Classification and Aggregation: An Application to Industrial Classification in CPS Data

R. Cotterman, F. Peracchi

Journal of Applied Econometrics, Volume 7, Issue 1 (Jan. - Mar., 1992), 31-51.

JOURNAL OF APPLIED ECONOMETRICS, VOL. 7, 31–51 (1992)

CLASSIFICATION AND AGGREGATION: AN APPLICATION TO INDUSTRIAL CLASSIFICATION IN CPS DATA

R. COTTERMAN
Unicon Research Corporation, 1640 Fifth Street, Santa Monica, CA 90401, USA

AND

F. PERACCHI
Department of Economics, New York University, 269 Mercer Street, New York NY 10003, USA

SUMMARY

In this paper we offer a method for deciding how to aggregate a set of elementary industries. The method is then applied to the problem of estimating a wage equation that allows for industry-specific effects. Our method explicitly formalizes the trade-off between goodness-of-fit and parsimony implicit in an aggregation problem. By varying the parameter of the assumed loss function, one obtains a whole sequence of aggregation levels. Further, the resulting sequence is consistent; that is, groupings formed at one level of aggregation will never be undone when one aggregates further.

1. INTRODUCTION

It is commonly assumed in empirical work that the conditional expectation of a worker's real wage is a function of worker's characteristics, including sex, education, experience, industry of employment, etc. Typically, however, there is uncertainty about the appropriate industrial classification to adopt. Few researchers have the inclination to define industries from scratch, so most adopt some variant of commonly used classifications, such as the Standard Industrial Classification (SIC) or those developed by the Bureau of the Census. The choice problem does not end there, however. In their most detailed forms both the Census and SIC classifications have hundreds of categories; thus, to facilitate understanding and communication, it is generally necessary to aggregate from the most detailed level, even though this may entail some loss of information. The questions then arise of how to aggregate and when to stop aggregating. That is, how does one aggregate so as to maintain important industry distinctions, and where does the information loss become great enough to dominate the desire for additional parsimony?

The first of these questions—how to aggregate—admits of several answers. One can appeal to the same authorities: both the Census and SIC classifications are available at various levels of detail, and the aggregation schemes are seemingly based on similarity of industry output or function. The implicit assumptions are that similarity of function is important information to retain, and that existing aggregation schemes faithfully reflect these functional differences and similarities. Alternatively, one may produce a new aggregation scheme through common sense

0883–7252/92/010031–21$10.50

Received September 1989
Revised June 1990

or intuition, usually with an eye towards the information that is crucial for the analysis to be conducted later.

A more systematic method is to set up an explicit information criterion and to aggregate so as to minimize the information loss. For example, a natural choice may be to aggregate so as to minimize the within-industry-group variance of wages.[1] This method, however, has two drawbacks. First, it can lead to grouping industries that seem completely unrelated in most dimensions other than wages. Second, it typically leads to inconsistent groupings: to minimize within-group variance, groups formed at a lower level of aggregation may have to be broken up at a higher level.

In this paper we offer an alternative method for deciding how to aggregate industries and when to stop the process. The method consists of two steps. First we use a measure of dissimilarity between industries to construct a hierarchical 'industry tree' which represents the set of admissible classifications. Then we produce a sequence of classifications by 'pruning' the tree using a loss function that trades off goodness-of-fit with the complexity of the classification. Each classification in the sequence is optimal for a range of values of this trade-off.

Our procedure offers two main advantages over other aggregation methods. First, we explicitly formalize the trade-off between goodness-of-fit and parsimony that is implicit in an aggregation problem. Second, because aggregation is constrained to follow the structure of the industry tree, the sequence of optimal classifications is consistent; that is, groupings formed at one level of aggregation will never be undone when additional aggregation takes place.

The plan of the paper is as follows. Section 2 presents the method adopted to construct and prune the industry tree. Section 3 describes the data set used for the empirical implementation, as well as the results obtained when using industry coding match rates to define a measure of dissimilarity. Section 4 reports the results obtained when using industry transition rates instead to form a measure of dissimilarity. Section 5 contains the conclusions.

2. TREE CONSTRUCTION AND OPTIMAL PRUNING

Given a set of p elements (in our application, elementary industries), a classification is a partition of this set into a number of mutually exclusive and totally exhaustive subsets. Given a classification P, the number of elements in the classification, denoted by $|P|$, is called the complexity of the classification. Given two classifications, P and P', we say that P is finer than P' (or P' is coarser than P), written $P > P'$, if all elements of P' consist of subsets of elements of P and $|P| > |P'|$.

An example is the Census classification. Elementary industries are first classified into 213 relatively fine groups (three-digit level). The three-digit industries are then further aggregated into 50 groups (two-digit level) and finally into 12 to 16 categories (one-digit level). Choosing the coarser classification facilitates understanding and communication and reduces the risk of misclassification. On the other hand, choosing the finer classification increases the goodness-of-fit and decreases the risk of missing important details. This trade-off between goodness-of-fit and parsimony may be represented by a loss function of the form

$$R(P) = S(P) + \psi(P) \tag{1}$$

where $S(\cdot)$ penalizes lack of fit and $\psi(\cdot)$ is a penalty for complexity. A classification P' is then preferred to P if $R(P') < R(P)$.

[1] For an application of this methodology to occupational aggregation, see Welch and Maclennan (1976).

Criteria such as (1) appear frequently in econometrics and statistics (e.g. Akaike information criterion, etc.). If we choose $\psi(P) = \alpha|P|$, $\alpha \geqslant 0$, we have the simple and mathematically convenient specification:

$$R_\alpha(P) = S(P) + \alpha\,|P|. \tag{2}$$

Since we are interested in the conditional expectation of a dependent variable y (log real wage) given a vector x of explanatory variables (schooling, experience, etc.), a reasonable choice for $S(P)$ is the total residual sum of squares

$$S(P) = \sum_{j \in P} \mathrm{SSE}(j), \tag{3}$$

where SSE(j) is the residual sum of squares in the least-squares (LS) regression of y on x using the observations corresponding to the jth industry group in the classification P.[2]

For a given value of the trade-off parameter α, an optimal classification P^* may be obtained by minimizing $R_\alpha(\cdot)$ over some set $\mathscr{P}$ of possible classifications. Formally

$$P^* = \operatorname*{argmin}_{P \in \mathscr{P}} R_\alpha(P). \tag{4}$$

In general, the smaller is α, the finer is the optimal classification P^*. For example, when $\alpha = 0$, P^* coincides with the set of p elementary industries. Further, there exists a value $\alpha_{\max}$ such that, for $\alpha > \alpha_{\max}$, P^* consists of only one element, namely the aggregate of all industries.

The nature of the choice set $\mathscr{P}$ is also important. If $\mathscr{P}$ consists of only a few elements (e.g. the one-, two-, and three-digit Census classifications), then problem (4) is straightforward. On the other hand, it is clear that such a choice of $\mathscr{P}$ is too narrow for many purposes. For example, one may be interested in levels of aggregation intermediate between the three- and two-digit level, or between the two- and one-digit level. Further, given the purposes of the analysis, it is not at all clear whether the two-digit (one-digit) Census classification is the best way of aggregating the three-digit industries into 51 (12 to 16) groups. At the other extreme, the alternative of letting $\mathscr{P}$ be the set of all possible partitions of a set of p elements is often not viable. Even for p moderately large, $\mathscr{P}$ would contain a very high number of classifications, and the solution to (4) would be difficult to characterize. Thus, to obtain an operational procedure, more structure must be placed on the problem.

