MAIZE YIELD ESTIMATION USING EAR SIZE - AN EXAMPLE FROM NORTHERN VERACRUZ

Larry W. Harrington

A) Background

Knowledge of current farm-level crop yields may often be of great assistance to researchers burdened with the responsibility of improving those yields. This knowledge may serve a variety of uses:

(1) Provide quantitative information on the extent to which yield improvement has been achieved,[1]

(2) Provide a check on yields obtained on "farmer technology" control plots in agronomic trials,

(3) Provide information on the performance of technological components under farm conditions. For instance if results of agronomic trials indicate that a recently released variety should provide increased yields, it might be wise to check this by comparing the actual yields of "users" with those of "non-users", at various levels of use of other relevant inputs.

(4) Provide useful data for economic planners, statisticians responsible for national accounts, etc.

To whatever use yield data is to be put, however, and at whatever level of aggregation,[2] those responsible for obtaining it face serious problems. There are several methods of estimating farm-level yields but each has its disadvantages. For example:

(1) Farm surveys: Questions are often included in farm surveys in an effort to estimate yields. Sometimes the farmer is requested to report yields directly. More frequently, the farmer is requested to report total production

---

[1] In practice, this is a challenging task, given the difficulty in separating weather effects on yield from those caused by technological change.

[2] This paper deals with estimating yields at the level of the "field". Weighting techniques for aggregating these estimates to higher levels are not treated here.

and the harvested area for which this production was taken, so that an implicit yield may be derived. It is not uncommon, however, for farmers to give relatively imprecise estimates of both variables. Indeed, the farmer may only have a rough idea of how large an area was harvested (in the absence of measured fields) or of the amount harvested (when fields are harvested bit by bit or when production has a variety of uses). Finally, it is not unknown for farmers to systematically report lower yields than those actually achieved.

(2) Whole-Field Harvesting: A far more accurate way of estimating yields would be harvesting whole fields of randomly selected farmers, directly measuring production and harvested area. Of all alternatives, however, this is by far the most expensive.

(3) Sample Harvesting: A method which should provide data almost as accurate on method (2), but which is sure to be far cheaper is one that makes greater use of sampling procedures. With this method, farmers' fields are selected at random. Within each selected field, however, several samples are cut. Field weight and harvested area of each sample are carefully measured and a subsample is taken for laboratory analysis so that proper discounts may be made for moisture content and shelling. The estimated per hectare yield for the whole field, then, is derived by taking the average (weighted by harvested area) of the various sample per hectare yield estimates. While this method is less expensive than method (2) it is nonetheless not cheap. Ears must be cut and weighed. Sub-samples must be sent for laboratory analysis. The farmer's permission must be obtained and he must be paid for maize carried from the field, or that maize must subsequently be returned.

In an effort to obtain farm-level estimates of maize yields that are more accurate than those obtained through surveys and significant cheaper than those obtained through sample harvesting, another method has been reported.

This method, versions of which are used in both Mexico and Kenya, infers yields from ear size.

A large scale effort was made by the Puebla Project in Mexico to select and fit a regression equation which would accurately predict grain weight (at 12% moisture) of an ear when given the length and diameter of that ear[3]. The purpose of the exercise was to measure, over time, the increase in regional maize yields and to determine what part of this increase was due to project intervention. With such an estimate of grain weight, along with an estimate of harvested density, per hectare maize yields may be calculated. Another equation was fitted for data aggregated to the "sample" level[4]; this equation was slightly less precise but possibly more appropriate for practical use in the field. The specified equation used for both ear-level and sample-level data was the same:

$$(Y)^{1/2} = a + b\,D + b_2 L + b_3 D^2 + b_4 L^2 + b_5 DL + b_6 D^2 L + b_7 DL^2 + e \qquad (1)$$

where Y = grain weight at 12% moisture,

D = ear diameter

L = ear lenghth

e = error

and b_i and a are parameters[5]. The ear-level equation achieved an R2 of .92. The equation did not seem to be sensitive to changes in fertilization level, variety of location.

