Entrepreneurship Research Through Databases: Measurement and Design Issues

Karl Wennberg

This article provides an account of how databases can be effectively used in entrepreneurship research. Improved quality and access to large secondary databases offer paths to answer questions of great theoretical value. I present an overview of theoretical, methodological, and practical difficulties in working with database data, together with advice on how such difficulties can be overcome. Conclusions are given, together with suggestions of areas where databases might provide real and important contributions to entrepreneurship research.

The purpose of this article is to outline the potential of secondary databases[1] in entrepreneurship research. I will describe different pros and cons of using databases for entrepreneurship research, and provide some suggestions on how to handle problems related to the analysis of such data. I will also discuss different ethical considerations of using databases and conclude with examples of areas where databases might provide answers to theoretically vital questions.

Entrepreneurship research has sparsely made use of databases compared to other fields such as economics or the organizational and managerial sciences. As Aldrich (1992) noted: "Given the increasing number of publicly available data sets, and the openness of governments and some private firms to make their records available, the low number of articles based on public data sets is somewhat surprising" (p.201). However, the past couple of years have seen a greater usage of databases in entrepreneurship research (Bouckenooghe et al. 2004; Grégoire, Meyer, and De Castro 2002). There are several reasons for this development. The quality of official records in many countries has improved, raising the promise of public databases as it applies to entrepreneurship research. In the European Union for example, intrastate cooperation between public bodies such as statistical bureaus has raised the general standard of data (e.g., Tronti, Ceccato, and Cimino 2004). Such cooperation has also facilitated the possibilities to make international comparison between public databases.

These improvements open up promising new paths to the entrepreneurship research community to address vital theoretical questions. For example, by combining data on new and emerging firms with labor market and tax data of individuals' education, working history, and personal finances, it should be possible to follow people's 'entrepreneurial careers' over time. In addition, longitudinal analysis of databases has the potential to contribute methodologically to the field of entrepreneurship. Because theories of entrepreneurship have increasingly come to stress the process nature of entrepreneurship, longitudinal methods offer much more promise than cross-sectional tools for improving our understanding of entrepreneurship and entrepreneurial processes (Davidsson 2004; Van de Ven 1992; Chandler and Lyon 2001).

This article focuses on how secondary databases can be effectively used in entrepreneurship research, addressing potential pros and cons compared to other types of data, and what can be done to overcome some of the methodological obstacles in working with databases. The article specifically addresses the following issues: First, the focus will be on general and large-scale databases such as those available from public authorities and organizations. I will not address, for instance, specific corporate databases, although many of the ideas put forward in the article are relevant to these as well. Second, I will not go into details on methods of analysis, although examples and suggestion will be given along the way. Third, the aim is to specify what has previously been found to work and what has been found not to, and to suggest interesting avenues for future research where databases might be gainfully applied.

Prior Usage of Databases

How has the field of entrepreneurship research made use of secondary databases up until now? Surveys of trends in research methodology indicate that until recently, most research has tended to use cross-sectional analyses of survey data (Aldrich 1992; Chandler and Lyon 2001). However, a trend during the latter years is that research to a greater extent makes use of longitudinal methods (Chandler and Lyon 2001). Since collection of survey data for longitudinal analysis is problematic for several reasons such as accumulating nonresponses (Samuelsson 2004; Wiklund 1998) and time expenditure (Chandler and Lyon 2001), secondary databases provides a feasible way to conduct analyses over longer periods of time.

The general features of secondary databases include long time series, often spanning several years or even decades. Another feature is large samples, often collected in real-time where cases of missing data are concurrently noted. The large samples often lead to demographic approaches in analyses of databases (Aldrich and Wiedenmayer 1993). This need not, however, be the case. Considering that entrepreneurship in its nature can be considered an outlier phenomenon (Schumpeter 1934), general population-like approaches might instead be inhibiting for theory development (Davidsson 2004). As will be indicated in the "How to Do It" section, a rarely used approach is to use databases to sample more specific *groups* of cases, strengthening the explanatory power of theories built from these groups.

Most commonly, official data from various countries and public bodies and census bureaus have been used. These are too many and too diverse to be described at length. In the United States, official data have been less accessible than in many other countries and therefore alternative, private sources have been employed. One of the first is Dun and Bradstreet Inc.'s Market Identifier Files (DMI), the source for Birch's (1979) seminal work on the job contribution of small firms. This database has been criticized because of its origin as a source of commercial credit information: Since the customers of such information are generally not very interested in small firms with little credit worthiness, the market identifier files have a bias against large- and middle-sized firms (Kalleberg et al. 1990; Phillips and Kirchhoff 1989; Storey and Johnson 1987), and problems with identifying the correct year a firm is closed (Williams 1993). Luger and Koo (2005) compared the DMI with several different sources of data, arguing that the Quarterly Unemployment Insurance files found in the U.S. Bureau of Labor Statistics reporting system is superior to the DMI files. However, the Unemployment Insurance files exclude self-employed individuals with sole proprietorships, and cannot distinguish between part-time and full-time workers.

To overcome the problems inherent in the DMI files, the U.S. Small Business Administration used the files to create the Small Business Data Base (SBDB), an extensive representative sample of all firms founded in the U.S. economy from 1976 until the late 1980s (Kirchhoff and Phillips 1992; Phillips and Kirchhoff 1989). Also the SBDB had a problematic sampling frame since (1) the underlying DMI files are based on self-reporting information, and (2) registration of firm foundings and firm discontinuances are often lagging one or two years (Phillips and Kirchhoff 1989). A more inclusive database is the Longitudinal Establishment and Enterprise Microdata (LEEM) developed by the U.S. Census Bureau (Acs and Armington 1998). This database includes all establishments in existence between 1989 and 1998, but only establishments with employees (i.e., excluding self-employed individuals). Acs and Malecki (2003) used the LEEM database to show that in contrast to what is often believed, the proportion of high growth firms in the United States is relatively larger within the smaller, nonmetropolitan labor market areas. Since the LEEM files can be linked to other census data, there might be possibilities to address other interesting questions.