There is yet another difficulty. Varying the value of α in problem (4) need not result in a nested sequence of classifications; that is $\alpha_1 < \alpha_2$ need not imply that $P_1^* \geqslant P_2^*$, where P_k^* denotes the optimal classification when $\alpha = \alpha_k$, $k = 1, 2$. To insist that the optimal classifications should form a nested sequence is a basic consistency requirement that prevents groups that have been formed at a lower level of aggregation from being broken at a higher level, and industries that have previously been merged with others from reappearing separately.

To construct a consistent sequence of optimal classifications, observe that a nested sequence of classifications may be obtained by progressively removing, or 'pruning', more and more branches starting from the bottom of a binary hierarchical tree[3] whose terminal nodes are the

[2] There are two problems with this choice. The first is that the SSE may not provide a good measure of the accuracy of prediction. A better measure may be obtained by splitting the set of observations into two subsets, the first used for estimation and the second for prediction. The second problem is the sensitivity of LS to the presence of outliers. In principle, LS could be replaced by other, more robust, alternatives.

[3] A binary hierarchical tree is a tree that has a unique root node and exactly two branches stemming from each non-terminal node. We adopt the conventional of orienting the tree with the root node at the top and the terminal nodes at the bottom.

p individual industries. Each classification in the sequence corresponds to the set of terminal nodes of one of these pruned subtrees. This suggests a two-step procedure in which we first construct a hierarchical 'industry tree', which represents the set of admissible classifications, and then we derive a sequence of classifications by optimally pruning the tree using the loss function (2)–(3). It is worth stressing that this procedure differs from a standard clustering algorithm: here the role of the industry tree is solely to provide a discipline for the pruning process. Further, our method represents an alternative to 'best subsets' procedures (see e.g. Leamer, 1990) which, by simply minimizing within-group variance for a given level of aggregation, cannot guarantee consistency of the sequence of optimal classifications.

2.1 Construction of a hierarchical tree

The basic idea in constructing a binary hierarchical tree is to group together the two industries or industry groups that, at each stage of the process, are closest to each other according to some measure of closeness. This construction requires specifying two elements: a measure of closeness between elementary industries, and a 'linkage method', that is, a way of measuring closeness between two disjoint industry groups. Different trees will result from using different measures of closeness or different linkage methods.

To measure closeness between elementary industries one may specify a distance based on a vector z of industry characteristics (e.g. cost and market structure, etc.) and a positive definite weighting matrix W that attaches different importance to the different characteristics. Closeness between two industries, i and j, may then be measured by the distance

$$d_{ij} = [(z_i - z_j)' W(z_i - z_j)]^{1/2}.$$

In practice, agreement on the choice of industry characteristics and on the weighting matrix W is unlikely. To bypass this problem we exploit the fact that, more generally, d_{ij} can be a symmetric dissimilarity measure that need not satisfy the triangle inequality. Our particular choice of dissimilarity measures will be discussed in Sections 3 and 4.

The second problem is to specify a method for measuring closeness between disjoint industry groups. If A, B and C are three disjoint industry groups, and A and B are merged together to form group D, then closeness between C and D may alternatively be measured by

$$\begin{aligned} d_{CD} &= \min\{d_{AC}, d_{BC}\rangle && \text{(single linkage)} \\ d_{CD} &= \max\{d_{AC}, d_{BC}\} && \text{(complete linkage)} \\ d_{CD} &= (w_A d_{AC} + w_B d_{BC})/(w_A + w_B) && \text{(average linkage),} \end{aligned}$$

where w_A and w_B are weights assigned to groups A and B respectively (for example, the weight may be the number of observations classified in each group). These linkage methods are widely used in automatic classification algorithms because they are easy to implement. Of the three methods, single-linkage is probably the least appealing because it tends to produce trees that are very unbalanced (see e.g. Lebart *et al.*, 1984.[4] Our experience in implementing these methods will be discussed in Sections 3 and 4.

[4] The maximal number of edges between the root and a terminal node of a tree is called the height of a tree. A balanced tree is one with minimal height, namely $[\ln_2 p]$ (see e.g. Zupan, 1982), where p is the number of terminal nodes and $[z]$ denotes the smallest integer greater than or equal to z. A completely unbalanced tree, on the other hand, has height equal to $p - 1$.

2.2. Pruning the tree

First we introduce some terminology. Let t be a node of a binary hierarchical tree T with root node $\{1\}$. A node t' is called a descendant of t if there exists a connected path down the tree leading from t to t'. The node t is then called an ancestor of t'. The tree T is entirely described by specifying the descendants (or, equivalently, the ancestor) of each of its nodes. A terminal node is one with no descendant. The root node is the unique node with no ancestor. The classification corresponding to T consists of the set of terminal nodes of T, denoted by $\tilde{T}$. Given a non-terminal node t, a branch T_t of T is the subtree consisting of t and all its descendants. Pruning a branch T_t from T consists of deleting from T all descendants of t. The resulting pruned subtree, denoted by $T - T_t$, has the same root node as T but fewer terminal nodes. Thus we write $T > T - T_t$.

By (2)–(3), the loss associated with T is given by

$$R_\alpha(T) = S(\tilde{T}) + \alpha \mid \tilde{T} \mid = \sum_{j \in \tilde{T}} \mathrm{SSE}(j) + \alpha \mid \tilde{T} \mid.$$

This implies that

$$R_\alpha(T) = R_\alpha(T - T_t) + R_\alpha(T_t) - R_\alpha(t)$$

for any non-terminal node t of T. Thus, the coarser classification corresponding to the pruned subtree $T - T_t$ is preferred to the classification corresponding to T whenever

$$R_\alpha(t) = S(t) + \alpha < S/\tilde{T}_t) + \alpha \mid \tilde{T}_t \mid = R_\alpha(T_t),$$

that is, whenever

$$\alpha > [S(t) - S(\tilde{T}_t)] / (\mid \tilde{T}_t \mid - 1) \equiv a^*(t).$$

Notice that $a^*(t) \geqslant 0$ since, by the algebra of LS, $S(t) \geqslant S(\tilde{T}_t)$.[5]

When $\alpha = 0$, no branch of T should be pruned off, and so the optimal classification is just $\tilde{T}$. As α increases from zero, the first branch of T to be pruned off is the one whose root node is the non-terminal node t for which $\alpha^*(t)$ is smallest. Thus let t^* be the first non-terminal node of T for which $R_\alpha(t) - R_\alpha(T_t) = 0$ as α increases. Clearly, $t^* = \mathrm{argmin}_{t \in T - \tilde{T}}\, \alpha^*(t)$. At this point the coarser classification corresponding to the pruned subtree $T^1 = T - T_t^*$ becomes preferable to $\tilde{T}$, and $\alpha^*(t^*)$ is the value of α at which this occurs. The entire process can now be repeated starting from the new tree T^1. This pruning method results in:

(a) A positive integer K.
(b) A strictly decreasing sequence of trees

$$T = T^0 > T^1 > \cdots > T^K = \{1\}.$$

(c) A corresponding strictly increasing sequence of α values

$$0 = \alpha_0 < \alpha_1 < \cdots < \alpha_K < \infty,$$

[5] The threshold value $\alpha^*(t)$ may be related to the classical F-test statistic for aggregation (see e.g. Grunfeld and Griliches, 1960). Since $T - T_t$ is nested within T and contains $\mid \tilde{T}_t \mid - 1$ terminal nodes less, the F-test statistic for aggregation is given by

$$\xi = \frac{[S(t) - S(\tilde{T}_t)]/k(\mid \tilde{T}_t \mid - 1)}{S(\tilde{T})/(N - k \mid \tilde{T} \mid)}$$

where k is the number of regressors and N is the total number of observations in the sample. Because ξ is proportional to $\alpha^*(t)$, choosing a value of α implicitly corresponds to choosing the significance level of an F-test for aggregation.

such that, for $\alpha_k \leqslant \alpha < \alpha_{k+1}$, $k = 0, \ldots, K$, the tree T^k is the smallest pruned subtree of T minimizing $R_\alpha(\cdot)$.

