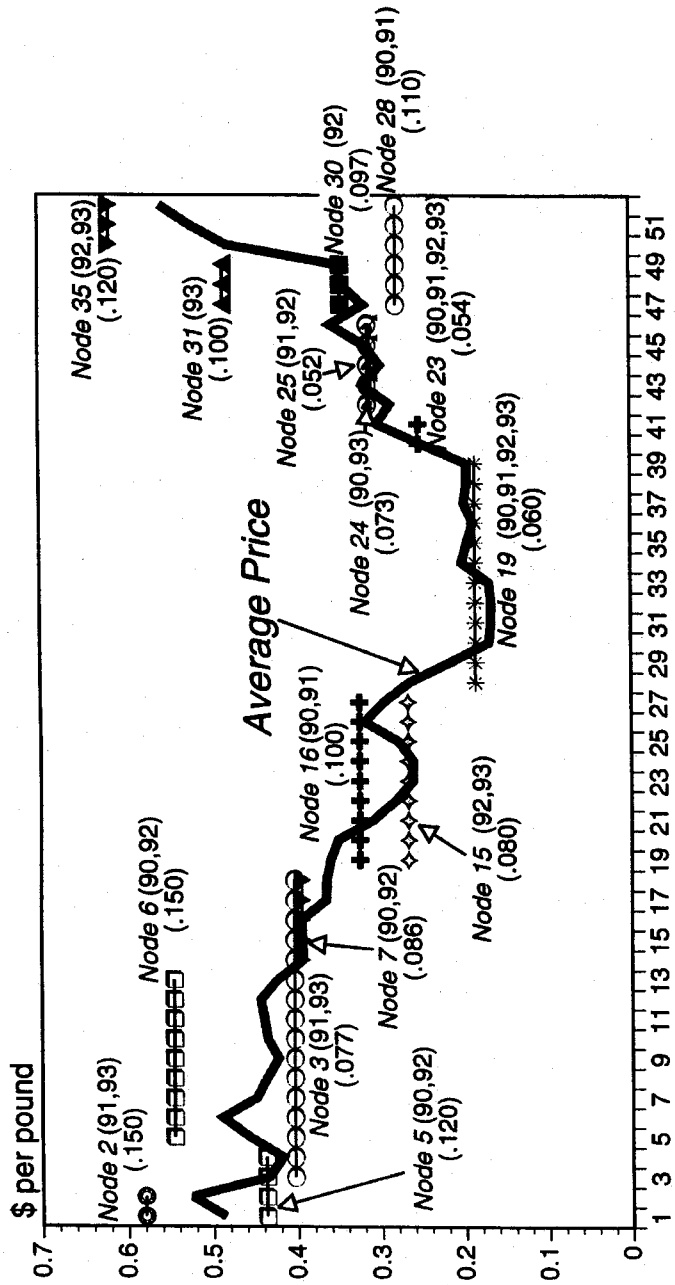


Figure 3. CART terminal nodes versus average honeydew prices, 1990–93

Note: Size number less than 8.80 and good quality. Sample standard deviations of the "learning sample" are given in parenthesis below each terminal node number. Applicable year(s) are given to the right of each node number.

