A Statistical Model to Estimate Potential Yields in Peach before Bloom

Carlos Miranda Jiménez and J. Bernardo Royo Díaz
Departamento de Producción Agraria, Universidad Pública de Navarra. Campus de Arrosadia 31006 Pamplona, España

ADDITIONAL INDEX WORDS. Prunus persica, crop density, average fruit weight, yield efficiency, trunk cross-sectional area, total shoot length

Abstract. Spring frosts are usual in many of Spain’s fruit-growing areas, so it is common to insure crops against frost damage. After a frost, crop loss must be evaluated, by comparing what crop is left with the amount that would have been obtained under normal conditions. Potential crop must be evaluated quickly through the use of measurements obtainable at the beginning of the tree’s growth cycle. During the years 1997 through 1999 and in 86 commercial plots of peach and nectarine [Prunus persica (L.) Batsch], the following measurements were obtained: trunk cross-sectional area (TCA, cm²), trunk cross sectional area per hectare (TCA/ha), estimated total shoot length per trunk cross-sectional area (SLT, shoot m/cm² TCA), crop density (CD, amount of fruit/cm² TCA), fruit weight (FW, g), yield efficiency (YE, kg of fruit/cm² TCA), yield per tree (Y, kg fruit/tree) and days between full bloom and harvest (BHP, days). CD and average FW were related to the rest of the variables through the use of multiple regression models. The models which provided the best fit were CD = SLT – TCA/ha and FW = SLT + BHP – CD. These models were significant, consistent, and appropriate for all three years. The models’ predictive ability was evaluated for 32 different plots in 2001 and 2002. Statistical analysis showed the models to be valid for the forecast of orchards’ potential yield efficiency, so that they represent a useful tool for early crop prediction and evaluation of losses due to late frosts.

Spring frosts are usual in many of Spain’s fruit-growing areas, so it is common to insure crops against frost damage. Frosts affect mainly flowers or young developing fruit, so that any damage must be estimated comparing the amount of fruit left on the tree to the number that would have been produced under normal conditions. Insurance technicians face two main difficulties in their work: the study must be performed after the frost has occurred (that is, must be performed using variables which can be measured at that point) and it must be performed in the field, quickly and without complex equipment.

Tree yield is a function of both the number of fruit on the tree and their weight. The number of fruit depends on the load after pruning (the number of flowering buds left), the tree’s size and the percentage of set. The percentage of fruit set decreases as the bud load increases, but this relationship is heavily influenced by climate (Chang et al., 1987; Monselise, 1986). When fruit set is too large for the tree’s size, thinning becomes necessary to ensure that the remaining fruit reach commercial size (Costa and Vizzotto, 2000).

Typical fruit size is different for each cultivar (Faust, 1989; Grossman and DeJong, 1995), although it is also highly dependent on the number of fruit left on the tree (Chang et al., 1987; Rowe and Johnson, 1992), climate (Lakso et al., 1995, Warrington et al., 1999) and cultural techniques (Berman and DeJong, 1996; Naor et al., 1999). Fruit growth rate is greater for early cultivars but, since growth time is shorter, final size tends to be smaller (Grossman and DeJong, 1995; Pavel and DeJong, 1993). The time between full bloom and ripeness depends on cultivar and accumulated heat units (Fischer, 1962), although, when the areas under study have similar climates, it can also be measured in days (Warrington et al., 1999).

Only tree size, plantation density and flower bud load after pruning are available at the beginning of the active period to estimate potential yield in a simple, rapid way. Potential yield increases with tree size, although not linearly since bigger trees are less efficient (Barrett et al., 1997; Faust, 1989; Robinson et al., 1991). There is a direct relationship between tree size and the canopy volume (Farmahan et al., 1999; Marini and Sowers, 2000; Robinson et al., 1991), but canopy volume is difficult to measure, requiring the use of image analyzers which are costly and require a long time to employ (Giuliani et al., 2000; Robinson and Lakso, 1991). Also, at the beginning of the growing period, shoots have just started to elongate, so canopy volume will increase during the season. Many studies have shown that tree size and trunk cross-sectional area (TCA) are closely related, so that TCA is used regularly to compare different plots’ vigor (TCA/ha), efficiency (kg/TCA), flower load (number of flowers/TCA), etc. (Barrett et al., 1997; Caruso et al., 1999; Dennis et al., 1996; Lombard et al., 1988; Strong and Azarenko, 2000; Westwood and Roberts, 1970).