Let t be a non-terminal node of T^k, and define

$$\alpha_k^*(t) = [S(t) - S(\tilde{T}_t^k)]/(|\tilde{T}_t^k| - 1),$$

where T_t^k is the branch of T^k consisting of t and all its descendants. A complete characterization of the process resulting in (a), (b), and (c) is given b the following:

Proposition (Breiman *et al.*, 1984, p. 289):

(i) $\alpha_{k+1} = \min_{t \in T^k - \tilde{T}^k} \alpha_k^*(t), \quad 0 \leqslant k < K.$

(ii) $T^{k+1} = \{t \in T^k : \alpha_k^*(j) > \alpha_{k+1}$ *for all ancestors* j *of* $t\}, \quad 0 \leqslant k < K.$

(iii) *The smallest subtree minimizing* $R_\alpha(\cdot)$ *is given by*

$$\begin{aligned} T^*(\alpha) &= T^k, \quad \text{if } \alpha_k \leqslant \alpha < \alpha_{k+1}, \qquad 0 \leqslant k \leqslant K-1, \\ &= T^K, \quad \text{if } \alpha \geqslant \alpha_K. \end{aligned}$$

This proposition implies that the nested sequence of classifications $\tilde{T}^0 > \tilde{T}^1 > \cdots > \tilde{T}^K$ is optimal in the sense that $\tilde{T}^k$ is the coarsest classification minimizing $R_\alpha(\cdot)$ for α in the interval $[\alpha_k, \alpha_{k+1})$ and $k = 0, \ldots, K$. The next two sections, illustrate the results obtained in our empirical example

3. AGGREGATION USING INDUSTRY MISMATCH DATA

3.1. Industry coding in the matched Current Population Survey

The Current Population Survey (CPS) is a rotating monthly survey of over 50,000 households with rotation scheme 4–8–4. That is, each household added to the survey is expected to be interviewed once per month for 4 consecutive months and, following an 8-month hiatus, for an additional four consecutive months. If the sample size in each rotation group were constant, then one-half of the households in any particular month could in principle be matched with one-half of the households in the survey for the same month of the following year.[6]

Although the CPS does not provide matched files, a statistical matching algorithm can be used to match individuals across adjacent years.[7] For some of the matched individuals it is then possible to get two independent codings of the same industry of employment, as follows. The basic survey instrument asks all employed persons the industry and other characteristics of the job held during the week preceding the survey.[8] Census personnel later encode this response. The supplement to the March survey also asks the number of employers for whom each individual worked in the preceding calendar year, as well as the industry and other characteristics of the longest spell of employment during the preceding calendar year.

[6] The survey follows a rooftop design, however, and makes no attempt to follow households who change residence after the initial interview. These households are simply replaced by the new residents, who are coded as new entrants to the survey. Thus, the ideal of two blocks of four interviews each with the same households is sometimes not met in practice. The differences between those who leave the survey and those who stay can lead to a selection bias. Later in this section we describe how we use the CPS data on industry of employment to examine the extent and distribution of inconsistent coding of industry. This use of the CPS data seems unlikely to suffer from selection bias.

[7] We are grateful to Shigeru Iwata and Alan Pitts for creating the matched March Annual Demographic Files that were used in this paper. For a detailed discussion of the matching algorithm, see Pitts (1988).

[8] When the individual worked at more than one job, the questions pertain to the job on which the most time was spent.

Consider, then, a matched individual who, during the second of the two March surveys, reports that he/she had only one employer during the preceding calendar year. This employer should be identical to the employer for the job, if any, held at the time of the preceding March survey, and hence industry codes should in principle be identical as well.

There are, or course, several possible reasons for these industry codes to disagree even when they should not. First, there may be pure coding errors or errors in recording the industry of employment or the number of employers in the preceding year.

Second, codes may not match because respondents are not in fact attempting to describe the same jobs. For example, the process of matching individuals across March surveys is subject to error, with the result that information may refer to two different people in the two different surveys. Even when individuals are matched correctly, the interviewee may differ in the two surveys. If some respondents are misinformed about the employment characteristics of other household members, they may give incorrect information on industry of employment, resulting in a mismatch of industry codes. In addition, some respondents may simply fail to recall correctly a job that was held last year.

The third, and in our view most important, source of differences is that the information available to Census coders may be insufficient to classify unambiguously the industry of employment. Even if industries are distinct at a purely conceptual level, the functional characteristics that define different industries may overlap in some dimensions. When the amount of overlap is large relative to the amount of information available, classification becomes inevitably ambiguous. This suggests that industry coding mismatches may be used as a measure of industry dissimilarity, provided that they can be attributed solely to coders' confusion in classifying industries. Unfortunately, we see no simple way to purge the mismatch data of the other sources of errors. We attempt to minimize the importance of recall errors on the part of respondents, however, by restricting the sample to those of age 16 and older who were working at the time of both surveys and for whom, for each of the two surveys, class of worker, three-digit industry, and three-digit occupation were the same for last week's job and last year's longest job.[9]

Using a pooled sample from 1977 to 1982,[10] we constructed a measure of mismatch rates for two-digit industry pairs as follows.[11] For those who met our sample selection criteria (a total of 71,449 workers), we computed a cross-tabulation of industry last week, as reported in year t, by industry last year, as reported in year $t+1$. Since the sample was restricted to those who reported in year $t+1$ that they had only one employer in year t, we should find, barring errors of the sort discussed above, a diagonal cross-tabulation matrix. Letting m_{ij} denote the number of persons who (in year t) report i as the industry last week and who (in year $t+1$) report j as the industry last year, we define the mismatch odds-ratio for industries i and j as

$$r_{ij} = (m_{ij} + m_{ji}) / (m_{ii} + m_{jj}). \tag{5}$$

[9] This procedure helps reducing the impact of recall errors by retaining persons with greater job continuity. In particular, for those persons in each survey for whom the job held last year is the same as the job held last week, the Census forces class of worker, occupation and industry codes to be identical for the two jobs.

[10] Two considerations are relevant in deciding which years to include. First, the question on the number of employers in the preceding calendar year is not asked prior to 1976. Second, only the years from 1971 to 1982 use the 1970 Industry Census classification scheme; earlier years use the 1960 scheme, and later years use the 1980 framework. Attempting to use surveys from 1983 onwards for this analysis runs the risk of introducing spurious errors in making the 1970 and 1980 coding schemes coincide.

[11] The reason for using the two-digit rather than three-digit coding level is discussed below.

