Overestimates of Returns to Scale in Agriculture—A Case of Synchronized Aggregation*

YOAV KISLEV

Past studies of the aggregate production function of American agriculture indicate very high returns to scale. These findings are not supported by farm-level analysis. It is suggested that the aggregate estimates are biased, even if the assumption that all farms operate on the same function is accepted, due to grouping synchronized with regional effects which are not included in the analysis. An algebraic analysis of synchronized grouping is presented, and the use of covariance analysis is suggested as at least a partial correction of the bias. Empirical findings, production functions fitted to the 1949 and 1959 Census of Agriculture data, support the hypothesis of overestimates of returns to scale.

AGRICULTURAL production functions estimated from aggregate data have indicated very high returns to scale. As reported by Walters [17, p. 26], Tintner estimated the sum of the coefficients in a Cobb-Douglas function at the division level (10 divisions in the United States) to be 2.51; and Johnson [7] reported a value of 1.26. Recently Griliches reported values of 1.352 to 1.362, estimated at the productivity region level (68 in the country) [3], and 1.192 to 1.282 at the state level [2].

At the farm level, however, the picture is quite different. Heady and Dillon [5, Table 20] summarized the results of eleven studies conducted at the farm level in the United States. In seven cases, the sum of the coefficients, in the Cobb-Douglas production function, was higher than unity; in only four cases were the sums significantly different from one. The values of those four estimates were 1.17, 1.27, 1.10, and 1.15. In private communication with Professor Griliches, J. G. Elterich of Michigan State University reported the results of 56 production function estimates at the farm level. In 20 cases, decreasing returns to scale were estimated; in 24 cases, the sum of the coefficients was between 1.0 and 1.2; and in 12 cases, it was larger than 1.2. The highest value reported was 1.256. There was no information available regarding the standard errors of the estimates.

In general there is no reason to expect farm-level and aggregate analysis to yield the same results. If farms operate on different microproduction functions, then aggregate estimates of the structural parameters will be biased, as Theil has shown [15], in unknown directions and to unknown extents. To

* I am indebted to the members of my thesis committee—Zvi Griliches, D. Gale Johnson, and T. W. Schultz—and to Giora Hanoch, Finis Welch, and Yair Mundlak, for guidance and help. The errors are my sole responsibility. This work was supported by a Ford Foundation grant for research in productivity and economic growth.

YOAV KISLEV *is a lecturer in the Department of Agricultural Economics, Faculty of Agriculture, The Hebrew University, Rehovot, Israel.*

be sure, such biased estimates are not necessarily useless. In a slowly changing world like ours, they can help to forecast future economic trends [11], and in many cases aggregate analysis will yield better predictions than individual-level data [4,]. But such biased estimates cannot tell anything about the structure of the farm-level production process.

There are, however, cases in which aggregate analysis will yield unbiased estimates of the micro parameters. One such case, discussed by Prais and Aitchison [13], is that of grouping—aggregation of individual observations operating on the same function. Here, however, the absence of bias depends on a completely specified and accurately observed model, conditions which are rarely met. In the circumstances under which most empirical work in economics is done, incomplete models will yield biased estimates at the individual as well as at the aggregate level. It is argued in this article that an important specification error in agricultural production function analysis is the omission from the regression of regional effects, and that grouping into regions, where the grouping process is synchronized with the omitted variables, can be the reason for the differences between aggregate and farm-level estimates.

In the terminology of covariance analysis [16], returns to scale in farming should be estimated from the "within group" regression. Omitting the regional effect, we estimate the overall (biased) regression at the farm level; the aggregate analysis is the "between group" regression. I will show here that the specification bias of the between-group regression is larger than that of the overall regression coefficients, but that, at the same time, aggregation eliminates biases due to the omission of variables which vary only within the groups. Empirical evidence presented seems to support the hypothesis that the net effect of aggregation, in agricultural production function estimates, is to increase the specification bias and, as a result, to overestimate returns to scale.

The Algebraic Analysis

Grouping as such does not introduce any bias into the estimated coefficients. Consider the simple model,

$$y_{ij} = \alpha + \beta x_{ij} + u_{ij}, \tag{1}$$

where

j is the group index and

i (i=1, 2 · · · I_j) is the index of the individual unit within the group.

Assume the classical regression conditions to prevail, that is, x_{km} is uncorrelated with u_{km} and

$$E(u_k u_m) = \begin{cases} 0 & k \neq m \\ \sigma_u^2 & k = m \end{cases}$$

(k and m run through all the observations).