Plant density also influences yield, since it affects the amount of light received by the trees. Robinson and Lakso (1991) found that the amount of light intercepted by a tree had a direct correlation with TCA per available area (TCA/ha), so that TCA/ha could be used to compare the productivity of plots of different cultivars and densities.

Flower density (expressed as the ratio between number of flowers and tree TCA) is a good estimate of bud load after pruning, another one of the main factors in a fruit tree’s productivity (Chang et al., 1987; Kappel, 1990; Lombard et al., 1988). In the case of stone fruit trees and due to their growth patterns, flower density is best expressed in terms of shoot length rather than flowers per TCA, which also expresses the number of leaves which are available to supply the fruit (Lombard et al., 1988; Marini and Sowers, 1994; Pérez-González, 1993).

The purpose of this study was the development of a mathematical model to estimate potential yield for peach plots, using parameters which can be easily measured at the beginning of the active period.

Materials and Methods

Data collection. Data were collected during the years 1997–99 in fully producing commercial plots, with trees ranging between 6 and 16 years of age. Orchards were located in the Ebro...
Yield efficiency (YE, kg fruit/cm² TCA)  

Fruit weight (FW, g)  

Crop density (CD, no. fruit/cm² TCA)  

Total shoot length per TCA (SLT, m shoot/cm² TCA)  

Trunk cross-sectional area (TCA, cm²)  

Space allocated per tree (ST, m²)  

Estimated total shoot length area per hectare (TCA/ha)  

Estimated space allocated per tree (ST, m²)  

Yield per hectare (T/ha)  

Table 1. List of cultivars sampled in 1997–99 to develop the prediction models. Each cultivar is identified as nectarine (NECT), freestone peach (FREE) and clingstone peach (CLING), and the period from full bloom to harvest and number of orchards sampled is given.

| Cultivar       | Type of cultivar | Full bloom to harvest (d) | No. of orchards sampled |
|----------------|------------------|---------------------------|-------------------------|
| Arinka         | NECT             | 95                        | 2                       |
| Spring Lady    | FREE             | 103                       | 2                       |
| Tierrina       | CLING            | 106                       | 2                       |
| Caldesi 2000   | NECT             | 113                       | 2                       |
| Snowqueen      | NECT             | 117                       | 6                       |
| Catherine      | CLING            | 120                       | 8                       |
| Queen Giant    | NECT             | 114                       | 1                       |
| Red Lady       | FREE             | 127                       | 2                       |
| Flavour Giant  | NECT             | 131                       | 2                       |
| Baby gold 5    | CLING            | 135                       | 3                       |
| Baby gold 6    | CLING            | 140                       | 10                      |
| Red Top        | FREE             | 138                       | 4                       |
| Fantasia       | NECT             | 148                       | 4                       |
| Andross        | CLING            | 150                       | 1                       |
| Sudanell       | CLING            | 160                       | 12                      |
| Baby gold 9    | CLING            | 161                       | 4                       |
| O’Henry         | FREE             | 163                       | 5                       |
| Amarillo de Septiembre | FREE | 176                       | 3                       |
| Fairlane       | NECT             | 182                       | 10                      |
| Miraflores     | CLING            | 184                       | 3                       |

River Valley in the provinces of La Rioja, Navarra, Zaragoza and Lérida in Spain. A total of 86 plots were included in the study (Table 1), out of which 14 were freestone peach [Prunus persica (L.) Batsch Peach Group] cultivars, 45 clingstone peach cultivars and 27 nectarine [Prunus persica (L.) Batsch Nectarine Group] cultivars. The plots represented a wide range of growth conditions. Means and standard deviations are shown for each of the cropping characteristics under study (Table 2), both globally and separated by cultivar family (freestone peach, clingstone peach and nectarine). Cultural practices, including thinning, were carried out as is normally done commercially in the area. Thinning was performed by the growers between 50 and 80 d after full bloom (DAFB), with early cultivars being thinned earlier.

For each plot and year, 30 trees were chosen randomly between winter pruning and bloom; their trunks cross-sections were measured 15 cm above the grafting point, as well as the number of productive shoots left after pruning and distance to neighboring trees. Twenty shoots were marked in each of the chosen trees, so that all of them were on the same main branch and that they represented all positions (interior, exterior, superior, inferior) and the ratio at which they occurred in the trees. The length of those shoots was measured. Total shoot length per tree was estimated as the average number of shoots per tree multiplied by the average length of the 20 shoots marked.