Table I. The 25 largest mismatch odds ratios in descending order

Industry 1	Industry 2	Odds ratio	Number in industry 1 both times	Number in industry 2 both times
5 Ordnance	14 Aircraft	0·0901	136	430
9 Primary metals	10 Fabricated metals	0·0724	879	779
32 Wholesale	34 Other retail	0·0570	1904	6812
38 Business	46 Other professional	0·0477	871	1561
42 Medical, except hospitals	43 Hospitals	0·0472	1918	2784
18 Food	32 Wholesale	0·0470	1246	1904
10 Fabricated metals	11 Machinery, except electrical	0·0452	779	1764
11 Machinery, except electrical	12 Electrical equipment	0·0425	1764	1527
3 Mining	25 Petroleum	0·0398	622	131
11 Machinery, except electrical	32 Wholesale	0·0360	1764	1904
17 Miscellaneous manufacturers	38 Business	0·0348	308	871
12 Electrical equipment	16 Instruments	0·0343	1527	338
37 Private household services	40 Personal services	0·0317	532	1518
42 Medical, except hospitals	44 Welfare/religion	0·0315	1918	1038
20 Textiles	21 Apparel	0·0277	516	822
44 Welfare/religion	46 Other professional	0·0262	1038	1561
10 Fabricated metals	13 Automobiles	0·0254	779	793
24 Chemicals	26 Rubber/plastics	0·0251	963	394
32 Wholesale	38 Business	0·0238	1904	871
32 Wholesale	39 Repair	0·0236	1904	723
44 Welfare/religion	50 State	0·0234	1038	759
29 Other transportation	32 Wholesale	0·0219	1886	1904
24 Chemicals	32 Wholesale	0·0216	963	1904
50 State	51 Local	0·0214	759	1483
34 Other retail	39 Repair	0·0207	6812	723

That is, the mismatch odds ratio between industries i and j is the number of individuals who are classified once as being in industry i and once as being in industry j, scaled by the number of individuals who are consistently classified in either industry i or industry j.[12]

Table I lists the 25 largest mismatch odds ratio in decreasing order, along with the size of the denominator. That coder confusion is an important source of mismatching seems fairly clear, for we generally find the industry pairs to be closely related on *a priori* grounds. The nature of the confusion seems to concern the precise product (e.g. the first entry in the list—ordnance versus aircraft) or the level of distribution (e.g. the third entry in the list—wholesale versus other retail).

We now illustrate the results obtained when coder confusion is used as the basis for constructing industry aggregates. To do so, we proceed in two steps: first, we use match rates to form an industry tree; then we optimally prune the tree using the method described in Section 2.2.

[12] Note that mismatching is not equivalent to miscoding. If an individual is classified in industry i on one occasion and in industry j on another, we label this event a mismatch regardless of whether either or neither of these is the true industry. Similarly, if an individual is classified in industry i on both occasions, we call this a match regardless of whether the individual is truly in industry i.

3.2. Construction of the industry tree

Construction of the industry tree requires a dissimilarity measure and a linkage method. We define the dissimilarity d_{ij} between industries i and j as the match rate

$$d_{ij} = (1 + r_{ij})^{-1} = (m_{ii} + m_{jj}) / (m_{ii} + m_{jj} + m_{ij} + m_{ji}). \tag{6}$$

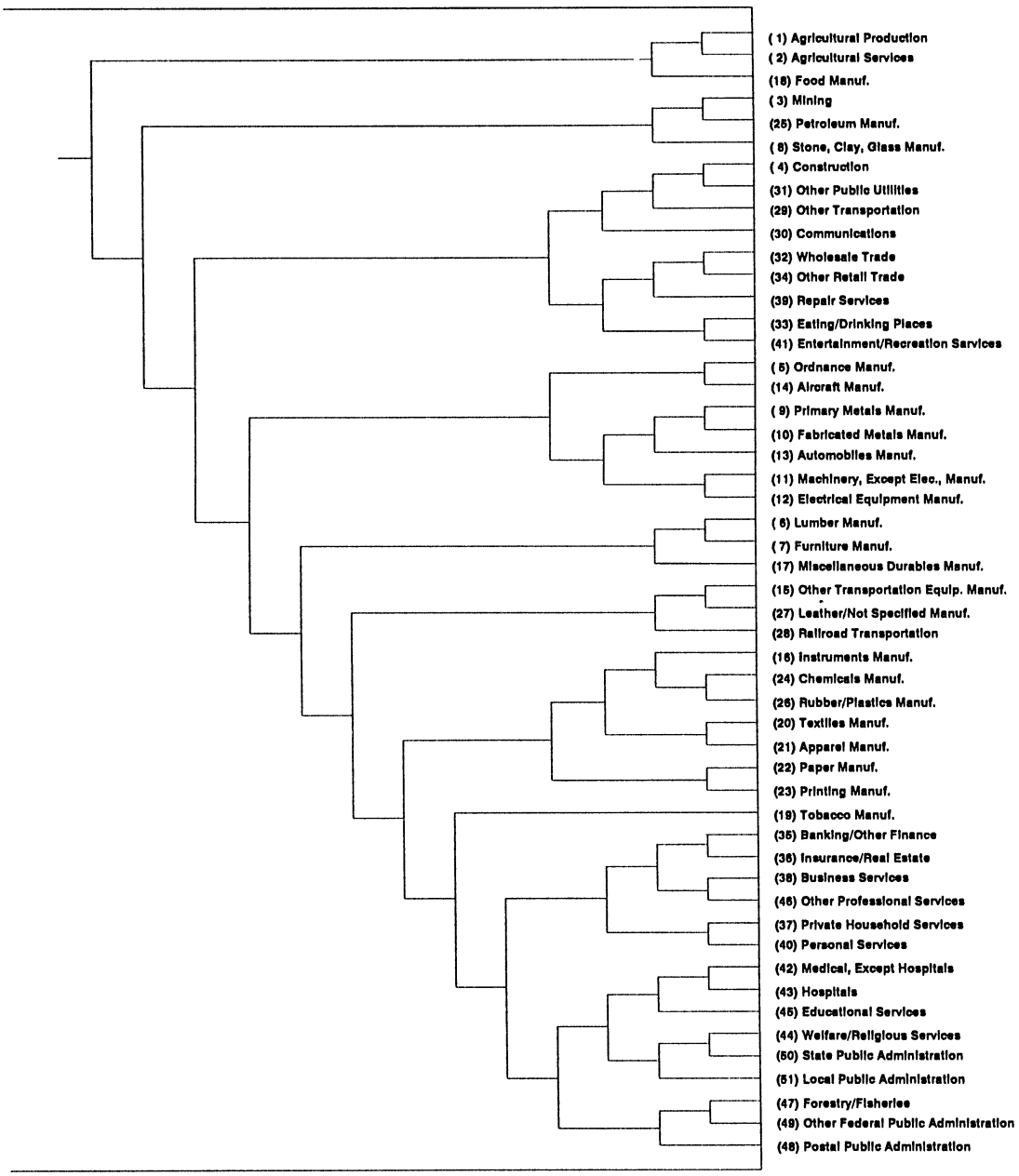

Figure 1. Two-digit industry tree formed from mismatch data

When the mismatch odds ratio is low, there is little confusion between the two industries, and in this sense they are more dissimilar. In the extreme case in which no errors are made in reporting and coding industries, the off-diagonal elements m_{ij} and m_{ji} are zero, and the dissimilarity measure attains its maximum value of 1. At the other extreme, if it is impossible to distinguish between the two industries or coding is entirely random, then in large samples the on- and off-diagonal elements will all be equal in size, resulting in $d_{ij} = 0.5$.

Having chosen a dissimilarity measure, we specify a linkage method. The results reported here use the complete linkage method defined in Section 2.1. We also experimented with other linkage methods, but virtually all of them were deemed much less satisfactory because they produced substantial 'chaining', that is, very unbalanced trees.

The industry tree thus obtained, starting from the two-digit Census industries,[13] is displayed in Figure 1. Each two-digit industry is listed on the right-hand side of the page, together with its code number. Starting from the right-hand side, the lines illustrate how the individual industries are joined together to form groups, which themselves merge later with other groups or individual industries. The only aspect of the tree that is relevant for our purposes is the nature of the industry groups, not in the order in which they are formed.