At the group level we estimate the between-group regression,

$$\bar{y}_j = \bar{a} + \bar{b}\bar{x}_j + \bar{u}_j \tag{2}$$

where a bar, here and throughout, indicates that averaging was done on the missing index. $\bar{b}$ is an unbiased estimate of β:

$$E(\bar{b}) = \frac{\Sigma(\bar{x}_j - \bar{x})(\bar{y}_j - \bar{y})}{\Sigma(\bar{x}_j - \bar{x})^2} = \frac{\Sigma(\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})\beta}{\Sigma(\bar{x}_j - \bar{x})^2} = \beta,$$

since

$$E(\bar{u}_j\bar{x}_j) = E(u_{ij}x_{ij}) = 0.$$

Homoscedasticity is maintained, however, only if the groups are of equal size, since

$$E(\bar{u}_j\bar{u}_k) = \begin{cases} 0 & j \neq k \\ \sigma_u^2/I_j & j = k. \end{cases} \tag{3}$$

This was the Prais and Aitchison case. But let us consider now the model

$$y_{ij} = \alpha + \beta x_{ij} + \gamma z_{ij} + u_{ij}. \tag{4}$$

Let us assume again that the classical conditions prevail but that the model is estimated by the regression

$$y_{ij} = a + bx_{ij} + v_{ij}. \tag{5}$$

As is well known, b is now biased:

$$E(b) = \beta + \gamma p, \tag{6}$$

where p is, in the terminology suggested by Theil [14, p. 326], the coefficient of the auxiliary regression of z_{ij} on x_{ij}.

At the group level, the estimated equation is

$$\bar{y}_j = \bar{a} + \bar{b}\bar{x}_j + \bar{v}_j, \tag{7}$$

and

$$E(\bar{b}) = \beta + \gamma\bar{p}, \tag{8}$$

where $\bar{p}$ is the coefficient of the regression of $\bar{z}_j$ on $\bar{x}_j$. Our goal is to compare the magnitudes of p and $\bar{p}$.

It is useful to assume that the groups are of equal size ($I_j = I$, for all j's) and to introduce two identities and efficient symbols:

$$\sum_i\sum_j (x_{ij} - \bar{x})(z_{ij} - \bar{z}) = \sum_i\sum_j (x_{ij} - \bar{x}_j)(z_{ij} - \bar{z}_j) + I\sum_j (\bar{x}_j - \bar{x})(\bar{z}_j - \bar{z}),$$

or, for brevity,

$$C = C_{ij} + IC_j.$$

Similarly,

$$\sum_i \sum_j (x_{ij} - x)^2 = Q = Q_{ij} + IQ_j.$$

In these symbols,

$$p = \frac{C}{Q} = \frac{C_{ij} + IC_j}{Q_{ij} + IQ_j} \tag{9}$$

and

$$\bar{p} = \frac{C_j}{Q_j}\,. \tag{10}$$

The ratio between the two coefficients is

$$\frac{\bar{p}}{p} = \frac{IC_j}{IQ_j} \cdot \frac{Q_{ij} + IQ_j}{C_{ij} + IC_j} = \frac{IC_j}{C_{ij} + IC_j} \cdot \frac{Q_{ij} + IQ_j}{IQ_j}\,. \tag{11}$$

In general, (11) will differ from unity, i.e., $\bar{p} \neq p$, and the bias in $\bar{b}$ will not be the same as the bias in b.

Several cases can be distinguished:

(a) $\bar{z}_j = \bar{z}$, for all j's. The unobserved variable varies only "within" the groups and has no "between" groups variance. Here $C_j = 0$ and $\bar{p} = 0$—a case of "good aggregation"; the grouping process eliminates the bias altogether.

(b) Grouping is done at random and the group lines are not correlated with the variables in the model. Here $E(Q_j) = Q/I$ and $E(C_j) = C/I$. Thus plim $(\bar{p})$ = plim (p). Random grouping into large groups does not change the specification bias.

(c) In our uncontrolled economic experiments, the omitted variables are usually correlated with the grouping method; we seldom, if ever, group randomly. An extreme case of synchronized aggregation will be the case in which the omitted variable is a group effect that varies only from group to group. Now (4) is a covariance model, in which

$$z_{ij} = \bar{z}_j, \quad \text{for all} \quad i\text{'s} \quad \text{and} \quad j\text{'s}, \tag{12}$$

which, in turn, means that $c_{ij} = 0$ and

$$\frac{\bar{p}}{p} = \frac{Q_{ij} + IQ_j}{IQ_j} = 1 + \frac{Q_{ij}}{IQ_j}\,. \tag{13}$$

Q_{ij} and Q_j, being sums of squares, are positive; (13) thus indicates that the omission of the group effect causes a larger bias in the "between group" regression coefficient than in the overall coefficient. This increase in the bias is due to aggregation synchronized with the omitted group effect and is, therefore, called here the Synchronization Effect. It is demonstrated graphically in Figure 1, in which auxiliary regressions of z on x are depicted. The dots mark three observations for each group at a different z level, and the between-group regression slope (line 1) is larger than the overall slope (line 2).

