Strategic Management Journal, Vol. 13, 511–524 (1992)

DIVERSIFICATION, VERTICAL INTEGRATION, AND INDUSTRY ANALYSIS: NEW PERSPECTIVES AND MEASUREMENT

RACHEL DAVIS
Leonard N. Stern School of Business, New York University, New York, New York, U.S.A.

IRENE M. DUHAIME
Fogelman College of Business and Economics, Memphis State University, Memphis, Tennessee, U.S.A.

Vertical integration, diversification, and industry analysis are fundamental topics in strategic management content research. We develop the conceptualization of these issues by exploring their nature as well as their correct measurement. Toward these ends, we conduct an extensive analysis of the COMPUSTAT II data base and the TRINET data base in all three research contexts. In addition to these two data bases, we also evaluate the Census of Manufacturers, S&P's Financial Dynamics, *S&P's* Industry Surveys *and Dun and Bradstreet's* Industry Norms and Key Business Ratios, *for purposes of industry analysis. Important contributions include our identification of the potential of COMPUSTAT II data to distinguish within-stage forward and backward vertical integration, and between-stage forward and backward vertical integration, as well as our recommendations for the protection of the integrity of studies based on the COMPUSTAT II and TRINET data bases.*

Strategic management research on multibusiness firms has long been hampered by the absence of suitable firm-specific data disaggregated to the business level. The COMPUSTAT II Industry Segment data base, disaggregated at the broader segment level rather than the more desirable business level, has been used by strategic management researchers to study multibusiness firms. This data base provides disaggregated annual (1978 onward) data on firms' multibusiness activities. A new and relatively untried source of establishment level data is the TRINET data base. The data base covers alternate (odd) years between 1981 and 1989.

Attractive and useful as publicly-available disaggregated data might seem to researchers, careful attention to the data's characteristics, understanding of its composition, and thus compensation for its limitations are necessary to protect the integrity of research using such data sets and the value of such research results. This paper evaluates the potential and pitfalls which COMPUSTAT II, TRINET and other data bases hold for strategic management research. First COMPUSTAT II and TRINET are described and discussed. Then, relationships between businesses in a segment, among segments and establishments in a firm, and across segments and establishments in the data bases are explored in order to evaluate the data bases' potential for studying vertical integration, diversification strategy and industry trends. In discussing each research issue, attention is also given to other data sources relevant to that issue. Certain

Key words: Diversification, vertical integration, industry analysis, data bases, measurement

0143–2095/92/080511–14$15.00

Received 25 January 1990
Final revision received 28 April 1992

common pitfalls which researchers should avoid are discussed and illustrated, and appropriate methods for correct use of the data are explained.

DATA FOR VERTICAL INTEGRATION, RELATEDNESS AND INDUSTRY ANALYSIS

The COMPUSTAT II Industry Segment data base is compiled from firms' annual reports and 10-K reports to the Securities and Exchange Commission (SEC). Disclosure of the financial information in this data base is required by FASB-SFAS No. 14 'Financial Reporting for Segments of a Business Enterprise.' Data items relevant for strategic management research include sales, operating income, identifiable assets, capital expenditures, and employees, for each segment of each firm.

A clear understanding of the scope of the term 'segment' as defined by the FASB is critical for the correct usage of the COMPUSTAT II line-of-business data base in strategic management research. The FASB defines a segment as:

> A component of an enterprise engaged in providing a product or service or a group of related products and services primarily to unaffiliated customers (i.e., customers outside the enterprise) for a profit (FASB 14, paragraph 10 a. and COMPUSTAT II, section 2, pg. 2).

FASB 14, paragraph 10 a. notes that:

> By defining an industry segment in terms of products and services that are sold primarily to unaffiliated customers, this Statement does not require the disaggregation of the vertically integrated operations of an enterprise.

Thus, by definition, if two businesses are vertically integrated they should be assigned to the same segment. However, a segment can consist of a single business or any *group or sub-group* of related businesses, therefore all businesses related at the 2-digit level need not be grouped in the same segment.

Paragraph 11 of FASB 14 clearly specifies the guidelines for determining segments, and states:

> The reportable segments of an enterprise shall be determined by (a) identifying the individual products and services from which the enterprise derives its revenue, (b) grouping those products and services by industry lines into industry segments, and (c) selecting those industry segments that are significant with respect to the enterprise as a whole.

These FASB guidelines are illustrated in Figure 1. From the two-by-two matrix in Figure 1, it is evident that either vertical integration *OR* relatedness are necessary conditions for assigning two businesses to a single segment; vertical integration is also a sufficient condition for assigning two businesses to a single segment, but relatedness is not. (Two businesses may be related on the basis of 2-digit SIC codes, but may be assigned to different segments for strategic reasons.)

Since the institution of these segment reporting requirements in 1978, Standard and Poor's (henceforth referred to as S&P) COMPUSTAT II Service has been compiling the segment information of more than 6,000 public companies traded on the NYSE, ASE, NASDAQ and OTC. Since FASB-SFAS 14 requires only that each company identify each of its segments *by name not by SIC*, S&P personnel assign a maximum of two 4-digit SICs to each segment (SSIC1 and SSIC2), in order to achieve comparable descriptions across segments. This further disaggregation of the data (identification of businesses *within* the FASB-required segments) has been viewed favorably by researchers interested in business-level strategic issues. However, herein lies the potential for misuse and abuse of this data base, which may lead to erroneous research results.

