An Assessment of the Multi-level Integrated Database Approach

Tom W. Smith
Jibum Kim

NORC/University of Chicago

GSS Methodological Report No. 116

November, 2009

The multi-level integrated database approach (MIDA) is an innovative procedure to improve survey-research in general and in particular to assess and adjust for non-response bias. This paper 1) describes MIDA, 2) outlines its utility for a) data collection, b) non-response measurement and adjustment, c) interviewer validation, and d) substantive, contextual analysis, 3) discusses how MIDA extends beyond existing practices, 4) demonstrates the construction of MIDA-enhanced sample frame for US households, and 5) indicates further steps for the use and testing of MIDA.

Outline of MIDA

The essence of MIDA is to use databases to collect as much information as practical about the target sample at both the household-level and at various aggregate levels during the initial sampling stage.

The first step in MIDA is to extract all relevant, public information at both the household-level and aggregate levels from the sampling frame from which the sample addresses are drawn. In European samples based on population registers, there is often very useful information on such matters as gender, age, and household composition (Bethlehem, 2002; Stoop, 2004; van Goor, Jansma, and Veenstra, 2005; Voogt and Van Kempen, 2002) and list samples (e.g. of employees and HMO enrollees) often have a wealth of sampling frame information (Fowler et al., 2002; Groves, 2006; Kennickell, 2005; Lessler and Kalsbeek, 1992; Moore and Tarnai, 2002; Smith, 1999). But in the US, general population samples of addresses are typically nearly void of household-level information. However, US address samples are rich in aggregate-level information. Address/location of course is the one known attribute of all cases, whether respondents or non-respondents. Moreover, address-based sampling frames are typically based on the US Census and as such the appropriate Census data from blocks, tracts, place, etc. are part of the sampling frame and linked to each address. (That is, the local sample points are selected based on the Census and then addresses within those sample points are obtained from the United States Postal Service Delivery Sequence File and/or special field listings, the later especially typical for rural areas (O'Muircheartaigh, 2003).)

The second step is to augment the sampling frame by linking all cases in the sample to other databases. As Groves (2005) has noted, "Collecting auxiliary variables on respondents and nonrespondents to guide attempts to balance response rates across key subgroups is wise."

At the household-level that means linking the addresses to such sources as telephone directories, credit records, property records, voter-registration lists, and other public sources (Berge et al., 2005; Brick et al., 2000; Cantor and Cunningham, 2002; Cox, 2006; Davern, 2006; Johnston et al., 2000; Marcus at al., 2006; Williams et al., 2006).[1] A number of special procedures have also been developed to use databases in ways not commonly expected and thereby extract much more information than available

[1] For a general discussion of record linkage involving surveys see Fair, 1996 and Jenkins et al., 2005. On linking surveys to administrative records see Obenski, 2006 and Davern, 2006.

from more limited and superficial applications (Cantor and Cunningham, 2002; Smith, 2006a; Traub, Pilhuj, and Mallet, 2005; Williams et al., 2006).[2]

The information obtained would include first of all whether a match was or was not found (e.g. listed in the telephone directory or not; has registered voter or not) and, if matched, whatever particular information is available (e.g. names, telephone numbers, voter registration status).

At the aggregate level, this means merging information from sources other than those in the sampling frame.[3] Examples of aggregate-level data beyond that from the Census that could be appended are consumer information from such sources as sociodemographics such as Claritas' PRIZM NE, voting information from national elections, and data on such other matters as vital statistics (Salvo and Lobo, 2003); crime rates (FBI, 2004), magazine subscriptions (Audit, 2005), religion (Jones, 2002), public housing (HUD, 1998), HIV/STD rates (CDC, 2004), and public welfare utilization (Salvo and Lobo, 2003).

The linked data would include information from multiple-levels of aggregation. The multi-level analysis will start with household-based data and include neighborhood-level data from Census track and zip code-based data sources, community-level data from the Census, election counts, crime rates, and other sources, and higher level aggregations (e.g. metropolitan areas and Census divisions).[4]

The third step in MIDA is to take information gained from the initial household-level linkages to secure additional information. For example, securing a name and telephone number from a telephone-directory search can lead to households being found in databases when a mere address was insufficient to allow a match. Also, once a respondent was identified, links to that person in addition to household-level matching could be carried out. Thus, the process of augmenting the sampling frame is iterative and continues during the data-collection phase.

The final step is to record, process, clean, and maintain a large amount of paradata for each case (Couper and Lyberg, 2005; Scheuren, 2000). This would include having interviewers systematically record information about the sample residence (e.g. dwelling type, condition of dwelling), contacts or call attempts, interactions with household members, and observations on the composition and demographics of the household) (Bethlehem, 2002; Cantor and Cunningham, 2002; Gfroerer, Lessler, and Parsley, 1997; Groves, 2006; Kennickell, 2005; Lynn et al., 2002; Safir et al., 2002; Smith, 1983; Stoop, 2004).[5] As Cantor and Cunningham (2002) note, surveys "should maintain the date and result of each contact or attempt to contact each subject (and each lead)... The reports

[2] Survey researchers already have considerable experience in using databases and other experts include data librarians, geographical information systems specialists, cyber-information technicians and data miners, and records searchers such as paralegals and investigators.

[3] When starting with addresses without prior Census information as part of the sampling frame, Census and other geographic-based information can be obtained by linking addresses to the geo units (e.g. Census tract, zip code, place/community, etc.) that they fall in. That is, the Census data are added as part of step two if they are not already available as part of the sampling frame. Address linkages to Census tract and higher geo units are possible for from 95-100% of cases (Geronimus, Bound, and Neidert, 1996; Groves and Couper, 1998; Kim, Smith, and Sokolowski, 2006).

[4] For multi-level analysis see Bryk and Raudenbush, 1988; DiPrete and Forristal, 1994; and Raudenbush and Bryk, 2002.

[5] This is obviously not possible for postal surveys.

should provide cost and hit data for each method to help manage the data collection effort. In the end it helps to determine those methods that were the most and least cost effective for searching for the population of interest, and this knowledge can be used for planning future surveys." For much of paradata a particular advantage is that information exists for both the non-respondents as well as the respondents and thus can be readily utilized to examine non-response bias.

The Utility of MIDA

Consider how the multi-level information in this greatly enriched, sampling frame can be used to advantage for data collection, non-response measurement and adjustment, interview validation, and substantive analysis.

Data Collection

First, more information on the target sample will make data collection both more efficient and more effective. For example, securing names and phone numbers can be very helpful in making contact with households and are particularly useful in the case of locked building, gated communities, and other hard to access residences. More information about households before the start of the data-collection phase can greatly ease making contact with households and thus allow efforts to be concentrated on gaining respondent cooperation. It is also very useful if a multiple-mode approach is used (e.g. data collection combining in-person + telephone).

Once contact is made, tailoring is very important in gaining cooperation (Couper and Groves, 1996; Groves and Couper, 1998; Smith, 2007). The more information that one has about the household (e.g. whether they have a listed phone number, home owner or renter, etc.), the better able one is to shape interviewers' approaches and to provide and highlight information most salient about the sampled household (Groves, 2006; Groves, Singer, and Corning, 2000). It is not that well-run surveys do not already make some use of databases to assist interviewers, but what is typically not done is the careful evaluation of various databases and the retention of the information for other than data-collection efforts.[6]

Non-Response Measurement and Adjustment

Second, while this added information will assist interviewers and decrease the overall non-response error, there will still remain a notable amount of non-response on even the better surveys. The information in the MIDA-augmented sampling frame will

[6]While databases have been used for some time to assist surveys, their use has been informal and underdocumented. For example, many RDD survey routinely run telephone numbers by business lists of phone numbers, but this is often not mentioned and providing details on the removal rate is even rarer (Merkle et al., 2009). On the General Social Survey, hard-to-contact households are routinely searched for to get a name and/or phone number to aid in making contact with and obtaining an interview from the cases. However, no systematic record of the searches or their outcomes has been maintained.

then be used to measure and adjust for non-response error.[7] Having a wide range of household-level and aggregate-level information is important both to test the representativeness of the achieved sample across as many variables as possible and because surveys covering different topics are likely to have different non-response profiles (e.g. non-voters under-represented in political surveys and the wealthy in the Survey of Consumer Finance – Kennickell, 1997; 2005). Having more relevant information on non-respondents allows for better modeling of non-response bias and the creation of weights that more fully account for the biases and has the particular advantage of having augmented data for all sample cases (Groves, 2005a). It also makes fresh, cross-sectional studies more like reinterview, panel studies where the bias from attrition can be well-modeled based on time 1 data (Lepkowski and Couper, 2002).