One reason for the construction of new databases in the United States is the nonexistence of any complete business register or inclusive individual-level databases. In smaller countries with extensive social welfare systems such as Denmark, Sweden, and the Netherlands, many large-scale databases have been constructed by public bodies to evaluate welfare programs and policies. These databases have just recently begun to be explored for entrepreneurship research. In a working paper called "Do Small Firms Produce Better Entrepreneurs?" Sørensen and Phillips (2004) tracked the employment history of all people in Denmark who engaged in entrepreneurship for the first time in 1995. They found that those who had been working for smaller firms prior to entering entrepreneurship were more likely to remain in entrepreneurship and have higher incomes than those who had been working for larger firms. In another recent paper, Giannetti and Simonov (2004) investigated a random sample of Swedes who became entrepreneurs between 1995 and 2000, finding that in social groups where entrepreneurship is more widespread, individuals are more likely to become entrepreneurs and invest more in their own businesses, even though their entrepreneurial profits are lower.

Since databases from public sources are typically quite coarse-grained and often provide only limited information on each case (individual, firm, region, etc.), the bulk of entrepreneurship research utilizing information from databases has focused on the industry, region, or national level of analysis. Examples of such research are when data of, for example, new business start-ups or patented innovations is used to provide an indicator of aggregate levels of entrepreneurship (i.e. *rates;* Aldrich and Wiedenmayer 1993). However, the theoretical setting of such studies often do not relate directly to "mainstream" entrepreneurship research such as the creation of new ventures (Gartner 1990) or the discovery and exploitation of entrepreneurial opportunities (Shane 2003). The field of entrepreneurship has therefore yet to address Aldrich's (1992) comment on the scant usage of database data.

Merits of Databases

As suggested earlier in this article, the longitudinal and often very comprehensive nature of secondary databases can be employed to answer theoretical questions where interrelated factors or the heterogeneous nature of firms and individuals necessitates large, unbiased samples with the possibility to simultaneously investigate a variety of factors. One example of where databases have been successfully employed to investigate important questions is in the case of the suggested "female underperformance hypothesis." This idea was built on the survey-based research findings that women-owned firms tend to exhibit lower growth levels (Fischer, Reuber, and Dyke 1993) as well as lower profits and higher failure rates (Carter, Williams, and Reynolds 1997). However, using a public database with data coverage exceeding 90 percent, Watson (2003) found no significant differences in failure rates between men- and women-owned firms after controlling for the type of industry that these firms are in. Du Rietz and Henrekson (2000) utilized a more comprehensive secondary database to investigate the hypothesis. After controlling for both type of industry and firm size, they concluded that, with the exception of sales, there were no significant differences in performance between men's and women's firms on any one of the three measures—growth, profit, and survival.

In addition to ease the testing and untangling of important concepts such as the female underperformance hypothesis, databases can help to facilitate the development of research design and methodology in entrepreneurship research as well. I will describe three such developments: improved sampling specification, correcting for endogenous effects, and multilevel methods of analysis. Looking first at sampling issues, it has been noted that a notoriously difficult issue in research on emerging organizations and activities has been different types of selection bias (Kalleberg et al. 1990). This is a problem both in quantitative and qualitative research designs, and most often these difficulties are related to a survival bias in the sampling frames. For example, if a study tries to explain the variance in performance among a set of firms, the results risk being overly inflated if the cases chosen are more common to what the study is looking for. Higher performance will be more common among surviving firms (e.g., Carroll and Hannan 2000). However, utilizing databases does not provide us with very good sampling frames per se. What is important is that sampling frames of databases are usually very *precise,* something which is still rare in selections of cases in entrepreneurship research (Aldrich 1992). As secondary databases are fundamentally based on a specific sampling frame, it is thus important that researchers using databases explicitly consider how such a sampling frame mirrors the population that is being investigated.

Another merit of databases is that the longitudinal nature of data facilitates drawing causal inference, as well as a coping with endogeneity problems. Endogeneity occurs when we try to explain an outcome where an independent variable—a predictor—in a statistical model is itself codetermined within the model (Wooldridge 2002). In other words, if we include an independent variable in our model that is potentially a choice variable that might be correlated with other unobservable variables, the variable is endogenous to the effect or choice we are trying to predict. This is a common and often underestimated problem in much of the managerial and organizational sciences since research often seeks to infer an event, such as firm performance, to prior actions taken by individuals or organizations (Hamilton and Nickerson 2003). As secondary databases usable in entrepreneurship research are often created by public bodies to assess the effects of political instruments and environmental changes on economic structure, such data provides a way to overcome the endogeneity problem. This can be accomplished by the inclusion of an exogenous instrument—a variable determined by something other than the system measured—which is correlated with the independent variable(s) but not with the error term (Hamilton and Nickerson 2003; for an example of endogeneity correction in entrepreneurship see Giannetti and Simonov 2004). There should also preferably be a theoretical rationale for such an exogenous instrument. As example, let us say that our goal is to determine the effect of some public entrepreneurship education program on the performance of a sample of small business managers. If we suspect that the more competent entrepreneurs would not participate in such a program but instead go directly into business, having participated in the education program could be seen as endogenous to performance in entrepreneurship and failing to control for this might yield a spurious estimated effect that program participation actually lowers performance. Having a longitudinal database can facilitate the inclusion of an exogenous instrument, which in this specific example would be a variable that we would expect to affect people's decision to engage in a short-term program but have a minor effect on their entrepreneurial performance (e.g., a measure of how many elective courses the individual took in college).