$$1.\quad \bar{z}_j = \bar{a} + p\bar{x}_j + \bar{u}_j$$

$$2.\quad z_{ij} = \bar{z}_j = a + px_{ij} + u_{ij}$$

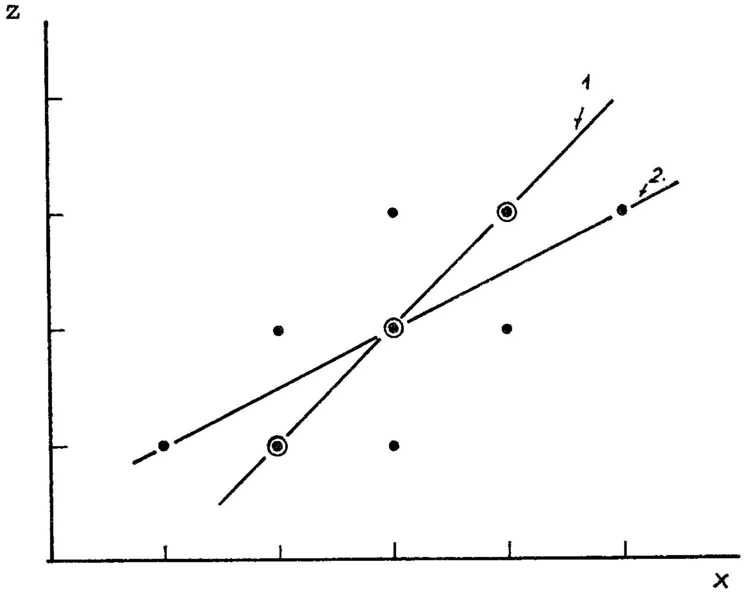

Figure 1. Auxiliary regressions, overall and between groups

(d) In most cases, however, the omitted variable, even if correlated with the group lines, is likely to vary within the groups too. Grouping combines, then, the effects of the cases discussed under (a)—elimination of within-group variance—and under (c)—synchronization effect; $\bar{p}$ may be smaller or larger than p. To make this point clear, we may rewrite (11) as

$$(14)\qquad \frac{\bar{p}}{p} = \frac{1}{1+\dfrac{C_{ij}}{IC_j}}\cdot\left(1+\frac{Q_{ij}}{IQ_j}\right).$$

Seldom will C_{ij} and C_j differ in sign. Seldom will one variable be positively correlated with another within groups and negatively correlated between them, or vice versa. If they are, then (14) may be larger than (13). More likely C_{ij} and C_j will have the same sign and (14) will be smaller than (13). Equation (14) may even be smaller than unity; if the elimination of the within-group variance dominates the synchronization effect, the specification bias will then be smaller at the group level than at the individual level of analysis.

In Appendix I it is shown that the conclusions for the multiple regression case are similar to those derived here. With respect to one point, however, the two cases differ. While in the simple regression the synchronization effect (equation 13) always increases the absolute value of the specification bias, the result is not predictable in a multivariant model. The reason is that the bias in any one variable is affected by the synchronization effects of all the variables included in the calculations, and of these variables some may be positive and some negative, a condition which makes the a priori determination of the result impossible. There is, however, one exception to this rule. The synchronization effect increases the absolute value of the bias in every one of the estimated coefficients if all the covariances among the included and the omitted variables are positive. In economic data, such positive covariance matrices are common.

Aggregate regressions would not have synchronization bias if models were completely specified and accurately and fully observed. This condition is, however, practically unattainable. If one is willing, on the other hand, to accept specification biases at the individual-unit level but wants to eliminate any changes in the biases resulting from aggregation, random grouping will be appropriate. In most cases, however, aggregation of data available in individual form has been suggested in order to economize in computations. With the prevalence of high-speed electronic computers today, it seems that the saving in computation effort is not worth the reduction in the efficiency of the estimates that results from any aggregation [13].

A partial correction of the specification errors is to allow for the group effects by a covariance analysis. This approach has already been suggested by Mundlak [12] and Hoch [6] for the elimination of specification biases at the individual level. The covariance analysis requires more than one observation per group; thus, at the aggregate level, at least two cross sections, from different time periods, will have to be utilized. Then, on the assumption that the structural parameters are constant over a period of time, the equation estimated, instead of (5), will be

$$\bar{y}_{jt} = \bar{a}_j + \bar{\bar{b}}\bar{x}_{jt} + \bar{v}_{jt}, \tag{15}$$

where t stands for the cross section.