Timing for harvesting was determined by measuring flesh firmness at the greenest part of 100 randomly chosen fruit in the plot, using a Penefel penetrometer (Agro-Technologie, Tarascon, France) with a 0.5 cm² point. Harvest took place when average firmness was 4.5 to 5 kg·cm⁻² for freestone peach and nectarine and 6 to 6.5 kg·cm⁻² for clingstone peach (Cemagref, 1988). At that moment, fruit were counted for each tree and the largest equatorial size was measured for the fruit on marked shoots. At the same degree of ripeness, 100 fruit were chosen randomly and their size and weight measured.

The following measurements were made: trunk cross-sectional area (TCA, cm²), space allocated per tree (ST, m²) trunk cross-sectional area per hectare (TCA/ha), estimated total shoot length per TCA (SLT, shoot m·cm⁻² TCA), crop density (CD, number of fruit/cm² TCA), average fruit weight (FW, g), yield efficiency (YE, fruit kg·cm⁻² TCA), tree yield (Y, fruit/kg/tree) and time between full bloom and harvest (BHP, days). For each plot, the values used in the regression analysis were the averages for all 30 sampled trees.

**Model building.** To obtain expected yield for a peach plot, yield was separated into its individual components (number of fruit and fruit weight) and multiple linear regression models linking each of those parameters with those that could be measured at the beginning of the activity cycle were examined. CD was measured as number of fruit/cm² TCA because this parameter is more useful than fruit/tree to compare different plots (Lombard et al., 1988). Therefore, two models were evaluated: in the first one CD was the dependent variable and the independent variables were TCA, ST, TCA/ha and SLT. In the second model, FW was the dependent variable and the independent variables were TCA, TCA/ha, SLT, CD and BHP. The relationships were evaluated by fitting multiple linear regression models with the Linear Regression Procedure of SPSS (SPSS Inc., Chicago, Ill.) and the Backward Stepwise Elimination Option. The backward stepwise elimination search procedure begins with a model containing all potential variables and identifies the variable with the smallest F value. If the minimum F value is less than a predetermined limit (usually a P ≤ 0.1), that variable is dropped. The model with the remaining variables is then fitted, and the next candidate for dropping is identified. This process continues until no further variables can be dropped. The stepwise term means that the procedure allows variables eliminated earlier to be added later if the F value is more than a predetermined limit (usually a P ≤ 0.05) (Neter et al., 1996). The internal validity of the models was tested by coefficient of determination (R²), mean square error (MSE) and...
Table 3. List of cultivars sampled in 2000–01 to validate the prediction models. Each cultivar is identified as nectarine (NECT), freestone peach (FREE) and clingstone peach (CLING), and the period from full bloom to harvest and number of orchards sampled is given.

| Cultivar          | Type of cultivar | Full bloom to harvest (d) | No. of orchards sampled |
|-------------------|------------------|---------------------------|-------------------------|
| Spring Lady       | FREE             | 102                       | 1                       |
| Crimson lady      | FREE             | 108                       | 1                       |
| Super Crimson     | NECT             | 108                       | 1                       |
| Caldesi 2000      | NECT             | 113                       | 1                       |
| Queen Giant       | NECT             | 114                       | 2                       |
| Big Top           | NECT             | 115                       | 1                       |
| Catherine         | CLING            | 120                       | 2                       |
| Elegant Lady      | FREE             | 137                       | 2                       |
| Baby Gold 6       | CLING            | 140                       | 3                       |
| Baby Gold 7       | CLING            | 147                       | 1                       |
| Baby Gold 8       | CLING            | 154                       | 1                       |
| Amarillo de Agosto| CLING            | 159                       | 1                       |
| Andross           | FREE             | 160                       | 1                       |
| Baby Gold 9       | CLING            | 161                       | 4                       |
| O’Henry            | FREE             | 163                       | 4                       |
| Tardibelle         | FREE             | 170                       | 1                       |
| Flamekiss         | NECT             | 174                       | 1                       |
| Caldesi 2020      | NECT             | 178                       | 1                       |
| Fairlane          | NECT             | 182                       | 1                       |
| Miraflores        | CLING            | 184                       | 2                       |

