THE JOURNAL OF INDUSTRIAL ECONOMICS 0021-1821
Volume XLIV March 1996 No. 1

ESTIMATING THE SIZE DISTRIBUTION OF FIRMS USING GOVERNMENT SUMMARY STATISTICS*

AMOS GOLAN, GEORGE JUDGE AND JEFFREY M. PERLOFF

Using a maximum entropy technique, we estimate the market shares of each firm in an industry using the available government summary statistics such as the four-firm concentration ratio (C4) and the Herfindahl-Hirschmann Index (HHI). We show that our technique is very effective in estimating the distribution of market shares in 20 industries. Our results provide support for the recent practice of using HHI rather than C4 as the key explanatory variable in many market power studies, if only one measure is to be used.

I. INTRODUCTION

USING PUBLISHED government summary statistics about the size distribution of firms and a maximum entropy approach, we can estimate the market share of each firm in an industry. For each of the 20 industries we examine, the average squared errors and the correlation between estimated and actual shares show a close fit, and Kolmogorov-Smirnov tests cannot reject the hypothesis that the estimated and actual distribution are the same at the 0.05 level. For recent years, where the US Bureau of the Census provides the Herfindahl-Hirschmann Index (HHI), our estimates are more accurate than for earlier years when only concentration ratios were available. As we discuss, even more accurate estimates could be obtained using our maximum entropy approach if the US Government published one or two additional summary statistics.

Knowing the size distribution of firms is important to economists who study industrial organization, to government regulators, and to courts. Courts use firm and industry measures of market share in a variety of antitrust cases, and structure-conduct-performance studies still commonly use firm-size measures (Carlton and Perloff [1994]). Under the merger guidelines of the US Department of Justice and the Federal Trade Commission, whether mergers are challenged depends on the relative sizes of the firms involved and the degree of concentration in the industry. In recent years, for example, the Department of Justice challenged mergers in railroads, soft drink, and airline industries using data on concentration and relative firm size. Concentration measures, such as the four-firm concentration ratio (C4) and the HHI, are used as explanatory variables in many industrial organization studies of market power and mergers.

* We are grateful to the editor, Larry White, and the anonymous referees for extremely helpful comments.

For many purposes, summary statistics are not as useful as detailed information on individual firms. For example, traditional structure-conduct-performance studies regress a measure of performance (price or profits) on a number of variables including a measure of concentration such as C4 or the HHI. If, instead of regressing the performance measure on C4, one uses the individual shares of the four largest firms, these four coefficients differ significantly and may have different signs. Using this approach, Kwoka [1979] shows that three (relatively equal-size) firms are much more competitive than two firms. Similarly, Isaac and Reynolds [1989] provide empirical and simulation evidence on the competitiveness of markets with two or four firms.

II. THE PROBLEM

Our problem is to determine the share, $0 < p_i \leqslant 1$, of the value of the total industry shipments of each firm $i = 1,\ldots,N$ in the industry using summary statistics such as were published by the US Department of Commerce, Economics and Statistics Administration, Bureau of Census in *Concentration Ratios in Manufacturing, 1987 Census of Manufactures*. The government knows the actual shares and orders the index of firms such that p_1 is the share of the largest firm and p_N is that of the smallest. Each share is positive, $p_i > 0$, and the shares sum to one:

$$\sum_{i=1}^{N} p_i = \mathbf{1}'\mathbf{p} = 1, \tag{1}$$

where $\mathbf{1}$ is a $(N \times 1)$ vector of ones and $\mathbf{p} = (p_1, p_2,\ldots,p_N)'$ is a vector of the market shares.

II(i). *Government Measures*

The Bureau of the Census publishes five summary statistics for most four-digit SIC manufacturing industries.[1] For decades, the Bureau of the Census has published four concentration measures — C4, C8, C20, and C50 — for most industries. Each of the measures *Ci*, $i = 4$, 8, 20, or 50, is the sum of the market shares of the first i firms in the industry. For example, $C4 = p_1 + p_2 + p_3 + p_4$.

In the two most recent surveys (1982 and 1987), the Bureau of the Census also published the Herfindahl-Hirschmann Index (HHI), which is the sum

[1] For industries with very few firms, some summary statistics are suppressed for reasons of confidentiality.

of the squares of the shares of each firm (up to a maximum of 50 firms):[2]

$$\sum_{i=1}^{50} (p_i)^2 = \text{HHI}. \tag{2}$$

For both the concentration measures and the HHI, higher values mean greater seller concentration.