Research has shown that neighborhood, community, and higher level attributes of areas are correlates of non-response. For example, non-response is consistently and notably higher in large cities than in small towns (Groves and Couper, 1998; Smith, 1983; Smith, 1984; Steeh et al., 2001), in some regions and metropolitan areas vs. others (Groves and Couper, 1998; Johnson and Cho, 2004; Lepkowski and Couper, 2002; Montaquila and Brick, 1997; Murray et al., 2003; Smith, 1983); and related to other aggregate-level attributes such as density, crime rate/fear of crime, social disorganization, geographic mobility, and family structure (Couper and Groves, 1996; Groves, 2006; Groves and Couper, 1998; Goyder, Lock, and McNair, 1992; Gfroerer, Lessler, and Parsley, 1997; Johnson and Cho, 2004; Johnson et al., 2006; Kim, Smith, and Sokolowski, 2006; Kojetin, 1994; O'Hare, Ziniel, and Groves, 2005; van Goor, Jansma, and Veenstra, 2005; Voogt and van Kempen, 2002). Thus, aggregate-level variables are very useful for assessing, understanding, and adjusting for non-response bias (Brick and Broene, 1997; Johnson and Cho, 2004; Kalsbeek, Yang, and Agans, 2002; Kennickell, 2005; Montaquila and Brick, 1997; Nolin et al., 2000; Turrell et al., 2003).

While MIDA is designed to address the matter of nonresponse bias in general, special attention can be focused on examining several prominent theories about the nature and source of nonresponse bias: social disorganization, social isolation, overextension, and structural barriers.

First, social disorganization theory holds that social structural conditions influence the social relations of people. Wirth (1938) notes that population size, density, and heterogeneity accompanying urbanization weaken individual, family, neighborhood, and social ties. Shaw and McKay (1969) show an association between certain structural conditions and the concentration of social ills such as delinquency. They attribute the higher prevalence of social ills in socially and economically disadvantaged areas to the differences in social organization in the community. Treating refusal rates in Primary Sampling Units (PSUs) "as a behavioral measure of interpersonal trust or helpfulness," House and Wolf (1978:1030) show a positive relationship between crime rate and refusal rate, and find that the total crime rate provides the strongest positive explanatory power on variation of refusal rates among different places. Groves and Couper (1998) show that, controlling for household characteristics, population density and the percentage of

[7] It is likely that some information will be most valuable at the data-collection stage and other at the non-response adjustment stage. For example, name and telephone number would be most useful to aid the field work and having a listed/unlisted telephone number, mobility history, and housing tenure would likely be more valuable for non-response adjustments.

individuals under 20-years of age are positively related to survey cooperation. The individual and especially the aggregate level data collected here will provide multiple measures of social disorganization (e.g. crime level, concentration of poverty, residential instability).

Related to social disorganization theory is the concept of collective efficacy which holds areas vary in the willingness of people to "intervene on behalf of the common good" (Johnson et al., 2006; Sampson, Morenoff, and Earls, 1997). Collective efficacy is related to such neighborhood traits as low population turnover, higher education, higher income, low density, fewer immigrants, and more intact families. Research has found that this propensity is related cooperation in surveys (Couper, Singer, and Kulka, 1998).

Second, social isolation theory argues that nonrespondents are likely to be poorly integrated members of society (Groves and Couper, 1998; Looseveldt and Carton, 2001; Stoop, 2005). According to this theory social isolates are likely to be non-respondents both because of personal misanthropy and because of social and civic disengagement. Personally, social isolates try to minimize inter-personal contacts with others and as such are disinclined to want to cooperate with and engage in an interview (i.e. a conversational interaction) with an interviewer (Converse and Schuman, 1974). Socially and civically, social isolates have little interest in general societal and community affairs and neither follow such matters nor are interested in discussing such topics in an interview. Thus, for these distinct, but associated, reasons social isolates are expected to be overrepresented among non-respondents. It will be possible to examine these expectations by both comparing households that are socially isolated (e.g. with no listed number, no members registered to vote nor belonging to large voluntary associations, etc.) to less isolated households and by comparing more engaged areas (e.g. higher voter turnout, more magazine/newspaper subscriptions) vs. less involved neighborhoods and communities.

Third, overextension theory argues that it is people leading busy lives that tend to be non-respondents (Campanelli, Sturgis, and Purdon, 1997; Groves and Couper, 1998; Lynn, 2002; Smith, 1984). This would include people working full time in general and especially those putting in over time, those with open-ended management responsibilities, and those whose work involves travel. It would naturally include people with multiple, major roles such as full-time employees and parents of small children or those providing in-home eldercare. Databases can often provide useful information on employment status and household composition that can be used to test this hypothesis.

Additionally, many structural factors such as gated communities, locked buildings, policies of gatekeepers, etc. influence contact rates and ultimately response rates and these can be observed and recorded by interviewers and examined by researchers. Including these structural impediments and other paradata will better specify the overall non-response model.

Interview Validation

Interviews are checked or validated through a combination of close supervision of field interviewers, the recontacting of respondents to verify that an interview had been conducted with the eligible respondent, and computer audio-recorded interviewing (Smith and Sokolowski, forthcoming). Invalid interviews are a relatively small

component of total survey error (Smith, 2005). MIDA can reduce it even further by allowing the information from the databases to be used along with recontacts to help corroborate that interviews were truly and correctly done.

Substantive Analysis

Finally, for respondents the household-level and aggregate-level data in the augmented sampling frame can be utilized for crucial substantive analysis. While most household-level information would come from the interviews with the respondents, household-level data would be supplemented with information from the augmented sample frame. Data from the database-augmented sample frame can be used to a) add information not covered by the survey, b) supply missing data for variables that are covered by the survey, and c) corroborate information reported by respondents.[8] Procedures for cross-checking information from different databases and between databases and surveys are discussed below.

Aggregate-level information is of great utility for research. Research has demonstrated that contextual, aggregate-level geographic effects in general and neighborhood characteristics in particular influence a wide range of attitudes and behaviors independent of the attributes of individuals. For example, research has shown that impacts exist on 1) political involvement (Bobo and Gilliam 1990; Cohen and Dawson 1993; Gilbert 1991), 2) residential and social mobility (Lee, Oropesa, and Kanan 1994; Massey and Eggers 1990; Massey et al. 1994; South, Baumer, and Lutz 2003), 3) the sexual and reproductive activities of youths and adults (Billy and Moore 1992; Brewster 1994a; Brooks-Gunn et al. 1993; Browning and Olinger-Wilbon 2003; Browning, Leventhal, and Brooks-Gunn, 2004; Cohen et al 2000; Crane 1991; South and Baumer 2001), 4) responses to poverty (Jencks and Mayer 1990; McLeod and Edwards 1995; Oreopoulos 2003), 5) racism and tolerance (Gibson 1995), 6) fear of and involvement in crime (Covington and Taylor 1991; Peeples and Loeber 1994; Sampson, Raudenbush, and Earls, 1997), 7) minorities politically (Cohen and Dawson 1993), economically (Lee et al. 1994; Massey and Eggers 1990), and in other ways (Brewster 1994b; Smith 1994a), 8) social capital and better health (Mellor and Milyo 2004), 9) group membership and economic improvement (Tolbert, Lyson, and Irwin 1998); 10) inequality and political trust (Rahn and Rudolph 2005); 11) religion and deviant behavior (Regnerus 2003), 12) drug use (Boardman et al. 2001; Ford and Beveridge, 2006;Galea, Ahern, and Vlahov, 2003; Snedker, Herting, and Walton, 2006), and 13) depression (Latkin and Curry, 2003).

Among the contextual effects that have been examined from the General Social Survey (GSS)(Davis, Smith, and Marsden, 2009) specifically are the following: 1) racial composition of the local population predicts levels of racial prejudice (Alesina and LaFerrara 2000; Charles 2003; Dixon and Rosenbaum 2004; Taylor 1998 and 2002) and class voting (Weakliem 1997), 2) higher collective levels of trust and civic engagement are associated with lower homicide rates (Rosenfeld et al. 1999 and 2001) and lower mortality in general (Kawachi et al. 1997b), 3) areas with greater aggregate happiness have lower mortality (Jencks 1999), 4) higher levels of anomia are related to higher local

[8]Examples of collaboration are the voter validations studies – Anderson and Silver, 1986; Burden, 2000; Silver, Anderson, and Abrahamson, 1986.

crime rates (Rosenfeld and Messner 1998), 5) community-level differences in attitudes on gender roles do not affect the demand for female labor (Cotter et al. 1998), 6) the prevalence of Fundamentalists reduces support for feminism (Moore 1999), 7) a higher level of people on welfare reduces support for welfare spending (Luttmer 1998), 8) living around gun owners increases one's likelihood of acquiring a gun (Glaeser and Glendon 1998), 9) lower income equality is associated with lower social trust and group membership (Kawachi et al. 1997a), 10) community heterogeneity influences civic engagement (Costa and Kahn 2002), 11) community norms shape attitudes toward capital punishment (Baumer, Messner, and Rosenfeld 2003), 12) state and regional differences may be declining over time (Weakliem and Biggert 1999), 13) voting and civic involvement vary by community as well as individual demographics (D'Urso 2003), 14) greater community acceptance of immigrants relates to more occupational achievement by immigrants (De Jong and Steinmetz 2004), 15) community religious beliefs and behaviors influence gender roles (Moore and Vanneman 2003), and 16) aggregate public opinion affects public policies on such as abortion laws, welfare payments, and AIDS-related funding (Brace et al. 2002).