Production Function Estimates

The empirical analysis proceeds on the assumption that the farms (the individual units) within regions (the groups) operate on a Cobb–Douglas production function of the form

$$y = A \prod_{r}^{k} x_r^{\beta_r} \qquad (r = 1, 2, \cdots, k), \tag{16}$$

where y stands for output and x_r for the input r. Farm and regional indices were omitted from (16), which was estimated by a linear regression in its logarithmic form.

Let the *k*th variable represent productivity characteristics of the region: soil and climatic conditions, availability of professional information, quality of roads, and similar factors. To apply the results of the previous analysis, we let the equation

$$\epsilon = \sum_{r}^{k-1} \beta_r \tag{17}$$

be the true indicator of returns to scale. The sum in (17) does not include β_k since the question is what change will be caused in farm output by an equiproportional change in all inputs, and not what change will occur as we move from one region to another.

If the only specification error is the omission of x_k, then b_r, the estimate of β_r, is biased:

$$E(b_r) = \beta_r + \beta_k p_r, \tag{18}$$

where p_r is the coefficient of the variable r in the auxiliary regression of the group variable on all the observed inputs. ϵ is estimated by

$$e = \sum_{r}^{k-1} b_r$$

and is biased:

$$E(e) = \sum_{r}^{k-1} \beta_r + \beta_k \sum_{r}^{k-1} p_r. \tag{19}$$

The hypothesis offered is that e is biased upward at the farm level and that synchronized aggregation increases the bias.

The estimate of returns to scale (e) will be biased upward at the farm level if β_k and the p_r's are positive. β_k is positive by definition. Positive p_r's mean that the omitted effect is complementary to the inputs in production, that the higher the level of this effect, the larger will be the doses of factors of production employed. The multicollinearity prevalent in farm production data (e.g., Table 3) indicates general complementarity among inputs, and there is no theoretical or practical reason to expect the group effect to be the exception to this rule.

Positive synchronization effects in the multifactor production function, which contribute to an increase in the bias at the group level, result from positive covariances among inputs and regions. Again, Table 3 (similar correlation coefficients were obtained at the regional level) and the complementarity of factors discussed above are consistent with such positive covariance matrices.

The resulting synchronization effects can be of substantial magnitudes. In Table 1, simple regression synchronization effects are reported for a sample of 351 farms in 16 counties in Texas. This is equation (13) applied successively to six inputs, assuming each time a single-factor production function and aggregation into counties. The values found range from 8.7 to 13.7. These effects, however, are likely to be overestimated, since the sample was taken from a rather homogeneous area where the between-counties sum-of-squares deviations are probably smaller than those for a sample of counties taken at random from the country as a whole.

The synchronization effect was also calculated for the number of acres per farm in 148 counties in the conterminous United States.[1] The ratio of the within- to the between-counties sum-of-squares deviation was 2.566, and the synchronization effect was 3.556. This is likely to be an underestimate, since the variable considered is the number of acres, instead of the value of land. The latter is the variable which should have been included in the production function, since the value of land probably varies less among counties than the number of acres.

These findings indicate the possibility that synchronized aggregation may increase the specification bias substantially. For example, let us assume a case of constant returns to scale; let us assume also, however, that the sum of the coefficients in a Cobb–Douglas farm-level regression, in which only the observable inputs are included, is not unity but 1.1—a bias of 0.1 in the scale coefficient. Let us assume further that the magnitude of the synchronization effect is 6.0—well within the range of our estimates—and that it is equal for all regression coefficients ($p_r = 6.0$, for all r's). Now, even if half the specification bias is eliminated in aggregation (case d on page 971) the aggregate estimate will still be of the order of 1.3. And if the bias at the farm level

[1] Data on frequency distribution of farm sizes are from the 1959 Census of Agriculture, County Table 2. All parts of Volume I of the Census were ordered by numbers and every nineteenth county was included. The calculations were made on the assumption that all farms in a class were the size of the class midpoint. The farms in the "less than ten" class were considered to be 5 acres each, and those classified as "two thousand or more" were taken as 3000 acres each.

Table 1. Synchronization effects[a]—Texas sample

Variables[b]	Sums of squares (in logarithms)		Synchronization effect
	within counties (Q_{ij})	between counties (IQ_i)	($1+Q_{ij}/IQ_i$)
Land	66.37	5.24	13.67
Farm expenditure	194.37	17.57	12.06
Hired labor	73.08	9.49	8.70
Equipment	401.24	39.75	11.09
Livestock	385.15	40.05	10.61

Source: part of the sample used in William G. Adkins, *Income of Rural Families in the Blackland Prairies,* USDA MP-659, College Station, Agricultural and Mechanical College of Texas, in cooperation with United States Department of Agriculture, May 1963.

[a] Equation (13). For explanation of computations, see text.