In noting that S&P's COMPUSTAT II Service, not the reporting companies, assigns the segment SICs (SSICs), the COMPUSTAT II Manual explains that:

> SSICs are based on the activities of the segments as described by the company in its annual report or 10-K. The first SIC should be considered the primary SIC of the segment. . .Standard and Poor's COMPUSTAT II Service will attempt to assign two SIC codes, for each industry segment (*COMPUSTAT II* Section 5-A, p. 26).

Clearly, two issues must be addressed to ensure researcher confidence in the value of this data: (1) the accuracy and consistency of firms' segment definitions and reporting practices, and (2) the

RELATIONSHIP TO CUSTOMER

RELATED AT 2-DIGIT SIC LEVEL	AFFILIATED (Vert. Integrated)	UNAFFILIATED (NOT Vert. Integrated)
YES	Vert. Integrated and Related Businesses SAME SEGMENT	Related Businesses SAME OR DIFFERENT SEGMENTS
NO	Vert. Integrated SAME SEGMENT	Neither Vert. Integrated Nor Related Businesses DIFFERENT SEGMENTS

Figure 1. Assignment of businesses to segments

accuracy and consistency of S&P's SIC coding procedures.

Assuming that companies in general do not violate FASB requirements and rules, we can surmise that most companies attempt to define segments using the above guidelines. Extensive discussions were conducted with COMPUSTAT II personnel to assess the manner in which firms defined their segments. The general consensus among the staff was that most firms exercised care and precision in their definition and identification of segments. However, in 5 to 10 per cent of cases, businesses which were neither related nor vertically integrated, from the perspective of COMPUSTAT II personnel, were segmented together by the firms. Segments identified as 'consumer products,' 'industrial products' and 'other' are the most likely to be catch-all segments; these segments bear close examination when used in research.

Knowledge of the basis on which S&P assigns SIC codes, and of the reliability of their code assignment procedure, is also important to researcher assessment of the value of this data. Interviews with S&P personnel who assign SSICs have provided evidence that SSIC code assignments are carefully executed. Segment SICs, SSIC1 (primary) and SSIC2 (secondary) are assigned based on each segment's primary products as identified in sales break-out data reported in the 10-Ks (about 45% of COMPUSTAT II companies report this data); if this data is not available, the ordering of products in 10-Ks is used to identify primary and secondary products. (COMPUSTAT II has canvassed companies to verify that the order of segment products reported in 10-Ks indeed reflects descending importance of the products to their companies.) In each succeeding year, the validity of SSIC1 as the primary activity, and SSIC2 as the secondary activity, is checked and updated by S&P personnel. Evidence of such updating is found in the many instances where the SSICs assigned to a segment by S&P changed from 1 year to the next, while the segment names reported by the company remained unchanged during the same period.

COMPUSTAT II has a staff of 14 people who work full time identifying and assigning segment SICs. SIC assignments are randomly rechecked by supervisors. Most of the staff have been with the division for over 2 years; the senior-most person has been in the division for 12 years. Staff members report that difficult SSIC assignments often lead to hours of discussion between staff members; there have been instances when an individual SSIC assignment took days of work and research. The accuracy and reliability of COMPUSTAT II's segment SIC classification is affirmed by the fact that the Securities and Exchange Commission (SEC) frequently consults with COMPUSTAT II on issues regarding company and/or segment industry affiliation.

As discussed above and described in Figure 1, when two SSIC codes within a segment *do not* belong to the same 2-digit SIC classification, those businesses could be either related or vertically integrated. Hence, it is critical that we

be able to identify and understand the nature of the relationship between the two SICs of a given segment. Some previous researchers have treated the businesses within segments as independent and separable, allowing the presence of different segment SIC codes to override the fact that the two businesses are reported by the firm as belonging to the same segment. For most questions of importance to strategic management, such as extent and type of diversification, or extent and type of vertical integration, researcher separation of SSICs would be a serious error. Thus, the second SIC code (SSIC2) may be assumed to denote an activity related to the manufacture or service of the activity in the primary SIC (SSIC1), indicating either relatedness or vertical integration within that segment of the end product.

The TRINET data base is yet another source of disaggregated data on multibusiness firms. Unlike the COMPUSTAT II segment data, the TRINET data base is not constructed from the 10-K reports, rather it is gathered from information maintained by each U.S. county on 'establishments' functioning within their boundaries. Although the TRINET data base was compiled for use as a market information source for identification of potential industrial customers, it has significant potential for business research as well.

The TRINET data base includes both public and private firms. The data base contains information on all establishments employing more than 20 people. Establishments are physically distinct locations of firm activity; an establishment could be a warehouse, a janitorial and maintenance unit, a plant, or any other physical location of company activity. Closer examination reveals that a number of these establishments are merely support services. For instance, Dow Chemicals has 143 establishments listed on TRINET, including special warehousing (SIC 4226) and cleaning and maintenance services (SIC 7349). Other establishments listed on TRINET for the *Fortune* 500 firms include such activities as pension and health insurance (SIC 6371) and educational and religious trusts (SIC 6732). Clearly, such establishments are not 'businesses' for strategic management research purposes, yet they are assigned a share of firms' sales volume on TRINET. Another problem is that some establishment activities are strategically ambiguous: activities such as computer facilities management (SIC 7376) or data processing and preparation (SIC 7374) would be administrative/support activities in most firms, but they would be strategic activities for firms involved in the computer industry. This ambiguity would not present a problem if the primary business of the firm in question is clearly computer-related; however, when this is not the case the researcher has no way of judging if these computer-related activities are strategic or nonstrategic. Unfortunately, TRINET does not provide us with any firm-reported information as to how the various establishments comprising a firm are vertically integrated or related in any other manner.