Work in Kenya by the Maize and Produce Board proceeded along similar lines[6]. As in Mexico, the measurement of average "regional" per hectare yields was of prime interest. Further work was carried out, however, to relate yields to "husbandry" levels. The ear level equation used in Kenya was:

---

[3] "Estimación de Rendimientos de Maiz en el Area de Trabajo del Proyecto Puebla Mediante un Modelo de Regresión en Base a Diámetro y Longitud de Mazorcas", por Heliodoro Díaz, Delbert T. Myren y Richard E. Lund.

[4] The sample consisted of a 10 meter long section of a row, randomly selected. Five such samples were selected per field.

[5] It is unclear in what units or in what fashion measurements of D and L were taken. With Calipers? With a tape measure? With the husk intact or removed? in cm or inches? The square toot transformation was used to reduce heterosedasticity.

[6] C. Hesselmark 'Maize Yields in Kenya, 1975" Maize and Produce Board, Nairobi, March 1976

ear-level equation used in Kenya was:

$$W = a + bD^2L + e \qquad (2)$$

when W = grain weight at 13% moisture

D = diameter of ear measured outside husk

L = Length of ear measured outside husk

e = error

and a and b are parameters. The fitted equation showed on R^2 of .93. Field-level per hectare yields were estimated by multiplying average per ear grain weight by an independent estimate of "harvested plant density"[7]. Results obtained in this manner were double-checked in a small number of cases by simultaneously carrying out a sample harvest. The two yield estimates were very close for "regular" fields (i.e., consistent density, little lodging) but inconsistent for very irregular fields.

The CIMMYT Economics Section, in support of the Maize Training Program, decided to field-test maize yield estimation methods that infer yield from ear size. Northern Veracruz, Mexico, was chosen as the test site given that this is where the Maize Training Program of CIMMYT carries out its field work.

B) Research Design and Field Procedure

The first question addressed was that of level of aggregation. The decision was made to use the sample, not the ear, as the basic unit of analysis. Thus reported diameter and length measurements (all of which are in cm) represent the respective sums over the ears contained in the sample.

Field work was conducted for one wet cycle and one dry cycle. Farmers were chosen from collaborators in on-farm agronomic experiments (dry cycle) or from respondents from a farm survey previously conducted in the area (wet cycle). One field was was chosen for each selected farmer. Nineteen fields were examined

---

[7] Thirty ears were chosen at random per field.

in all, seven in the dry cycle and twelve in the wet cycle, with a total of three varieties represented. For each field, four samples were "randomly" selected[8]. A total of 98 samples were examined. For each sample, the following was collected:

(1) Sample row length: The sample was begun exactly between two maize hills and extended 5-6 meters, terminating exactly between two more hills.

(2) Row width, averaged over a 5 row measurement

(3) Number of plants in the sample (including lodged plants producing acceptable ears)[9].

(4) Number of ears

(5) Field weight of unhusked ears (grams)

(6) Sum of sample ear circumferences measured at the base of the unhusked ears (cm, using a tape measure); this was later converted to diameter by algebraic means.

(7) Sum of sample ear lengths measured over grain-fill of the unhusked ears (cm).

(8) A sub-sample of 3-4 ears, from which sample-specific husking rates, moisture content and shelling rates were derived.

Per sample yields (in per hectare terms) were calculated by means of the following

$$Y = (FWP)\left(\frac{FWX}{GWX}\right)\left(\frac{10,000}{HA}\right)\left(\frac{100}{85}\right) \qquad (3)$$

where Y = per hectare yield at 15% moisture

FWP = field weight of sampled ears

FWX = sub-sample fresh weight (with husks)

GWX = sub-sample dry grain weight

---

[8] Six samples per field were used in the dry cycle. Selection of samples was not carried out using random numbers but rather in "pseudo-random" fashion, e.g. beginning a sample after closing one's eyes and walking a pre-determined number of steps.

[9] Acceptable in the sense that a farmer would carry it in from the field.

HA = sample harvested area[10].

Average field yields were calculated by averaging per sample yields, weighted by harvested area.

C) Descriptive Results

Due to the fact that neither farmers nor fields were randomly chosen, the following results should not be used to infer population averages. Nonetheless, they should be of some interest. Major results are summarized in Table 1.

TABLE 1

Descriptive Statistics over 94 samples

Variable	Average	CV
Husking Rate	.833	11%
Moisture %	28.6	44%
Shelling Rate	.833	6%
Yield (kg/ha)	2888	58%
Density (pl/ha)	25,497	29%

Further results, disaggregated by cycle and variety, are shown in Tables 2 and 3.