The coding of a rich array of aggregate-level data from the sampling frame and a wide range of databases can facilitate such contextual analysis and make it a regular part of survey analysis rather than an occasional approach carried out only when special multi-level data are added, often after the fact, to standard surveys. Rather than involving an extensive, extra, post-hoc effort, the contextual data would be pre-collected for the entire frame and thus automatically available for contextual analysis. In brief, the information in the augmented sampling frame that can be used to assist data collection and adjust for non-response bias can in turn be used for multi-level, contextual analysis.[9]

MIDA Expansion over Existing Practices

While all of the elements of MIDA have been used in some way or another in some existing surveys, the use of household-level and aggregate-level linkage to databases has not be used in an integrated, systematic manner. The use of databases has been quite limited in terms of both what sources are used and how the linked information is utilized, and the databases and the information from them have not be assessed and evaluated.

One of the limitations of existing approaches is that databases are not used in a systematic manner. For example, telephone directories are often used to try and find the name and number associated with a sampled address or to track a respondent in a panel who has moved. The telephone-directory searches are often quite helpful for these purposes, but their use is purely operational. The information gathered is used by interviewers to help locate respondents, but seldom, if ever, systematically analyzed, used for non-response adjustment, or retained as part of the final analysis file. Conversely,

[9] What is important information will depend in part on at what stage and how it is being used. At the initial interviewing stage finding a name of the likely resident and his/her phone number can be very useful in making contact and gaining cooperation. In a panel survey fining a current address for a mover or the person's place of employment is valuable for locating them. For assessing non-response bias, name and actual telephone number are of little use, while information on age, gender, voting status, etc. can be very useful to check the representativeness of the interviewed sample.

linkage data are sometimes collected for substantive purposes (e.g. to see if graduates of a job-training program end up on welfare), but this information is not used for field operations or non-response adjustment purposes.

A second limitation is that the use of different databases has apparently never been systematically assessed. Different practitioners use different data sources (e.g. telephone directories, credit records, various public, governmental files) based on their familiarity with data sets and/or the data providers and other general preferences. Apparently no rigorous comparisons of the ease-of-use, cost, and yield of various databases have been conducted and none have closely examined the cumulative gain from the use of multiple data sets (Smith, 2006a).

A third limitation has been that few databases have been typically utilized. Telephone directories are the only commonly used database. Other databases such as credit records, property records, and voter registration have been used only occasionally (and only for limited purposes when used at all). Many other potentially valuable databases have apparently never been used (e.g. political contribution lists, membership lists, subscription lists).

A final limitation is that the uses of databases have generally focused on only information obtained about respondents who are found in particular sources. Typically, searches in telephone directories are deemed useful when the target individual or household is located and as not useful when no match occurs (as is the case with the large proportion of households with unlisted numbers plus those with no telephone). But being found or not found in a database is in itself a useful piece of information and should be recorded for comparing respondents and non-respondents. For example, those listed in the telephone directory are much more likely to be respondents than those not included (Brick et al., 2003; Brick, Montaquila, and Scheuren, 2002; Harvey et al., 2003; Kennedy, Keeter, and DiMonk, 2008; Minato and Luo, 2004; O'Hare, Ziniel, and Groves, 2005).

MIDA is designed to overcome each of these standard limitations by comparing and evaluating data sources, flagging both matched and unmatched records, and retaining data for use in all phases of research.

Constructing a MIDA-enhanced Sample Frame

Sample

To test MIDA NORC's sampling statisticians drew a sample of 400 addresses clustered in 40 segments that was a) a representative sample of addresses from NORC national sample frame and b) similar to the type of sample used in NORC's General Social Survey and other national samples selected by NORC. These 40 segments included one or more addresses from 54 different zip codes. NORC's sample frame utilizes a multi-stage, area probability design in which national frame areas (NFAs) are selected, segments are selected within the NFAs, and addresses are selected within the segments (Davis, Smith, and Marsden, 2009). For sampling strata representing urban areas covered by city-style addresses (i.e. with street name and number), addresses were selected from the Delivery Sequence File (DSF) compiled and maintained by the United States Postal Service (USPS). Access to the DSF is provided to users via USPS Certified

DSF[2] Licensees. In NORC's case they obtained the sample from ADVO (now Valassis). For the strata with more rural population and much less complete coverage by city-style addresses, NORC sent field enumerators to the segments to compile their own address/location lists. The strata covered by the DSF list represented about 85% of the US population and the NORC-listed areas covered the other 15% of the population (Davis, Smith, and Marsden, 2009; Harter, Eckman, English, and O'Muircheartaigh, 2008).

Augmenting the Sample Frame

The first step is to extract all useful information from the sample frame. In this case the sample frame is constructed from two sources, the US Census and the United States Postal System's Delivery Sequence File (DSF) as augmented by NORC's own address listings. The Census provides a wide range of demographic data from block groups, tracts, places, counties, and metro areas. What is available depends on the geographic unit with more detailed information being released for larger units. For Census tract hundreds of demographic breakdowns by such variables as age, race, ethnicity, gender, marital status, household size, income, labor-force status, education, etc. are available. (Gatewood, 2001).

From the DSF provided by ADVO there are 41 variables for each address. About 16 deal with the addresses themselves (zip code, street number, walk sequence, etc.) Many other variables deal with the status of the address such as seasonal delivery, vacancy, "do not deliver", "address type", and business/residential. Still other variables cover how the mail is actually delivered to the address (e.g. OWGM_Indicator, Record Type Code, Delivery Point Type Code). Besides defining the location of the address, these auxiliary variables provide important information about the nature of the unit at the address such as whether it is a vacant HU (7 so listed), a season HU (6), or a throwback address in which mail is delivered to a PO box instead of the street address (none). These and other characteristics are likely to correlate with and predict final disposition status of sample addresses used in an actual survey.

Aggregate-Level Data

Since the address and geographic location of the sampled cases are known, all of them can be linked to aggregate-level data that are tied to location. Table 1 lists the aggregate-level sources that cases were linked to. While they represent a wide range of sources and variables, they are not exhaustive, but rather are illustrative of the type of information what can be compiled. What can be linked and at what geo-level depends on how the aggregate sources are geographically organized and coded. The sources may have information coded according 1) to exact location in terms of either a) longitude and latitudes or b) address or 2) aggregated into various units such as zip code, Census categories such as block group and tract, and political units such as place/community and county.

Global Information Systems (GIS)

One way to link aggregate-level information to addresses is via GIS. The longitude and latitude (L/L) of all sample addresses is known and this enables the addresses to be linked to any other data source that has L/L coded. A large and growing amount of information is available in GISs. It is possible to code Euclidean distances between sampled addresses and various targets (e.g. nearest school or superfund site) and to categorize these into discrete categories (e.g. within a mile, 1-9 miles, 10+ miles). Alternatively, instead of using Euclidean distance, travel-times via the road network can be calculated and used as either continuous or categorized variables.

Among the L/L linked facilities are hospitals, trauma centers, schools/colleges/ universities, places of worship, government offices, cemeteries, golf courses, cultural centers such as museums and zoos, major retails centers, transportation hubs, airports, prisons, military installations, parks and recreational areas, Superfund sites, public housing units, power plants, and rivers/lakes (Table 1). Examples of studies doing this include Branas et al. (2005) on trauma centers, Downey (2006) and Holmes (1999) on employers, and Slvo and Lobo (2003) on various governmental measures.

Of course GIS-based mapping programs such as Google Maps and MapQuest have a wide range of commercial and non-commercial sites that can be linked to given addresses and stratified by distance. For example, tests of dermatologists, drug stores, churches, synagogues, tattoos, firearms, pizzas, schools, and nails all produced reliable results. (It is noteworthy that "nails" located nail salons (as intended) and not hardware stores.) However, no way is known of using GIS-based mapping programs in batch mode to search across all target addresses for a given type of site. Nor does one get estimated travel times or distances unless one does a follow-up search on each initial hit.

Addresses

Other information is identified by addresses, but no GIS data are included. As long as the addresses are in city-style, they could be converted to L/L using ArcGIS or a similar routine and then handled as other GIS-based data. Examples are a national list of correctional facilities, political contributions under federal election law, and the not-for-profits list maintained by the IRS and provided to users by the National Center for Charitable Statistics (Table 1).

Small-Level Census Categories

As discussed above, the Census was one of the original sources of information in the construction of the sample frame and thus all public Census data are available.[10] What is released depends on geographic level with less detail available at lower level like block and progressively more for larger geographic units such as tract, place/community, county, metropolitan area, etc. (www.census.gov).

Zip Codes

Many databases provide data aggregated at the zip code level. These include www.zip-codes.com, www.zipcodeworld.com, www.melissadata.com, and

[10] This includes the American Community Survey as well as the decennial Census.

www.zipcodestogo.com. Most provide only basic location information and a limited range of Census-based demographics. Limited extra data are appended to a few databases (e.g. on income tax refunds and returns in www.zipcodes togo.com). In addition, to these zip-code-centric databases, many other sources also aggregate by zip code (often along with other geo-units like block group, Census track, community/place, and county). From the government the Census itself has regrouped the 2000 population data into zip codes (censtat.census.gov) and also organized the 2002 economic Census by zip code (http://censtat.census.gov; www.census.gov/geo/ZCTA/zcta.html ; www.census.gov.epcd/www/zipstats.html). The Internal Revenue Service makes income tax information available by zip code and higher aggregations (www.irs.gov/taxstats/indtaxstats/article/0,,id=96947,00.html). The Environmental Protection Agency has some zip-code-level data and more for larger geo-units (www.epa.gov/epahome/commsearch.htm; www.epa.gov/air/datahelp/hziploc.html. Political contributions are aggregated by zip code by number of Democratic and Republican contributors and total amount donated (fundrace.huffingtonpost.com) (Table 1).