[b] The definition of variables (according to the source):
Land—value of land operated.
Hired labor—total man-work equivalents of hired labor.
Livestock—value of livestock on hand, December 1959.
Farm expenditure—current farm expenditures.

is not 0.1 but 0.2, the sum of the coefficients at the aggregate level will, under the same assumption, be 1.6.

Actual production function estimates were consistent with the hypothesis offered. Such estimates were conducted on two cross sections, one for 1959—Tables 2 and 3—and one for 1949—Table 4. County data for the 1959 study were from the Census of Agriculture for the conterminous United States. The production function was calculated once by the factor-share method,[2]

Table 2. Simple correlation coefficients at the county level, 1959 data

Variables	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Y Value of output	1.00							
X_1 Machinery	.805	1.00						
X_2 Farmers	.627	.653	1.00					
X_3 Hired labor	.589	.241	.204	1.00				
X_4 Livestock	.654	.546	.340	.204	1.00			
X_5 Fertilizers	.373	.239	.237	.384	.141	1.00		
X_6 Land and buildings	.861	.787	.503	.492	.531	.236	1.00	
X_7 Other	.612	.538	.404	.304	.431	.183	.595	1.00

For source and definition of variables, see Appendix II A.

[2] For reference to the factor-share method, see Klein [10, pp 193–194]. It can be shown [1] that factor-share estimates are biased, but in a study such as the present one, with close to 3000 observations, the bias, which is inversely proportional to the size of the sample, can be safely ignored.

Table 3. Cobb–Douglas production function, 1959 cross section[a]

Variables	Geometric average	Factor shares	County		Productivity region		
			Regressions[b]				
			1	2	3	4	5
R^2			.907	.934	.999	.973	.974
Y (Output)	7,446						
X_1 (Machinery)	1,935	.260	.241 (.011)	.239 (.014)	.401 (.059)	.333 (.053)	.258 (.051)
X_2 (Farmers)	3,080	.415	.254 (.012)	.230 (.013)	.051[c] (.078)	.219 (.085)	.276 (.073)
X_3 (Hired labor)			.176 (.005)	.170 (.006)	.201 (.029)	.224 (.027)	.194 (.021)
X_4 (Livestock)	982	.125	.171 (.005)	.178 (.006)	.134 (.034)	.141 (.031)	.128 (.028)
X_5 (Fertilizers)	187	.025	.045 (.004)	.048 (.005)	.085 (.036)	.102 (.024)	.076 (.024)
X_6 (Land and buildings)	27,390	.283	.254 (.010)	.166 (.012)	.425 (.064)	.093 (.055)	.153 (.055)
X_7 (Other)	768	.103	.026 (.003)	.022 (.003)	−.085 (.043)	.157 (.044)	.153 (.041)
Sum of coefficients		1.211	1.167 (.028)	1.053 (.033)	1.212	1.269	1.238 (.056)

For source and definitions of variables, see Appendix II A.

[a] Values in parentheses are standard errors of estimates.

[b] Regressions:

1. Without regional dummy variables.
2. With regional dummies.
3. Logarithmic aggregation.
4. Arithmetic aggregation, unweighted.
5. Arithmetic aggregation, weighted by the number of farms in the region.

[c] Not significant at usually accepted levels.

and regressions were computed at the *county* and the *productivity region* level.[3]

The sum of the coefficients at the county level (regression 1) is 1.167. Covariance analysis—the inclusion of dummy variables for the 68 productivity regions—contributed significantly to the explanatory power of the regression ($F^{67}{}_{2885} = 18.33$ for the test of the hypothesis that the dummy variables did not contribute to the explanation) and reduced the sum of the coefficients to 1.053.

The 1949 analysis (Table 4) was conducted at the productivity-region level only. Two sets of variables were prepared. One is the set used by Griliches [3] in his study of the same data (regression 6 is, apart from

[3] The delineation of the 68 productivity regions was done according to the source of the 1949 data. See Appendix II B.

Table 4. Cobb-Douglas production function, 1949 cross section—productivity-region level[a]

Variables	Regressions				
	6	7	8	9	10
R^2	.976	.975	.976	.973	.986
X_1 (Machinery)	.359 (.051)	.306 (.050)	.298 (.042)	.196 (.064)	.157 (.047)
X_2 (Family labor)				.296 (.180)	.349 (.133)
X_3 (Hired labor)				.134 (.027)	.117 (.020)
X_4 (Livestock)	.168 (.023)	.128 (.024)	.154 (.022)	.131 (.025)	.175 (.020)
X_5 (Fertilizers)				−.064[b] (.040)	−.093 (.030)
X_6 (Land and buildings)				.380 (.046)	.281 (.036)
X_7 (Other)	.122 (.033)	.102 (.034)	.111 (.029)	.229 (.071)	.277 (.053)
X_8 (Land)	.174 (.033)	.232 (.029)	.186 (.026)		
X_9 (Buildings)	.102 (.045)	.116 (.045)	.093 (.038)		
X_{10} (Workers)	.441 (.074)	.472 (.076)	.380 (.066)		
X_{11} (Regional effect)			.472 (.093)		.603 (.085)
Sum of coefficients	1.366	1.356 (.070)	1.222 (.018)	1.302 (.062)	1.263 (.045)

For source and definition of variables see Appendix II B.