TRINET data items relevant for strategic management research include the number of employees, volume of sales, market share, and SIC codes associated with each establishment. (If an establishment has more than one SIC ID, data for each establishment SIC is reported separately.) However, sales and market share data are not actually reported by each establishment, rather they are estimated by multiplying the number of employees (reported) by the shipments per employee (figures from industry census reports); these estimates are then reconciled to match the overall company sales reported in the 10-K. Thus, TRINET data are problematic when the issues being researched require accurate sales data which reflect individual differences between companies, both within and across industries. The TRINET sales and market share data tend to mask differences between companies, as they only reflect average employee productivity within each industry, and not differences across firms.

The most accurate, and therefore promising, TRINET data items are those which are reported rather than estimated; these are the establishment SIC identities and the number of employees at each establishment. Therefore, in research questions for which number of employees is a reasonable and relevant proxy for establishment size, the TRINET data base can be effectively used. Researchers should note, however, that there is some doubt about future availability of his data: (1) 1989 may be the last year published, and (2) while 1981, 1983, 1985 and 1987 are also currently available, in the future it is likely that only subsets of the current year's data may be available, and only to DIALOG subscribers.

Following this introduction of the COMPUSTAT II and TRINET data bases, we now examine their use in three research contexts: vertical integration, diversification strategy, and industry trends.

VERTICAL INTEGRATION RESEARCH

For many vertical integration questions of interest to strategic management researchers, business or at least segment level data are necessary. Before FASB 14 was enacted, researchers were forced to make subjective judgements about the existence of vertical integration relationships in firms and to allocate firm level data accordingly. In the COMPUSTAT II line-of-business data base, the dual SSICs assigned to each segment provide information on the vertical integrative relationship; there is no data equivalent to the dual SSIC data item in data bases such as TRINET. If research using TRINET identifies vertically integrative relationships, those results reflect researchers' judgements, which might not necessarily represent the actual strategy enacted by firms.

With the guidelines imposed by FASB 14, firms are required to provide some information about their actual vertical integration relationships. Although this data is available on COMPUSTAT II, its effective usage for vertical integration research is not immediately obvious. Understanding the relationship between businesses within each of firms' segments is critical not only to vertical integration but also to the other research issues addressed in this paper; therefore, we conducted an analysis of COMPUSTAT II's industry segment data base to explore the nature of those relationships. The procedures and results of that analysis are described next.

In our analysis, we compared SSIC1 and SSIC2 for all segments in the data base for the years 1984–90 (availability of COMPUSTAT II data begins with 1978). As shown in Table 1-A, 29.7 percent of the segments had only a primary SIC (SSIC2 was 0—COMPUSTAT II assigns a *maximum* of two SIC codes per segment); these may be described as single-business segments. In segments for which both the primary and secondary businesses (SSIC1 and SSIC2) were in the same 2-digit SIC, relatedness can reasonably be assumed (Jacquemin and Berry, 1979; Montgomery, 1982); this group accounted for 28.8 percent of all segments, as shown in Table 1-B.

We then examined the remaining 41.5 percent of segments. In these segments, primary and secondary businesses did *not* fall in the same 2-digit SIC. Erroneous classification of a segment's businesses as unrelated seemed most likely to occur within this (fairly large) group. In computing the entropy index using COMPUSTAT II data, Palepu (1985) defined a segment's SSICs that were dissimilar at the 2-digit level as representing unrelated businesses. However, the concerned firms had explicitly identified those businesses as being related by assigning them to the same segment. S&P personnel confirmed that breaking up company-defined segments on the basis of differences in segment SIC codes (SSICs) would be inappropriate. In order to prevent further misuse of the COMPUSTAT II data, we considered it imperative to fully explore and understand the types of relatedness which were typically represented by SSIC1 and SSIC2 in a segment.

The FASB definition specifies that segments should be formed such that they provide 'products or services to *unaffiliated* customers;' therefore, any vertical integration (products and services provided to *affiliated* customers) should be aggregated within the segment of the end product for purposes of reporting and disclosure. In light of this stipulation we examined the 41.5 percent of segments whose SSICs were dissimilar at the 2-digit level, for vertical integration.

Between-stage vertical integration—forward and backward

Vertical integration in a segment may be of two types: between stages or within stages of the value-added chain. Using COMPUSTAT II, it is possible to identify whether a business is forward or backward integrated by identifying the upstream or downstream location of SSIC2 (the secondary business of the segment) relative to SSIC1 (the primary business of the segment).

Between-stage vertical integration occurs between stages in the value-added chain, such as between mining and manufacturing, or manufacturing and distribution. A segment might include, for example, iron ore extraction (SSIC1 1010) and iron and steel foundries (SSIC2 3320), or iron ore extraction (SSIC1 1010) and cold