the prediction sum of squares (PRESS), and adjusted coefficient of multiple determination ($R^2_a$), defined by Neter et al. (1996) as the division of each of the sums of squares that constitute the $R^2$ by its associated degrees of freedom. Residuals were analyzed to determine the presence of outliers and nonconstant error variance. Variance inflation factor (VIF) for the parameter estimates was calculated to detect intercorrelation between variables. The best models were, for CD, CD = SLT – TCA/ha and, for FW, FW = SLT + BHP – CD. Analysis of covariance was then performed with the GLM Procedure of SPSS (Marini, 1999) to test equality of models for different years. The categorical variable (year) was included as the indicator or dummy variable and the continuous variables for the CD model (SLT and TCA/ha) and for FW model (SLT, BHP and CD), were included as the covariates. Type III sums of squares were used to evaluate the significance of the indicator variable and an interaction term involving the indicator variable (year $\times$ SLT or year $\times$ TCA/ha, for example). Because the indicator variable and the interaction terms containing the indicator variable were not significant ($P > 0.05$) for the indicator variable, data from all the orchards were pooled to develop two models: one for crop density estimation and another one for fruit weight estimation.

Model validation. To validate the models obtained, a validation data set was obtained using 32 plots different from the previous ones in years 2000 and 2001. Cultivars used and the number of plots for each one are listed (Table 3). Validation data were obtained in the same way as described earlier for the model building data set. Two techniques were used to validate the fitting models.

1) The validation data set was used to produce a validation model by reestimating the model parameters using the Backward Elimination Option approach to develop the estimation model and the models were compared for consistency.

2) Regression parameter estimates from the estimation models were used to predict outcomes for observations in the validation data set and then the mean squared prediction error (MSPE) was calculated and compared with the mean square error (MSE) of the regression fit to the model-building data set (Neter et al., 1996).

Results

Parameter estimates and statistics obtained from SPSS output are presented for the estimation and validation models developed for crop density (Table 4) and fruit weight (Table 5). Parameter estimates, $R^2$, and $R^2_a$ values are similar for the estimation and validation models, providing some assurance about the applicability of the model to data beyond those on which the model is based (Neter et al., 1996). In multiple regression analysis, it is very important to test the predictor variables for intercorrelation, because, when predictor variables are correlated, their regression coefficients do not reflect the inherent effect of the particular predictor variable on the response variable but only a marginal or partial effect. Calculation of the VIF of the parameter estimates is a formal method of detecting the presence of intercorrelation. This is a more accurate and useful method than the examination of the pairwise correlations method (Neter et al., 1996). When VIF value is above 10, it is an indication that there is a correlation between variables. The VIF for all parameter estimates were below 2.5 in both estimation models, so the coefficients obtained are reliable and the standardized regression coefficients (the regression coefficients divided by their respective standard deviations) can be used as a valid indicator of the relative influence of each parameter on the dependent variable (Neter et al., 1996). For the crop density model, SLT had the greatest impact, due to a standardized coefficient of 0.630, greater than TCA/ha (standardized coefficient = –0.309). For the fruit weight model, CD was the most influential factor (–0.798), whereas SLT and BHP had lesser influence and their weights were similar (0.525 and 0.444, respectively).

A means of measuring the actual predictive capability of the estimation regression models is to use the model to predict each case in the validation data set and then to calculate the mean of the squared prediction errors, MSPE. If the MSPE is fairly close to the MSE based on the regression fit to the estimation data set, then the MSE for the selected regression model is not seriously biased and gives an appropriate indication of the predictive ability of the model. In this case, the MSPE from the validation data set for the crop density model (0.104) does not differ greatly from the MSE of the estimation data set (0.103). The same case was true for the fruit weight model (MSPE = 256.39, MSE = 687.47). This implies that the MSE based on the estimation data set is a reasonably valid indicator of the predictive ability of the estimation regression model (Neter et al., 1996).

PRESS criterion is a measure of how well the use of the fitted values for a subset model can predict the observed values of the response value $Y_i$. The error sum of squares (SSE) is another such measurement. PRESS measure differs from SSE in that each fitted value for the PRESS criterion is obtained by deleting the $i^{th}$ case from the data set. The regression function is then estimated for the subset model from the remaining $n–1$ cases, and then the fitted regression function is used to obtain the predicted value of the $i^{th}$ case. PRESS was calculated for the models developed for crop density and fruit weight by summing the squared prediction errors (the difference between the observed and predicted values when one observation is deleted). Some evidence of the internal validity of the fitted model is to compare PRESS and SSE. PRESS value is always larger than SSE because the regression fit for the $i^{th}$ case, when this case is deleted in fitting, can never be as good as that when the $i^{th}$ case is included.
Table 4. Statistics and parameter estimates from regression models for crop density (CD) estimation. The estimation model was developed from 86 peach and nectarine orchards sampled in 1997, 1998, and 1999. Validation model was developed from 32 orchards sampled in 2000-01.