[a] For explanation of regressions, see text. Values in parentheses are standard errors of estimates.

[b] Not significant at usually accepted levels.

rounding errors, his U17); the other is as close as possible in its definition to our 1959 set. Regressions 6, 7, and 8 belong to the first set, 9 and 10 to the second. Since there was only one observation per region, an ordinary covariance analysis could not be made. Instead, the coefficients of the regional dummies from regression 2 in Table 3 were included in regressions 8 and 10 of Table 4. This procedure is identical to a covariance analysis under the assumption that the structure of the regional effect did not change from 1949 to 1959 and that it was estimated correctly in the 1959 study. The coefficients of the regional effects are significant, and again the estimates of the scale coefficients in regressions 8 and 10 (1.222 and 1.263) are smaller than their counterparts in regressions 7 and 9 (1.356 and 1.302), in which the regional effects were not included.

The inclusion of regional dummies in regression 2 is not a full-scale covariance analysis and cannot be expected to eliminate all the synchronization biases. To clarify this statement, let us consider the separation of the omitted effects into farm, county, and regional effects. The county effect is so defined as to be measured from the regional effect, and the farm effect from the county effect. Thus there are no within-region variations in the regional effect and no within-county variations in the county effect. In this form there are three sources of bias at the farm level—the omission of each of the three effects. A covariance analysis at the county level, in which the regional effect is allowed for, eliminates the biases due to the farm effects (case a) and to the omission of the regional effect. The county effect is not allowed for and the part of the bias due to it is even increased—compared to the farm-level regression—as a result of synchronized aggregation. The availability of more than one observation per county will permit us to allow for the county effect, as well as for the other two.

Two additional points should be noted. First, since counties and regions do not have equal numbers of farms, homoscedasticity is not maintained (equation 3 and Appendix I), and the regressions have to be weighted by the number of farms in the county or region [8, pp. 207–211]. Regression 5 in Table 3 is weighted, and in Table 4, regressions 7 and 10 are weighted. If we compare regression 5 with regression 4, and regression 7 with regression 6, we can see that the weighting process reduces the coefficients of machinery and livestock and increases those of land and family labor in 1959 and of all workers in 1949.

Another point that has not been discussed hitherto is the form of aggregation. In the framework of the Cobb–Douglas function, arithmetical averaging constitutes an aggregation error. Logarithmic aggregation is tried, from the *county* to the *regional* productivity level, in regression 3, but the coefficients are "unreasonable," perhaps because of the fact that aggregation into counties was arithmetic. The differences between logarithmic and arithmetic aggregation are functions of the variance. Furthermore it can be shown [9] that, if the distribution of the farm data is log-normal, then arithmetic aggregation can be corrected by inclusion of county variances in the regression. The log-normal hypothesis was tested on the Texas sample of Table 1 and had to be rejected, but further attempts to fit a theoretical distribution to farm data are warranted, particularly since the Census of Agriculture contains information on frequency distributions of factors of production which can be used to approximate the variances.

Summary

Aggregation, even grouping—aggregation of economic units operating on the same function—alters the specification bias if, as is usually the case, it is synchronized with the errors in the model. This effect can be quite substantial and is perhaps the reason for the very high aggregate estimates of returns to scale in agriculture. Covariance analysis was suggested as a solution and tried on two cross sections. This method reduced the estimated scale coefficients.

Appendix I
The Multiple Regression Case

To express grouping as a matrix operation, define $G(m\times n)(m\leq n)$ as a grouping matrix, exemplified by

$$\begin{pmatrix}\bar{y}_1\\ \bar{y}_2\end{pmatrix} = \begin{pmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0\\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2}\end{pmatrix}\begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\end{bmatrix} = GY.$$

There are $n!((m!(n-m)!)$ ways to group n observations in m groups. Consider the family of all the distinct G matrices possible (each one is a permuted version of every other). Random grouping can be represented as a selection at random of one matrix G (or, alternatively, as a selection of a permutation of matrix M, of the order n, followed by the multiplication operation GMY). The complete model at the ungrouped level is

$$Y = X\beta + U. \tag{I.1}$$

Estimating β (now a vector), we may write

$$b = (X'X)^{-1}X'Y$$

and

$$E(b) = (X'X)^{-1}X'X\beta = \beta. \tag{I.2}$$

Grouping, we get

$$GY = GX\beta + GU. \tag{I.3}$$

$E(GU)=0$ since $E(U)=0$, if the classical regression conditions are maintained in (I.1), so that

$$\begin{aligned} E(\bar{b}) &= (X'G'GX)^{-1}XG'GX\beta \\ &= \beta. \end{aligned} \tag{I.4}$$

Grouping as such does not bias the estimated coefficients.