The third and final merit of databases to be addressed is the potential to conduct studies on different levels of analysis—and also to link these to each other. This is an important issue since entrepreneurship research has long been troubled with confusion on levels of analysis (Aldrich 1992; Davidsson and Wiklund 2001; Sarasvathy 2004). One example of such confusion is the effect of the founder(s)'s level of education on the performance of new ventures. A considerable amount of research has stressed founders' education to have a positive relationship with venture performance. However, studying individuals' characteristics and trying to draw inferences to the outcomes of their venture can be problematic as some ventures are founded by one person and others by several individuals. In addition, some individuals are simultaneously active in several ventures and might put differing amounts of effort into each one of these. Accordingly, the length and type of education of a group of founders might very well affect firm performance in other ways than the education of a single owner-managed firm. *If venture level outcome* is studied, *venture level resources* such as human, financial, or social forms of capital should be the natural inputs (Davidsson 2004). Here, data on individuals' length and type of education, together with their personal finances and occupational experiences, could be used to assess the importance of such inputs for the performance of venturing activities.These types of databases have been greatly exploited on aggregate levels of analysis in, for instance, labor economics but have yet to see applicability in more fine-grained studies of, for example, the creation and development of new ventures. Such analyses can be especially powerful if ventures can be linked to their individual founder(s) (Scott and Rosa 1996). This can be achieved by using multilevel research methods (DiPrete and Forristal 1994; Kozlowski and Klein 2000). The possibilities of using different levels of analysis are important considering theorists' arguments that entrepreneurship researchers have been focusing on a rather narrow set of outcomes. Venkataraman (1997) argues that entrepreneurship research should move from focusing on firm-level outcomes of entrepreneurship to focus on *societal-level outcomes,* whereas Sarasvathy (2004) argues that entrepreneurship research should focus more on *individual-level outcomes* from entrepreneurial acts. Addressing Venkataraman's call for society-level outcomes, databases on firms can be used to investigate how technological shifts, for example, affect the number of new firms, products, or activities, as well as the productivity and profitability of certain industries. In this case, *industry-level factors,* such as changes in demographics, legislation, or technological inputs should be used to infer the outcomes from entrepreneurship. Addressing Sarasvathy's call for individual-level outcomes, databases on individuals—possibly linked with data on the firms where they are active as employees or entrepreneurs—can be used to investigate how participating in different types of entrepreneurial activities affect the subsequent careers and wealth levels of these individuals. In this case, *individual-level resources,* such as education, personal finances, or social network, should be used to infer the outcomes from entrepreneurship (Davidsson 2005).

Potential Problems with Databases

As argued in the onset of this article, databases have been underutilized in entrepreneurship research. There are some likely reasons for this: secondary databases differ from research methods such as experiments or surveys where researchers themselves can choose a sampling frame to study a population they are interested in. Most databases build fundamentally on organized sets of control systems used by authorities to record the existence of, for example, taxes paid by firms and individuals. Alternatively, databases might be based on census information used by authorities to gain knowledge of the demographics of firms and individuals. In either case, secondary databases are not designed to easily accommodate researchers' demands on theory-driven definitions or types of measurement (Phillips and Kirchhoff 1989). This section outlines some of the problems inherent in using data from databases. Most notable are sampling problems and problems related to how, and for what purposes, variables in a database were assembled initially. The section concludes with a critical assessment of the validity of such variables. Specifically, I will address internal and construct validity.

Regarding the issue of sampling of cases from databases, it has been pointed out that data collected for purposes other than research often show severe undercoverage of parts of the population that might be the most relevant to entrepreneurship researchers, such as young and/or small ventures (Aldrich et al. 1989). For example, lists of new firms provided in industry directories or government records often exclude new ventures that fail very early in their existence (Aldrich and Wiedenmayer 1993; Katz and Gartner 1988). Furthermore, statistical authorities are often lagging in creating identification codes for new types of industries or organizational populations (Aldrich 1999).This leads to problems in applying such data to entrepreneurship research if we accept the principle that entrepreneurship is comprised of new and *emerging* economic activities (Schumpeter 1934). A consequence of this is that secondary databases can seldom be straightforwardly utilized in entrepreneurship research; researchers need to select or combine data carefully from different databases to reach a data sample that is theoretically useful. Another problem is that the kind of data found, although comparatively consistent and reliable, is often quite coarse and might not be a feasible approximation of more complex theoretical concepts (Davidsson 2004). For example, information on an individual's type and amount of human capital (e.g., education and work experience) in databases is often limited to *levels* of education and job tenure at the current workplace—more seldom on the *type* of education and work experience. It is doubtful what such crude approximations actually tell us about an individual's human capital. Also, official statistics on individuals' employment at a specific location (firm) is often estimated at a single point in time (Acs and Armington 1998; Delmar, Sjöberg, and Wiklund 2003). Such data will underestimate employment flows and small firm processes in dynamic or seasonal industries.

Secondary databases are generally considered to be more reliable than data collected in surveys. However, this is a "truth" with modifications since information found in databases are generally collected (1) automatically, or, (2) through survey-like methods. The first type of information, for instance, provides demographic details, such as household composition, which is generally very reliable with few (systematic) errors. The second kind of information, however, suffers from the same type of problem as any type of survey (i.e., internal and/or external nonresponses). This is especially the case for SIC-codes that in many European countries are based on "mandatory" information regarding the new firm's (will-be) line of business. However, disregarding this information will not prohibit the new firm from being registered.

Census authorities will use the mail and occasionally telephone calls to remind the firm to submit information on its line of business. This procedure is akin to most type of surveys—with one exception: Very rarely will the SIC-codes in a database tell whether the information was obtained through voluntary registration or in one of the subsequent reminders.

