Small Business Economics **24**: 17–28, 2005.

Defining and Tracking Business Start-Ups

Michael I. Luger
Jun Koo

ABSTRACT. This paper addresses a curious disjuncture between one aspect of regional development theory and the empiricism required to test its implications. On the one hand, researchers long have argued that firm births fuel the growth and development of regional economies. Just as long, however, they have employed different, often *ad hoc*, approaches to the definition and measurement of key concepts and relationships. The inconsistency among the studies in this literature creates a validity problem. We begin by providing an omnibus definition of a start-up that applies to some degree to all the articles we reviewed, namely, that it is new, active, and independent. We explain why all three criteria should be applied, rather than a subset. Second, we review the data sources that are commonly used to identify start-ups, and compare them using seven criteria. We conclude that ES202 data is the best source. Third, we develop a step-by-step tracking system for identifying new firms. By matching ES202 files from two different years and applying direct enumeration techniques, it is possible to identify newly created establishments during that time period with accuracy. This article serves both to explain the differences among the published studies of new firms and economic development, and to provide a common standard that can enhance the validity of future work on the topic.

Final version accepted on January 8, 2003

Michael I. Luger
Office of Economic Development
University of North Carolina at Chapel Hill
CB #3435
Chapel Hill, NC 27599-3435
U.S.A.
E-mail: mluger@email.unc.edu

Jun Koo
Maxine Goodman Levin College of Urban Affairs
Cleveland State University, U.S.A.
E-mail: jkoo@urban.csuohio.edu

1. Introduction

Numerous authors have attempted to show that new small firms (start-ups) are a major source of job creation, technological innovation, and consequent regional growth (Schumpeter, 1934; Birch, 1981; Kirchkoff and Philips, 1988; Reynolds and Maki, 1990; Davidsson et al., 1994; and Reynolds, 1994).[1] While the message these studies convey is generally the same, they tend to employ different, often *ad hoc*, approaches to the definition and measurement of key concepts and relationships. The inconsistency among the studies in this literature creates a generalizability problem and weakens their external validity. That diminishes the impact of the message regarding start-ups on policy-makers. Therefore, it is imperative to establish a widely accepted definition of start-ups and to develop a consistent tracking system for future research.[2]

This paper attempts to do that. First, we address definitional issues of start-ups and propose an omnibus definition for future researchers to adopt. Second, we evaluate the databases commonly used for research on start-ups. And third, we suggest a systematic method to track new firm births. We provide a summary of our tracking approach in the body of the paper, illustrating its value for three sample counties in North Carolina. We provide more details about the method in a technical appendix.

2. The problem of definition

The literature on start-ups uses three *different* definitional criteria: "new," "active" and "independent." To make studies comparable and reproducible, *all* three criteria should be employed together when classifying businesses as start-ups.

New

Virtually all studies of start-up firms use "new" as the main discriminator. To Keeble (1976), for example, that means "the creation of an entirely new enterprise which did not formerly exist as an organization." For Gudgin (1978), a new manufacturing firm is one "which beg[a]n production for the first time." "New" by these definitions includes every newly created firm in a given time period except the ones created by changes in name, ownership, location, or legal status.

The operationalization of the "new" criterion necessarily requires knowledge of the start-up date. That is commonly considered to be the date of a firm's registration as a legal entity (usually a limited liability company (LLC), limited liability partnership (LLP), or corporation). This approach is popular because registration records are readily accessible. So, for example, "new" firms created in 1995 can be identified by searching firms registered in 1995 for the first time.

Active

Hadden (1977) criticizes this approach because some registered companies exist only on paper – often created as a legal contrivance for such purposes as tax avoidance. That suggests "active" as a second criterion. Thus, to be considered a start-up, a firm should not only be new, but also should engage in the trading of goods or services. Dun & Bradstreet, for example, defines start-ups as "newly opened active establishments."

The distinction between "newly created" and "newly created and active" is important because of a high prevalence of "paper" firms in many regions. Scott (1980), for instance, found that about 23 percent of all companies created in Scotland in 1969 had never traded. Since most "paper" firms do not actually hire workers or invest in capital, they have little economic impact, and therefore, can reasonably be excluded from the population of start-up firms in a region, at least for economic development research purposes.

Once we add "active" as a discriminator, registration records are no longer an appropriate data source. Alternatively, we can define the start-up date according to when the firm takes on its first full-time paid employee (Johnson and Cathcart, 1979). Obviously, some home-based businesses without any paid employees will be omitted under this approach. However, according to Mason (1982), the resulting bias is only around 7.7 percent, considerably less than the overestimation using registration records. Moreover, research indicates that most home-based businesses never develop into major employers (Birley, 1984), further reducing the consequence of their omission.

Independent

So far we have defined a start-up as "a new and active business entity which did not formerly exist." We still have not gone far enough. This definition includes not only newly created firms by one or a group of individual founders, but also non-founder new firms (branches) created by existing businesses (Mason, 1983).

As Johnson (1978) argues, non-founder new firms are very different from other new firms established by original founders in terms of size, capitalization, and economic stimuli. Start-ups are more likely to bring new ideas into a region, which can affect existing structures and stimulate the regional economy at a larger scale. We can account for this by adopting Johnson and Cathcart's (1979) definition of a new firm as "one which has no obvious parent in any business organization." This justifies the third (additive) criterion: a firm should be not only newly created and active but also independent to be considered a start-up.