TABLE 2

Descriptive Statistics, by Cycle

	Dry Cycle (78A)		Wet Cycle (78B)	
Variable	Average	CV	Average	CV
Moisture %	33.4	41%	23.9	37%
Yield (kg/ha)	2397	78%	3369	37%
Density (pl/ha)	NA	-	25,497	29%

[10] Multiplication of the other terms by 100/85 converts dry yield to yield at 15% moisture.

TABLE 3

Descriptive Statistics, by Variety

Variable	Tuxpeñito Average	Tuxpeñito CV	Criollo Average	Criollo CV	"Olotillo" Average	"Olotillo" CV
Moisture %	21.8	43%	30.8	40%	26.3	35%
Shelling Rate	.847	3%	.833	7%	.87	2%
Yield (kg/ha)	4439	32%	2621	64%	2785	14%
Density (pl/ha)	29,751	34%	24,431	26%	22,489	26%

Several items appear to be of immediate interest, non-random data not withstanding. Some farmers complain of overly large cobs for Tuxpeñito; the above data indicates however, that Tuxpeñito shows a higher shelling rate than that of the local variety. As is expected, the short and early Tuxpeñito is found with a lower moisture content and a higher harvested density than that of Criollo.

D) Regression Analysis

Four equations were examined that relate sample grain weight to ear size. The functional forms were chosen in light of Mexican and Kenyan experience. These equations were:

$$(1)\quad GW = a + b\,(D^2L) + e, \tag{4}$$

$$(2)\quad (GW)^{1/2} = a + b\,(D^2L) + e_2 \tag{5}$$

$$(3)\quad GW = a + b, D + b_2L + b_3D^2 + b_4L^2 + b_5\,DL + b_6D^2L + b_7\,DL^2 + e_3 \tag{6}$$

$$(4)\quad (GW)^{1/2} = a + b, D + b_2L = b^2\,D^2 + b_4\,L^2 + b_5\,DL + b_6\,D^2L + b_7\,DL^2 + e_4 \tag{7}$$

where GW = sample dry grain weight (grams)

D = sum of unhusked sample ear diameters (cm)

L = sum of unhusked sample ear lengths (cm)

e_i = error

a & b_i = parameters

Each equation was fitted three times: once only with dry cycle data, once only with wet cycle data and once with the combined data. This allows the performance of a Chow test, the purpose of which is to determine if data from both cycles may be joined to fit one single equation, i.e., it tests the "equality between sets of coefficients in two linear regressions".[11]

The Chow test takes the form of an F test:

$$F = \frac{Q_3/K}{Q_2/(m + n - 2K)} \tag{8}$$

when $Q_3 = Q_1 - Q_2$ (9)

Q_1 = sum of squared residuals, combined data

Q_2 = sum of squared residuals, first data set, plus sum of squared residuals, second data set

K = number of parameters to be estimated

m = number of observations, first data set

n = number of observations, second data set

A large F statistic implies the need for two separate equations. Results are summarized in Table 4:

TABLE 4

Chow Test Results

Equation	F	Accept/Reject Null Hypothesis of equal coefficients
1	40.41	reject
2	42.33	reject
3	2.26	accept (10% level)
4	2.14	accept (10% level)

[11] G.C. Chow, "Tests of Equality between Sets of Coefficients in Two Linear Regressions," Econometrica, vol. 28, no. 3, pp. 591-605, July 1960.

It seems that only equations (3) and (4) may be used for fitting over cycles. Equations (1) and (2) are constrained to cycle - specific cases; they furthermore suffer from relatively low R^2 values, as may be seen in Table 5:

TABLE 5

Equation R^2 Values

Equation	Combined Cycle	Dry Cycle	Wet Cycle
1	.72	.85	.70
2	.63	.69	.68
3	.91	.97	.80
4	.93	.96	.81

For the reasons discussed above, then, equations (1) and (2) were discarded.

An examination of the residuals of equations (3) and (4) was carried out. Simple correlations between the absolute value of the residual and the corresponding observation were .42 for equation (3) and - .03 for equation (4). Equation (3), then, is characterized by increasing disturbance variance as observations increase in size, i.e. heteroscedasticity. Equation (4) is not subject to this problem; it furthermore envoys a higher R^2. Equation (4) was thereby chosen for further work.