The groupings in Figure 1 are generally intuitively appealing. For example, the top of the diagram shows agricultural production grouped with agricultural services, and the latter group later merging with food manufacturing. Further down the diagram, grouping textile manufacturing with apparel manufacturing seems entirely reasonable, as does grouping printing and paper manufacturing.

3.3. Optimal pruning

The industry tree obtained as described in the previous section provides the discipline for the process of optimal aggregation. The within-node SSEs that appear as part of the loss function (2)–(3) were obtained from weekly wage regressions estimated in logarithmic form over unmatched CPS data for the years 1971–1982. We made the regression sample fairly homogeneous by restricting attention to white males, not enrolled in school (major activity last week), civilians, aged 19–71, who were full-time workers and were not self-employed or working without pay in the previous year.[14] A total of 331,046 persons were included in these regressions. Weekly wages were computed as wage and salary income last year divided by weeks worked last year. The regression specification included years of schooling, years of experience and its square,[15] time, a time–experience interaction, and dummies at 5-year

[13] The reason for basing the industry tree on the two-digit Census classification is the relative prevalence of empty cells with the three-digit Census classification. At the two-digit level, 45·5 per cent of the cells in the cross-tabulation of industry last week (as reported in year t) by industry last year (as reported in year $t+1$) are empty, as opposed to 89·5 per cent at the three-digit level. The inability to order pairs with observed mismatch rates of zero has an important effect on tree construction. The reason is that complete linkage takes the distance between two groups to be the maximum distance between pairs of industries, one from each group. When an industry pair has an observed mismatch rate of zero, any two groups each of which contains one member of that industry pair will be at maximal distance from each other. At this point, tie-breaking procedures are the sole determinant of group formation.

[14] The sample was further restricted to those who were not living in group quarters and who reported weekly wages of more than \$10 and less than \$1700, after deflating (using the GNP implicit price deflator with 1972 as base year) and correcting for truncation due to top coding. Truncation corrections for top coding assumed a gamma distribution for the upper tail of the yearly income distribution. The relevant parameter was estimated from the CPS. We are grateful to Kevin Murphy for supplying the corrected wage and salary data.

[15] More precisely, we used Z and Z^2, where Z is defined as $Z = E - (E^2/60)$, where E is Mincer experience (age minus schooling minus 6). The latter transformations of experience and experience squared appear to provide a better fit for the early career portion of the life cycle wage profile. See Murphy and Welch (1988).

intervals indicating whether the individual was present in the market at the start of that interval.[16]

The outcome of the pruning process is illustrated in Figure 2. When read from right to left the numbers at the bottom and top of the diagram show how many industries or industry groups are left at each step of the aggregation process. For example, the process starts with all 51 of the two-digit industries, and the first three pruning steps are successive mergings of individual industries that reduce the number of industry groups to 48. The place at which lines are merged in the diagram shows which industries or groups are merged at each step. The first pruning step merges mining with petroleum manufacturing to reduce the number of industry groups from 51 to 50, the second merges ordnance with aircraft, reducing the number of industry groups from 50 to 49, and so on.

A missing integer at the bottom of the diagram indicates that more than one merging operation was completed at that stage of the pruning process. Moreover, when multiple merging operations are performed in a single step, the order of merging illustrated by the diagram reflects only the binary nature of the tree. For example, when the scale jumps from 48 to 46, the diagram shows that three industries are merged in one step: lumber, furniture, and miscellaneous durables manufacturing. The fact that lumber and furniture are merged first (that is, before miscellaneous durables are merged in as well) does not indicate that these industries are pruned first, but only that they descend from the same parent node. Similarly, when further along in the pruning process the scale shows a reduction from 15 to 8 industry groups in a single step, the diagrammatic formation of groups within this step reflects only the binary relationships of the industry tree.

Figure 3 illustrates other aspects of the pruning procedure. Panel (a) shows the number of industry groups that result when the trade-off parameter α is varied. When there is no penalty for complexity ($\alpha = 0$), it is optimal to retain the set of all 51 elementary industries. Starting from that point, small increases in α initially cause rapid reductions in the complexity of the optimal classification because only modest increases in residual variance occur at the early stages of the aggregation process. Eventually, however, large increases in α are required to further reduce the number of industry groups, a reflection of the rapid increments in residual variance associated with higher levels of aggregation.

Panel (b) of Figure 3 reveals the same pattern from a different perspective. When going from right to left, we track the increases in the total SSE as aggregation leads to coarser and coarser classifications. The initial stages of aggregation are associated with relatively minor increases in SSE, but eventually aggregation of increasingly dissimilar industry groups causes more substantial increases in SSE. This result is reminiscent of the findings of Welch and Maclennan (1976), who report only modest increases in within-group wage variation when they begin to aggregate across Census occupational groups.

In panel (b) a vertical line is traced at the 15-industry mark. This particular level of aggregation is of interest because we can make comparisons with a standard 15-industry Census classification.[17] As indicated by the triangular symbol, the total SSE for the Census classification is larger than for our 15-industry classification and is roughly the same as the total SSE for our eight-industry classification.

[16] For these purposes we assumed that each individual entered the market at age 25.

[17] The 15-industry Census classification is as follows: agriculture, forestry, and fisheries; mining; construction; durables manufacturing; nondurables manufacturing; transportation; communication; utilities and sanitary services; wholesale and retail trade; finance, insurance, and real estate; business and repair services; personal services; entertainment and recreation services; professional and related services; and public administration.