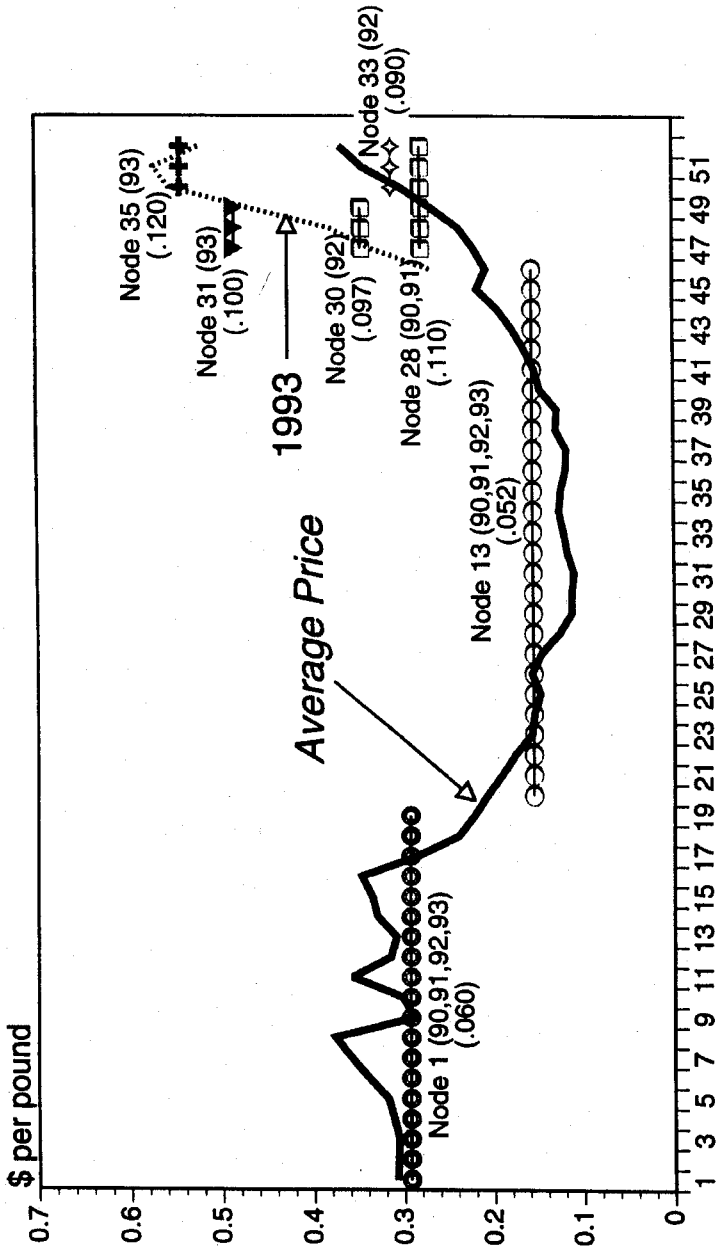


Figure 4. CART terminal nodes versus average red-flesh watermelon prices, 1990–93

Note: All grades, sizes, and varieties. Sample standard deviations of the "learning sample" are given in parenthesis below each terminal node number. Applicable year(s) are given to the right of each node number.

CART classified honeydew and cantaloupe prices in the same terminal nodes, except for a few cases, as shown in figures 1, 2, and 3. In May and June of 1992 and 1993, honeydew prices are estimated slightly higher than cantaloupes (i.e., nodes 15 vs. 14). But in the late fall of 1990, 1991, and 1992, cantaloupe prices are estimated higher (i.e., nodes 26 vs. 25, and nodes 29 vs. 28). The first diversion is probably from the earlier maturity of some cantaloupe varieties, while the latter probably reflects a shorter post-harvest life of cantaloupes than honeydews.

Price splits for grade occur only for "one-label" versus all "other grades." Splits occur when week is less than 18.9, or between weeks 27.2 and 39.7. This result suggests that "one-label" grades receive a premium only for limited time periods, and no detectable discount is received for grades that are worse than good quality. Because little grade variation occurs (recall that only 3.8%, or 197 price quotes make up the three poorest grades), price sensitivity to poorer grades may have been overwhelmed by sheer numbers. But CART can isolate a few price quotes, as in terminal node 12, suggesting that price quotes for the poorer grades were not overwhelmed by absolute numbers.

Following all "one-label" versus "other-grade" splits is a binary split associated with melon type. Crenshaw, orange-flesh, and sharlyn show a significant premium for "one-label" over other grades. A premium for "one-label" orange-flesh appears for weeks less than 18.9 but not for weeks between 27.2 and 39.7. No "one-label" sharlyn melon price quotes were given for week less than 18.9. CART arbitrarily places these and other categorically missing melon types "left." Arbitrarily placing absent categories from the learning sample to the left causes no harm as long as the specified categories continue to remain nonexistent. But this is a shortcoming of the data and CART approach if melon production were targeted for "one-label" sharlyn melons when week is less than 18.9. The data have no "one-label" sharlyn when week is less than 18.9, and the CART algorithm fails to associate the price premium for weeks 27.2 to 39.7 with week less than 18.9. Due to this lack of association and nonlinear flexibility inherent in the CART method, special caution needs to be given for making extrapolations beyond the data. Price predictions are denoted in brackets, in figure 1 for node characteristics specified that have no price quotes in the learning or test samples. This happens only for specialty melon types during specific time periods in the late fall, winter, and early spring months.