Private-sector products are also available (Table 1). Claritas (www.claritas.com), for example, offers a wide range of Census-based demographic data and commercial data aggregated by zip codes and other geo-units. Likewise, MarketPlace by Dunn and Bradstreet (www.sales-tools.com) codes over 10 million US businesses on over 40 variables by zip code and county (Powell et al., 2006).

Zip codes are widely used in epidemiological and medical research (Grubesic and Matisziw, 2006; Krieger et al., 2002a; Krieger et al, 2002b), most commonly to relate the distribution of diseases and medical diagnoses by socio-economic levels (Pappas et al., 1997; Thomas et al., 2005) or population levels (Peel et al., 2005).

When zip code data are based on matches to the US decennial Census, the assignment of zip code values depends on the matching of block-group data to the larger zip code areas. While reliable procedures from such assignment have been developed, there is an element of estimating and approximating in aggregating Census data by zip codes or what the Census designates as ZIP Code Tabulation Areas (ZCTAs). On such matching and the error associated with same see (www.census.gov/geo/ZCTA/zcta.html ; www.census.gov.epcd/www/zipstats.html; Grubesic and Matisziw, 2006; Krieger at al., 2002a, 2002b; Thomas et al., 2006). When the original data are collected by zip code or by using address or GIS data that can be definitively assigned to zip codes, the assignment error that occurs with retrofitting Census data into zip codes does not occur.

County/Metropolitan Areas

A very wide range of information is available by county and metropolitan area. Of course any information collected at a more detailed level such as Census tract or zip code can be aggregated at the county- or metro-level. These include information from both the decennial population and economic censuses. Additional data available at the county/metro area include denominational adherents, circulation levels for periodicals, reported crimes and arrests, election returns, and various variables relating to governing such as about taxes, public expenditures, Medicare enrollment, and building permits (Table 1).

Regarding election returns, while votes at the country level are easily assessable, votes for communities, precincts, and other smaller units are much less available (Committee, 2008). The situation various greatly from state to state. In Indiana the Board of Elections maintains sub-county results only on hard copy and nothing is available on the web. Pennsylvania has detailed figures on-line, but one needs to go separately to each county's site to gather the results. Iowa has precinct level results by county at a centralized site. It appears that no national compilation of community- or precinct-level votes has been compiled. Even when available, the sub-county data are challenging to work with. While it is relatively easy to match addresses at the community level with voting results, identifying what precinct an address is in is much more difficult.

Geodemographic Segmentation

Geodemographic segmentation, sometimes referred to as geodemography, is a multivariate classification procedure for dividing areas into distinctive socio-demographic types. It is especially used in marketing to define lifestyle groups that would help in the targeting of particular types of consumers. The most widely used systems are Prizm NE by Claritas, Mosaic by Experian, and Community Tapestry by ESRI.com. Prizm NE has 66 cluster types, Community Tapestry has 65, and Mosaic 60. For example, Prizm NE assigns types to individual households based on a combination of Census data down to the block group level and the unspecified use of other non-Census information.

Address-/Household-Level Data

While aggregate-level data are available for all cases, information for particular addresses or households depends on them being linked to various databases. Household-level linkages are only possible when dwellings have city-style addresses (a unit number and street name). Traditionally, these are absent in many rural areas. Fortunately in part to ease the tasks of first responders in general and as part of the enhanced 911 effort in particular, city-style addresses are being assigned to more and more dwellings (DiSogra, Callegaro, and Hendarwan, 2009). This effort is coordinated by the LACSLink system of the United States Postal Service (Key and Miracle, 2009). In our sample of 400 addresses, 95.8% had searchable/street-style addresses. It is anticipated that this level will be even greater in the future.

It is also noteworthy that initially an additional 3% of addresses could not be linked. These were all from one rural segment which had a place name associated with it that differed from that used in other sources. While it was easy to determine that an alternative place name applied for these cases and they were then linkable, this cautions that alternative or changed names may be used across different sources and one needs to be sensitive and flexible to such variation when matching addresses.

General Record-Linkage Issues

Searches of and linkages across databases depend in part on the search algorithms that are employed. These are rarely, if ever, explicitly specified and one learns about

some of their characteristics from usage. Most limiting are those that require an exact match across all elements of the search terms and the targets in the database. For example, a restrictive system would be one that used the full word “Street” in its records and would not recognize “street” or “St.” as matching terms to it. Other examples would be systems that would not match the search for “234 Maple” when its records contained “234 Maple St.”. Alternative or misspellings are another example. Many systems require that place names be spelled exactly the same so that “Olde Boalsburg Road” would not match “Old Boalsburg Road” or “Pittsburg” would not be linked to “Pittsburgh”. But other systems locate phonetic equivalents and would have matched the two previous examples. Of course if a misspelling is too divergent even algorithms allowing for phonetic equivalence/similarity will not find matches. Additionally, very permissive algorithms will increase false positives (e.g. confusing “234 Maple [Lane]” with“234 Maple Drive”). However, given the procedures actually used, false negatives appear much more common than false positives.

When a match is made, the next issue is whether information on the current resident(s) can be identified. Multiple listings per address are the norm and they may or may not include the current resident(s). Dates in which a named person is associated with an address are frequently available, but they often end short of the current date, are often incomplete, and are sometimes contradictory. In most cases the likely current resident(s) can be identified. For additional cases two or more possible current residents are suggested. For a small number of cases, no likely current resident(s) appears (i.e. at least one name is associated with the address, but there is no evidence that that person currently resides there).

Then for the current resident(s) the issue becomes what additional information is available. For the reserve-directory searches, the situation is simple; a found address generates a name and phone number. For many other databases, a wide range of potential information may be available. Some such as gender may exist for almost all linked names, others such as age/date of birth may be available for substantially fewer people, and still others like SSN may be relatively rare.

Finally, when information is available, the issue is whether it is accurate. The legal disclaimer by one database illustrates this point, “Important: The Public Records and commercially available data sources used in this system have errors. Data is [sic] sometimes entered poorly, processed incorrectly, and is generally not free from defect. This system should not be relied upon as definitively accurate. Before relying on any data this system supplies, it should be independently verified. For Secretary of State documents, the following data [sic] is for information purposes only and not an official record. Certified copies may be obtained from that individual state’s Department of State.” Errors can consist of information that is simply wrong (e.g. transposed numbers; misreports in original records used by the databases) or out-of-date (e.g. employment information that refers to a previous job; marital status before a recent marriage/divorce).

Multi-unit addresses

Addresses with multiple housing units at the same street address present special challenges. These include everything from duplexes to very large apartment complexes and represent all situations when two or more different units have the same street number

for their address. (For brevity sake these various types of multi-units will be referred to as “apartments”.) First, some multi-unit addresses, especially smaller units like duplexes, do not have official or regular unit designations such as numbers, letters, or number-letter combinations for the units. They may have no designation or only informal and irregular designations (e.g. “rear”, “basement”). Second, many databases do not recognize apartment numbers as a searchable fields and thus it is often not possible to search for specific apartments. Third, even when regular unit designations exist and such are a field in a particular database, the full information may not exist for units or people associated with a particular street number. Thus, for people associated with a particular street number, some may have an apartment number listed and others have no specific listing. When an apartment is searched for in various databases, one will often get a list of all people associated with that street address regardless of their apartment number or one name will be reported following some default heuristic. For example, Accurint will report the name of the person with the surname that comes first in the alphabet.

In the sample of 383 searchable addresses, 24.2% involved multi-unit addresses.

Using Record Linkage Outcomes

The initial result of any database search is that the target address is either found or not found. Except for addresses in unsearchable, non-city-style addresses, not found would generally indicate that the particular address is not in the database (e.g. has no listed telephone number or no registered voter). Of course errors or variations in the listing of addresses in either the sampled addresses or the target database will also cause some non-matches (false negatives) and conceivably even some false positives. The outcome of each search (found/not found) should be recorded and can become a valuable measure for studying non-response bias. Since a found/not found code will be generated for all sample addresses and all linked databases, one can make an inter-database measure of how often an address appears across records. This aggregate measure may prove to be a good predictor of non-response in general and in particular may indicate households that are socially disengaged or “flying below the social radar.” That is, the more databases an address is found in, the more engaged is that ”address” (or people living at that address).