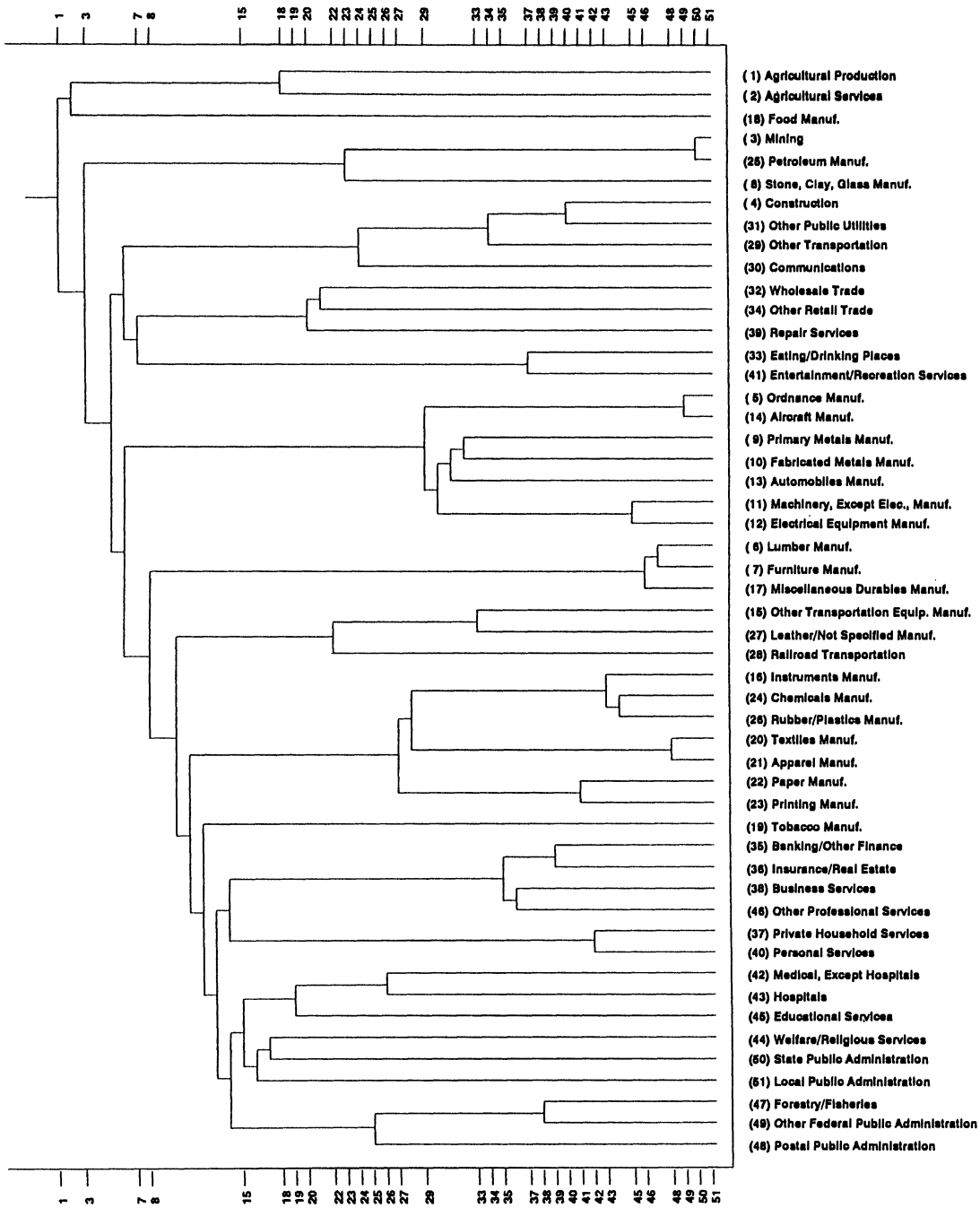

Figure 2. Optimal aggregation scheme using mismatch data

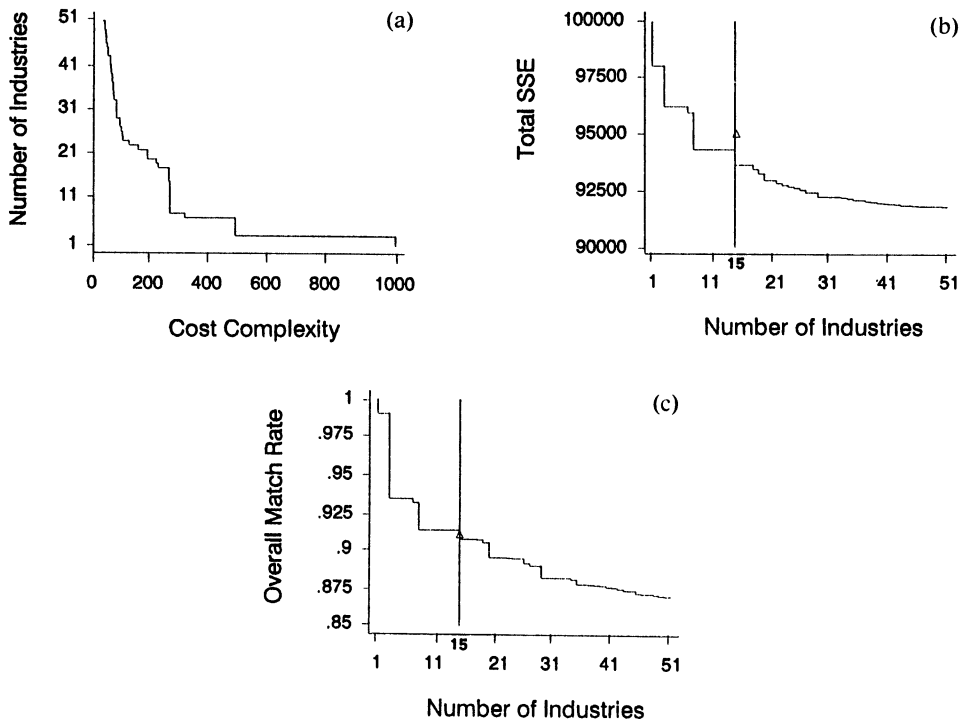

Figure 3

Panel (c) of Figure 3 shows how the level of aggregation influences the overall match rate, which, while not a part of the pruning criteria, is of interest nonetheless. The overall match rate is here defined as the fraction of cases lying along the diagonal of the match matrix $M = [m_{ij}]$ at each level of aggregation. That is, at each level of aggregation we form a new match matrix M and compute the total fraction of the sample that lies on the diagonal. Because the sample size is fixed, the growth in the overall match rate is solely a result of moving off-diagonal mass onto the diagonal as aggregation proceeds. In this light it may be somewhat surprising that the match rate appears to increase most rapidly at the late stages of the aggregation process. One might instead expect to find that the overall match rate rises most rapidly at the initial pruning stages when more closely related industries are being aggregated. The key here is the fact that the closely related pairs of industries that are aggregated initially tend to be small in size, and thus aggregating them in the course of the pruning process adds little to the total mass on the diagonal. Panel (c) also provides a comparison with the overall match rate obtained for the 15-industry Census classification (shown by the triangular symbol). The Census match rate is very slightly lower than the match rate for our 15-industry classification.

4. AGGREGATION USING INDUSTRY TRANSITION DATA

In this section we consider an alternative basis for constructing an industry tree. Rather than using coders' confusion as an indicator of dissimilarity, we instead use workers' transitions between industries. The assumption is that individuals will move more frequently between

industries that are close in the sense that skills generated or used in one are also useful in the other, than between industries that utilize very different skills.

To measure transitions between industries, unmatched CPS files for the years 1971–1982 were used to cross-tabulate industry last year by industry last week at the two-digit Census level[18] for those who worked both years.[19] Denoting by q_{ij} the (i, j)th element of the empirical transition matrix, transition odds ratios and dissimilarity between industries were defined using (5) and (6) respectively, with the m_{ij}'s now replaced by the q_{ij}'s.[20]

Table II lists in descending order the 25 industries with the highest measured transition odds-ratios. The lists in Tables I and II are very different. Table II is dominated by a few industries—eating/drinking places and construction, for example—that tend to be heavy users of low-skilled, entry-level personnel. The suggestion is that transitions are founded on inter-industry differences in occupational structure; that is, mobility is occupation rather than industry driven.[21]

Table II. The 25 largest mismatch odds ratios in descending order

Industry 1	Industry 2	Odds ratio	Number in industry 1 both times	Number in industry 2 both times
33 Eating/drinking places	40 Personal services	0·0148	6,772	3,940
33 Eating/drinking places	41 Entertainment/recreation	0·0123	6,772	2,459
42 Medical, except hospitals	43 Hospitals	0·0121	3,434	4,087
1 Agricultural production	4 Construction	0·0095	11,059	27,049
4 Construction	34 Other retail	0·0092	27,049	58,418
40 Personal services	41 Entertainment/recreation	0·0083	3,940	2,459
33 Eating/drinking places	34 Other retail	0·0080	6,772	58,418
4 Construction	33 Eating/drinking places	0·0080	27,049	6,772
32 Wholesale	34 Other retail	0·0071	12,602	58,418
4 Construction	32 Wholesale	0·0069	27,049	12,602
33 Eating/drinking places	38 Business	0·0067	6,772	4,974
3 Mining	4 Construction	0·0065	4,593	27,049
4 Construction	38 Business	0·0065	27,049	4,974
33 Eating/drinking places	45 Education	0·0065	6,772	13,373
4 Construction	39 Repair	0·0063	27,049	5,437
34 Other retail	39 Repair	0·0061	58,418	5,437
1 Agricultural production	2 Agricultural services	0·0061	11,059	1,475
38 Business	40 Personal services	0·0059	4,974	3,940
29 Other transportation	39 Repair	0·0059	10,693	5,437
35 Banking/finance	36 Insurance/real estate	0·0058	3,625	6,841
29 Other transportation	32 Wholesale	0·0058	10,693	12,602
32 Wholesale	38 Business	0·0057	12,602	4,974
1 Agricultural production	18 Food	0·0057	11,059	6,089
18 Food	33 Eating/drinking places	0·0056	6,089	6,772

[18] With the transition data there was again an empty cell problem that was more severe at the three-digit level. While 20·9 per cent of the cells were empty at the two-digit level, 68·7 per cent were empty at the three-digit level.