Table 1. Relationship between segments' SICs by year

Segment SIC relationships		1984	1985	1986	1987	1988	1989	1990	Total
A. Single business segments									
SSIC2 EQ 0	NO.	2133	2396	2634	2810	2926	3052	2319	18270
	%	28.7	29.5	29.8	30.2	29.9	30.3	28.9	29.7
B. Related at 2-digit SIC									
SSIC1 EQ SSIC2	NO.	2144	2335	2530	2679	2827	2854	2375	17744
	%	28.8	28.7	28.6	28.7	28.9	28.3	29.6	28.8
C. Vertical integration									
i. Between stage—Forward integration									
raw material	NO.	38	40	43	42	43	45	38	
W/ manufacturing	%	0.5	0.5	0.5	0.5	0.4	0.4	0.5	
raw material	NO.	221	227	240	251	259	247	190	
W/ service	%	3.0	2.8	2.7	2.7	2.6	2.5	2.4	
manufacturing	NO.	540	595	628	649	673	700	579	
W/ service	%	7.3	7.3	7.1	7.0	6.9	6.9	7.2	
Total	NO.	799	862	911	942	975	992	807	6288
	%	10.8	10.6	10.3	10.2	9.9	9.8	10.1	10.3
ii. Between stage—Backward integration									
manufacturing	NO.	82	85	81	85	87	98	79	
W/ raw material	%	1.1	1.0	0.9	0.9	0.9	1.0	1.0	
service	NO.	106	105	108	118	133	136	110	
W/ raw material	%	1.4	1.3	1.2	1.3	1.4	1.3	1.4	
service	NO.	190	212	236	276	273	296	220	
W/ manufacturing	%	2.6	2.6	2.7	3.0	2.8	2.9	2.7	
Total	NO.	378	402	425	479	493	530	409	3116
	%	5.1	4.9	4.8	5.2	5.1	5.2	5.1	5.1
iii. Within stage—Bwd. & Fwd. Integration									
raw material	NO.	58	53	60	55	68	72	44	
W/ raw material	%	0.8	0.7	0.7	0.6	0.7	0.7	0.5	
manufacturing	NO.	1017	1064	1130	1132	1147	1144	9161	
W/ manufacturing	%	13.7	13.1	12.8	12.1	11.7	11.3	11.4	
service	NO.	904	1018	1143	1222	1346	1437	1167	
W/ service	%	12.2	12.5	12.9	13.1	13.8	14.3	14.5	
Total	NO.	1979	2135	2333	2409	2561	2653	2127	16197
	%	26.7	26.3	26.4	25.8	26.2	26.3	26.4	26.4
Grand total	NO.	7433	8130	8833	9319	9782	10081	8037	61615
	%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Data Source: COMPUSTAT II

rolled steel (SSIC2 3313). (Each of the above combinations represents a different type of forward integration, though both combinations start from iron ore extraction.) This type of vertical integration is relatively simple to discern and use in the COMPUSTAT II data base. When one of the SSICs belongs to one of the value-added stages (raw materials (SIC 0100–1999), manufacturing (SIC 2000–3999), or service (SIC 4000–9999)), while the other SSIC belongs to one of the other two stages, it may be assumed that the two businesses' presence in the same segment is an indication of vertical integration. For instance, a segment having SSIC1-2020 and

SSIC2-5143 is forward integrated, because SIC-2020, the primary SIC, is the manufacture of dairy products and SIC-5143, the secondary SIC, is the wholesale of dairy products.

In order to test the validity of the above assumption we examined a random sample of 75 segments, whose SSICs fell in different value-added stages, for the presence of between-stage vertical integration. We examined the SIC code definitions of the two businesses within each segment. Only six segments (7%) of the randomly selected 75 were not found to be vertically integrated. Using the normal approximation of the binomial distribution (Siegel, 1988), we can infer that at least 93 percent ($p < 0.001$) of these segments are between-stage vertically integrated.

Our analysis, summarized in Table 1 (C-i. and C-ii.), indicates that for 15.4 percent of the segments in question, the primary and secondary businesses were related by this first type of vertical integration. Forward integration was identified in those instances (10.3%) where SSIC1 was in raw materials and SSIC2 was in manufacturing or service, or SSIC1 was in manufacturing and SSIC2 was in service. Similarly, backward integration was identified in those instances (5.1%) where SSIC1 was in service, and SSIC2 was in manufacturing or raw materials, or SSIC1 was in manufacturing and SSIC2 was in raw materials.

So far, Table 1 has shown that in the cases of multibusiness segments, the relationship between SSIC1 and SSIC2 could be one of relatedness (28.8 percent) or one of between-stages vertical integration (e.g. mining and manufacturing or manufacturing and distributing) 15.4 percent (10.3% forward integrated and 5.1% backward integrated); 26.4 percent of segments have yet to be classified, and will be addressed next.

Within-stage vertical integration—forward and backward

Based on FASB's segment definition, we expected that the unclassified 26.4 percent of segments would exhibit within-stage vertical integration (vertical integration within a single stage of the value-added chain). For example, a segment might include activities such that manufacturing output from one 2-digit SIC (e.g. 2,200, textile mill products) becomes input for manufacturing in a different 2-digit SIC (e.g. 2330, women's apparel). Thus, segment SICs would tend to portray within-stage (e.g. manufacturing-stage-based) vertical integration. Similarly, most auto and aircraft companies have in-house production of most of the component parts necessary for their end-products. A segment with SSIC1 3721 and SSIC2 3664 appears to consist of unrelated manufacturing activity, but SIC-3721 is aircraft manufacturing and SIC-3664 is the manufacture of search, detection, navigation and guidance systems and equipment. Activity in SIC-3664 provides critical instrumentation used in all types of aircraft. This type of vertical integration is to be expected in the manufacture of complex products and services. However, establishing the presence of this second type of vertical integration is more complex for researchers. As part of our study, we examined the remaining 26.4 percent of segments (those for which SSIC1 and SSIC2 were dissimilar at the 2-digit SIC level and were not in different industry stages), for which both the primary and secondary businesses were in raw materials, manufacturing or service.

We randomly selected 75 of the yet unclassified segments from the data base and examined the relationship between SSIC1 and SSIC2 of those segments for the presence of vertical integration. Only 11 segments (15%) of the random sample of 75 were not found to be vertically integrated. Using the normal approximation of the binomial distribution (Siegel, 1988), we can infer that at least 85 percent ($p < 0.001$) of these segments are within-stage vertically integrated. The above analysis reconfirms that activity reported by firms as being associated with a single product or group of related products is indeed so despite the fact that the varied SSICs contained within segments can give the appearance of unrelated businesses.