II(ii). Formal Statement of the Problem

We start by stating the problem of estimating market shares using data prior to 1982, a period for which we only have the concentration measures. These data are linear functions of the market shares of the firms. Later we generalize the problem to include the HHI measure, which is a nonlinear function of the shares.

In vector notation, the four concentration measure identities are

$$\text{C4} = \sum_{i=1}^{4} p_i = (\mathbf{1}_1, \mathbf{0}_1)'\mathbf{p} = (1,1,1,1,0,\ldots,0)'\mathbf{p}, \tag{3.a}$$

$$\text{C8} = \sum_{i=1}^{8} p_i = (\mathbf{1}_2, \mathbf{0}_2)'\mathbf{p}, \tag{3.b}$$

$$\text{C20} = \sum_{i=1}^{20} p_i = (\mathbf{1}_3, \mathbf{0}_3)'\mathbf{p}, \tag{3.c}$$

$$\text{C50} = \sum_{i=1}^{50} p_i = (\mathbf{1}_4, \mathbf{0}_4)'\mathbf{p}, \tag{3.d}$$

where, for example, $\mathbf{1}_1$ is a vector of 4 ones, and $\mathbf{0}_1$ is a vector of $N - 4$ zeros. If the number of firms in the industry, N, is less than or equal to i, the Ci concentration measure is 1.

These definitions can be summarized as

$$\begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_1 \\ \mathbf{1}_2 & \mathbf{0}_2 \\ \mathbf{1}_3 & \mathbf{0}_3 \\ \mathbf{1}_4 & \mathbf{0}_4 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \\ \vdots \\ p_N \end{bmatrix} = \begin{bmatrix} \text{C4} \\ \text{C8} \\ \text{C20} \\ \text{C50} \end{bmatrix}$$

or written compactly as

$$Z\mathbf{p} = \mathbf{c}. \tag{4}$$

[2] For simplicity, our proportions p_i lie between zero and one whereas the Bureau of the Census uses percentages. As a result, our concentration measures equal the Bureau's divided by 100, and our HHI measure equals the Bureau's measure divided by $100^2 = 10{,}000$.

Our problem is to determine $N > 4$ unknown parameters (the market shares, p_i) using only four data points (**c**). Because the unknown linear operator Z is noninvertible, Equation 4 is an ill-posed, underdetermined pure inverse problem that cannot be solved by direct mathematical inversion. If we were to attempt to recover **p** by traditional mathematical or statistical procedures, we must consider the entire class of solutions that contain $(N - 4)$ arbitrary parameters. Without a criterion function, we have no basis for picking a particular solution vector for **p**.

II(iii). *A Maximum Entropy Formulation*

One way of solving underdetermined problems is to use the maximum entropy approach proposed by Jaynes [1957a, b] and Levine [1980]. Jaynes suggests using the Shannon [1948] entropy measure,

$$(5) \qquad H = -\sum_{i=1}^{N} p_i \ln p_i = -\mathbf{p}' \ln \mathbf{p},$$

which represents a measure of the uncertainty of a collection of events, as the criterion for choosing a solution. To recover the unknown probabilities (market shares), p_i, Jaynes proposes an optimization problem where we maximize this entropy measure, Equations, subject to our consistency conditions or data (the concentration constraints), Equation 4, and the adding up constraint (the market shares sum to one), Equation 1.

Under this formulation, we have converted our problem from one of deductive mathematics to one of inference involving the use of an optimization procedure. Through the use of the principle of maximum entropy, we have a basis for using the data (the restrictions implied by the concentration ratios) to estimate the market share distribution. There are an infinite number of such possible solutions. A rationale for using the maximum entropy approach is that it chooses the particular set of shares that is the one that could be generated in the greatest number of ways consistent with what we know (Jaynes [1957a, b]). One desirable property of the maximum entropy approach is that it is a fully-efficient information processing rule (Zellner [1988]).