| Statistic or parameter estimate | Estimation model | Validation model |
|--------------------------------|-----------------|-----------------|
| Intercept                      | 1.446           | 0.773           |
| Standard error of intercept    | 0.337           | 0.380           |
| Regression coefficient for CD† | −0.079          | −0.038          |
| Standard error for CD          | 0.018           | 0.018           |
| Standardized regression coefficient for CD | 1.028 | 1.015 |
| Variance inflation factor for CD | 1.548 | 2.321 |
| Regression coefficient for SLT† | 3.791           | 4.743           |
| Standard error for SLT         | 0.431           | 0.588           |
| Standardized regression coefficient for SLT | 0.630 | 0.775 |
| Variance inflation factor for SLT | 1.548 | 2.321 |
| Prediction sum of squares (PRESS) | 42.752         |                  |
| Error sum of squares (SSE)     | 39.830           | 3.231           |
| Mean squared prediction error (MSPE) | 0.480         | 0.111           |
| Mean square error (MSE)        | 0.480           | 0.111           |
| P > F                          | <0.0001         | <0.0001         |
| Coefficient of multiple determination $R^2$ | 0.725 | 0.884 |
| Adjusted coefficient of multiple determination $R^2$ | 0.710 | 0.876 |
| Total DF                       | 85              | 31              |

$^a$TCA/ha = Trunk cross-sectional area per hectare (m²).
$^b$SLT = total shoot length per TCA (m of shoot length/cm² TCA).

PRESS values are reasonably close to SSE, as for CD model PRESS = 42.75, SSE = 39.83 and for FW model PRESS = 60862.39, SSE = 56372.21, and support the validity of the fitted regression models and of the MSEs as an indication of the predictive capability of these models (Neter et al., 1996).

PRESS residuals can be used to generate another $R^2$-like statistic that reflects prediction capabilities (Myers, 1990). For the model developed for crop density, this statistic is calculated as $R^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\text{total sums of squares}} = 1 - \frac{42.7(144.6)}{60862.4/129396.7} = 0.530$. The loss in $R^2$ is very small in both models, so there is a reasonably strong evidence that the least squares models will be satisfactory predictors.

**Discussion**

Total length of productive shoots per TCA (SLT) had the greatest weight in explaining crop density, probably because there is a...
close correlation between SLT and number of flowers and leaves on a tree, which is supported by previous research (Marini and Sowers, 1994; Pérez-González, 1993). Also, SLT is much easier to control than the number of flowers or leaves. Using TCA/ha instead of TCA provided a better fit for crop density. This matches the results reported by Robinson and Lakso (1991), who studied the efficiency of two apple cultivars on four rootstocks and three training systems and determined that TCA/ha showed the best correlation with amount of light intercepted by the tree and that TCA/ha is the factor which best explains tree efficiency.

Crop density has the greatest influence on fruit weight, although the influences of SLT and BHP are also significant. CD’s negative influence on fruit weight has been shown in previous studies (Blanco et al., 1995; Chang et al., 1987; Johnson and Handley, 1989; Kappel, 1990; Rowe and Johnson, 1992). Marini and Sowers (1994) studied the influence of CD and total shoot length on trees of ‘Redhaven’ and found that although CD was the most important factor, shoot length also had an important effect on FW because shoot length has a close relationship with the number of canopy leaves, thus is a good estimate for the tree’s photosynthetic capability. The influence of the period between full bloom and harvest (BHP) on FW is similar to that of SLT, so that this varietal parameter can be used to evaluate yield potential in a large number of peach and nectarine cultivars. For the FW model, the coefficient of determination ($R^2 = 0.56$) is not very high but, since it is consistent and has good predictive ability, and since the purpose of this research was to estimate the yield of peach and nectarine orchards in a very large area and for a very wide range of cultivars through the use of parameters which can be measured at the beginning of the cycle, it can be considered valid within reasonable error limits.

Figure 1 presents the relationship between observed and predicted yields, showing that global performance for the model (prediction $= 0.96$ observed, $R^2 = 0.84$) is very good, thus the models are valid to forecast crop efficiency, thus providing a useful tool for early forecast of yield and evaluation of crop losses due to late frosts.

The results of this study show that the models developed are good predictors of crop density and fruit weight in the Ebro River Valley, so that a orchards’ potential yield can be determined reasonably well from the knowledge of TCA/ha, SLT and BHP for the tree’s cultivar.