Homoscedasticity is not maintained. If, at the ungrouped level, $U'U = \sigma^2 I$ is assumed, then after grouping,

$$(GU)'(GU) = U'G'GU = \sigma^2 GG' \tag{I.5}$$

In case of specification error, a matrix, R, is observed instead of the true data matrix, X. Estimating, we write

$$E(b) = (R'R)^{-1}R'X\beta = P\beta \tag{I.6}$$

and, at the group level, we have

$$E(\bar{b}) = (R'G'GR)^{-1}R'G'GX\beta = \ \ \beta. \tag{I.7}$$

If the specification error is an omission of some variables, the correct matrix X can be written, without loss of generality, as $X = [X \vdots Z]$. Now,

$$\begin{aligned} P &= (R'R)^{-1}R'[R \vdots Z] \\ &= [I \vdots (R'R)^{-1}R'Z] \\ &= [I \vdots \Sigma], \end{aligned}$$

and

$$\begin{aligned} P &= (R'G'GR)^{-1}R'G'G[R \vdots Z] \\ &= [I \vdots (R'G'GX)^{-1}R'G'GZ] \\ &= [I \vdots \bar{\Sigma}]. \end{aligned}$$

In comparing P to $\bar{P}$, it will be sufficient to compare Σ to $\bar{\Sigma}$. This comparison shows the following:

(a) No between-group variations in Z; that is, $GZ = 0$, $\Sigma = 0$, $P = [I:0]$. No bias at group level.

(b) Random grouping; that is, plim $(\bar{P}) =$ plim (P). In large samples and groups, random grouping does not alter the specification bias.

(c) No within-group variations in Z. $R'G'GZ = R'Z$; that is, all the variations are between the groups; but $(R'R)^{-1} \neq (R'G'GR)^{-1}$. Moreover, in multiplying the matrices which create Σ, the elements in the rows of $(R'G'GR)^{-1}$ operate as weights in the summation of the columns of $R'G'GZ$. Some of the weights can be negative and some positive, and the net result cannot be predicted.

(d) All covariances in Σ and $\bar{\Sigma}$ are positive. Then, since $R'R = (R'R - R'G'GR) + (R'G'GR)$, and all weights are positive, $R'R < R'G'GR$ and therefore $\Sigma < \bar{\Sigma}$, a clear synchronization effect.

(e) Nonzero within- and between-group variances in Z; that is, the results are inconclusive, as in the simple regression case in the text.

Appendix II
The Data for the Production Function Estimates

A. 1959 cross section

Source: 1959 Census of Agriculture county tables (on punched cards), unless otherwise specified.

Y (Output) is the total value of farm products sold or intended to be sold in 1959, plus state average of rental value of farm dwellings and value of home consumption, plus value of farm products sold, times state ratio of government payments and changes in inventory to cash value of farm marketing. Source for data on nonmarket output: U. S. Department of Agriculture, Agricultural Marketing Service, *Farm Income Situation*, FIS-179 supplement (1961), Table 4.

X_1 (Machinery) is defined as 22 percent of the value of machines and equipment on farms, plus expenditures on gasoline and oil. The values of the different pieces of machinery were taken from unpublished data prepared by Professor Griliches.

X_2 (Farmers) is the number of farmers by occupation as given by the Census of Population (on magnetic tapes) for 1960. Females are counted as 65 percent.

X_3 (Hired labor) is the expenditures on hired labor divided by the state wage rate per day. Source: U. S. Department of Agriculture, Agricultural Marketing Service, Crop Reporting Board, *Farm Labor*, LA-1, 1959. Quarterly data on wage rates per day, without room and board, were used to calculate simple annual averages for 1959. For Washington, Oregon, and California, the per hour wage rate was multiplied by 90 percent of the number of hours worked.

X_4 (Livestock) is defined as 10 percent of the value of livestock on farm plus expenditures on feed.

X_5 (Fertilizers) is the value of fertilizers and lime, weighted by state prices. Prices were calculated by dividing the 1954 expenditures on fertilizers by the quantities, and correcting for changes in price level. Source: 1959 Census of Agriculture, State Table 5, and U. S. Department of Agriculture, Statistical Reporting Service, *Agricultural Prices*, 1961 Annual Summary, Pr 1-3 (62).