This "independent" criterion is closely related to the "new" criterion and excludes subsidiaries and branches created by existing firms. Strict definition of new firms often makes a significant difference in terms of the number of new firms identified. When Mason (1983) applied Johnson and Cathcart's strict definition of new firms to his earlier (1982) study he excluded almost 31 percent of the companies he originally identified as start-ups.

The "independent" criterion raises an important question in conducting research on start-ups: should start-ups and spin-offs be distinguished? Spin-offs are defined as "new firms created by individuals breaking off from existing ones to establish competing companies of their own" (Garvin, 1983). This definition implies that some spin-offs may not meet the "independent" criterion applied to new firms. Whether a new firm can satisfy the "independent" criterion depends on the relationship between a newly created start-up company and the firm for which its founder(s) originally worked. If a new firm is independent from the mother firm legally, financially, and functionally, and focuses on a niche market different from the mother firm's, it should be considered an independent start-up company. However, if a new firm operates more like a branch of the mother firm, even if it is legally independent, it does not satisfy the "independent" criterion. So, unlike the "new" and the "active" criteria which are dichotomous, the "independent" criterion is continuous. Therefore, the decision on whether a new firm is independent requires judgment by the researcher based on his/her research objective.

Taking the three criteria together, a start-up can be defined as a business entity:

> "which did not exist before during a given time period (new),which starts hiring at least one paid employee during the given time period (active),[3] and which is neither a subsidiary nor a branch of an existing firm (independent)."

3. Measurement and data issues

The definitional problem in start-up research is compounded by measurement and data issues. If researchers begin with a narrow definition of start-ups, they can apply a single source of data. The converse is also true – if researchers first choose a convenient data source, they necessarily narrow their definition of a start-up firm. The choice of data is important because each of the available sources has its own strengths and weaknesses. We demonstrate that below for the data sources used in the start-up literature: license records, business name registration records, Dun Marketing Identifier data, ES202 data, the *Census of Manufactures*, and enumeration.

License records[4]

Most businesses require a city or county license. The remainder (doctors, dentists, lawyers, and accountants) require a state license. Therefore, license records constitute a very comprehensive database.

License records, however, have serious drawbacks as a data source for identifying start-up companies. First, license records do not provide enough information to identify new firms that satisfy the "new" criterion. Firms have to renew their license on a regular basis (normally every year). Therefore, license records at a certain point have information about both new and existing firms. To identify new firms by comparing data from two different years, a researcher should be able to separate firms registering for the first time from firms renewing their registrations. In addition to that, a researcher should be able to pick out any firms created by changes in name, ownership, location, and legal type of existing firms to meet the "new" criterion. However, none of this information is available in license records.

Second, license records do not indicate whether a registered firm is actually involved in a transaction, which is required to satisfy the "active" criterion. We noted previously that employment status (whether the business has at least one employee) is typically used for that purpose; license records do not include that information.

Third, license records also do not satisfy the "independent" criterion. Subsidiaries and branches of an existing company require separate business licenses, so the list of start-ups constructed with license record data would be inflated.

The last problem of license records is that it is not a fully computerized database. A researcher who wants to utilize license records, therefore, will experience a great deal of inconvenience and cost in retrieving and analyzing the data.

Business name registration records

Every sole proprietorship and partnership has to register its business name properly with the Register of Deeds office in each county where business is conducted. In addition, business names for limited liability corporations and partnerships (LLCs and LLPs) are required to be registered with the Secretary of State, Corporations Division. Therefore, business name registration records are also a very comprehensive data source.

Business name registration records have several advantages over license records. Because the database maintains information about status changes it can be used to eliminate duplicate entries by the same business that may have registered twice. That pertains, as well, to branches of registered businesses that may have registered independently. In addition, sole proprietorships, partnerships, LLCs, and corporations have to register their name only once as long as they keep the same name. LLPs are required to renew their name registration every year, but it is possible to track the first year of their registration. Therefore, the first and the second feature of business name registration records satisfy the "independent" and "new" criteria.

There are, however, two drawbacks in utilizing this database in start-ups research. First, there is no way to identify "paper" firms which register their names but never get involved in transactions (i.e., it does not satisfy the "active" criterion). Second, unlike the DMI or ES202 file (which are discussed below), business name registration records provide very little firm- and industry-specific information, such as the number of employees and SIC code. Therefore, their usage is limited.

Dun Marketing Identifier (DMI)

One of the most frequently used sources of data in start-up research is the Dun Marketing Identifier (DMI) file, which is created by a commercial credit rating and marketing service, Dun and Bradstreet (D&B). The U.S. Small Business Administration has modified the DMI file for public use, creating the U.S. Enterprise and Establishment Micro database (USEEM). Since the USEEM is developed based on the DMI, this section focuses only on the DMI file.