To recapitulate the findings from Table 1, from an analysis of all 6,007 firms on the COMPUSTAT II industry segment data base: 29.7 percent of segments were single business; 28.8 percent of segments had SSIC1 and SSIC2 in the same 2-digit code suggesting relatedness; at least 93 percent ($p < 0.001$) of 15.4 percent of the segments were between-stage vertically integrated (10.3% forward integrated, 5.1% backward integrated); and at least 85 percent ($p < 0.001$) of the remaining 26.4 percent of segments were either backward or forward within-stage vertically integrated. These findings are in accordance with the definition of segment under

FASB-14 (above). Of equal, if not greater, importance to researchers studying vertical integration is the fact that, by definition, integration activity cannot take place *between* segments, as this would not fulfill the requirement that segments' sales be to unaffiliated customers.

Research on vertical integration that raises questions about the type or extent of vertical integration both within and across industries can thus make use of COMPUSTAT II segment data very effectively. Some of the possibilities for use of the data base are suggested by the between- and within-stage forward and backward vertical integration that we were able to identify (Table 1).

Limitations for use in vertical integration research

Two limitations are associated with using COMPUSTAT II data to study vertical integration. First, each segment has only two SIC codes, therefore only one link, or two levels, of the value-added chain can be identified, even if other levels are actually present in a segment. COMPUSTAT II unmasks data on *some* vertical integration within firms (the two levels of the value-added chain in which the firm's activities are most significant), but does not provide data on *all* stages of firms' participation in the value-added chain, thus the extent of vertical integration that can be addressed with COMPUSTAT II is constrained.

A second limitation is that evaluation of the presence and direction of vertical integration activity for the 26.4 percent of segments identified as 'Within Stage—Bwd. & Fwd. Integration' in Table 1 is a subjective process; without more detailed information from the firms themselves, researchers can only assert that vertical integration activities are *probably* present in those firms.

As noted above, TRINET data has little value for research on vertical integration, because *presence* of vertical integration is strictly a researcher's judgement call with the TRINET data base, and *nature* of vertical integration cannot be discerned at all with TRINET data. Although COMPUSTAT II data should be preferred to TRINET data for vertical integration questions, researchers should note that many vertical integration questions require primary information obtained directly from firms. In such research, COMPUSTAT II data should be used only for sample selection or for preliminary data collection.

In addition to highlighting the value of the COMPUSTAT II data for study of relatedness and vertical integration, our analysis shows that treating segments' two SSICs which are either related or vertically integrated as separable unrelated businesses, as has been done in some studies, inappropriately increases the measure of strategic diversity of that firm's activity. In the next section we will explore relatedness and diversity *between* segments in a firm, by building on this knowledge of the relationship between a firm's segments and businesses within a firm's segments.

BUSINESS RELATEDNESS IN DIVERSIFICATION RESEARCH

A significant stream of research in Strategic Management has been concerned with the relationship between diversification strategy and firm performance. In much of that work, researchers have attempted to measure firms' diversity, focusing on one or both of the central issues of diversity measurement: degree of relatedness among a firm's businesses, and type of relatedness among the businesses (Ramanujam and Varadarajan refer to these as 'diversity status,' 1989: 527). The value of COMPUSTAT II and TRINET for addressing these critical research questions will be discussed next.

Degree of relatedness

Rumelt (1974) and many researchers following him have used methods that differentiate between related and unrelated diversification in categoric terms. Berry (1974), Jacquemin and Berry (1979), Montgomery (1982), and Palepu (1985) have all used continuous measures or indices to evaluate total diversification. Berry (1974) and Montgomery (1982) used a variant of the Herfindahl index of industry concentration to measure firms' total diversification. Jacquemin and Berry (1979) developed an entropy-based measure of diversification, later used by Palepu (1985). The entropy measure used by these researchers measured total diversification (DT) as the sum of two indices (DR + DU), where DT (total diversification) =

DR (related diversification) + DU (unrelated diversification).

The COMPUSTAT II industry segment data base lends itself to measuring a firm's degree of relatedness by any of the above indices. However, researchers addressing relatedness issues should note that firms have already defined each segment as comprising of single, related or vertically integrated activities, therefore, it would be erroneous for researchers to split up segments, even if the SSICs differ greatly. Therefore, for computing the Herfindahl and entropy index measures, the segment identity should be based on SSIC1, the primary SIC of the segment. Although this approach may lose (or not use) some segment description detail, it would be more conservative (and for many relatedness questions, more accurate) than splitting a firm's segment data into the two by segment SIC.

There are three considerations for researchers when choosing between TRINET and COMPUSTAT II data bases to study diversification. The first two considerations, top-down vs. bottom-up research perspective, and researcher- vs. firm-defined evaluation of relatedness, are matters of researcher judgement according to the research questions being addressed. The third consideration involves establishing criteria for evaluating relatedness between units with *dissimilar* SIC codes.

The TRINET and COMPUSTAT II data bases provide the researcher with two very different perspectives on firms' diversity. TRINET, based on establishment-level reports, provides a bottom-up view of firms; each establishment stands separate and independent. COMPUSTAT II, based on corporate-level reports, provides a top-down view of firms; establishments are grouped into segments reflecting their interrelationship. Thus researchers have to first judge which perspective of firms' diversity is appropriate for their research question. For example, Farjoun (1990) used TRINET sales data when he studied firm diversity by examining employee skill diversity, a bottom-up view; for many research questions, however, the central issue, assessment of strategic business diversity and relatedness, requires a top-down view for which COMPUSTAT II data is more appropriate.