To what extent can we then assess if a secondary database is valid for our specific purpose? Validity problems with databases are often attributed to internal and construct validity. In regards to internal validity, the proliferation of large sets of databases increase the risk that (any) available data that seems somehow fit for the purpose might be used to test a theoretical model—although the data in practical terms are very distal proxies of the theoretical concept in question (Davidsson 2004). In other words, despite a seemingly consistent model and significant relationships, there are either no, lacking, or faulty theoretical underpinnings for why one or several independent variables should affect the dependent variable in a model.

In regards to construct validity, an inherent problem of using data assembled by someone else is that it is impossible to design specific measurements in ways we would like. Consequently, there is a risk that what seems apparent in data assembled in a database is not what was actually measured.

How to Do It: Design and Measurement Issues

As outlined in the earlier sections, there exist some specific problems on the successful usage of databases in entrepreneurship research. A main problem is that most databases are just designed for purposes other than (entrepreneurship) research. Simply looking for associations in a large enough data set could bring results in one or two finds. However this is probably not the ideal way to conduct exploratory research. It would be more preferable to start out with a careful research design—considering the questions *why* we choose to work with a certain type of data, and *how* this relates to the theoretical problem that is being investigated. It is difficult to improve upon research efforts when one has simply used a database and tried to do something with it. Also, using data collected by someone else is problematic for two quite different reasons: First, there is a fairly large risk that the type, number, and specification of variables are not well suited to the theoretical framework that one wants to use. Minor model adjustments in the design of a study is of course not unusual, but there is an apparent risk that many small adjustments in the end means major "squeezing of the model" to fit the data. The second problem is that even if the data seems suitable to our theoretical framework, not having participated in the first-hand outlining, sampling, and collecting means that there could be significant difficulties in becoming familiar with the data. Specifically, the great number of variables often found in secondary databases means that detailed definitions of variables and how these were collected are crucial. Such definitions are often inadequate for the simple reason that statistical bodies work primarily with collecting data, not analyzing it.

Combining Databases

As opposed to using databases assembled elsewhere, theoretically derived sampling frames might actually help to *create* new databases by drawing upon different types of publicly or privately available data. To conduct entrepreneurship research using databases in such a way, Davidsson (2004; 2005) argues that success to a large extent is dependent on how much influence the researcher can have on the type of sampling frame, variables, units, and time span that is used:

> *…the trick behind this [success] was careful and thorough work in close collaboration with experts at the statistical organization in order to use and combine the best available data for creating reliable, customized data sets that could actually answer the research questions that we were asking….* (Davidsson 2005, p. 26 in manuscript).

For example, databases that maintain identification keys to firms, individuals, or workplaces might at a later date be used by researchers to match against other databases with complementary information (Linder 2004). This means that the researcher has access to both contacts within such relevant statistical authorities as well as the ability to fund the extraction of customized data. If we assume that the state of affairs is somewhere between this "ideal" put forward by Davidsson and that of exploiting a preexisting database, what kinds of problems are we then likely to encounter, and what can be done to handle them?

Theory-Driven Research

A fundamental requirement for successful research is that key variables in a database are actually theoretically relevant. If the data does not seem to be suitable to the kind of theory we intend to test, it is recommendable to go back to the drawing board to reconsider the study. Frost and Stablein (1992) argue that being immersed in the data is a fundamental requirement for conducting exemplary research. If the database should prove to be unsuitable for a particular purpose, getting "immersed" does not necessarily mean a waste of time. Explore alternative paths! Is it possible to change the level of analysis? Did you unsuccessfully look for approximations of behavior variables but instead found data more suitable as sociodemographic variables? Theories other than the one(s) you originally relied on might prove useful. By getting "immersed" in the data, you might actually discover something that existing theories cannot readily explain. Chance and surprise account for many good ideas in science. For example, Acs and Audretsch's (1989) original findings that small firms account for the relative majority of innovations in competitive industries originated while the researchers were investigating other questions, using a large secondary database (Acs 2004).

Defining and Sampling

Let us now turn to how databases can be related to definitions of entrepreneurial activities. Take for example individual level data that usually denote people's occupation as their "main" activity (e.g., employee, homemaker, self-employed, etc.) These kinds of definitions easily clash with our theoretical concepts, since an employee or a homemaker can very well make a stab at entrepreneurship by starting a business "on the side." Even if this new business is something the individual spends most of her mental energy and resources on, it might not be registered in official data as her "main" activity. In addition, occupation is often defined in census-like data as "the place where an individual receives her largest earnings from'. The result is that an unemployed person will be considered to have a full-time income even if only making $5,000 a year, but an investment banker with a firm on the side *that she strives to expand* might be excluded from the new firm definition, even if his business's turnover is $100,000 a year (Aldrich 1999). When using databases, one should consequently be careful not to accept definitions that might exclude some of the most relevant cases. If the cases we are looking at are not suitable for theory-testing, it is quite irrelevant how many, how good, or how valid variables we have at our disposal. The results will still be of very little value. This problem might be alleviated by validating a measure by comparison with other types of data. In regards to individuals' occupation for example, one could compare how an individual's labor market activities are denoted in one type of database compared to another. If data in a public labor market database defines occupation as the activity from where the individual receives the largest earnings, this can be weighed against, for example, tax registers that list an individual's total income and its sources. Thus, it is possible to circumvent the limitations imposed by a particular data source to better fit our theoretical definition of a concept. It has been pointed out that oddly, such cross-validations seem to be lacking in entrepreneurship research (Chandler and Lyon 2001).