[19] The sample was restricted to 404,727 males age 17–65.

[20] Notice that systematic coding errors may create problems in using such measure for industry grouping, for what appear to be transitions may merely reflect coding errors. The opposite problem may have occurred in using mismatches data. Some apparent mismatches may have been actual transitions between industries, accompanied by a mistake in recording the number of industries of employment last year.

[21] We thank one of the referees for raising this point.

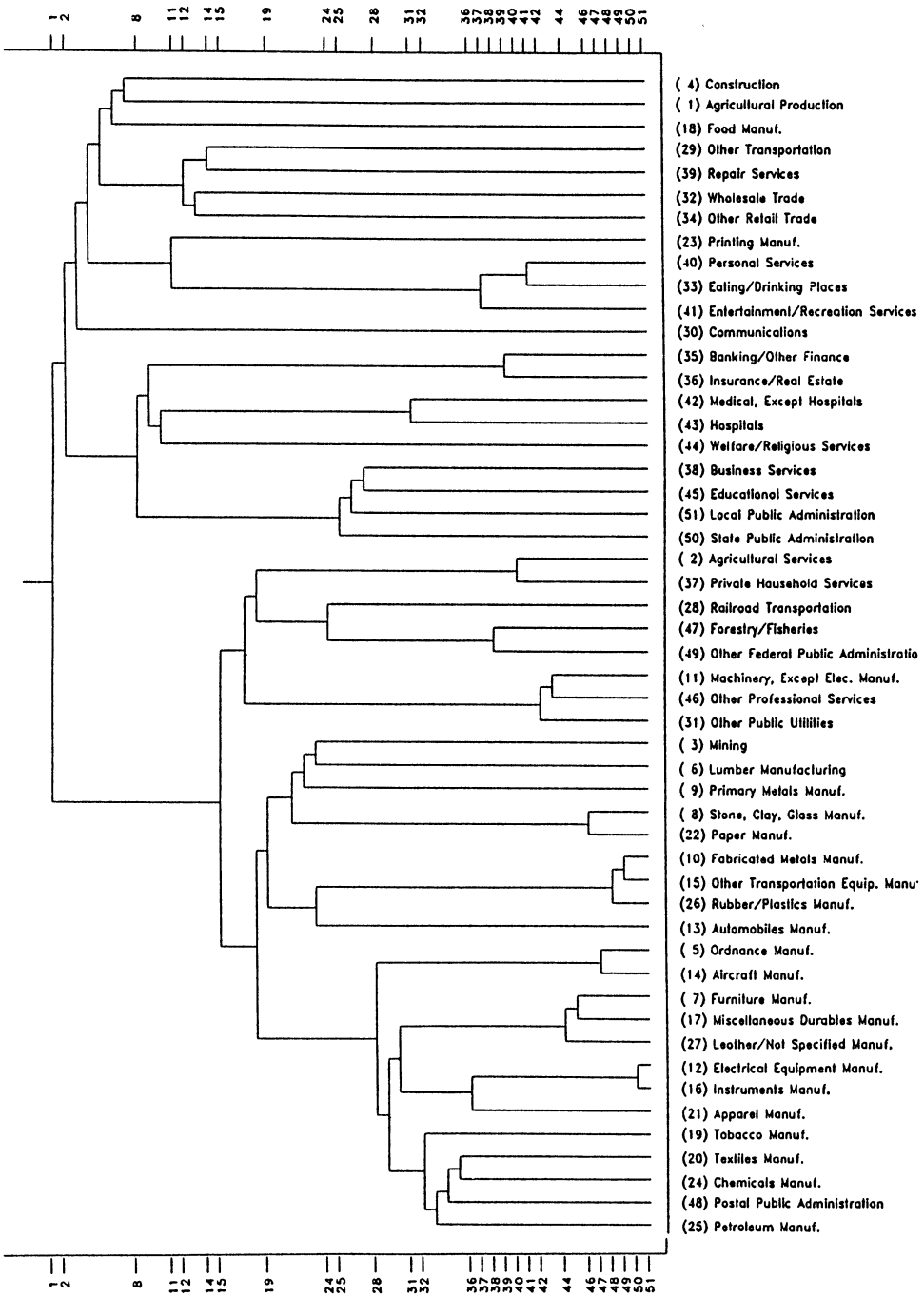

Figure 4. Optimal aggregation scheme using transition data

Measuring dissimilarity with transition rates and using the complete linkage method yields the industry classification tree displayed in Figure 4. This figure serves the dual purpose of illustrating the nature of the industry tree and the process of optimal pruning. As in Figure 1, the joining of lines indicates the formation of groups. Although most groupings in Figure 4 make sense, they appear less appealing than those in the corresponding tree in Figure 1. As in Figure 2, the scales on the top and bottom of Figure 4 show the progress of the pruning process that occurs diagrammatically when industry lines merge.

Figure 5 provides for the transition-based aggregation scheme analogous information to that provided in Figure 3. Since the qualitative results are similar, we need only summarize the main findings. Panel (a) shows that, beginning with a full array of 51 disaggregated industries, relatively small increases in the trade-off parameter α initially generate substantial reductions in the number of distinct industry groups, but eventually relatively large increases in the importance of parsimony are required to effect aggregation. Correspondingly, panel (b) shows that the aggregation process initially increases total SSE only modestly, but more sizeable increases in SSE are occasioned by the later stages of the pruning process. The triangular symbol on the vertical line in panel (b) again illustrates the larger total SSE associated with the more standard Census categorization at the 15-industry level. Panel (c) shows that aggregation initially produces only minor increases in the frequency of staying in the same industry (measured as the fraction of all sample members whose longest industry last year [at a particular level of aggregation] is the same as the industry last week [again at that level of aggregation]), but increases in the frequency of staying become sizeable toward the end of the aggregation process. The triangular symbol on the vertical line in panel (c) is again the corresponding frequency for the 15-industry Census classification. The value is virtually identical to that obtained from the 15-industry classification constructed from transition data.