The next issue that researchers have to decide is whether their research question requires firms' evaluation of their own diversity/relatedness or whether the researchers themselves, as informed observers, want to independently evaluate firms' diversity/relatedness using the most disaggregated data available. COMPUSTAT II provides firms' evaluations; with TRINET, researchers impose their own judgements.

The third issue relates to measurement validity and concerns the rules and criteria for evaluating relatedness between units with dissimilar SIC codes. COMPUSTAT II segments are structured so as to provide information on vertical integration: vertical integration activities must be grouped within the segment of the end activity. TRINET establishments, on the other hand, are simply physically separate locations of a firm's activity. Vertically integrated activities within a firm might or might not be physically separated; however, TRINET provides no information about the presence or absence of vertical integration activities, or any other interdependency, among physically separated establishments.

What then, should researchers assume about relatedness between firms' establishments with dissimilar 2-digit SICs on TRINET? Given that an SIC code is the only identification information available to us on a TRINET establishment, can it be assumed that establishments with dissimilar 2-digit SICs are unrelated? In order to evaluate the validity of this assumption, a series of correlation analyses were conducted.

As the first step in our analysis we set out to establish the comparability of diversity measures computed from each of the two data bases. Entropy measures for the *Fortune* 500 (1987) were computed from each data base and a correlation analysis was done; the results are presented in Table 2. Because TRINET reports more 4- and 2-digit SIC establishments than COMPUSTAT II does segments, we would not expect entropy measures calculated from the two data bases to be perfectly correlated. However, as the same construct (diversity) is being measured using the same operationalization (entropy measure) in both data bases, we would expect to see a systematic relationship (resulting in multicollinearity) between the two sets of entropy measures, if the data from the two sources is inherently comparable and the only difference is the level of aggregation at which diversity is being measured (higher for COMPUSTAT II

Table 2. Correlations between TRINET and COMPUSTAT II entropy measures (1987) for the *Fortune* 500

COMPUSTAT II (C) vs. TRINET (T)	*Fortune* 500	*Fortune* 250	*Fortune* 251–500
(C) DU – (T) DU[1]	0.5598**	0.5613**	0.5294**
(C) DR – (T) DR[2]	0.3635**	0.3710**	0.2598**
(C) DT – (T) DT[3]	0.5230**	0.5050**	0.4548**

** $p < 0.01$

NOTE:

[1] $DU = \sum_{i=1}^{G} x_i \ln \frac{1}{x_i}$ [2] $DR = \sum_{i=1}^{G} \sum_{k=1}^{S} x_{ik} \ln \frac{x_{ik}}{x_i}$

where: x = % Sales
i = 1, 2....., G (Groups—2-digit SIC code)
k = 1, 2....., S (Segments—4-digit SIC code)

[3] DT = DU + DR

than for TRINET). Instead, our analysis (see Table 2) shows relatively low correlation between entropy measures derived from the two data bases. In order to explain these unsatisfactory results, we sought external validation of the data bases by comparison of entropy measures from each data base to a common standard.

A data base's effectiveness for research purposes can be judged by comparability of diversity measures which could be computed from the data bases to standards already established in the research literature. Rumelt's (1974) related ratio-based measures of diversity are broadly accepted, thus we decided to compare estimates of firms' diversity derived from COMPUSTAT II and TRINET with related-ratio estimates of firms' diversity. Two recently published studies, Michel and Shaked (1984) and Dubofsky and Varadarajan (1987), were particularly appropriate for our comparison test because their authors independently derived related-ratios on the same sample of firms, using Rumelt's (1974) methodology, for the period 1975–81.

We computed entropy measures for 1981 (1981 is the first year for which TRINET data is available) from COMPUSTAT II and TRINET data for the sample of firms used by Michel and Shaked (1984) and Dubofsky and Varadarajan (1987). Next, we ran correlations with our entropy measures and the related ratios calculated by Michel and Shaked (1984) and Dubofsky and Varadarajan (1987). Both studies used the same sample of 51 firms from the *Fortune* 250; Dubofsky and Varadarajan (1987) excluded three firms that had specialization ratios above 70 percent; we used each of these studies' samples for our comparisons with them, but excluded two firms, Sperry Rand, and Bendix, because they were undergoing significant strategic changes during the period 1975–81 which rendered their 1981 related ratios noncomparable to their mean related ratios for 1975–81. Because there are significant computational differences between the two diversity measures (entropy vs. related ratios), we expected moderately high rather than perfect correlations. Slight time frame differences also led us to expect only moderately high correlations: related ratios from Michel and Shaked (1984) and Dubofsky and Varadarajan (1987) were computed as the mean of the period 1975–81, but we used only 1981 data for both the COMPUSTAT II- and TRINET-derived entropy measures, as 1981 is the first year for which TRINET data is available.