While some addresses could not be linked to particular databases, this does not mean that no non-response bias assessment is possible. The unlinked cases become a category for analysis. For example, after linking all addresses with city-style addresses to a reverse directory database, one can classify all cases as having a listed phone number vs. not having a listed phone number and examine response outcomes by this dichotomy. After linking to a database with information on the gender of adults, there would be the unlinked households and those for whom gender information is known (e.g. only males, only females, an opposite gender couple, other both gender combinations) and the response rate for each of these five categories could be examined. Thus, all cases in the sample are covered and retained even when no linkage occurs.

Databases

There are a large and expanding number of data sources that can be linked to addresses either on the HU or aggregate level. Data sources come in complex and ever changing types. While the major types are separately discussed below, there is both considerable variation across databases within categories and considerable overlap across databases in different categories.

Reverse Directories

There are various types of "reverse directories" (e.g. phone number to name, phone number to address). Our interest is in reverse directories from address to name and/or phone number. From the master, national directory of listed phone numbers maintained by the phone companies, a number of providers offer, on-line address look ups. Principal examples are 411.com, address.com, infoUSA411.com, phonenumber.com, whitepages.com, and yb.com. All of these and several others were tested. Given that they all depend on the national phone directories as a primary source, it is not surprising that they yield similar, but not identical, results. A comparison across various reverse directories sometime produced virtually identical results, indicating that they were using the same editions of the telephone directories and had apparently not enhanced them in any way. Across other reverse directories, more differences appeared, For example, addresses that had yielded no hits in 411.com were run against a combination of yb.com, infoUSA411.com, and whitepages.com. For 27% a hit was found, for 7% a link at the apartment address, but without a specific apartment number, was found, and for 3% a possible alternative description of the address was suggested. It is unclear whether these additional links came because different editions of the telephone directories were being used, because the lists were enhanced from other sources,[11] because the address matching algorithms differed, or due to a combination of factors. The results do indicate that multiple sources should be consulted to maximum the linkages.

One common limitation is that reverse directories do not accept apartment number as a predefined, searchable field and this complicates searches involving apartments. However, it was discovered that some would recognize apartment numbers if inputted as part of the street address field. Moreover, in the output from the search, the apartment number of found people was often included and thus could be used to identify the correct unit among those at the same street address. While it might be possible to develop a way of searching the output automatically for the apartment matches, at present this must be done manually.

List Providers

List providers offer both general lists and many types of specialized lists (from boaters to voters). In general, each list is a standalone and typically there is no or limited attempts to compile a dossier on individuals or addresses by merging across lists, but even within this category there are exceptions. Among the more prominent list providers

[11] Some reverse directories offer links to allied databases where follow-up searches for unfound addresses can be conducted. For example, whitepages.com links to peoplefinder.com. In other cases, additional information on a found address can be accessed within the reverse directory itself. It also appears that some reverse directories also augment the information they receive from the telephone directories.

are Century List Services, e-merges.com, InfoUSA.com, and U.S. Data Corporation. Specifically covering voter registration lists are Catalist, e-merges.com, and registeredvoterslists.com.[12] These differ from some other lists in more frequently having data from outside the public, registration information appended to them in what is commonly called "data enhancement". Lists may also be "groomed," updated, verified, or pruned of "deadwood". This is done in various ways by the list providers from incorporating address changes, cross-checking with other lists, and checking death records.

General, Address-Searchable Databases

A wide variety of sources allow searches by addresses. In general these sources include more information than the reverse-directory databases, but less credit and financial information than the credit reports do. Often one can obtain minimal information about an address for a lower price and needs to pay a premium for more complete information. The exact information available and its format varies considerably across providers. Some produce only a list of what appears to be of people currently/recently associated with the address while others are more inclusive and list even people associated with an address only many years ago. They do not document their criteria for inclusion. Some allow individual addresses to be searched by investigators and others require batch processing of addresses by the data provider themselves. The sources consulted include Accurint.com, AutoTrackXP at atxp.choicepoint.com, Donnelleymarketing.com/infoUSA.com, govdmvrecords.com, government-records.com/public-records-search.com, infoUSA411.com, Intelius.com, Peachtreedata.com, Peoplefinders.com, peoplesearchnow.com, and Targusinfo.com. As an example, information on using Accurint is presented.

Accurint Example

Link searches can be done either in batch mode in which a list of addresses is processed by Accurint or in individual mode in which the investigators enter each address one at the time. To describe the capabilities of Accurint, examples using individual mode will be presented. One starts with the Person Search template and since one has only address information to start with, one fills in the location fields and leave the name, SSN, telephone number, date of birth, and age range fields blank. One can then select Person Search or Advanced Person Search.[13] This will produce a list of people associated with the queried address. The lists generally includes any names associated with the address and will include some quite old and outdated linkages. It will often include multiple listings of the same person under slight variants of his/her name (e.g. full name, nickname, middle initial, middle name, etc.). Generally, the most recent linked person is listed first and probable current resident is often indicated. For each linked person it will report name and as much of the following information as it has for each named person: gender, date/year of birth and/or age, age at death, telephone number, address, dates at

[12] On state voter databases and efforts to improve same see Committee on State Voter Registration Databases, 2008 & 2009.
[13] When starting with addresses only, these two search routines are essentially the same.

residence, and SSN. After this initial search, several possibilities exist. Especially if it is unclear who is likely to be the current resident, one may want to do a second Person Search for one or more of the linked names. In particular, one can enter the name and see if that person seems to reside at the sampled address or lives elsewhere. This can help to clarify if the named person is still at the target address. Then either directly from the initial linkages or after the intermediate search to clarify who is the current resident, one can download various extended reports for each linked individual. One is the Summary Report which will indicate such matters as state and approximate time SSN was issued, "others associated with SSN," and a yes/no indication for bankruptcy, property, and corporate affiliations. Another is the Comprehensive Report which includes information on who owns the property, probable current and past residents, land use/zoning, names and other information on neighbors, listings in numerous public records such as hunting/fishing permits, professional licenses, aircraft registration, concealed weapons permits, property assessments, utility hookups, voter registrations, etc., place of employment and position there, possible relatives, possible associates, and possible former residences. As part of the Comprehensive Report, but also available separately, is People at Work which reports on place of employment and position. Some of this information can also be requested in a diagram format, Relavint, that "will help you visualize the relationships between people and their possible relatives and associates, vehicles, property, and even businesses."

Credit Reports

There are three major providers of credit report (Equifax, Experian, and TransUnion). These companies are oriented towards providing credit scores and relating financial information about individuals to lenders and businesses in general. While focusing on proving this inform for named individuals, they all allow searches based on more limited information such as address alone. Credit reports are generally not accessible by researchers with new samples, but under some circumstances these could be utilized in panel studies to follow-up respondents.

Specialized Change Data Bases

Several types of databases specialize in people changing statuses such as moving or dying. Using these specialized databases can help clarify who is the current resident at an address and of course are especially useful in panel studies.

New movers databases are one change type. The two main sources for such information are the change of address listings from the USPS known as NCOA^{Link} and telephone number changes and connects (e.g. New Connect/New Movers from Telematch.com and New Movers from infoUSA.com; Hotline List of New Telephone Listings from newmoverslist.com). In some cases these are supplemented by listings from utilities, governmental property records, magazine subscriber move data, and/or other "proprietary sources" (e.g. Hotline PLUS from newmoverlists.com, New Home Owners Lists from directmail.com, New Movers from infoUSA.com, New Movers from cas-online.com).

Another specialized source on changes are various death records. While used more commonly in longitudinal, panel studies and epidemiological research, death records can be consulted to see if names associated with an address are still living. This can be especially useful when the dates associated with the persons lined to an address are several years old and/or when the identified person is elderly. While several databases do indicate if a person matched to a particular address is deceased and will often report a date/year of death, such notations are not universal and may not be up-to-date. The best source is the National Death Index (NDI) maintained by the National Center for Health Statistics (http://www.cdc.gov/nchs/ndi.htm). This source is "available to investigators solely for statistical purposes in medical and health research." Another source is the Social Security Death Master File (Hill and Rosenwaike, 2001/2002). For people 65+ it covers 93-96% of deaths in the NDI. The Social Security list is accessible via various commercial sites (e.g. http://ssdi.rootsweb.ancestry.com and http://www.genealogybank.com). Other sources include state death records (e.g. http://death-records.net).

Extended Searches

If the initial address search comes up with the name of residents and/or phone numbers for the address, then follow-up searches can be made using these identifiers. These follow-up searches can serve several purposes: 1) to determine who are current residents when either multiple people are associated with a particular address because of a) people from different apartment numbers not being distinguished and/or b) turnover in tenancy, 2) to confirm that the located person does currently reside at the sampled address (i.e. hasn't moved away), and 3) to add additional information about the located person.

Most of the databases used in the initial, address-only search will permit further searches with the added information on name and/or telephone number (see Accurint example above). In addition, there are numerous other databases that do not allow address-only searches, but can be utilized once the additional identifiers have been obtained. These include Claritas.com, in-foquest.com, statewidegovrecords.com, telematch.com, and theultimates.com.