Measurement

An important measurement issue is that while good research requires consistent definitions and measurements of theoretical concepts, this might not be the case for data assembled for other purposes. In any case, it is necessary to ensure whether the variables in a specific database are consistently defined and measured; if not there is no way to control for differences in measurement. To ensure consistent measurement procedures, discussions with statistics experts in charge of assembling and updating databases are crucial. Such discussions will probably also reveal important details of how a certain database was actually created. For example, most individuals and firms are obliged to report certain types of financial information to the authorities for taxation and other reasons. For one reason or another, both individuals and firms might over- or underreport their financial statements (Gentry and Hubbard 2004), thus creating biases in database information on, for instance, net sales of small firms. Therefore, researchers relying on databases need to consider questions similar to that of survey design, namely, *what is it* that people actually report? when asked to provide certain information. Such questions could be posed to the experts in charge of the database, or researchers with prior experience of working with the same data set. If the biases are random in nature, it might be possible to disregard them as measurement errors. If the biases are systematic and consistent in nature, it might be possible to control for this if we know the direction of the bias.

Causal Directions and Effect Size

As pointed out in the previous section, the nature of secondary data often tells us less about the absolute number or quantity of something that we wish to know. On the other hand, longitudinal databases can, with a high degree of reliability, help us to assess how changes in one (set of) factor(s) affect another factor. What we learn is primarily about *effects,* and then secondarily, the exact magnitude of these. Therefore, despite the fact that research using databases state specific magnitudes as outcomes of their studies, more important are the general causal directions that can be determined through changes in variables over periods of time. From this perspective, findings such as Hamilton's (2000) conclusion that self-employed entrepreneurs in general pay a 25 percent premium in terms of lower long-term income is less important than the more general fact that entering self-employment has a negative effect on subsequent personal income. Levels of earnings are often measured through tax registers in ways that make it impossible to determine whether the salary came as a lump-sum payment for a short period of work or as regular wages, or if the wages came from one or several different sources. Salary levels and other observable attributes should therefore be considered "indicators" of personal earnings instead of actual levels comparable across time and individuals, a common procedure in much of sociology research (Eckhardt and Ermann 1977). This does not mean that we need to stop at simple analyses when utilizing census-like databases. For instance, after testing theoretical models and mapping causal factors on macro or meso levels, it is often worthwhile to break down more general samples of individuals based on, for example, age, sex, education, job tenure or number of firms founded. If the objects of the study are firms, these can be grouped based on industry or geographical belonging, ownership structure, age, etc. Investigating more homogenous groups of cases mean that the actual level of variables will be much more informative and comparable across individuals and time.

Validity

This section addresses the possible validity problems that were earlier described. Usually, low internal and construct validity can be dealt with through multimethod measures (Chandler and Lyon 2001). To validate information drawn from databases, Carroll and Hannan (2000, p.166) offer three suggestions. First, external information of the population (of firms, individuals, etc) that the data is drawn from can be used to authenticate the database. If external information covering the population in a database is not available, Carroll and Hannan suggests that publicly available information about a set of well-known cases might be used. However, such validation is much weaker since looking only at well-known cases will lead to undersampling of smaller or newer cases already failed or disbanded (Denrell 2003). A second approach offered by Carroll and Hannan is to compare the aggregate numbers—or marginal distribution—tabulated from a data set with numbers reported elsewhere. From the author's own experience, I would specifically suggest that distribution and rough means of key variables should be cross-checked against other sources whenever possible. In regards to firm-level databases, similar information on some or all cases might be found in industry registers, trade magazines or other types of public or semi-public sources. In regards to individual-level databases, similar information on some or all cases can often be found in public censuses. Even if census data overlapping the time period covered in a database is available only for one or a few years, the information from census data is generally broad and accurate enough to validate most individual-level data in other secondary databases. The third way of validation offered by Carroll and Hannan (2000, p.167) is to evaluate a potential data source prior to actually collecting the data by asking experts on this type of data regarding its credibility and usefulness. Such experts can be statisticians or other researchers in the field, for individual-level databases they can be sociologist or demographists, and for firm-level databases they can be historians or industry experts with a general overview of the population in question. One of the strongest validating methods would be to contact some of the cases covered in a specific database, for example, through surveys or "embedded" interviews. However, since databases with information on identifiable units (i.e., individuals or registered firms) are often anonymous and classified, contacting a few persons from the sample is not a straightforward issue. In addition, the data might be several years old and thus make validating questions of time-specific events or concepts unfeasible.

Combining Databases and Surveys

One way to obtain validating information is to use a database in combination with surveys of the same cases. In, for example, epidemiological or social medical research, there are long traditions of using established databases on a specific population or a set of patients and then combine this information with surveys sent directly to all or a (random) set of individuals drawn from the database. Linder (2004) describes how surveys can be micro-linked to administrative databases. This not only provides more detailed and specific information, but also the information is more reliable and complete when there are two or more sources with respect to the same subject. Such a procedure might be possible even if the cases in a database cannot be directly identified. Data providers such as statistical authorities can frequently administer and distribute surveys in conjunction with providing a certain data set (Petersen et al. 2004). How can this then be useful to entrepreneurship research? If we, for instance, return to the case of social medical research, it is not uncommon to use databases to identify sociodemographic conditions, such as family and labor market status and then combine this with attitude or behavior variables measured through surveys. The same kind of approach might also be productive in entrepreneurship research where, for example, economic, sociological or cognitive theories might be aligned and tested within the same empirical setting. A word of caution is required here if databases are combined with surveys measuring behavioral constructs *at the end* of the measurement period. One problem with the inclusion of behavior variables where attributes and potential outcomes are measured at different points is that since behavior is not a stable psychological construct, a person's behavioral style might have changed from the time it is measured to the time an outcome is measured (Wiklund, Davidsson, and Delmar 2003). This problem might be alleviated by using theoretically more valid operationalizations of how behavioral variables (e.g., perceptions, intentions, or self-efficacy) relate to actual actions taken by entrepreneurs (Delmar 2000; Krueger 2003).