The analytical solution to the problem of maximizing Equation 5 subject to Equations 1 and 4 can be obtained by specifying the Lagrangian function

$$(6) \qquad \mathscr{L} = -\mathbf{p}' \ln \mathbf{p} + \boldsymbol{\mu}'(\mathbf{c} - Z\mathbf{p}) + \lambda(1 - \mathbf{p}'\mathbf{1}),$$

and deriving the corresponding optimal conditions:

$$(7.a) \qquad \frac{\partial \mathscr{L}}{\partial \mathbf{p}} = -\ln \hat{\mathbf{p}} - \mathbf{1} - Z'\hat{\boldsymbol{\mu}} - \hat{\lambda}\mathbf{1} = \mathbf{0},$$

$$(7.b) \qquad \frac{\partial \mathscr{L}}{\partial \boldsymbol{\mu}} = \mathbf{c} - Z'\hat{\mathbf{p}} = \mathbf{0},$$

(7.c) $$\frac{\partial \mathscr{L}}{\partial \lambda} = 1 - \hat{\mathbf{p}}'\mathbf{1} = 0.$$

Based on the system of Equations 7, we can solve for $\hat{\boldsymbol{\mu}}$, $\hat{\lambda}$, and then $\hat{\mathbf{p}}$ to obtain

(8) $$\hat{\mathbf{p}} = \exp(-\mathbf{1} - Z'\hat{\boldsymbol{\mu}} - \hat{\lambda}\mathbf{1})$$

or

(9) $$\hat{\mathbf{p}} = W^{-1}\exp(-Z'\hat{\boldsymbol{\mu}}),$$

where $W = \mathbf{1}' \exp(-Z'\hat{\boldsymbol{\mu}})$. The second-order Hessian is negative definite, which assures a unique global solution.

II(iv). *More Powerful Alternatives*

This approach gives a simple method for calculating market shares. Lacking any additional information (constraints), our maximum entropy criterion forces the estimates toward equal shares. That is, the estimate of the shares of each of the first four firms equals C4/4; the estimate of the share of the next four firms equals (C8 − C4)/4; and so forth. Because researchers are usually interested especially in the shares of the first four or eight firms, we want estimates that allow the shares of (especially the first few) firms to differ.

To achieve this objective, we impose two additional restrictions. First, we add the information contained in the nonlinear HHI restriction. Our new problem is to maximize entropy, Equation 5, subject to the adding up constraint, Equation 1, the concentration ratio restrictions, Equation 4, and the HHI restriction, Equation 2. We find that imposing this HHI restriction always results in a more accurate estimate according to the criteria we discuss below.

Second, we restrict the estimated distribution (the probabilities) to be a smooth function, $p_j = f(j)$, of the firm index, j. From inspection of a large number of industries, we find that the $f(\cdot)$ function is almost always well approximated by a negative exponential or a low-order polynomial. Any structural form may be easily imposed in our framework. In the work discussed below, we use a cubic

(10) $$p_j = \alpha_0 + \alpha_1 j + \alpha_2 j^2 + \alpha_3 j^3.$$

Below, we develop two criteria that allow us to infer whether this restriction will help.

II(v). *Noise*

In some cases (mainly when we require that the estimated distribution be smooth), we cannot obtain a pure solution to these maximum entropy problems. In those cases, we follow Judge and Golan [1992] and generalize the pure consistency relations, Equations 2 and 4, to include a noise component:

(11.a) $$\mathbf{c} = Z\mathbf{p} + Z\mathbf{e},$$

(11.b) $$\sum_{i=1}^{50} (p_i + e_i)^2 = \text{HHI},$$

where $\mathbf{e}$ is a noise vector of dimension $(1 \times N)$. If we are to use the information in the inverse relations (11a) and (11b) with the entropy framework, the unknown $\mathbf{p}$ and $\mathbf{e}$ must have the properties of probabilities. The elements of $\mathbf{p}$ are already in the form of probabilities, and the e_i may range over the interval no larger than $[-C4/4, C4/4]$. Because the noise is not in a probability form, we reparameterize the elements of $\mathbf{e}$. To do so, we define a set of $J \geqslant 2$ discrete points $\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iJ})'$ such that $e_i \in [v_{i1}, v_{iJ}]$, where $e_i = \mathbf{v}_i'\mathbf{w}_i = \Sigma_j v_{ij} w_{ij}$, and $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{iJ})'$ is an unknown vector of probabilities such that