X_6 (Land and buildings) is the value of land and buildings.

X_7 (Other) is the purchases of livestock and poultry, machines hired, seeds, bulbs, plants, and trees.

All observations are county totals. A few counties with fewer than 100 farms or with otherwise insufficient information were not used. Variables in the regressions were logarithms of the per farm averages in each county.

B. 1949 cross section

Source (except X_{11}): E. G. Strand, E. O. Heady, and J. A. Seagraves, *Productivity of Resources Used on Commercial Farms*, USDA Tech. Bul. 1126, Nov. 1955.

Y (Output) is the sum of the values of farm products sold and those used in household (Table 4 in source).

X_1 (Machinery) is the interest on machinery, depreciation of machinery, expenditures on gasoline and oil, repairs of machinery, and machine hire (Table 27).

X_2 (Family labor) is the expenditures (imputed) on unpaid family and operator labor (Table 27) divided by the annual wage rate (Table 7).

X_3 (Hired labor) is the expenditures on hired labor (Table 27), divided by average annual wage rate (Table 7).

X_4 (Livestock) is the livestock purchased, interest on livestock, and expenditures on feed (Table 27).

X_5 (Fertilizers) is the expenditures on fertilizers and lime (Table 27).

X_6 (Land and buildings) is the value of land and buildings (Table 14).

X_7 (Other) is the expenditures on seeds and plants, fertilizers and lime (Table 27), and irrigation (Table 28).

X_8 (Land) is the interest on land (Table 27).

X_9 (Buildings) is the interest on buildings (Table 27).

X_{10} (Workers) is the average number of workers per farm (Table 7).

X_{11} (Regional effect) is the regional coefficients of regression 2, Table 3.

Variables in the regressions were logarithms of the per farm averages in each productivity region.

References

[1] DHRYMES, PHOEBUS J., "On Devising Unbiased Estimators for the Parameters of the Cobb–Douglas Production Function," *Econometrica* 30:297–304, April 1962.

[2] ———, "The Sources of Measured Productivity Growth: United States Agriculture, 1940–60," *J. Pol. Econ.* 71:331–346, Aug. 1963.

[3] GRILICHES, ZVI, "Research Expenditures, Education, and the Aggregate Agricultural Production Function," *Am. Econ. Rev.* 54:961–974, Dec. 1964.

[4] GRUNFELD, YEHUDA, AND ZVI GRILICHES, "Is Aggregation Necessarily Bad?" *Rev. Econ. and Stat.* 42:1–13, Feb. 1960.

[5] HEADY, E. O., AND J. L. DILLON, *Agricultural Production Function*, Ames, Iowa State University Press, 1961.

[6] HOCH, IRVING, "Estimation of Production Function Parameters Combining Time-Series and Cross-Section Data," *Econometrica* 30:34–53, Jan. 1962.

[7] JOHNSON, D. GALE, *Forward Prices for Agriculture*, Chicago, The University of Chicago Press, 1947.

[8] JOHNSTON, J., *Econometric Methods*, New York, McGraw-Hill Book Company, 1963.

[9] KISLEV, YOAV, "Estimating a Production Function from the 1959 Census of Agriculture Data," unpublished Ph.D. thesis, The University of Chicago, 1965.

[10] KLEIN, LAWRENCE R., *A Textbook of Econometrics*, Evanston, Ill., Row Peterson, 1953.

[11] MARSCHAK, JACOB, "Economic Measurements for Policy and Predictions," in *Studies in Econometric Methods*, Cowles Commission Monograph 14, ed. W. C. Hood and T. C. Koopmans, New York, John Wiley & Sons, Inc., 1953.

[12] MUNDLAK, YAIR, "Empirical Production Function Free of Management Bias," *J. Farm Econ.* 43:44–56, Feb. 1961.

[13] PRAIS, S. J., AND J. AITCHISON, "The Grouping of Observations in Regression Analysis," *Revue de l'Institute International de Statistique*, Vol. XXIII, 1954.

[14] THEIL, HENRI, *Economic Forecasts and Policy*, Amsterdam, North-Holland Publishing Co., 1958.

[15] ———, *Linear Aggregation of Economic Relations*, Amsterdam, North-Holland Publishing Co., 1954.

[16] SCHEFFE, HENRY, *The Analysis of Variance*, New York, John Wiley & Sons, Inc., 1959.

[17] WALTERS, A. A., "Production and Cost Functions: An Econometric Survey," *Econometrica* 31:1–66, Jan.–April 1963.

Reprinted from JOURNAL OF FARM ECONOMICS
Vol. 48, No. 4, Part I, November, 1966