The DMI is a relatively comprehensive database covering almost every firm that has engaged in any type of financial transaction. Therefore, companies in the file can be assumed to be active. The database includes very detailed information about individual firms, including name, address, number of employees, revenue, 4-digit SIC, type of operations, and business history. Its unique D-U-N-S Number System[5] makes it possible to identify headquarters and branches, and accordingly, helps to build corporate family relationships that can be used to satisfy the "independent" criterion.

Many researchers (Birch, 1983; Harris, 1983; Reynolds and Miller, 1988; and Reynolds et al., 1995) have used the DMI file for their research due to its relative comprehensiveness and convenience. Nonetheless, DMI data suffers from several drawbacks. First, the start-up date it uses is a significant source of bias. Normally, the editors of Dun and Bradstreet make an informed judgment on start-up date whenever ownership changes occur. If a firm would have gone out of business had it not been for ownership changes, the start-up date is changed to the date of ownership change. However, if a firm would have continued without an ownership change, the previous start-up date is retained (Reynolds et al., 1985). This methodology can create significant confusion in identifying new firms because D&B defines a business as a start-up for the current year if its birth date is within the most recent three calendar years. Reynolds et al. (1988) report that about one in four firms on the DMI list cannot be considered "new." In another study, Birley (1984) verified that about 29 percent of firms on the DMI file had existed too long to be defined as new firms.

Second, Reynolds et al. (1988) also report that the DMI file failed to identify "independent" and "active" start-ups accurately. Through telephone interviews, they found that about 2 percent of firms in the DMI file were subsidiaries or branches, and about 6 percent of firms were not active. Reynolds et al. (1985) report, as well, that the DMI file overestimates new firms on average, and after the adjustment, only about 50 percent of firms on the file are "new", "active", and "independent" actual start-ups.

Third, the accuracy of the DMI data varies by sector and time period. Mokry (1988), drawing on a comprehensive review of the literature, argues that the DMI file is more representative for manufacturing than it is for the wholesale and service sectors. He also shows that there is greater year-to-year fluctuation in new firm verifications in the DMI file, compared to ES202.

The last problem with the DMI data is its cost. In 1999, Dun and Bradstreet charged 44 cents for providing basic information about a company. Therefore, even though the rate is discounted for large volumes, DMI would be a very costly source unless a study were focused on a small region. For example, the total cost to get basic information about newly created firms in North Carolina in 1997-1998 (57,017 new firms were identified) is \$10,662 (\$187 per 1,000 firms). In addition, accessing archives older than two years incurs an additional \$750 per year service fee.

ES202

Researchers also have used data found in the Quarterly Unemployment Insurance (QUI) files that are part of the ES202 reporting system. That system is a joint effort of the Bureau of Labor Statistics (BLS), the U.S. Department of Labor, and the state Employment Security Agencies (ESAs). The database, updated quarterly, contains wage and employment levels for all establishments within the unemployment insurance system; that includes 96 percent of nonagricultural employment in the United States.[6]

As with other data sources, the ES202 system has strengths and weaknesses. On the positive side of the ledger, the Employer Status Report contains the name, location, and ownership status of respondents, and thus, can be used to track the date and direction of any ownership transfer. That makes it possible to separate firm births *de novo* from false births (with predecessors), which is necessary to satisfy the "new" criterion for start-ups. Second, the system assigns different identification numbers to headquarters and branches, allowing them to be separated as well, thereby satisfying the "independent" criterion. Third, since the only businesses required to file (and pay the tax) are those with at least one paid employee, the 'active" criterion is also satisfied.[7]

The ES202 system also has some weaknesses. First, as noted above, it is not as comprehensive a database as the license records or the business name registration records databases because sole proprietorships and agricultural establishments with less than 10 employees are not included. Second, the start-up date in the ES202 file is not the day a business actually started but the day employees begin to work. Therefore, if a firm starts without any employee, and then hires someone later, the start-up date can be biased. Third, the ES202 fails to distinguish full-time and part-time workers. That can create a bias if the data are used to generate the number of workers employed in start-up businesses (White et al., 1990). Fourth, and perhaps most important, the ES202 data are not always accessible to researchers due to confidentiality agreements between the state ESAs and businesses. In many states, the ES202 file is available only at the aggregate level, and therefore, individual firms are not identifiable.[8]

Despite these limitations, there is a consensus in the literature that ES202 is better than DMI as a source to track start-ups does (Jacobson, 1987).

The census of manufactures

If research focuses only on the manufacturing sector, the *Census of Manufactures* can be another good data source. It provides comprehensive statistics about all domestic establishments, disaggregated by 4-digit SIC at the plant level. The census has been conducted by the U.S. Census Bureau every five years since 1967. By comparing two points in time, one can identify entering firms in each industry in each census year. These underlying plant-level data allow a researcher to aggregate up to the firm level and to identify any ownership changes.[9] Therefore, the Census of Manufactures data satisfy the "new" and the "independent" criteria. The "active" criterion is also satisfied because the database only covers firms with at least one employee.

The *Census of Manufactures* has four major weaknesses. First, because the census is conducted at five year intervals rather than annually, a considerable amount of information is lost, including start-ups and deaths in the intercensal years. Second, bias is introduced due to the failure to assign unique identification numbers to plants with less than 250 employees.[10] Third, as for the ES202 system, the *Census of Manufactures* underestimates self-employed businesses. Fourth, access to the *Census of Manufactures* database is often restricted for confidentiality purposes. With permission, a researcher may be able to tap into the database on site at the Center for Economic Studies in Maryland or at one of regional census micro data centers. However, the cost for accessing the database is substantial: about \$4,000 per month.