Equation (4), fitted to the combined data from both cycles, is as follows:

$$(GW)^{1/2} = \underset{(2.02)}{.16029271} + \underset{(0.16)}{.00258255\,D} + \underset{(1.01)}{.00611948\,L} + \underset{(1.87)}{.00070112\,D^2}$$

$$+ \underset{(2.3)}{.00013281\,L^2} - \underset{(-2.61)}{.00065767\,DL} + \underset{(0.72)}{.00000089\,D^2L} - \underset{(-0.78)}{.00000036\,DL^2} \qquad (10)$$

where variables are defined as before[12].

[12] Number in parentheses are "t" values.

The combinations of high R^2 and low "t" values leads one to suspect high multicollinearity, a problem in some circumstances but one which does not harm the predictive capacity of the equation.[13]

E) Use of the Equation in Yield Prediction

The fitted equation predicts sample-level grain weight from ear size values from that sample. Of more practical interest, however, is an estimate of yield per hectare for a given field from which samples have been taken. This may be calculated by averaging the sample yields for the field of interest, weighting sample yields by sample harvested area, as follows:

$$Y_{ij} = GW_{ij} \left(\frac{10{,}000}{HA_{ij}}\right) \left(\frac{100}{85}\right) \qquad (11)$$

where Y_{ij} = estimated yield per hectare for sample i and field j, 15% moisture

GW_{ij} = dry grain weight, sample i and field j

and HA_{ij} = harvested area, sample i and field j

Furthermore:

$$Y_j = \sum_i Y_{ij} \left(\frac{HA_{ij}}{\sum_i HA_{ij}}\right) \qquad (12)$$

where Y_j = estimated yield per hectare, field j (weighted by harvested area)

Whenever sample estimates of grain weight are available, the above two equations may be used to calculate the estimated yield per hectare for the field from which the samples were drawn.

The present study has used two estimators of sample grain weight: the first uses sample crop cuts but the second is inferred from the selected regression equation. Both sets of grain weight estimates were then used to estimate per hectare yields for each sampled field.

---

[13] Puebla researchers, using the same equation at the "ear level", achieved "t" values of around 50.0 - but they were using 20,000 + observations!

Of major interest is the difference between these field-level yield estimates. Assuming that the estimates based on crop cutting are accurate, one would like to examine how closely these are approached by the inferred estimates. The complete distribution of these differences is given in Table 6:

TABLE 6

Differences between Sample Cut, and Regression Estimates of Plot Yields* (kg/ha)

Plot #	Difference	Plot #	Difference
1	- 599	11	373
2	- 413	12	393
3	66	13	- 108
4	- 70	14	- 667
5	- 92	15	- 201
6	- 60	16	971
7	- 28	17	- 7
8	47	18	- 48
9	- 160	19	- 94
10	96		

The differences noted in Table 6 have a mean of - 31 kg/ha and a standard error $(\frac{S}{\sqrt{n}})$ of 82.1 kg/ha. A 95% confidence interval around this mean = $\bar{X} \pm t_{.975,18} (\frac{S}{\sqrt{n}}) = (-31) \pm 2.101 (\frac{358}{4.36})$ or from - 221 to 151 Kg/ha, assuming a normal distribution.

* i.e., in plot ___, the difference between the field yield estimate based on crop cutting and that based on the selected equation equals _____ kg/ha.

F) Summary

In an attempt to be in the position to estimate farmers' yields without physical cutting, a method was examined that infers sample grain weight from sample ear size. One equation (10) was identified that provided a close fit and could be generalized to more than one cycle. Yields were then calculated (based on inferred grain weight) and compared with yields estimated by crop cutting methods. In most cases, the difference in field-level yield estimates was less than ± 100 kg/ha. However, in three cases, the difference came to more than 500 kg/ha. It was impossible to determine why these outlines were present.[14]

The prescense of these outliers, however, makes it inadvisable to use the method presented to estimate field-level yields for individual fields. The method may better be used to obtain average field-level yields over several fields.

[14] There does not seem to be any important relation between the "differences" reported in Table 6 and either yield level or field heterogeneity.