The results of that second correlation analysis are presented in Table 3. The negative sign of the correlation coefficients in Table 3 is due to the fact that the related ratio decreases with increasing diversity while the entropy measures increase with increasing diversity. The high correlation coefficient for DU is due to the fact that the related ratio captures firms' diversity in

Table 3. COMPUSTAT II & TRINET entropy measures—1981 correlated with Dubofsky & Varadarajan's (1987) and Michel & Shaked's (1984) related ratios—1975–81

Comparison studies	Related ratios W/ entropy measures	Correlation coefficients: RR W/ COMPUSTAT	Correlation coefficients: RR W/ TRINET
Dubofsky and Varadarajan (1987)	RR-DU	−0.6440**	−0.4547**
	RR-DR	0.3222*	0.0635
	RR-DT	−0.4443**	−0.2694
Michel and Shaked (1984)	RR-DU	−0.6116**	−0.4572**
	RR-DR	0.3493*	0.1366
	RR-DT	−0.3728*	−0.2199

* $-p < 0.05$
** $-p < 0.01$

terms of related and unrelated businesses, which is conceptually closer to the DU component (diversification *across* 2-digit SIC categories) than to the DR component (diversification *within* 2-digit SIC categories). The results presented in Table 3 show that entropy measures computed with COMPUSTAT II data are more highly and more significantly correlated with the independently-computed related ratios from the previous studies, than are the entropy measures computed with TRINET data. It is quite clear from Table 3 that entropy measures of diversity computed from COMPUSTAT II data are closer to researchers' subjective assessments of diversity than are entropy measures computed from TRINET data.

Limitations for use in business relatedness research

A well-known limitation of any index measure of relatedness based on SIC codes is that SIC code assessments cannot capture all the possible sources of relatedness (production, marketing, R&D, etc.). SIC codes tend to capture shared resources in production and raw material usage better than shared resources in marketing or R&D.

While the TRINET data base provides greater detail, the SIC code of every establishment in the firm, compared to SIC codes of two businesses for up to 10 segments per firm in COMPUSTAT II data, TRINET supplies no information about how those establishments are grouped together in the firm's view. Thus, assessments of possible relatedness among businesses using TRINET data must rely on 2-digit SIC similarity and/or researcher judgement. Although COMPUSTAT II is not comprehensive (reflecting at most the 20 largest and most important businesses of each firm), it has the advantage of providing insights about firms' perceptions or intentions of relationships among their business units. Thus, COMPUSTAT II segment data has been and can continue to be effectively used to measure relatedness or strategic diversity.

INDUSTRY ANALYSIS IN STRATEGIC MANAGEMENT RESEARCH

Many research questions of interest in strategic management require that industry trends be studied in conjunction with firms' data, or with firms' data disaggregated at the segment level. When such research questions are addressed, compatibility of data bases is (or should be) a critical consideration.

In past studies requiring industry level information, data from the Census of Manufacturers were most commonly used. However, this data set has limitations which render it inappropriate for use in conjunction with COMPUSTAT II industry segment data. First, the Census cautions that the value of shipments is not accurate at the 3-digit and 2-digit levels (Comments on Statistical Measures and Tables, Nos. 18, 19, *Census of Manufacturers*, 1982):

> Multiunit companies were instructed to report for each establishment as if it were a separate economic unit and, in particular to report interplant transfers at their full economic value (pg. xxi).

Therefore, any shared resources and/or vertical integration would be double-counted in the Census of Manufacturers data, as acknowledged in this caveat from the Census of Manufacturers documentation:

> The aggregates of the cost of materials and value of shipments figures for industry groups and all manufacturing industries includes large amounts of duplication since the products of some industries are used as materials by others. . .Because the amount of duplication of the cost of materials in the value of products figures cannot be measured with any degree of precision, caution is urged with the use of the value of shipments total at the two- and three-digit industry group levels (pg. xxiii).

By contrast, COMPUSTAT II data avoids such double-counting, thus are more effective than Census data at the 2-digit level of aggregation for industry trend study. At the 4-digit SIC level, the Census of Manufacturers data are preferable.

Second, the census is conducted only every 5 years (1977, 1982, 1987). Data for the intervening years are estimated by surveying a sample of one-fourth of the population. Of the years covered by COMPUSTAT II industry segment data base (1978 on), only 1982 and 1987 are census years; Census of Manufacturers data for all other years are estimates.

A third limitation is that the Census of Manufacturers covers only firms in the SIC range 2000–3999, and does not provide comparable data for nonmanufacturing activities in SICs 0100–1999 and 4000–9999. Researchers requiring data on those nonmanufacturing SIC groups would have to obtain that data from a variety of sources (Census of Mining, Census of Agriculture, etc.), thus comparability of definitions and of time periods would be problematic.

We therefore argue that for many research questions requiring industry trend data, using the COMPUSTAT II industry segment data aggregated to the industry level is preferable to using Census of Manufacturers data. Among the advantages of COMPUSTAT II are that data are reported annually for all firms in the data base, that data are readily available online, that duplication (double-counting of sales) as found in the Census data is avoided, and that trends at the business, firm and industry levels can be studied with confidence that the variables' definitions are the same at all levels.

TRINET's establishment data are also useful for industry studies; these SIC IDs are correct to the 4-digit level. However, the associated TRINET data have to be used with caution: while data on the number of employees are accurate, the establishment sales data are not measured, they are estimated using industry census data, therefore the two can be expected to be highly correlated. By virtue of this fact, industry figures computed from TRINET data will share many of the problems associated with industry census data, outlined in this section. Another TRINET data item of interest is 'percent of industry,' a market share-type data item. Percent of industry figures indicate the fraction of sales accounted for by each establishment relative to all establishments within a given SIC category area in the TRINET data base. Although it is an advantage that TRINET provides a market share measure at such a disaggregated level (individual establishments), those market share measures are questionable because sales figures are estimated in TRINET by applying industry-average productivity to establishment employee numbers.

Industry data provided by S&P's *Financial Dynamics*, S&P's *Industry Surveys* and Dun and Bradstreet's *Industry Norms and Key Business Ratios* are all flawed for purposes of strategic management research because in each of those sources the entirety of firms' sales are assigned to firms' dominant businesses. In light of the fact that almost all large firms are multibusiness firms, this type of computation results in distorted industry data.