Enumeration

The last source is a personal listing built by enumeration. Because that is unsystematic and often very costly, it is not regarded as a reasonable primary approach for identifying start-ups. However, enumeration helps to reduce the biases inherent in the other, more systematic, approaches, and therefore, is considered an important complementary procedure.

A wide variety of sources can be used for enumeration: direct site visits, the phone book, newspaper clippings, personal collections, and so on. To compile a comprehensive listing of businesses, Aldrich et al. (1985) hired observers to record information about all visible businesses on the streets in three English cities. Buss et al. (1991) utilized telephone directories, newspaper articles reporting the opening of new firms, and the Chamber of Commerce membership list. Ko (1993) also utilized newspaper articles and personal collections. The phone book seems to be the most popular and widely used source due to its wide availability, convenience, and relatively low cost. In addition, phonebooks are available in electronic format. Either Yellow or White Pages can be used. However, Kalleberg et al. (1990) argue that White Pages are more reliable because they include all firms with a telephone, while Yellow Pages only cover firms that decide to advertise there. One good electronic source is a commercial national phone number database called "Select Phone." It covers every residential and business telephone number listed in printed directories. Select Phone is updated annually and includes such detailed information as the firm's name, street address, city, state, zip code, phone number, SIC code, and year in which its listing first appeared in the database.

Summary

Table I summarizes the discussion of the six data sources. It illustrates the superiority of the ES202 as a source for start-up data.

We used our professional judgment to assign a "poor," "fair," or "excellent" score for each of seven criteria, for all six data sources. The attributes include the sources' consistency with the "new," "active," and "independent" definition criteria, as well as their scope of use, comprehensiveness, convenience, and affordability. Scope of use refers to the amount of information contained in the database. Comprehensiveness refers to the database's inclusiveness. The last column sums up the score, assuming equal weights for all attributes. The ranking would change for all sources except ES202 if the different attributes were weighted unevenly. Because ES202 received as many or more asterisks as the other sources, for each attribute, it would be ranked first under any weighting scheme.

4. Designing a tracking system

The problems of definition and measurement have compromised the quality of empirical work on start-ups. Most studies in the 1970s relied upon published industry-level census data (McGuckin, 1972; Deutsch, 1975; Gorecki, 1976). The limitation of that approach is that researchers could identify only *net* changes in the number of firms in an industry. Since the 1980s, researchers (Birch,

TABLE I
Summary of data sources

	New	Active	Independent	Scope of use	Comprehensiveness	Convenience	Affordability	Total asterisks
License	*	*	*	*	***	*	*	9
Business name	**	*	***	*	**	**	***	14
DMI	*	**	***	***	**	**	*	14
ES202	**	**	***	***	**	**	***	17
Census of Manufactures	**	**	**	***	*	**	*	13
Enumeration	**	**	**	*	*	*	*	10

* = poor; ** = fair; *** = excellent.

1983; Highfield and Smiley, 1987; Reynolds et al., 1988; Buss et al., 1991; and Reynolds et al., 1995) have begun utilizing more detailed firm-level data, such as the DMI and ES202 files. Although methodologies to track new firms have developed over time, recent empirical studies still largely depend on only one data source, with related measurement errors.

In this section, we summarize a procedure to identify start-up firms. Further details are provided in the Appendix. To employ our approach researchers must be able to access the ES202 file by signing a data-sharing agreement with their state's ESA. As indicated in Table I, the ES202 files contain enough data to make them a useful source for researchers wishing to conduct economic analyses. Compared to the DMI file, which is the most frequently used data source in start-ups research, the ES202 file is less likely to create bias.

Identification of new businesses using the ES202 files requires the matching of data from two different years. Businesses appearing only in the later year file are likely to be start-ups, but need to be screened to make sure they are "active" and "independent." The matching process we have developed consists of four steps: data screening, an initial merge, iterative merging, and testing/ verification. We then fine tune the list of matched businesses using an enumeration technique (based on Select Phone data) Each of our steps can be programmed into a statistical or database software program such as SAS or Fox Pro.

Summary of the tracking approach

The first step in the procedure is to conduct a *preliminary data scan* to eliminate double counting of establishments. That would occur when a parent company and individual branches both identify the same establishments, or because of inconsistencies in year-to-year reporting by firms.

The second step is an *initial merge of different years' data*, based on a set of "matching criteria." A merge that successfully matches records year-to-year allows us to interpret unmatched valid observations as new (if present only in the later year) or closed (if present only in the earlier year). For several reasons (discussed in the Appendix) it will take several iterations to achieve a successful merge.

Iterative merges constitute the next step of the procedure. We subdivide the unmatched observations from each earlier attempted merge into two datasets, one with the earlier year record and the other with the later year record. We then attempt to match them using a succession of criteria (plant name, employer account number, employer federal ID, ZIP code, 4-digit SIC code, and initial liability date). That is intended to correct errors caused by inconsistent or omitted information. The more merges one attempts, the more certain he can be that unmatched observations indeed represent either new or closed plants.