Multilevel Analyses

Databases provide an ideal empirical setting for multilevel entrepreneurship research. Methods for such research have been utilized and discussed at length in, for instance, organizational behavior (Kozlowski and Klein 2000) and sociology research (DiPrete and Forristal 1994). Many have argued that confusion has existed in the entrepreneurship field between firm and individual levels of analysis (Aldrich 1992; Davidsson and Wiklund 2001; Sarasvathy 2004). Acknowledging this prior confusion and also the methodological difficulties in conducting multilevel analyses, Davidsson (2005) suggests thinking of an entrepreneurship research project as a single design level before starting to make crossovers to other levels. The starting point of such a design demands the predictor variables and the criterion variable(s) should refer to the same level of analysis. For example, instead of using the education of an entrepreneur (individual level) to infer the financial performance of her firm (firm level), we should use the total amount of human capital in a firm (firm level) to infer financial performance (firm level). Alternatively, we could use the education of an entrepreneur (individual level) to infer her earnings from self-employment (individual level). The cautious or less-experienced researcher would thus be suggested to start out with a more straightforward single-level research design before moving on to more advanced methods of combining and analyzing data.

Ethical Considerations

A delicate issue with regards to large-scale databases of firms, and especially individuals, is the ethical dimension. In most countries, individuals and firms are obligated to report certain types of financial information to the authorities for taxation and other purposes. This information is largely dependent on individuals' conviction that the information will not be used for purposes they disagree to. For example, census authorities often ask or demand newly registered businesses to report their current or planned "line of business," which is subsequently transformed into SIC-compatible codes by the census office. If people believe that information they give out is being used in ways they do not agree to, they might be disinclined to give out information in the future, or worse—when reporting is mandatory, they might provide inaccurate information. Careful consideration of how the subjects featured in a database would consider being part of the current research project is thus an important question. Worst case scenario: a sloppy or unethical project might damage the usefulness of important databases.

A final word on the ethics of using databases concerns the risk of "data mining" empirical material. Since database research often carries large investments in time and costs for acquiring and learning about data, researchers might be pressed to show that this was a justified investment. As discussed previously, data might not readily be used as proxies of theoretical concepts. Researchers using secondary databases therefore need to obtain information on sampling details and variable specification. Failing to do so, the researcher might find herself standing with a large amount of data with little value for the original objective. Hence the ethical dilemma: Vacuuming the material for significant correlations might eventually reveal some variable(s) that can be used to explain something vaguely related to something entrepreneurial. With larger sample sizes, the t-values used for statistical inference testing becomes larger, making it easier to reject a null hypothesis of no relationship between two variables. It is also possible to omit a variable that is found to interfere with the theoretical model, causing the variables in the tested model to be inflated and thus overestimating the effect of our model. From an ethical standpoint, all such procedures are of questionable value.

Discussion

I have argued in this article that entrepreneurship research has yet to make use of the possibilities inherent in databases. I suggested several ways to cope with the problems and practicalities of database research: using theory-driven sampling specification and variable definitions, discussing the data with experts and those familiar with it, and getting immersed in the data to learn about its possibilities and inherent limitations. To ensure validity, I particularly argued for combining different types of databases with each other or with other types of data. It is also important to uphold the higher norms of research and resist the temptation of "data mining." So, what good can these details, arguments, and suggestions put forward, do us as researchers in entrepreneurship? I will round off by giving three examples of theoretically important questions where databases might provide some answers.

As first example, the possibilities to provide analyses on entrepreneurs and nonentrepreneurs with similar skills and experiences over time might help us to pinpoint the elusive concept of "opportunity costs," which is often put forward in theoretical and empirical work as well as in public policy documents. Although opportunity cost is frequently mentioned as a possible explanation for empirical findings, there has been little research to date that explicitly investigate the existence, magnitude, and effects of opportunity costs of engaging in entrepreneurship. The studies in existence (notably Amit, Muller, and Cockburn 1995) have been relying on somewhat crude proxies, such as prior salary before engaging in entrepreneurship, as a measure of opportunity costs.

As a second example, by using database on many individuals over a longer period of time, it would be possible to look at entrepreneurs' *career performance* instead of trying to infer variables related to the individual entrepreneur to the performance of his or her *firm* (Sarasvathy 2004). This can be done by using both "long" and "broad" research designs: With a long design, data on a comprehensive set of individuals' characteristics and resources can be looked at to see how (periods of) entrepreneurship affect individual level outcomes such as long-term wealth and earnings, as well as personal health and other nontangible affects. With a broad design, individual-level data are combined with firm-level data to determine the workplaces where individuals are active as employees or entrepreneurs—and thus to test how participating in different types of entrepreneurial activities affect the long-term careers, social standing, and wealth levels of these individuals. Such an approach also has the potential to examine the long-term differences between novice and serial entrepreneurs (Westhead and Wright 1998).

As a third example, databases of (new) firms might be combined with data on patented innovations to assess how technological opportunities affect the development and performance of new firms. Save for a few studies making use of survey (Klevorick et al, 1995) or qualitative data (Shane 2000), empirical work on how different types of opportunities affect the establishment and development of new ventures is still lacking (Shane 2003; McMullen and Shepherd 2005). Based on the suggestions given in this article, secondary databases provide a source of great yet untapped value that can help us to expand the depth and scope of entrepreneurship research.

Acknowledgments

An earlier version of this article benefited from valuable comments from Per Davidsson, Johan Wiklund, Johanna Levallius, and an anonymous referee. All errors, as well as viewpoints presented in this article, belong entirely to the author. Generous funding was provided by Handelsbanken research foundations and the Swedish Institute for Labour Market Policy Evaluation (IFAU).