These extended searches can fill-in information that is missing from the initial linkages. For example, an initial source may turn up a name, but no demographic information. A follow-up search using name may then add certain demographic information such as age. Or it might turn up names, but no clear evidence on who is the current resident. Searching on the listed names can often clarify who is likely to be currently at the address in question. Similarly, an initial search may turn up a phone number, but no name and that phone number could lead to name and other information emerging from subsequent searches. If a follow-up search yields new information that information could lead to another search which in turn could yield more leads and further fruitful searches. Pooling information across databases and from a combination of initial and extended sources will collect more complete information on the residents of sampled addresses and lead to a better assessment of non-response bias.

Database Search Results

To demonstrate MIDA's use of multiple, address-level sources, results from searching five databases were merged together. These included three general, address-based sources (Accurint, Peachtree, and infoUSA), one reverse directory (411.com), and one list of registered voters (Catalist).

Of the 383 cases with street-styles address, some linkage to databases was found for all but 9 addresses. These unlinked addresses were searched for in various other maps and address-based databases (e.g. Google Maps, Mapquest, American Fact Finder). Two were found to represent an area that had changed both its place name and zip code and in addition one has a misspelled street name. One could not be located in any source which may indicate some problem with the address such as an error in the post office list or a change in place name. Whether this represents an errant address for a residence that actually exist or a non-existent residence is uncertain. Four were located and appeared to designate actual residences. These could represent long-term vacancies, very recent construction, or residences occupied by people who have managed to avoid inclusion in a wide range of databases. Two were recognized as legitimate addresses, but showed either a vacant lot or open country at the point associated with the address. Of course since the mapping programs only show approximate locations, this may reflect their limitations rather than the absence of a residence linked to the address. These handful of uncertain cases would be resolved if the sampled addresses were utilized as part of an actual survey rather than merely at the sample frame augmenting stage as in this preliminary study.

When the unlinked and non-street-styles addresses are considered together, only 6.5% of addresses had no information at the household-level outside the sample frame. This is much lower than the 26% unlinked by Raghunathan and Van Hoewyk (2005). This difference mostly reflects their use of a single database and may also reflect changes from 2005 to 2009 and more thorough search strategies.

A last name was obtained for 97.4% of the city-style addresses, age for 93.7%, social security number for 88.8%, phone number 83.1%, gender for 78.6% (and when first names were used to infer gender for about 95.1%), race/ethnicity for 78.1%, income for 66.1%, occupation for 56.0%, and education for 37.0%. Using multiple databases of course notably increases the proportion of addresses with information. For example, in individual sources age was available for 24.5-78.9% of addresses and from at least one source for 93.7% of addresses.

Of course when particular information comes from a specific, public record, one needs to use a database that accesses that source rather than consult multiple databases. For example, from the voter-registration database (Catalist) and the federal, political-contributions database (FUNDRACE), one can determine if the sampled address has one or more registered voter and whether any federally-regulated political contributions were made. One or more voters were present at 67.3% of addresses. 44.6% had a voter in the 2004 general election and 43.2% that a voter in the 2008 general election. Voting status was related to such other variables are resident type. For example, 75.3% of the non-apartment addresses and 43.2% of the apartments had at least one active voter. For those addresses with a registered voter their party is of course known. 46.4% have a Democrat, 28.1% a Republican, and others include independents and third party registrants.
For the 2008 campaign cycle none of the addresses have a confirmed federal political contribution.

Besides using the aggregate- and HU-level data separately, they can also be joined together and analyzed. A few examples will illustrate the type of comparisons that can be made: 1) using data from the Census on track-level racial composition and from the USPS' DSF on housing type (apartment/not apartment building) showed no association between these variables, 2) linking the Superfund database with information on having a phone number also showed little association. In areas with no Superfund sites 39.2% had a listed phone number and in areas with one or more sites 37.9% had a phone number, and 3) the aggregate data on public housing units showed an association with having a listed phone number. It was 40.2% in areas with few units and 33.3% in areas with higher public housing density.

Limitations

A number of factors do limit or complicate the use databases for MIDA. First of all, there are legal restrictions. Even when there is a database that permits address searches and has content of interest, use may be restricted. For example, the various providers of national voting and registration lists are limited by state laws as to both what they can provide to users and what type of users may access the records for that state. For example, to access the Illinois records one needs a letter from the State Board of Elections or Illinois Attorney General authorizing the search. For California, the records may be used only for a political purpose. Information from driver's license records are restricted by the provisions of the Driver's Privacy Protection Act (DPPA) and that generally excludes any research use. The Gramm-Leach-Bliley Act (GLBA) regulates access to other types of records.

Second of all, several general factors complicate the extraction of information from databases. First, most databases are limited in the documentation that they supply about their data. The original source of information is often not indicated nor is its recency. Likewise, quality-control procedures are never detailed. Definitions, data-procurement procedures, and quality-control procedures are often not indicated. Even obtaining limited information usually involves considerable digging and/or special requests from the providers. However, database providers will usually clarify matters and often supply additional documentation upon request. Also, additional information comes from using the databases and becoming familiar with their features through application and comparison. The impediment was usually not that information was being withheld to cover up poor procedures or serious flaws, but that the information and documentation had never been compiled, had not been prepared for dissemination, or was restricted to protect proprietary interests.

Second, most databases collect specific information from various sources and include it when found and omit it otherwise. There are thus many gaps in the data matrix. Thus, a given source will typically include certain information (e.g. last name) while omitting other data (e.g. race or marital status).

Third, much information is obtained from state-level, public records. The content, form, and availability of this information (e.g. voter registration, property records, professional licenses) varies from state-to-state and thus the information is often not uniform across states. Databases however do harmonize data to reduce this problem.

Fourth, changes across products and firms are fairly frequent. The nature of databases and who maintains them often varies over time and one needs to continually update information about databases and their providers.

Fifth, many databases are frequently updated. This is generally a positive situation since it means that new information is being added. But it also means that using the same database for the same addresses just a few months apart can produce appreciable differences in results.

Sixth, information in the databases may be out-of-date. While accurate when compiled, it may not reflect more recent changes in statuses. This may involve alterations in a person's personal characteristics (e.g. a marriage or divorce or a job change) or changes in official information such as a new place name and/or zip code for an address. But while occurring, out-of-date data does not seem to be a major problem.

Seventh, errors in either the original source record or in the databases compiled from original sources occur. Some errors were detected in all records. Even the master DSF list wrongly classified a few apartment buildings as non-apartments. However, by cross-checking across multiple records the accurate status was usually determinable. Consulting multiple sources notably reduced the level of error.

Eighth, information about addresses are about the households that residence there and the individuals who residence in those household, not about a specific respondent. The household-level information is one level of aggregation above that of an individual household resident selected as a respondent. As such it is very useful in determining the attributes of households that yield respondents vs. non-respondents, but not a direct comparison of respondents and non-respondents. However, in most cases, it is relatively easy to match an individual from the household-level databases to a respondent selected from the household members. Based on the 2002-2008 General Social Surveys, 18% of respondents live in a household with no other adults and 53% live in a household with two adults of opposite gender. Thus, for 71% of respondents, the identity of the respondent can be readily matched between the residents listed in the database and the respondent selected in the survey. For other households, some matching is possible, but it becomes more difficult and less certain.

Ninth, apartments are especially challenging since most databases do not allow explicit searches for units. The extra complication of apartments is illustrated by looking at the number of names listed from the reverse-directory search. For non-apartments there were an average 2.9 residents listed per address. For apartments an average of 19.0 people per address are listed. Most of the large excess results from the fact that the reverse-directory database, like most, did not distinguish between individual units in an apartment building. That is, they represent people associated with the street address of the building and not the individual unit in the sample. In addition, the greater turnover in apartment renters vs. homeowners also contributes somewhat to the higher number of people associated with apartment addresses.

When dealing with apartments one needs to a) determine if there is a away to get units recognized, b) try alternative input formats both within and across databases, c) examine the output to see if unit is indicated at this stage, and d) do follow-up searches across databases to clarify who resides in a particular unit. By taking these steps, links can be made for most apartment units.

Tenth, even among street-styles addresses, a few will not link to residents even across multiple databases (9 of 383). The existence and nature of addresses in general and unlinked addresses in particular, can be examined by running addresses in various mapping programs such as Google Maps and Mapquest. These could often clarify if an address exists or not, indicate whether it was a residence, business, or other, and show if it was a single family residence or a multi-unit dwelling. Of course these sources are also invaluable for interviewers in the field actually locating the address to conduct an interview.

Finally, there are several challenges involving using the names found in various databases and associated with a particular address. First, many databases will list multiple names associated with an addresses. [14] This often includes many names only associated with an address at some point in the past. Many databases do indicate dates associated with a particular person and often designate the most recent resident as the likely current occupant. As noted above, follow-up searches of names associated with an address can be done to figure out who are current residents. Second, many databases list the same individual multiple times under slight variations of his/her name. While it is usually obvious that these are the same person, that is not always clear. In some cases the records are intentionally unclear. For example, it used to be common for women living alone to use only their initials or even their late husbands' first name to disguise that theirs was a female-headed household. Fourth, shared family names often make it possible to infer which named individuals are members of the same households, but the same family names are not used by some married couples, for most cohabitators, and for members of some mixed/blended families. Fourth, some databases such as telephone directories only list a single name for a household and thus ignore other household members. In addition, most databases do not list the names of minors in the household. Careful attention to details and comparison across sources is needed to clarify the household composition of many households.