Following the iterative merges we *test the accuracy of the corrected dataset* by applying a program that consists of a series of logical statements, also described in the Appendix. That test is based on the assignment of scores indicating the strength of matches between pairs of observations.

Application to North Carolina data

We tested our tracking system using three North Carolina counties: Durham, Orange, and Wake. Those are the core counties of the Research Triangle metropolitan statistical area.[11] We used 1996 and 1998 ES202 data for the matching.[12]

There were 2,577 and 2,902 observations in the original 1996 and 1998 datasets, respectively. Data screening reduced the numbers to 2,262 and 2,586 for the two years. If there were no matches between the years there would then be 4,848 different records (establishments). We conducted an initial merge to create a new dataset with 3,453 observations. That included 1,395 matched businesses (combining 2,790 into half the number due to the matches) and 2,058 unmatched (note that 2,790 plus 2,058 equals 4,848).

For reasons discussed above, we could not be certain that the 2,058 unmatched observations represented new or closed plants, or errors in variables, so we conducted iterative merges. The first iteration (using plant name) found eight new matches. The second iteration (on employer account number) yielded six additional matches. The third (on 3-digit SIC) produced nine new matches. Merges on liability date and ZIP code added two and fifty new matches.

Following the iterative matches using single criteria, we attempted to match observations using combinations of criteria. Those further iterations created 126 more matches. In total, iterative merges identified 201 new matches, or almost 10 percent of the total.

We next applied our test program to the corrected dataset. Of the 201 new matches in our three county illustration, only three were identified as incorrect (twelve were judged to be inconclusive).

5. Conclusion

The lack of consensus about how to define start-ups, and the application of ad hoc methods to measure them, have created an external validity problem. In this paper, we proposed an omnibus definition of start-ups, as businesses that are "new," "active," and "independent." We then reviewed the six commonly used sources of data for tracking start-ups, and concluded that the ES202 file is the best alternative, but enumeration techniques must be used to verify and augment the ES202 data. Finally, we developed a system for improving the completeness and accuracy of the start-ups identification process.

The application of our method to three sample counties in North Carolina illustrates how a consistent, systematic approach to tracking business start-ups can significantly improve the reliability of data. The wide adoption of our proposals would strengthen the body of start-ups research and allow more direct comparisons to be made across different studies.

Technical appendix

This Appendix presents our tracking procedure in more detail. Each of the steps – data screening, initial merge, iterative merges, and testing/validation – is discussed in turn.

Data screening

A preliminary data screening procedure is necessary to eliminate double counting. When filing employment status reports, some multi-establishment companies count all branch employment in their headquarter employment statistics at the same time that their branches report their own employment separately. That practice will not affect the count of new start-ups, but would bias upward the amount of employment associated with start-ups. To address that problem, one must visually inspect the headquarters and branch plant employment numbers and make appropriate adjustments if double counting is found.

A second reason to screen data at the outset is to eliminate duplicate records for a single firm that might have been created at different points in time. While individual branch plants are required to report (or be reported) separately, that is not always done. Some firms report individual branches separately in one year and all branches together in another year. That practice affects the accuracy of the data merge. For example, suppose that firm A reported each branch separately in 1996 and all branches together in 1998. The matching procedure will simply duplicate the entries for firm A in 1998 to match the entries for firm A in 1996 because the entries in 1996 and 1998 would be identified as the same establishment; however, the number of entries in 1996 would be greater than the number of entries in 1998. To avoid that problem, the 1996 data would have to be aggregated to be consistent with the 1998 observation.

For our three county illustration, there were 2,577 and 2,902 observations in the original 1996 and 1998 ES202 datasets, respectively. By making the adjustments described above we counted 2,262 and 2,586 businesses in 1996 and 1998, respectively.[13]

Initial merge

Our merge procedure matches two files from different years based on a set of "matching criteria." In the first merge we apply a full list of matching criteria (employer account number, employer federal ID, plant name, ZIP code, 4-digit SIC code, and initial liability date).[14] In subsequent merges, we remove each of the six criteria in turn as a way to control for possible errors-in-variables, and then remove pairs of criteria to correct for errors caused by the combination of two factors. In all, we conduct twenty-one merges (one with all six criteria, six with five criteria, and fourteen with pairs of criteria (5+4+3+2)).

In our three county illustration we begin with a total of 4,848 records (2,262 plus 2,586). If there were no matches, we would conclude that we had 4,848 different businesses. By applying the matching criteria we create a new dataset with 3,453 observations. That includes 1,395 matched businesses (combining 2,790 into half the number due to the match) and 2,058 unmatched. (Note that 2,790 plus 2,058 equals 4,848.)

If the initial merge procedure worked perfectly, we could interpret the unmatched observations as new (in 1998, not in 1996) or closed (in 1996, not in 1998) plants. However, that is not likely to be the case, for several reasons:

- the names of some establishments may well have been misspelled, abbreviated, or changed, preventing a match from being made that should have been made;
- some employment identification numbers are likely to have been altered following changes in establishments' ownership, name, or accounting procedures;
- 4-digit SIC code may not have been assigned consistently because employers often change their industry identifier over time;
- the liability date may have been changed as a consequence of ownership or status change, merger, or the discovery of an employer by the state ESC who was supposed to have reported in an earlier year, but did not; and
- ZIP codes are not always recorded consistently. Some plants report five-digit ZIP code while others report nine-digit ZIP code. It also can be altered when a firm changes its location.