Endnotes

[1] By "secondary databases" I mean databases that were not collected as primary data by researchers (e.g., not data such as the PSED). Since there are much secondary (or "archival") data that could be used in research, the focus in this article is mainly on large-scale databases such as those available from public authorities and organizations. The terms "database" and "secondary database" are used interchangeably.

References

Acs, Z.J. 2004. Keynote speech at the seminar on Innovation, Entrepreneurship and Growth. November 18. Royal Institute of Technology, Stockholm, Sweden.

Acs, Z.J., and Audretsch, D. 1989. *Innovation and small firms.* Cambridge, MA: MIT Press.

Acs, Z.J., and Armington, C. 1998. Longitudinal establishment and enterprise microdata (LEEM) documentation. *Washington DC: Center for Economic Studies,* U.S. Bureau of the Census, CES 98–99.

Acs, Z. J., and Malecki, E. J. 2003. Entrepreneurship in rural America: The big picture. In T. M. Hoenig, ed., *Main streets of tomorrow: Growing and financing rural entrepreneurs.* Kansas City: Center for the Study of Rural America.

Aldrich, H.E. 1992. Methods in our madness? Trends in entrepreneurship research. In D.L. Sexton and J.D. Kasarda, eds., *The state of the art of entrepreneurship.* Boston, MA: PWS-Kent, 191–213.

Aldrich, H.E. 1999. *Organizations evolving.* London: Sage Publications.

Aldrich, H.E., Kalleberg, A.L., Marsden, P.V., and Cassell, J.W. 1989. In pursuit of evidence: strategies for locating new businesses. *Journal of Business Venturing* 4, 6: 367–386.

Aldrich, H.E., and Wiedenmayer, G. 1993. From traits to rates: An ecological perspective on organizational foundings. In J.A. Katz and R.H. Brockhaus, eds., *Advances in entrepreneurship, firm emergence, and growth* 1. Greenwich, CN: JAI Press, 145–195

Amit, R., Muller, E., and Cockburn, I. 1995. Opportunity costs and entrepreneurial activity. *Journal of Business Venturing* 10, 2: 95–106.

Birch, D. 1979. *The job generating process.* A report prepared for the Economic Development Administration, U.S. Department of Commerce.

Bouckenooghe, D., Buelens, M., De Clercq, D., and Willem, A. 2004. *A review of research methodology in entrepreneurship: Current practices and trends* (1999--2003). Paper presented November 24 at the XVII RENT Conference. Copenhagen, Denmark.

Carroll, G. R., and Hannan, M.T. 2000. *The demography of corporations and industries.* Princeton, NJ: Princeton University Press.

Carter, N.M., Williams, M. and P.D. Reynolds. 1997. Discontinuance among new firms in retail: the influence of initial resources, strategy and gender. *Journal of Business Venturing* 12, 2: 125–145.

Chandler, G.N., and Lyon, D.W. 2001. Issues of research design and construct measurement in entrepreneurship research:The past decade. *Entrepreneurship Theory and Practice,* 27, 3:101–113.

Davidsson, P. 2004. *Researching entrepreneurship.* New York: Springer.

Davidsson, P. 2005. Method challenges and opportunities in the psychological study of entrepreneurship. In J.R. Baum, M. Frese, and R.A. Baron, eds., *The psychology of entrepreneurship* (Chapter 13). Mahway, NJ: Erlbaum.

Davidsson, P., and Wiklund, J. 2001. Levels of analysis in entrepreneurship research: Current research practice and suggestions for the future. *Entrepreneurship Theory and Practice* 27, 3:81–99.

Delmar, F. 2000.The psychology of the entrepreneur. In S. Carter and D. Jones-Evans, eds., *Enterprise and small business: Principles, practice and policy.* Harlow: Financial Times, 132–154.

Delmar, F., Sjöberg, K., and Wiklund, J. 2003. *The involvement in self-employment among the Swedish science and technology labor force between 1990 and 2000.* The Swedish Institute for Growth Policy Studies (ITPS). http://www.itps.se/pdf/A2003_017.pdf.

Denrell, J. 2003.Vicarious learning, undersampling of failure, and the myths of management. *Organization Science* 14, 3: 227–243.

DiPrete,T.A., and Forristal, J.D. 1994. Multilevel models: Methods and substance. *Annual Review of Sociology* 20, 1: 331–357.

Du Rietz, A., and Henrekson, M. 2000.Testing the female underperformance hypothesis. *Small Business Economics* 14, 1: 1–10.

Eckhardt, K., and Ermann, M. 1977. *Social research methods: Perspective, theory and analysis.* New York: Random House.

Fischer, E.M., Reuber, A.R., and Dyke, L.S. 1993.A theoretical overview and extension of research on sex, gender and entrepreneurship. *Journal of Business Venturing* 8, 2: 151–168.

Frost, P.J., and Stablein, R.B. 1992. *Doing exemplary research.* Newbury Park, CA: Sage.

Gartner, W. 1990.What are we talking about when we talk about entrepreneurship? *Journal of Business Venturing* 5: 15–28.

Gentry, W.M., and Hubbard, R.G. 2004. "Success taxes," entrepreneurial entry, and innovation. NBER Working Paper #10551. U.S. National Bureau of Economic Research.

Giannetti, M., Simonov, A. 2004. *Social interactions and entrepreneurial activity.* Working Paper. Stockholm: Stockholm School of Economics.

Grégoire, D., Meyer, D.G., and De Castro, J.O. 2002. *The crystallization of entrepreneurship research DVs and methods in mainstream management journals.* In W.D. Bygrave et al., eds., Frontiers of Entrepreneurship Research 2002. Babson Park, MA: Babson College, 663–674.

Hamilton, B.H. 2000. Does entrepreneurship pay? An empirical analysis of the returns of self-employment. *Journal of Political Economy* 108, 3: 604–631.