While it is of course desirable that all information be complete and accurate, data can be useful without being perfect. When the databases are being used to identify a respondent at the target address or to follow-up a mover in a panel study, then complete accuracy is needed. One needs to identify the respondent and/or figure out where the respondent now resides. But when non-response bias is being examined, data with some error can still be useful. For example, due to moves and other factors voter registration information will never be completely up-to-date. Voter information for an address may not reflect recent registrations and/or moves. But most voter information for sampled addresses will be accurate (i.e. reflecting the current registration status of the present residents). In addition, the information can be seen as reflecting information on the address at the most recent time the voting records were updated and that should be useful in accessing non-response bias even when individual changes have occurred for some addresses. First, it is likely, but not established by this research, that addresses tend to be

[14] A large number of names per address is most often associated with it being a multi-unit dwelling. Among single unit dwelling more names are associated with larger household size (i.e. more adults; not minor children since they are not listed), more turnover which would be associated with it being a rental property rather than owner occupied, and an older dwelling for which there could be more former residents associated with it.

occupied over time by people with similar socio-demographics. Second, the address or housing unit is an element in the sample, above that of the respondent, but below the other aggregate-level units (e.g. block group, Census tract, zip code). As long as the address-level information is accurate, it can be used to identify the sample and to discriminate between addresses yielding respondents vs. non-respondents, even when the occupancy of some HUs has changed since the records were compiled. When the address-level information is complete, accurate, and up-to-date, then one can compare the characteristics of respondents and non-respondents. When the address-level information is less perfect (but still generally accurate), then one can still compare the characteristics of addresses yielding respondents to those with non-respondents just like one can look at areas (e.g. tracks, zip codes, etc.) with higher and lower response rates.

Further MIDA Test and Full Application

Having shown that MIDA can be used to augment a national sample frame, the next step is to test its utility in an actual survey. The General Social Survey (GSS) is prepared to implement MIDA once support for its use is secured. The GSS is a national, in-person, full-probability sample of adults living in households (Davis, Smith, and Marsden, 2009). Except for data from the Census, it is the most widely used data source in the social sciences (Smith, 2002b). Its high quality, wide use, and extensive content make the GSS especially appropriate for the application of MIDA. Specific application of MIDA to the GSS regarding non-response bias detection and correction and substantive analysis are described below.

Assessing and Adjusting for Non-response

The MIDA dataset will contain much more data about non-respondents than are usually available. The full dataset will have household and aggregate-level data for both respondents and non-respondents. Such a rich dataset is uncommon in nationally representative demographic/attitudinal surveys. It provides an opportunity to explore different approaches to estimating and adjusting for non-response bias. For the many variables for which the dataset contains values for both respondents and non-respondents, it will be possible to explore the effects of non-response by comparing estimates from these variables for the full dataset with estimates on the respondent cases only. These analyses will suggest which estimates would be most vulnerable to non-response bias. This knowledge will then inform our understanding of the error implicit in estimates from the survey variables themselves, for which non-respondent data are not available. (Gelman and Carlin, 2002; Geronimus, Bound, and Neidert, 1996; Groves, 2005a; Groves, 2006; Marker, Judkins, and Winglee, 2002; Meng, 2002; Zanutto and Zaslavsky, 2002).

In addition, the availability of data on non-respondents can improve weighting techniques. In recent rounds, the GSS has incorporated a non-response adjustment at the level of the Primary Sampling Units (PSU) which are metro areas or non-metro counties. It assumes that the non-respondents in a given area are more like the respondents near them than other respondents. This assumption has been empirically verified and is probably the most common type of non-respondent adjustment used in national, in-person

surveys But the use of PSU to form non-response adjustment cells is limited in the improvement it can provide and is based primarily on a heuristic of availability rather than relying on specific theoretical connection with the study variables. The MIDA-enriched dataset, by providing data on both respondents and non-respondents on many variables, will allow for more discretion in creating non-response adjustment cells and for more sophisticated weighting adjustments. (Bethlehem, 2001; Kalton and Kasprzyk, 1986).

Response-propensity weighting is a common method of adjusting for non-response. The theory behind this approach is that all cases, both responding and non-responding, have a non-zero propensity to respond which can be. The dependent variable is a dummy variable indicating response and the independent variables are those that predict response: urbanicity, region, household size and composition, interest in the survey topic, etc. Responding cases are then weighted by the inverse of their response propensity to account for the non-responding cases, with low-propensity cases given more weight than high. Like the non-response weighting adjustment discussed above, this method often suffers from a lack of frame variables: the right hand side variables are usually those that are available for all cases rather than those that would be most appropriate. MIDA will permit more thoughtful choices in the independent variables and should improve the response-propensity weighting adjustment. (Ekholm and Laaksonen, 1991).

In addition to giving one a wider selection of variables with which to adjust the weights, MIDA will also provide data with which to compare and evaluate the adjustment methods. These results will greatly improve the weighting methods used on the GSS and for surveys in general.

Similarly, having more variables in the MIDA dataset will improve imputation techniques. Hot-deck imputation fills in values that are missing due to item-non-response by matching cases with missing data to cases without missing data. MIDA will allow better matches and should thus improve the imputation. Also, if the imputation technique chosen involves modeling (e.g., mean regression or multiple imputation), the MIDA dataset will allow better models to be formed with the additional variables. Either way, MIDA will improve the imputation techniques available to surveys in general and the GSS in particular (Marker et al., 2002).

Substantive Analysis

The wide range of content in the GSS is well-suited for examining the utility of the multi-level, aggregate data. Contextual analysis has already been shown to be very valuable in analysis using the GSS (see references cited above) and the wide content will provide a broad test of the value of such contextual data. In addition, the high use of the GSS will insure that the contextualized data will also be widely utilized by researchers. The public data file will be constructed to insure that no deductive disclosure of respondents will be possible. Files with more detailed information, but not personal identifiers or information readily-allowing deductive disclosure, will be made accessible to researchers following standard, limited-access protocols to insure confidentiality.

Conclusion

MIDA has the potential to advance social-science research in general by notably improving survey-research methodology. Moreover, it does so by drawing on one of the major societal changes in recent decades, the development of large-scale, computerized databases that hold extensive information about individuals, households, neighborhoods, and other societal units.

Methodologically, it should help to increase response rates, allow for a much more comprehensive assessment of non-response bias, and facilitate the calculation of weights and imputations to adjust for the detected non-response bias. Besides providing for a general approach to deal with non-response, it will in particular permit the testing of several prominent theories and hypotheses explaining non-response: social disorganization theory, social isolation theory, overextension theory, structural impediments, etc. In addition, the auxiliary data from the databases will permit an examination of general, non-response models (Groves and Couper, 1998).

Substantively, MIDA will improve analysis by easily and automatically making multi-level, contextual variables as ready for analysis as data directly collected in surveys. As the list of examples cited above attest, geographic context has notable impacts on many aspects of people's lives. The contextual data from sampling frames and augmented from multiple databases will provide a rich, contextual array of data for analysis across scores of central substantive topics.

This partial, pilot study demonstrates the riches of both aggregate and household-level data sources and the practicality of linking both to a national sample of addresses. But it also shows that using multiple databases and linking them to both each other and a national sample of addresses is a complex and challenging task that must be carried out carefully. When available databases are rigorously utilized to augment sample frames with aggregate- and household-level data, then survey research benefits by the enhancement of data-collection efforts, the measurements and adjustment of non-response bias, and the routine addition of contextual data to substantive analyses.

References

Alesina, Alberto and Eliana LaFerrara, "Participation in Heterogeneous Communities," Quarterly Journal of Economics, 115 (2000), 847-904.

Anderson, Barbara A. and Silver, Brian D., “Measurement and Mismeasurement of the Validity of Self-Reported Vote,’ American Journal of Political Science, 80 (1986), 771-785.

Audit Bureau of Circulations, “Turning Information into Inspiration: Annual Report, 2005,” (http://www.accessabc.com)

Baumer, Eric P.; Messner, Steven F.; and Rosenfeld, Richard, “Explaining Spatial Variation in Support for Capital Punishment: A Multilevel Analysis," American Journal of Sociology, 108 (2003), 844-875.

Berge, Keith H. et al., “Resource Utilization and Outcome in Gravely Ill Intensive Care Unit Patients with Predicted In-hospital Morality Rates of 95% or Higher by APACHE III Scores,” Mayo Clinic Proceedings, 80 (2005), 166-173.

Bethlehem, Jelke G., “Weighting Nonresponse Adjustments Based on Auxiliary Information,” in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002.

Billy, John O. G. and Moore, David E., "A Multilevel Analysis of Marital and Nonmarital Fertility in the U.S," Social Forces, 70 (1992), 977-1011.

Boardman, Jason D.; Finch, Brian Karl; Ellison, Christopher G.; Williams, David R.; and Jackson, James S., “Neighborhood Disadvantage, Stress, and Drug Use Among Adults," Journal of Health and Social Behavior, 42 (2001), 151-165.