Due to these problems, unmatched observations are not necessarily new or closed plants, and additional iterative merges need to be done.

Iterative merge

The next step is to subdivide the unmatched observations from the initial merge file into two data sets (one with the earlier year record and the other with the later year record) to be remerged using a different set of matching criteria. For the first iterative merge we drop plant name and use the remaining five initial matching criteria (employer account number, employer federal ID, ZIP code, 4-digit SIC code, and initial liability date). This step is intended to correct errors caused by inconsistent or omitted plant name record. When we did that with the data from our sample counties, we found eight new matches.

We then repeat the step for another merge, retaining plant name, but dropping employer account number in an attempt to correct errors caused by noise in that criterion. This procedure found six new matches. The unmatched observations from the second iterative merge file are subdivided into two data sets again. This time, we drop the 3-digit SIC code but retain the other criteria to correct errors caused by inconsistent SIC code data. We found nine new matches in our three county example. In the fourth iterative merge we drop liability date, and in the fifth iterative merge we drop ZIP code. That added two and fifty new matches respectively.

Following the merges with *n-1* criteria (where *n* is the total number of matching criteria), we perform merges with *n-2* criteria, deleting every possible unique combination of two criteria. Those fourteen additional merges yielded another 126 matches. In total, iterative merges identified 201 new matches, or almost 10 percent more of the total.

Test/verification

One could test the accuracy of the iterative merges by directly contacting establishments, but that is infeasible with a large number of observations. Alternatively, we can apply the test procedure described below and summarized in Table A-1.

The test is based on the assignment of scores indicating the strength of a match between two observations. If two observations have identical values for narrowly defined criteria such as employment ID or address, we assign 12 points (the highest possible score). If they have identical values in more broadly defined criteria such as city or ZIP code, we assign fewer points, 5 or 6. In general, the assigned score gets lower as shared criteria get broader. If two observations were perfectly identical in all eleven criteria, the total score would be 100. The test procedure is to be applied to all matched observations from iterative merges. The higher the score, the more likely

TABLE A-1
Scoring system for testing the result of iterative merges[15]

If employer ID 96 = employer ID 98, then score 1 = 12;
    Else score 1 = 0;
If primary name 96 = primary name 98, then score 2 = 12;
    Else if fifteen substrings of primary name 96 = fifteen substrings of primary name 98, then score 2 = 8;
    Else if ten substrings of primary name 96 = ten substrings of primary name 98, then score 2 = 5;
    Else if five substrings of primary name 96 = five substrings of primary name 98, then score 2 = 3;
    Else score 2 = 0;
If primary name 96 = secondary name 98, then score 3 = 12;
    Else if fifteen substrings of secondary name 96 = fifteen substrings of secondary name 98, then score 3 = 8;
    Else if ten substrings of secondary name 96 = ten substrings of secondary name 98, then score 3 = 5;
    Else if five substrings of secondary name 96 = five substrings of secondary name 98, then score 3 = 3;
    Else score 3 = 0;
If physical address 96 = physical address 98, then score 4 = 12;
    Else if fifteen substrings of physical address 96 = fifteen substrings of physical address 98, then score 4 = 8;
    Else if ten substrings of physical address 96 = ten substrings of physical address 98, then score 4 = 5;
    Else if five substrings of physical address 96 = five substrings of physical address 98, then score 4 = 3;
    Else score 4 = 0;
If mailing address 96 = mailing address 98, then score 4 = 12;
    Else if fifteen substrings of mailing address 96 = fifteen substrings of mailing address 98, then score 5 = 8;
    Else if ten substrings of mailing address 96 = ten substrings of mailing address 98, then score 5 = 5;
    Else if five substrings of mailing address 96 = five substrings of mailing address 98, then score 5 = 3;
    Else score 5 = 0;
If physical city 96 = physical city 98, then score 6 = 6;
    Else score 6 = 0;
If mailing city 96 = mailing city 98, then score 7 = 6;
    Else score 7 = 0;
If telephone number 96 = telephone number 98, then score 8 = 12;
    Else score 8 = 0;
If ZIP 96 = ZIP 98, then score 9 = 5;
    Else score 9 = 0;
If SIC4 96 = SIC4 98; then score 10 = 6;
If SIC3 96 = SIC3 98; then score 10 = 3;
    Else score 10 = 0;
If employment 98-employment 96 (50), then score 11 = 5;
    else score 11 = 0;

matches are to be correct. Researchers establish a threshold score for correct and incorrect matches, based on their research objectives. For example, the test using a 60-point threshold failed 25 matches out of 201 matches.

Triage

Cases with very high (matched) or very low (not matched) scores are easier to interpret than those with mid-level scores. On the 100 point scale, for example, scores ≥ 70 and ≤ 50 might be judged to be conclusive, while those in the 50–70 range might warrant further consideration.