A price split for size was identified only when week was less than 18.9 and for crenshaw, mayan, and orange-flesh melon types. The two size splits identified were both for a size number that was almost one standard deviation (0.792 and 1.33) above their respective average sizes. This indicates that no price premium was identified for melons larger than average size, since smaller numbers reflect larger melons. A binary split associated with a size number less than zero in figure 1 would be necessary for melons larger than average to receive a premium. A discount for small size does not occur until some specialty melons are about one standard deviation in size less than average. Size appears not to have a major influence on melon prices.

The relative importance of variables for explaining melon prices is as follows: week (100), type of melon (76), year (45), size (40), grade (34), and container (18). As depicted in the predictor tree and relative importance ranking, seasonality or week is the most important factor that determines melon prices with melon type not far behind. Year is next, reflecting that plantings and weather vary significantly from one year to the next. Both size and grade splits occur only twice in the price-predictor tree. But the relative importance variable is higher for size than grade, indicating that size has better surrogate splits than

grade. Shipping container ranks at the bottom, and this is consistent with no container splits identified in the price-predictor tree.

How good is the overall fit of the nonparametric CART procedure? The coefficient of determination or R^2 (Kmenta) was calculated at 0.6795 for all melon prices using the learning sample predictors from figure 1 or table 1.[9] This compares favorably with an R^2 of 0.6061 calculated from a parametric ordinary least squares (OLS) regression equation with dummy variables for melon type, year, week rounded to the nearest integer (0 to 53), grade, container, and the continuous size variable. In total, 73 dummy variables, size, and a constant term were included in the OLS regression.[10] The number of dummy variables in the OLS regression are more than double the binary splits (34) or terminal nodes (35) in the CART predictor tree. The mean absolute percent error was calculated at 22.71 for CART and 27.27 for the OLS regression. CART performed better since it allows for interactions between variables. For example, the premium for seedless watermelons relative to other melon types can be higher for some weeks than other weeks. Without interactive dummy variables in the OLS regression, the premium for seedless watermelons is fixed constant for every week of the year.[11]

Similarly, if the overall price level for cantaloupes starts out high as in January 1991, this is no indication that weeks to follow for 1991 will be at a seasonally higher or lower price than previous years. Prices for a year are never always above or below the seasonal price quotes of other years (figs. 2, 3, and 4). Supply shifts from one week to the next, reflecting the perishable nature of the crop, changes in geographic production, and relatively inelastic supply for a given week. Prices are grouped together for all years only between the weeks of 27.2 and 41.2 (terminal nodes 19, 23, and 13). Dummy variables are much more amenable for capturing yearly effects of crops that are on an annual production and storage cycle rather than a commodity like fresh melons. Thus, a strength of the CART approach appears to lie with its ability to identify interactions between discrete variables without requiring an unduly large number of dummy variables as may be the case for a parametric regression equation.

Concluding Comments

CART, a computer intensive nonparametric regression procedure, was used to determine how melon type, size, grade, shipping container, week, and year influence melon prices. The "relative importance" of variables calculated by CART ranked week (100), type of melon (76), year (45), size (40), grade (34), and shipping container (18) as the most-to-least important factors. Given the importance of time variables, parametric procedures may be justified in focusing only on seasonality for a longer series of data, even if the data represent just one grade, size, and type of melon.

A price-predictor tree constructed by CART with 34 binary splits, or 35 terminal nodes, explained melon prices favorably to an OLS regression equation with 73 dummy variables, size, and a constant term. The mean absolute percent error was 22.71 for CART and 27.27 for the OLS regression. Similarly, the R^2, or coefficient of determination, was 0.6795 for CART and 0.6061 for the OLS regression. CART performed better since it allows for

[9] R^2 values are 0.6051 and 0.7600 for predictor trees with 19 (SER of 7) and 89 (SER of 3) terminal nodes, respectively.

[10] When size specifications were missing from watermelon quotes, the average size for all watermelons was used instead. Results were the same as regressing size on all other variables and then using these estimates of size for missing values.

[11] Interaction dummy variables (e.g., week multiplied by year) would have been unmanageable if all were considered as variables in the OLS regression.

interactions between discrete variables. Allowing for these interactions in the OLS regression would have required an unduly large number of interaction dummy variables.

Crenshaw, orange-flesh, and sharlyn melons exhibited the most consistent premiums for melon type, while red-flesh watermelon and mayan melons generally received a discount. No premium was found for melons larger than average. Discounts for small sizes were found only for limited melon types and time periods. The only grade price differential detected was for a "one-label" grade associated with limited melon types and time periods. Year was an important factor for all time periods except between 11 July and 6 October.