(12) $$\mathbf{1}'\mathbf{w}_m = 1.$$

That is,

(13) $$\mathbf{e} = \begin{bmatrix} \mathbf{v}_1' & & & \\ & \mathbf{v}_2' & & \\ & & \ddots & \\ & & & \mathbf{v}_N' \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_N \end{bmatrix} = V\mathbf{w}.$$

Consequently, we may rewrite Equations 11 as

(14.a) $$\mathbf{c} = Z\mathbf{p} + ZV\mathbf{w},$$

(14.b) $$\sum_{i=1}^{50} (p_i + \mathbf{v}_i'\mathbf{w}_i)^2 = \text{HHI},$$

and the generalized maximum entropy problem becomes

(15) $$\max_{\mathbf{p}, \mathbf{w}} - \mathbf{p}' \ln \mathbf{p} - \mathbf{w}' \ln \mathbf{w},$$

subject to equations 1, 12, 14a, and 14b. Forming the Lagrangian and maximizing with respect to $\mathbf{p}$ and $\mathbf{w}$ yields the optimal solution $\tilde{\mathbf{p}}$ and $\tilde{\mathbf{w}}$, where the second-order Hessian is, again, negative definite, which assures a unique global solution.

III. A MEASURE OF DIVERSITY

The solution $\hat{\mathbf{p}}$ (or $\tilde{\mathbf{p}}$) to the maximum entropy problem suggests a measure of firm-size diversity (cf. Jacquemin and Berry [1979]). If all the firms were of the same size, each would have a market share of $1/N$, which maximizes the Shannon entropy measure H. Consequently, the ratio of the estimated Shannon entropy measure, $H(\hat{\mathbf{p}})$, to the Shannon measure if each firm is the same size, $1/N$, is

$$(16) \qquad S(\hat{\mathbf{p}}) = \frac{-\sum_{i=1}^{N} \hat{p}_i \ln \hat{p}_i}{-\sum_{i=1}^{N} \frac{1}{N} \ln \frac{1}{N}} = \frac{-\sum_{i=1}^{N} \hat{p}_i \ln \hat{p}_i}{\ln N}.$$

If the firms are equal in size, $S(\hat{\mathbf{p}}) = 1$. As the size distribution of firms becomes less uniform (more diverse), $S(\hat{\mathbf{p}})$ approaches 0. Thus, $0 < S(\mathbf{p}) \leqslant 1$.

IV. APPLICATIONS

How well can we recover firm shares using the maximum entropy approach? To answer this question, we apply our method to actual data. Because we do not know the true distribution of shares for the four-digit SIC industries that the Bureau of the Census analyzes, we conduct our experiments using data from industries for which we do have data. Many trade journals collect data on market shares. This information is summarized in *Market Share Reporter* (Detroit, MI: Gale Research, 1992, 1994). For our experiments, we restrict ourselves to 20 industries for which there are at least 8 firms and the shares reported constitute a substantial majority of the total industry.

IV(i). *An Example*

We start by using one industry, Telecommunication Equipment Firms Worldwide, to illustrate the difference among our various estimates, as shown in Figure 1. The actual shares lie along a moderately smooth, downward-sloping line.

According to our "simple" estimator that uses information only about concentration measures, shares of the first four firms are equal, the shares of the next four firms are equal, and the shares of the last two firms are equal. Although this estimate is uninformative about the relative shares of the largest firms, it fits the overall distribution surprisingly well. The correlation, ρ, between the actual shares, $\mathbf{p}$, and the estimated shares, $\hat{\mathbf{p}}$, is 0.84. The average squared errors, $\text{ASE} = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - p_i)^2$, is 0.0007.

We can do substantially better, however, by also making use of information about the HHI as well as the concentration measures, which we call our "base" estimate. The base estimate of the share of the first firm is almost exactly right ($\hat{p}_1 = 0.229$, $p_1 = 0.213$). The estimated shares of the next three firms are roughly (but not exactly) equal, the shares for the next four firms are roughly equal, and the shares for the last two firms are roughly equal.