We propose applying an enumeration technique to those mid-level cases, whereby Select Phone or other telephone directory data are used to determine matches or non-matches, either by inspecting published entries and/or by phoning establishments directly.

As in the iterative merges described above, we would subdivide the incorrect matches that emerge from this procedure and establish a new threshold and repeat the steps until we converge on the final cut-off (or expend fully our budget).

Of the 201 new matches in our three county illustration, 195 had 51 or more points. Scores range from 44 to 100, with the 51–70 interval containing two-thirds of all observations. The distribution of scores is shown in Table A-2.

Our matching procedure appears to have worked well. Only one match in the 30–50 range

TABLE A-2
Distribution of scores

Score	# observations	Correct	Incorrect	Unidentified
< 30	0	0	0	0
30–50	6	5	1	0
51–70	135	123	0	12
71–90	37	37	0	0
> 90	23	23	0	0
Total	201	188	1	12

was incorrect, while 12 in the 51–70 range were not identified. This suggests 40 as the threshold if accuracy is paramount, and 60 as the threshold if efficiency is valued most.

When we performed the tests described above, and dropped all establishments which appear only in 1998 but have an initial liability date earlier than 1996, and cases with an initial liability date before January 1st, 1996, we ended up with a final tally 831 start-ups.

Acknowledgements

The authors acknowledge helpful comments from Paul Reynolds and an anonymous referee. An earlier version of the paper was presented at the annual research conference of the Western Regional Science Association.

Notes

[1] Other researchers have focused on the factors that affect new firm births (for example, Mason, 1989; Hart and Gudgin, 1994; Garofoli, 1994; and Keeble and Walker, 1994).

[2] The literature on closing firms is subject to similar definition and measurement inconsistencies. A longer-run agenda is to explore firm births and deaths within the same framework in order to understand the dynamics of regional economic change.

[3] We use a broad definition of "active" (i.e., paid employee rather than a full-time paid employee) because it is very hard to separate full- and part-time workers with currently available data sources.

[4] Because each state has slightly different legal requirements for starting new businesses, we use North Carolina as the basis of our discussion of license record and business name registration data.

[5] That is D&B's unique 9-digit business identification system.

[6] All employers hiring at least one paid employee for 20 or more weeks during the year or paying an employee $1,500 or more during any quarter are required to submit the Employer Status Report. Agricultural establishments with less than 10 employees and sole proprietorships are exempt.

[7] Businesses employing unpaid family member of the founder(s) are exempt.

[8] In some cases, however, a researcher can get access to plant-level micro data dolely for his/her research after signing a data-sharing agreement.

[9] An accuracy issue arises when dealing with small plants (with less than 250 employees). The original plant identification number is assigned only to large plants. For that reason, researchers often drop small plants to avoid possible bias (Dunne et al., 1988).

[10] See Dunne and Roberts (1986) for more detail.

[11] The MSA also includes somewhat less dense Johnston, Granville, and Chatham counties. The three counties we include have a large and vibrant high tech industrial base, with considerable start-up activity.

[12] The data matching methodology using the ES202 database was developed by Bergman et al. (1996). We modify that approach in three ways: by using a unique identification number for branch plants in the match, allowing us to aggregate or disaggregate those branch plants according to research objectives; by applying a scoring system we developed to test the results of our iterative merges; and, by following-up the matching with a telephone directory-based enumeration to check and refine the results.

[13] To avoid double counting, we only included single establishments and multi-establishments reporting as a single unit due to small employment (< 10). Although most start-ups are single establishment units, some new companies do begin with more than one facility. We found only 117 of those in 1996 and 102 in 1998, which is approximately 4 to 5 percent of the total.

[14] Matching procedures using character variables (especially business name) as criteria can be tricky. Some plant names are recorded with the article "the" in one year, but not in another. Sometimes spaces between words cause trouble. To avoid those problems we omitted the word "the" and all spaces between words in plant names.

[15] This is not an actual computer program, but simply logical statements that could easily be translated into a program. Note that physical and mailing addresses are different in the UI files.

References

Aldrich, H. E., J. Cater, T. Jones, D. McEvoy and P. Velleman, 1985, 'Ethnic Residential Concentration and the Protected Market Hypothesis', *Social Forces* **63**, 996–1009.

Bergman, E., E. Feser and S. Sweeny, 1996, 'Targeting North Carolina Manufacturing: Understanding the State's Economy Through Industrial Cluster Analysis.' Institute fore Economic Development. University of North Carolina at Chapel Hill.

Birch, D. L., 1981, 'Who Creates Jobs?', *The Public Interest* **65**, 3–14.

Birch, D. L., 1983, *Share of Jobs vs. Share of Job Creation: A Flow Analysis.* Cambridge, MA.

Birley, S., 1984, *Finding the New Firm.* Academy of Management Proceedings, 44th Annual Meetings, pp. 64–68.

Buss, T. F., X. Lin. and M. G. Popovich, 1991, 'Locating New Firms in Local Economies: A Comparison of Sampling Procedures in Rural Iowa', *Journal of Economic and Social Measurement* **17**, 45–55.

Deutsch, L., 1975, 'Structure, Performance, and the Net Rate of Entry into Manufacturing Industry', *Southern Economic Journal* **41**, 450–456.