Hamilton, B.H., and Nickerson, J.A. 2003. Correcting for endogeneity in strategic management research. *Strategic Organization* 1, 1: 51–78.

Kalleberg, A.L., Marsden, P.V., Aldrich, H.W, and Cassell, J.W. 1990. Comparing organizational sampling frames. *Administrative Science Quarterly* 35, 3: 658–688.

Katz, J., and Gartner, W. B. 1988. Properties of emerging organizations. *Academy of Management Review* 13, 3: 429–441.

Kirchhoff, B.A., and Phillips, B.D. 1992. Research Applications of the Small Business Data Base of the U.S. Small Business Administration. In D.L. Sexton and J.D. Kasarda, eds., *The state of the art of entrepreneurship.* Boston, MA: PWS-Kent, 243–267.

Klevorick, A., Levin, R., Nelson, R., and Winter, S. 1995. On the sources and significance of interindustry differences in technological opportunities. *Research Policy* 24, 2: 185–205.

Kozlowski, W.J., and Klein, K.J. 2000.A multilevel approach to theory and research in organizations. In K. J. Klein and W. J. Kozlowski, eds., *Multilevel theory, research, and methods in organizations.* San Francisco: Jossey-Bass: 3–80.

Krueger, N.F. 2003.The cognitive psychology of entrepreneurship. In Z. Acs and D. Audretsch, eds., *Handbook of entrepreneurship research:An interdisciplinary survey and introduction.* Dordrecht, NL: Kluwer, 105–140.

Linder, F. 2004 The Dutch virtual census 2001:A new approach by combining administrative registers and household sample surveys. *Austrian Journal of Statistics* 33, 1–2: 69–88.

Luger, M.I., and Koo, J. 2005. Defining and tracking business start-ups. *Small Business Economics* 24, 1:17–28.

McMullen, J.S., and Shepherd, D.A. 2005. Toward a theory of entrepreneurial action: Detecting and evaluating opportunities. *Academy of Management Review.* In press.

Petersson, F., Petersen, J. K., Schnor, O., and Husted, L. 2004. *Microdata for research and analysis: Potential and problems.* Paper presented at the Siena Group meeting on Social Statistics February 9–11, Helsinki, Finland.

Phillips, B. D., and Kirchoff, B.A. 1989. Formation, growth and survival; small firm dynamics in the U.S. economy. Small Business Economics 1, 1: 65–74.

Samuelsson, M. 2004. *Creating new ventures: A longitudinal analysis of the nascent venturing process.* Doctoral dissertation. Jönköping: Jönköping International Business School.

Sarasvathy, S. 2004. The questions we ask and the questions we care about: Reformulating some problems in entrepreneurship research. *Journal of Business Venturing* 19, 5: 707–720.

Schumpeter, J.A. 1934. *The theory of economic development: An inquiry into profits, capital, credit, interest, and the business cycle.* Cambridge, MA: Harvard University Press.

Scott, M., and Rosa, P. 1996. Opinion: Has firm level analysis reached its limits? Time for a rethink. *International Small Business Journal* 14, 4: 81–89.

Shane, S. 2000. Prior knowledge and the discovery of entrepreneurial opportunities. *Organization Science* 11, 4: 448–469.

Shane, S. 2003. *A general theory of entrepreneurship: The individual-opportunity nexus.* Aldershot, UK: Edward Elgar.

Sørensen, J., and Phillips, D. 2004 Do small firms produce better entrepreneurs? Working Paper. Sloan School of Management, Massachusetts Institute of Technology.

Storey, D.J., and Johnson, S. 1987. *Job generation and labour market changes.* London: Macmillan.

Tronti, L., Ceccato, F., and Cimino, E. 2004. Measuring atypical jobs: Levels and changes. Suggestions for a new classification of non-standard employment arrangements. OECD Working paper 2004/01, OECD statistics.

Van de Ven, A. 1992. Longitudinal methods for studying the process of entrepreneurship. In D.L. Sexton and D. Kasarda, eds., *The state of the art of entrepreneurship.* Boston, MA: PWS-Kent, 214–242.

Venkataraman, S. 1997. The distinctive domain of entrepreneurship research: An editor's perspective. In J. Katz and R. Brockhaus, eds. Advances in entrepreneurship, firm emergence, and growth 3. Greenwich: JAI Press: 119–138

Watson, J. 2003. Failure rates for female-controlled businesses: Are they any different? *Journal of Small Business Management* 41, 3: 262–277.

Westhead, P., and Wright, M. 1998. Novice, portfolio and serial founders: Are they different? *Journal of Business Venturing* 13, 3: 173–204.

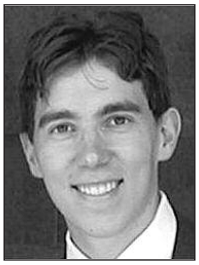
Wiklund, J. 1998. *Small firm growth: Entrepreneurship and beyond.* Doctoral dissertation. Jönköping: Jönköping International Business School.

Wiklund, J., Davidsson, P., and Delmar, F. 2003. What do they think and feel about growth? An expectancy-value approach to small business managers' attitudes toward growth. *Entrepreneurship Theory and Practice* 27, 3: 247–270.

Williams, M. L. 1993. Measuring business starts, success and survival: Some database considerations. *Journal of Business Venturing* 8, 4: 295–299.

Wooldridge, J. M. 2002. *Econometric analysis of cross section and panel data.* Cambridge, MA: The MIT Press.

About the Author



KARL WENNBERG (karl.wennberg@hhs.se) is a Ph.D. candidate at the Center for Entrepreneurship and Business Creation, Stockholm School of Economics. His dissertation work combines individual- and firm-level databases to investigate entrepreneurial performance and exit from a behavioral perspective.