Melon prices show their first big price drop in the spring after 13 May. Most melon prices drop further until they reach bottom at around 11 July. Prices are the lowest and most predictable between 11 July and 6 October. After 6 October, prices start to increase, but more gradually than they drop in the spring. Most melon prices peak in mid-December.

[*Received June 1994; final version received January 1995.*]

References

Bowman, K. R., and D. E. Ethridge. "Characteristic Supplies and Demands in a Hedonic Framework: U.S. Market for Cotton Fiber Attributes." *Amer. J. Agr. Econ.* 74(1992):991–92.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Belmont CA: Wadsworth Publishing Co., 1984.

Brorsen, B. W., W. R. Grant, and E. M. Rister. "A Hedonic Price Model for Rice Bid/Acceptance Markets." *Amer. J. Agr. Econ.* 66(1984):156–63.

Collette, A. W., and G. B. Wall. "Evaluating Vegetable Production for Market Windows as an Alternative for Limited Resource Farmers." *S. J. Agr. Econ.* 10(1978):189–93.

Efron, B., and R. Tibshirani. "Statistical Data Analysis in the Computer Age." *Science* 253(1991):390–95.

Epple D. "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products." *J. Polit. Econ.* 95(1987):59–80.

Ethridge, D. E., and B. Davis. "Hedonic Price Estimation for Commodities: An Application to Cotton." *West. J. Agr. Econ.* 7(1982):293–300.

Goodwin, H. L., Jr., S. W. Fuller, O. Capps, Jr., and O. W. Asgill. "Factors Affecting Fresh Potato Price in Selected Terminal Markets." *West. J. Agr. Econ.* 13(1988):233–43.

Hanemann, W. M. "Quality and Demand Analysis." In *New Directions in Econometric Modeling and Forecasting in U.S. Agriculture*, ed., G. C. Rausser. New York: Elsevier Science Publishing Co., 1982.

Horowitz, J. K., and R. T. Carson. "A Classification Tree for Predicting Consumer Preferences for Risk Reduction." *Amer. J. Agr. Econ.* 73(1991):1416–421.

Jones, L. E. "The Characteristics Model, Hedonic Prices, and the Clientele Effect." *J. Polit. Econ.* 96(1988):551–67.

Kmenta, J. *Elements of Econometrics.* New York: Macmillan Publishing Company Incorporated, 1971.

Lancaster, K. J. *Consumer Demand: A New Approach.* New York: Columbia University Press, 1971.

Lenz, J. E., R. C. Mittelhammer, and H. Shi. "Retail-Level Hedonics and the Valuation of Milk Components." *Amer. J. Agr. Econ.* 76(1994):492–503.

Lucas, R. E. B. "Hedonic Price Functions." *Econ. Inquiry* 13(1974):157–78.

Rosen, S. "Hedonic Prices and Implicit Market Product Differentiation in Pure Competition." *J. Polit. Econ.* 82(1974):34–55.

The Packer, *Produce Services Sourcebook '93.* Vance Publishing Corp., 1993.

Tronstad, R., and R. Gum. "Cow Culling Decisions Adapted for Management with CART." *Amer. J. Agr. Econ.* 76(1994):237–49.

Tronstad, R., L. S. Huthoefer, and E. Monke. "Market Windows and Hedonic Price Analyses: An Application to the Apple Industry." *J. Agr. and Resour. Econ.* 17(1992):314–22.

U. S. Department of Agriculture, Federal State Market News Service. *Los Angeles Wholesale Fruit and Vegetable Report*, Weekly, 3 January 1990 through 28 December 1993.

Unnevehr, L. J., and S. Bard. "Beef Quality: Will Consumers Pay for Less Fat?" *J. Agr. and Resour. Econ.* 18(1993):288–95.

Veeman, M. "Hedonic Price Functions for Wheat in the World Market: Implications for Canadian Wheat Export Strategy." *Can. J. Agr. Econ.* 35(1987):535–52.
Wahl, T. I., H. Shi, and R. C. Mittelhammer. "A Hedonic Price Analysis of the Quality Characteristics of Japanese Wagyu Beef." Selected paper, WAEA annual meetings, Edmonton, Alberta, 11–14 July 1993.
Waugh, F. V. "Quality Factors Influencing Vegetable Prices." *J. Farm Econ.* 10(1928):185–96.
Wilson, W. W. "Hedonic Prices in the Malting Barley Market." *West. J. Agr. Econ.* 9(1984):29–40.