Davidsson, P., L. Lindmark and C. Olofsson, 1994, 'New Firm Formation and Regional Development in Sweden', *Regional Studies* **28**, 395–410.

Dunne, T. and M. J. Roberts, 1986, *Measuring Firm Entry and Exit with Census of Manufactures Data.* Department of Economics, Pennsylvania State University.

Dunne, T., M. J. Roberts and L. Samuelson, 1988, 'Patterns of Firm Entry and Exit in U.S. Manufacturing Industries', *RAND Journal of Economics* **19**(4), 495–515.

Garvin, D. A., 1983, 'Spin-offs and the New Firm Formation Process', *California Management Review* **2**, 3–20.

Garofoli, G., 1994, 'New Firm Formation and Regional Development: The Italian Case', *Regional Studies* **28**, 381–393.

Gorecki, P. K., 1976, 'The Determinants of Entry by Domestic and Foreign Enterprises in Canadian Manufacturing Industries: Some Comments and Empirical Results', *Review of Economics and Statistics* **58**, 485–488.

Hadden, T., 1977, *Company Law and Capitalism.* London: Weidenfeld and Nicholson.

Harris, C. S., 1983, *Small Business and Job Generation: A Changing Economy or Differing Methodologies?* U.S. Small Business Administration.

Hart, M. and G. Gudgin, 1994, 'Spatial Variations in New Firm Formation in the Republic of Ireland, 1980–1990', *Regional Studies* **28**, 367–380.

Highfield, R. and R. Smiley, 1987, 'New Business Starts and Economic Activity: An Empirical Investigation', *International Journal of Industrial Organization* **5**, 51–66.

Jacobson, L., 1987, *Texas Duns-UI Comparability Analysis-Phase II.* Kalamazoo, Michigan: W.E. Upjohn Institute for Employment Research.

Johnson, P. S., 1978, 'New Firms and Regional Development: Some Issues and Evidence.' Center for Urban and Regional Development Studies. Univ. of Newcastle-upon-Tyne, Disc. Paper 11.

Johnson, P.S. and D. G. Cathcart, 1979, 'New Manufacturing Firms and Regional Development: Some Evidence from the Northern Region', *Regional Studies* **13**, 269–280.

Kalleberg, A. L., P. V. Marsden, H. E. Aldrich and J. W. Cassell, 1990, 'Comparing Organizational Sampling Frames', *Administrative Science Quarterly* **35**, 658–688.

Keeble, D., 1976, *Industrial Location and Planning in the United Kingdom.* London.

Keeble, D. and S. Walker, 1994, 'New Firms, Small Firms and Dead Firms: Spatial Patterns and Determinants in the U.K.', *Regional Studies* **28**, 411–427.

Kirchkoff, B. and B. Phillips, 1988, 'The Effect of Firm Formation and Growth on Job Creation in the U.S.', *Journal of Business Venturing* **3**, 361–372.

Ko, S., 1993, *The Incidence of High-Technology Start-Ups and Spin-Offs in a Technology Oriented Branch Plant Complex: A Case of Research Triangle Region.* University of North Carolina.

Mason, C., 1982, *New Manufacturing Firms in South Hampshire: Survey Results.* Discussion Paper 13. Department of Geography, University of Southamton.

Mason, C., 1983, 'Some Definitional Difficulties in New Firms Research', *Area* **15**, 53–60.

Mason, C., 1989, 'Explaining Recent Trends in New Firm Formation in the U.K.: Some Evidence from South Hampshire', *Regional Studies* **23**, 331–346.

McGuckin, R., 1972, 'Entry, Concentration Change, and Stability of Market Share', *Southern Economic Journal* **38**, 363–370.

Morkry, B., 1988, *Entrepreneurship and Public Policy.* New York: Quorum Books.

Reynolds, P., S. West and M. Finch, 1985, 'Estimating New Firms and New Jobs: Considerations in Using the Dun and Bradstreet Files', *Frontiers of Entrepreneurship Research* **1985**, 384–399.

Reynolds, P. and B. Miller, 1988, *1987 Minnesota New Firms Study: An Exploration of New Firms and Their Economic Contributions.* Center for Urban and Regional Affairs: MN.

Reynolds, P., 1994, 'Autonomous Firm Dynamics and Economic Growth in the United States, 1986–1990', *Regional Studies* **28**, 429–442.

Reynolds, P. and W. R. Maki, 1990, *Business Volatility and Economic Growth.* Project Report prepared for the U.S. Small Business Administration.

Reynolds, P., B. Miller and W. R. Maki, 1995, 'Explaining Regional Variation in Business Births and Deaths: U.S. 1976–88', *Small Business Economics* **7**, 389–407.

Scott, M. G., 1980, *Mythology and Misplaced Pessimism: A Look at the Failure Record of New Small Businesses.* Paper to UKSBMTA 3rd annual conference.

Schumpeter, J., 1934, *The Theory of Economic Development.* Cambridge: Harvard University Press.

White, S. B., J. F. Zipp, W. F. McMahon, P. D. Reynolds, J. D. Osterman and L. S. Binkley, 1990, 'ES202: The Database for Local Employment Analysis', *Economic Development Quarterly* **4**, 240–253.