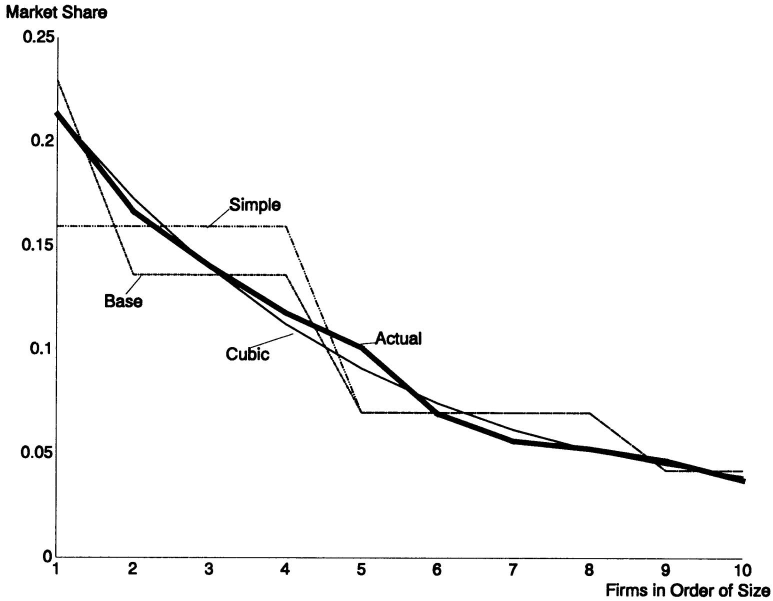

Figure 1
Size Distribution of Telecommunication Equipment Firms Worldwide

The fit is substantially better than for the simple estimate. The correlation, $\rho = 0.95$, is 11 points higher than for the simple estimate, and the average squared errors, ASE = 0.0003, is less than half as large. Moreover, by including the HHI measure, we insure that the variance of the estimated distribution equals the variance of the actual distribution.

Our third, "cubic", estimate uses the restrictions from the concentration and HHI measures and imposes smoothness by requiring that the shares lie along a cubic curve. We use the noise formulation (Equations 15, 1, 12, 14) with $v_i = (-0.01, 0, 0.01)$ for all $i = 1, 2, \ldots, N$. The estimated parameters of the cubic are $\hat{\alpha}_0 = 0.261$, $\hat{\alpha}_1 = -0.053$, $\hat{\alpha}_2 = 0.0043$, $\hat{\alpha}_3 = -0.00013$. The cubic estimate shows greater diversity of firms, $S(\hat{\mathbf{p}}) = 0.93$, than does the basic estimate, $S(\hat{\mathbf{p}}) = 0.94$.

The cubic estimate is almost identical to the actual distribution: $\rho = 0.996$ and ASE = 0.00002. The estimate is very accurate for the first four shares — the ones that are of particular interest. The difference between the estimated and actual share of the first firm, $D_1 = |p_1 - \hat{p}_1|$, is only 0.00002; and the sum of the absolute differences for the first four firms, $D_2 = \sum_{j=1}^{4} |p_j - \hat{p}_j|$, is 0.013, or only a little more than a percentage point.

IV(ii). *Twenty Industries*

Is our method's ability to fit the size distribution for the Telecommunications Equipment Manufacturing industry unusual? To answer this question, we look at the base and cubic estimates for this industry and for 19 other industries, as shown in Table I. The first column of the table shows the