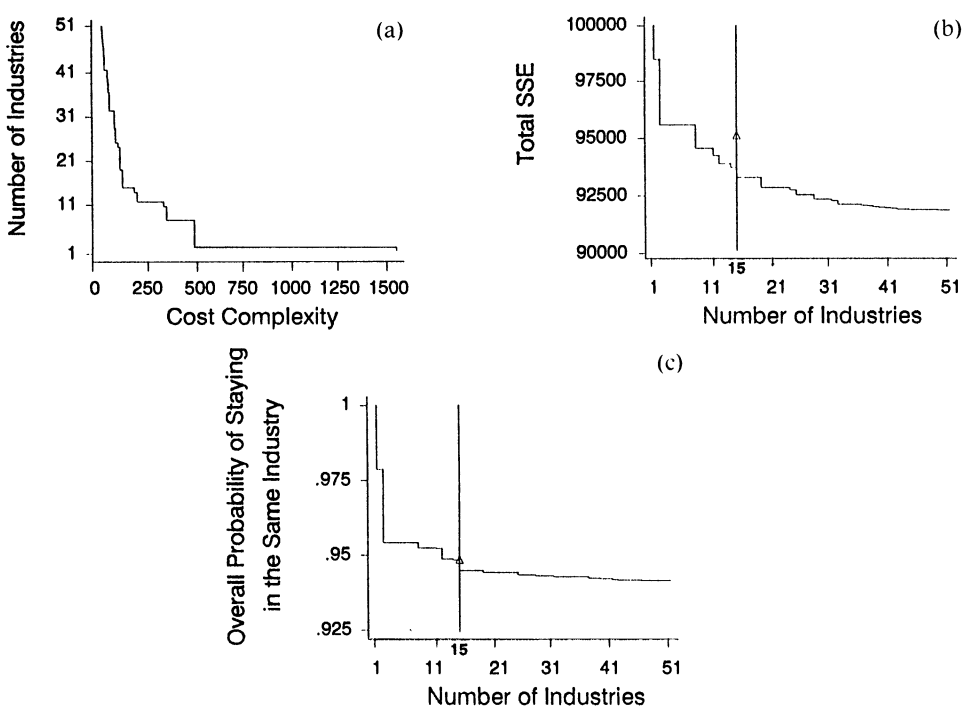

Figure 5.

Table 3. Fifteen industry aggregates using alternative classification schemes

	Census	Mismatch data	Transition data
Agriculture, forestry and fish	Agricultural production Agricultural services Forestry and fisheries	Agricultural production Agricultural services	Agricultural production
Mining	Mining	Mining Stone, clay, glass manufacture Petroleum manufacture	
Construction	Construction		Construction
Durables manufacture	Ordnance Lumber Furniture Stone, clay, glass Primary metals Fabricated metals Machinery, excluding electrical Electrical equipment Automobiles Aircraft Other transport equipment Instruments Miscellaneous manufacturing	Ordnance Primary metals Fabricated metals Machinery, excluding electrical Electrical equipment Automobiles Aircraft Lumber Furniture Miscellaneous manufacturing	Agricultural services Forestry and fisheries Mining Ordnance Lumber Furniture Stone, clay, glass Primary metals Fabricated metals Machinery, excluding electrical Electrical equipment Automobiles Aircraft Instruments Miscellaneous Tobacco Textiles Apparel Paper Chemicals Petroleum Rubber and plastics Leather and not specified manufacturing Railroad and railway Other transport services Other public utilities Private household services Other transport equipment

(Continued)

Table 3. (*continued*)

	Census	Mismatch data	Transition data
Nondurables manufacture	Food	Food	Food
	Tobacco		
	Textiles	Tobacco	
	Apparel		
	Paper	Instruments	
	Printing	Textiles	
	Chemicals	Apparel	
	Petroleum	Paper	
	Rubber and plastics	Printing	Printing
	Leather and not specified manufacturing	Chemicals	
		Rubber and plastics	
Transportation	Railroads and railway	Other transport equipment	
	Other transportation	Leather and not specified manufacturing	
		Railroads and railway	
Communications	Communications	Construction communications	
		Other transport services	
		Communications	
		Other public utilities	
Utilities	Other public utilities		
Wholesale/retail	Wholesale	Wholesale	Wholesale
	Eating/drinking places	Other retail	
	Other retail	Repair	
			Other retail

Finance, insurance	Banking/other finance Insurance/real estate	Banking/other finance Insurance/real estate Business services Other professional services	Banking/other finance Insurance/real estate
Business	Business services Repair services		Repair services
Personal services	Private household services Personal services, excluding private household	Private household services Personal services, excluding private household	Eating/drinking places Personal services, excluding private household Entertainment/recreation services
Entertainment	Entertainment/recreation services	Eating/drinking places Entertainment/recreation services	
Professional services	Medical, excluding hospitals Hospitals Welfare/religious Educational Other professional	Medical, excluding hospitals Hospitals Welfare/religious Educational State public administration Local public administration Forestry and fisheries Postal Other federal public administration	Medical, excluding hospitals Hospitals Welfare/religious
Public administration	Postal Other federal public administration State public administration Local public administration		Business Educational State public administration Local public administration

5. CONCLUSIONS

The tree construction and pruning procedures explored in this paper produce industry aggregates that differ from each other and from those proposed by the Census. Table III illustrates the differences between the various aggregation schemes at the 15-industry level. The 'Census' column describes the composition of each of the Census groups. Under the columns 'Mismatch data' and 'Transition data' we attempt to compare each aggregate of our two 15-industry classifications with the Census category that seems most closely related. Given the differences among the various aggregation schemes, this effort is of limited success. When more than one aggregate seem to fit in the same Census category, they are all listed next to that particular Census category and spaces are used to delimit the individual groups. For example, the classification based on mismatch data yields three industry groups that seem most appropriately viewed as nondurables manufacturing: the individual industries food and tobacco, and the group composed of instruments, textiles, apparel, paper, printing, chemicals, and rubber and plastics.

The emphasis of this paper, however, is not on producing a new aggregation scheme but rather on methods for deriving such an aggregation scheme. The outcome of much applied work may hinge on the aggregates employed, and thus procedure for classification and aggregation are legitimate and important subjects for inquiry.

Our method is an attempt to formalize two aspects that are often neglected by standard aggregation procedures. First, one has generally available at least some information on the degree of dissimilarity between the elements that are to be aggregated. Second, choosing a level of aggregation involves a trade-off between goodness-of-fit and parsimony. In this paper, summaries of observed behaviour, namely coders' confusion and workers' transitions, are first used to construct measures of dissimilarity between industries. This eliminates the need of arbitrarily specifying a list of industry characteristics and a set of weights assigned to them. The prior information on the patterns of dissimilarity between industries is then exploited to construct an industry tree that constrains the set of admissible classifications. Finally, a sequence of optimal classifications is produced by minimizing a loss function that trades off goodness-of-fit and parsimony. By forcing the aggregation process to obey the relationships embodied in the industry tree, no inconsistency in groupings can arise: coarser aggregates are always formed by merging groups of finer aggregates.

ACKNOWLEDGEMENTS

We thank Finis Welch and two anonymous referees for their helpful comments. This research was supported by the US Department of Health and Human Services, National Institute on Aging (Grant Number R01 AG07112). Opinions and conclusions contained herein are solely those of the authors and should not be construed as representing opinions or policy of any agency of the Federal Government.

REFERENCES

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

Grunfeld, Y., and Z. Griliches (1960). 'Is aggregation necessarily bad?', *Review of Economics and Statistics*, **42**, 1–13.

Leamer, E. E. (1990). 'Optimal aggregation of linear net export systems', in T. Barker, and M. H. Pesaran (eds), *Disaggregation in Econometric Modelling*, Routledge, London, pp. 150–170.

Lebart, L., A. Morineau, and K. M. Warwick (1984). *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*, Wiley, New York.

Murphy, K., and F. Welch (1988). 'Empirical age–earnings profiles', Los Unicon Research Corporation, Los Angeles (mimeo).

Pitts, A. (1988). 'Matching adjacent years of the current population survey', Unicon Research Corporation, Los Angeles (mimeo).

Welch, F., and I. Maclennan (1976). 'The U.S, Census occupational taxonomy: how much information does it contain?', R-1849-HEW, Rand Corporation.

Zupan, J. (1982). *Clustering of Large Data Sets*, Research Studies Press, New York.