In summary, for researchers using firm and/or segment level data from TRINET, it is defensible to use Census data for industry benchmarks, as both data sources are at the establishment level. However, for researchers using firm and/or segment level data from COMPUSTAT II, it is preferable to use industry data aggregated from the segment data in the COMPUSTAT II data base. While the other data sources are useful for absolute industry comparisons, COMPUSTAT II is most useful for assessing industry trends.

Limitations for use in industry analysis

A major limitation on use of COMPUSTAT II segment data for industry analysis is that the

data base includes only the companies traded on the NYSE, ASE and OTC exchange (about 6,000 firms) while the Census of Manufacturing covers more than 220,000 public and private firms. Therefore, COMPUSTAT II would not provide accurate data for research requiring absolute number data (e.g. industry size), especially for industries in which a significant proportion of the industry consists of small firms.

However, many research studies require assessments of industry *trends*; for such research, COMPUSTAT II can be very satisfactory. The population of publicly traded companies, of which COMPUSTAT II is composed, represents almost all the large U.S. firms, and those in turn represent a significant proportion of the output of U.S. business enterprise. (The Census of Manufacturers has estimated that the 200 largest manufacturing firms account for 43 percent of value added by manufacture. Therefore, 6,000 of the largest firms would certainly represent the bulk of output in goods and services.) Furthermore, if research questions require a referent population for large to medium firms, industry membership among 6,007 of the largest firms could be safely assumed to constitute the relevant industry reference groups.

Problems with Census data relate to the duplication of shipments and 5-year data collection frequency; problems with COMPUSTAT II data relate to limited company coverage. Researchers choosing one data base or the other may also be interested to know that industry growth rates (as measured by change in sales, 1978–84) calculated from COMPUSTAT II and from the Census of Manufacturers showed a high degree of correlation (more than 0.70, $p < 0.001$).

CONCLUSION

We have considered the use of industry segment and establishment data for three issues of significant interest to strategic management practitioners and researchers: assessment of strategic diversity, analysis of industry trends, and evaluation of the presence of vertical integration. Based on the analysis reported in this paper, we now summarize some guidelines and caveats for making choices among data sets.

First, COMPUSTAT II is an unexploited resource for research on vertical integration. As explained in this paper, with proper use of COMPUSTAT II data, researchers can detect forward and backward vertical integration, not only within firms but also within segments of firms.

Second, for such questions as measurement of diversity, a COMPUSTAT II segment should not be considered unrelated beyond the level declared by the firm, despite the presence of two seemingly diverse segment SICs. With this caveat observed, the data base can be quite effectively used to calculate Herfindahl, entropy, and other measures.

Third, we find COMPUSTAT II industry segment data to be quite satisfactory for the study of industry trends, assuming the above caveat is observed. With an increasing proportion of industry output originating from highly diversified firms, accurate data for industry-level questions has been difficult to obtain. COMPUSTAT II provides data disaggregated from diversified firms to the segment level of those firms, permitting that segment data to then be reaggregated by industry.

Fourth, the TRINET data base lends itself to use in the calculation of diversification indices as well as in the calculation of industry benchmarks. However, TRINET has certain limitations: (1) TRINET-estimation rather than company-reporting of critical variables (sales and market share); (2) years of coverage (1981, 1983, 1985, 1987 and 1989 only); and (3) likely termination of publication of the data base.

We have delineated the composition, strengths, and weaknesses of two important sources of disaggregated firm data, COMPUSTAT II and TRINET, for various issues critical to strategic management research. Armed with the results of our analyses, researchers can now make informed choices on which data set is appropriate for the questions they are addressing.

REFERENCES

Berry, C. H. 'Corporate diversification and market structure', *Bell Journal of Economic and Management Science*, Spring 1974, pp. 196–204.

Dubofsky, P. and P. Varadarajan. 'Diversification and measures of performance: Additional empirical evidence', *Academy of Management Journal*, September 1987, pp. 597–608.

Farjoun, M. 'Beyond industry boundaries: Human expertise, diversification and resource-related industry groups'. Unpublished dissertation, Northwestern University, 1990.

Financial Accounting Standards Board. *Statement of Financial Accounting Standards No. 14: Financial Reporting for Segments of a Business Enterprise*. FASB, Stamford, CT, December 1976.

Jacquemin, A. P. and C. H. Berry. 'Entropy measure of diversification and corporate growth', *Journal of Industrial Economics*, June 1979, pp. 359–369.

Michel, A. and I. Shaked. 'Does business diversification affect performance?' *Financial Management*, Winter 1984, pp. 18–25.

Montgomery, C. A. 'The measurement of firm diversification: Some new empirical evidence', *Academy of Management Journal*, June 1982, pp. 299–307.

Palepu, K. 'Diversification strategy, profit performance and the entropy measure', *Strategic Management Journal*, May–June 1985, pp. 239–255.

Ramanujam, V. and P. Varadarajan. 'Research on corporate diversification: A synthesis', *Strategic Management Journal*, November–December 1989, pp. 523–551.

Rumelt, R. P. *Strategy, Structure and Economic Performance*, Harvard University Press, Cambridge, MA, 1974.

Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*. 2nd edn. McGraw-Hill, New York, 1988.

Standard and Poors COMPUSTAT II Services, Inc. *Business Information: COMPUSTAT II,* Standard and Poors, New York, 1986.

U.S. Bureau of the Census, *Census of Manufacturers, 1982*. U.S. Government Printing Office, Washington, DC, 1982.