TABLE I
ACCURACY OF TWO-FIRM SIZE ESTIMATES FOR 20 INDUSTRIES

	Number of Firms, N	*Base Estimate* 1,000 × ASE	*Base Estimate* ρ	*Cubic Estimate* 1,000 × ASE	*Cubic Estimate* ρ	D_1	D_2
AC and Refrigeration Worldwide 1990	12	0.16	0.94	**0.09**	**0.97**	0.004	0.029
Airlines 1991	12	0.45	0.94	**0.04**	**0.99**	0.011	0.038
Apples and Pears 1990	21	0.03	0.97	**0.02**	**0.98**	0.014	0.030
U. S. Builders 1992	15	**0.03**	**0.997**	2.40	0.75	0.0008	0.014
General Building Contractors 1992	10	0.56	0.88	**0.13**	**0.97**	0.018	0.063
European Electronics Producers 1991	15	0.30	0.96	**0.06**	**0.99**	0.005	0.028
Facsimile Machines 1990	12	0.90	0.92	**0.23**	**0.98**	0.011	0.060
Urban Fiber Systems Producers 1990	10	1.30	0.93	**0.16**	**0.99**	0.011	0.073
Furniture Makers 1991	25	**0.15**	**0.98**	0.56	0.92	0.010	0.043
Oil and Gas Producers 1992	15	0.53	0.90	**0.19**	**0.97**	0.005	0.073
Paper Companies in Japan 1990	11	0.20	0.94	**0.01**	**0.998**	0.003	0.014
Ethical Pharmaceutical Worldwide 1990	20	0.37	0.92	**0.004**	**0.99**	0.003	0.007
Polyethylene 1990	11	0.31	0.95	**0.07**	**0.99**	0.001	0.020
Shampoo 1990:4	21	0.29	0.91	**0.07**	**0.98**	0.007	0.043
Software 1989	47	**0.02**	**0.99**	0.45	0.79	0.006	0.038
Stone, Clay, Glass and Concrete 1992–3	8	1.50	0.88	**0.18**	**0.99**	0.021	0.043
Telecommunications Eqp. Worldwide 1988	10	0.30	0.95	**0.02**	**0.996**	0.000	0.013
Tissues 1990	13	0.66	0.93	**0.14**	**0.99**	0.020	0.048
Turkey Meat Products, 1993 (est.)	28	0.04	0.97	**0.03**	**0.98**	0.021	0.047
Veterinary Drug Worldwide 1990	19	0.09	0.95	**0.07**	**0.96**	0.025	0.050

Note: $ASE = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i)^2$; $D_1 = |p_1 - \hat{p}_1|$ and $D_2 = \sum_{j=1}^{4}|p_j - \hat{p}_j|$, for the superior estimate.

number of firms for which we have data for each of the industries.[3] The next two columns show summary statistics—1,000 × ASE and ρ—for our base

[3] If the shares do not add to 100 percent, we renormalize them so that they do.

estimate. The next two columns show the same summary statistics for our cubic estimate. The final two columns show D_1 and D_2 for the best estimate (the cubic for all but three industries). Which of the two estimates fits better for a given industry is indicated by bold numbers (a lower ASE is better and a higher ρ is better).

The ASE, ρ, D_1, and D_2 measures show that the fit is very close for at least one of the two estimates for each industry. Using a Kolmogorov-Smirnov test of the hypothesis that the actual and estimated distribution are the same, we cannot reject the hypothesis at the $\alpha = 0.05$ level for any of our 40 estimates.

For all industries, the base estimate is reasonably good: $\rho \geqslant 0.88$ for all 20 industries, and $\rho \geqslant 0.91$ for 18 industries. In 17 out of 20 (85 percent) of our industries, the cubic estimate is superior to the base estimate. If we use the best estimate for each industry, the minimum correlation is 0.96, the correlation is at least 0.98 for 17 industries (85 percent), and the correlation is at least 0.99 for 11 of our industries (55 percent). Our best estimate of the share of the largest firm is never off by more than 2.5 percentage points, and for 11 industries, the estimate is well within one percentage point. Similarly, for 16 of the industries, the sum of the error for the first four firms is five percentage points or less.

How do we know whether to use the base estimate or the cubic? We find that the cubic dominates the base estimate when a pure cubic solution exists or we only have to add less than one percent or less noise [that is, $v_i = (-0.01, 0, 0.01)$ for all i] to obtain a solution. For the three industries where the cubic was inferior, we had to allow for a substantial amount of noise (4 percent). Thus, if we only use the cubic estimate when little noise is required and a solution exits, the cubic estimator is superior to the base estimator (at least for our sample of 20 industries).

We also conducted Monte Carlo experiments where we allowed the total number of firms to grow. We find that the accuracy of the firm share estimates for the first 10 firms (the ones we care most about) does not change with the total number of firms. When there are many firms, we always estimate the shares of the small firms accurately, of course, because those shares are very small and virtually identical.

IV(iii). *Extra Measures of Information (Constraints)*

As we illustrated for the Telecommunications Industry, when the Bureau of the Census started providing the HHI, our method's ability to fit the distribution of firm sizes accurately increased substantially. If the government supplied a few more measures (without violating confidentiality), we could estimate the firm sizes even more accurately.

We experimented with several measures, including C2, the share of the top two firms, and a measure of the third moment of the distribution, HHI3 $= \Sigma\, p_i^3$ (the concentration measures provide information about the first moment, and the HHI is a measure of the second moment). For two-thirds of the industries, the HHI3 improves the fit (as measured by the ASE and ρ), markedly. If the government would supply C2 and HHI3, we could recover virtually all the information about the size distribution of firms; however, we doubt if the government would ever provide C2 for reasons of confidentiality. The less uniform the shares, the higher the contribution of these additional measures.

As a final experiment, we removed information. Using the three summary measures as criteria, dropping either C4 or HHI substantially reduces our ability to fit the distribution. Dropping C8 and higher concentration measures, however, has relatively little effect on our ability to fit the distribution.

If we drop all measures save one, we find that the HHI contains more information [in terms of ASE, ρ, and $S(\hat{\mathbf{p}})$] than C4 or the other concentration measures. Many industrial organization researchers have recently switched from using C4 to the HHI as an explanatory variable in their performance equations, which is consistent with our results concerning using a single measure. We recommend, of course, using all the available measures.

V. SUMMARY AND CONCLUSIONS

Using a maximum entropy approach, we can recover estimates of the market shares of individual firms based on only a handful of government summary statistics. Because the number of unknown parameters exceeds the number of data points provided by the Bureau of the Census, this problem is ill-posed, and traditional statistical approaches are not feasible. Nonetheless, our method allows us to use all the available information to predict market shares that are highly correlated with the actual shares. For the 20 industries we studied, the minimum correlation was 0.96; and, the correlation was 0.99 or above for 11 industries. This technique is easy to implement with any optimization software (such as GAMS or MATLAB).

Our results have implications for industrial organization research. They suggest that the recent switch of most researchers to using the HHI instead of C4 is reasonable (if only one measure is to be used). Both measures, however, contain useful information. If the government would provide even one or two pieces of additional information, the two firm concentration measures or a variation on the Herfindahl-Hirschmann Index, HHI3 $= \Sigma_i p_i^3$, a nearly perfect fit of the distribution is possible for virtually every industry using our maximum entropy approach.

AMOS GOLAN,
Department of Agricultural and Resource Economics,
207 Giannini Hall,
University of California,
Berkeley, CA 94720,
USA.

ACCEPTED APRIL 1995

and

GEORGE JUDGE,
Department of Agricultural and Resource Economics,
207 Giannini Hall,
University of California,
Berkeley, CA 94720,
USA

and

JEFFREY M. PERLOFF,
Department of Agricultural and Resource Economics,
207 Giannini Hall,
University of California,
Berkeley, CA 94720,
USA

REFERENCES

CARLTON, D. W. and PERLOFF, J. M., 1994, *Modern Industrial Organization*, 2nd Edition, (HarperCollins Publishers, New York).

ISAAC, R. M. and REYNOLDS, S.S., 1989, 'Two or Four Firms: Does it Matter?' Working paper, University of Arizona.

JACQUEMIN, A. P. and BERRY, C. H., 1979, 'Entropy Measure of Diversification and Corporate Growth,' *Journal of Industrial Economics*, 27, pp. 359–369.

JAYNES, E. T., 1957a, 'Information Theory and Statistical Mechanics,' *Physics Review*, 106, pp. 620–630.

JAYNES, E. T., 1957b, 'Information Theory and Statistical Mechanics, II' *Physics Review*, 108, pp. 171–190.

JUDGE, G. and GOLAN, A., 1992, 'Recovering Information in the Case of Ill-Posed Inverse Problems with Noise,' Working paper, University of California, Berkeley, Department of Agricultural and Resource Economics.

KWOKA, J. E., 1979, 'The Effect of Market Share Distribution on Industry Performance,' *Review of Economics and Statistics*, 61, pp. 101–9.

LEVINE, R. D., 1980, 'An Information Theoretical Approach to Inversion Problems,' *Journal of Physics*, A, 13, pp. 91–108.

SHANNON, C. E., 1948, 'A Mathematical Theory of Communication,' *Bell System Technical Journal*, 27, 379–423.

ZELLNER, A., 1988, 'Optimal Information Processing and Bayes Theorem,' *American Statistician*, 42, pp. 278–84.