Bobo, Lawrence and Gilliam, Franklin D., Jr., "Race, Sociopolitical Participation, and Black Empowerment," American Political Science Review, 84 (1990), 377-393.

Brace, Paul; Sims-Butler, Kellie; Arceneaux, Kevin; and Johnson, Martin, “Public Opinion in the American States: New Perspectives Using National Survey Data,” American Journal of Political Science, 46 (2002), 173-189.

Branas, Charles C. et al., “Access to Trauma Centers in the United States,” Journal of the American Medical Association, 293 (June 1, 2005), 2626-2633.

Brewster, Karin L., “Neighborhood Context and the Transition to Sexual Activity among Young Black Women,” Demography, 31 (1994a), 603-604.

Brewster, Karin L., "Race Differences in Sexual Activity Among Adolescent Women: The Role of Neighborhood Characteristics," American Sociological Review, 59 (1994b), 408-424.

Brick, J. Michael and Broene, Pam, "Unit and Item Response, Weighting, and Imputation Procedures in the 1995 National Household Education Survey (NHES:95)," Working Paper No. 97-06. Washington, DC: National Center for Education Statistics, 1997.

Brick, J. Michael et al., "Evaluating Secondary Data Sources for Random Digit Dialing Samples," Proceedings of the American Statistical Association, 2002. Washington, DC: ASA, 2000.

Brick, J. Michael et al., "Increased Efforts in RDD Surveys," Paper presented to the American Association for Public Opinion Research, Nashville, May, 2003.

Brick, J. Michael; Montaquila, Jill; and Scheuren, Fritz, "Estimating Residency Rates for Undetermined Telephone Numbers," Public Opinion Quarterly, 66 (2002), 18-39.

Brooks-Gunn, J.; Duncan, Greg J.; and Klebanov, Sealand N., "Do Neighborhoods Influence Child and Adolescent Development?" American Journal of Sociology, 99 (1993), 353-395.

Browning, Christopher R. and Olinger-Wilbon, Matisa, "Neighborhood Structure, Social Organization, and Number of Short-Term Sexual Partnerships," Journal of Marriage and the Family, 65 (2003), 730-745.

Browning, Christopher R.; Leventhal, Tama; and Brooks-Gunn, Jeanne, "Neighborhood Context and Racial Differences in Early Adolescent Sexual Activity," Demography, 41 (2004), 697-702.

Bryk, Anthony S. and Raudenbush, Stephan W., "Toward a More Appropriate Conceptualization of Research on School Effects: A Three-level Hierarchical Linear Model," American Journal of Education, 97 (1988), 65-108.

Burden, Barry C., "Voter Turnout and the National Elections Studies," unpublished report, Harvard University, 2000.

Campanelli, Pamela; Sturgis, Patrick; and Purdon, Susan, "Can You Hear Me Knocking: An Investigation into the Impact of Interviewers on Survey Response Rates," SCPR, 1997.

Cantor, David and Cunningham, Patricia, "Methods for Obtaining High Response Rates in Telephone Surveys," in Studies of Welfare Populations: Data Collection and Research Issues, edited by C. Citro et a.. Washington, DC: National Academy Press, 2002.

Center for Disease Control (CDC), "Sexually Transmitted Disease Surveillance, 2004," Atlanta: Centers for Disease Control and Prevention, 2004.

Charles, Camille Zubrinsky, "The Dynamics of Racial Residential Segregation," Annual Review of Sociology, 29 (2003), 167-207.

Cohen, Cathy J. and Dawson, Michael C., "Neighborhood Poverty and African American Politics," American Political Science Review, 87 (1993), 286-302.

Cohen, D.; Spear, S.; Scribner, R.; Kissinger, P.; Mason, K.; and Wildgen, J., "Broken Windows and the risk of gonorrhea," American Journal of Public Health, 90 (2000), 230-236.

Committee on State Voter Registration Databases, Improving State Voter Registration Databases: Final Report. Washington, DC: The national Academies Press, 2009.

Committee on State Voter Registration Databases, State Voter Registration Databases. Washington, DC: The National Academies Press, 2008.

Converse, Jean M. and Schuman, Howard, Conversations at Random: Survey Research as Interviewers See It. New York: John Wiley & Sons, 1974.

Costa, Dora L. and Kahn, Matthew E., "Civic Engagement and Community Heterogeneity: An Economist's Perspective," Paper presented to the Conference on Social Connectedness and Public Activism, Cambridge, 2002.

Cotter, David A. et al., "The Demand for Female Labor," American Journal of Sociology, 103 (1998), 1673-1712.

Couper, Mick P. and Groves, Robert M., "Social Environmental Impacts on Survey Cooperation," Quality & Quantity, 30 (1996), 173-188.

Couper, Mick P. and Lyberg, Lars E., "The Use of Paradata in Survey Research," Proceedings of the International Statistical Institute, Sydney, April, 2005.

Couper, Mick P.; Singer, Eleanor; and Kulka, Richard A., "Participation in the 1990 Decennial Census: Politics, Privacy, Pressures," American Politics Quarterly, 26 (1998), 59-80.

Covington, J. and Taylor, R. B., "Fear of Crime in Urban Residential Neighborhoods: Implications of Between- and Within-Neighborhood Sources for Current Models," Sociological Quarterly, 32 (1991), 231-249.

Cox, Christine S., "Integrating Data: Policy and Legal Issues," Report prepared for the Federal Committee on Statistical Methodology, November, 2006.

Crane, Jonathan, "The Epidemic Theory of Ghettos and Neighborhood Effects on Dropping Out and Teenage Childbearing," American Journal of Sociology, 96 (1991), 1226-1259.

D'Urso, Victoria T., "Internet Use and the Duration of Buying and Selling in the Residential Housing Market, Economic Incentives, and Voting," Unpublished Ph.D. Thesis, MIT, 2003.

Davis, James A.; Smith, Tom W.; and Marsden, Pater V., General Social Survey Cumulative Codebook, 1972-2008. Chicago: NORC, 2009.

Davern, Michael. "Evaluating Linked Survey and Administrative Data for Policy Research," Paper presented to the Federal Committee on Statistical Methodology, Washington, DC, November, 2006.

De Jong, Gordon F. and Steinmetz, Michele, "Receptivity Attitudes and the Occupational Attainment of Male and Female Immigrant Workers," Population Research and Policy Review, 23 (2004), 91-116.

DiPrete, Thomas A. and Forristal, Jerry D., "Multivariate Models: Methods and Substance," Annual Review of Sociology, 20 (1994), 331-357.

DiSogra, Charles; Callegaro, Mario, and Hendarwan, Ellina, "Recruiting Probability-Based Web Panel Members Using an Address-Based Sample Frame: Results from a Pilot Study Conducted by Knowledge Networks, Paper presented to the American Statistical Association, Washington, DC, August, 2009.

Dixon, Jeffrey C. and Rosenbaum, Michael S., "Nice to Know You? Testing Contact, Cultural, and Group Threat Theories of Anti-Black and Anti-Hispanic Stereotypes," Social Science Quarterly, 85 (2004), 257-280.

Downey, Liam, "Using Geographic Information Systems to Reconceptualize Spatial Relationships and Ecological Context," American Journal of Sociology, 112 (2006), 567-612.

Ekholm, Anders and Seppo Laaksonen. 1991. "Weighting via Response Modeling in the Finnish Household Budget Survey." *Journal of Official Statistics* 7(3): 325-337.

Fair, Martha E., "Record Linkage in an Information Age Society," Proceedings of the Census Bureau's Conference and Technology Interchange. Washington, DC: Bureau of the Census, 1996.

Federal Bureau of Investigation (FBI), "Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense Data, 2004," Washington, DC: Federal Bureau of Investigation, 2004.

Ford, Julie and Beveridge, Andrew A., "Neighborhood Crime Victimization, Drug Use, and Drug Sales," Paper presented to the American Sociological Association, Montreal, 2006.

Fowler, Floyd J., Jr. et al., "Using Telephone Interviews to Reduce Nonresponse Bias to Mail Surveys of Health Plan Members," Medical Care, 40 (2002), 190-200.

Galea, Sandro; Ahern, Jennifer; and Vlahov, David., "Contextual Determinants of Drug Use Risk Behavior: A Theoretic Framework," The Review of Economics and Statistics, 83 (2003), 257-268.

Gatewood, George, A Monograph on Confidentiality and Privacy in the U.S. Census. Washington, DC: US Census, 2001.

Gelman, Andrew and Carlin, John B., "Poststratification and Weighting Adjustments," in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002.

Geronimus, Arline T.; Bound, John; and Neidert, Lisa J., "On the Validity of Using Census Geocode Characteristics to Proxy Individual Socioeconomic Characteristics," Journal of the American Statistical Association, 91 (1996), 529-537.

Gfroerer, Joseph; Lessler, Judith; and Parsley, Teresa, "Studies of Nonresponse and Measurement Error in the National Household Survey of Drug Abuse," In L. Harrison & A. Hughes (Eds.), The validity of self-reported drug use: Improving the accuracy of survey estimates (NIH Publication No. 97-4147, NIDA Research Monograph 167, pp. 273-295). Rockville, MD: National Institute on Drug Abuse, 1997.

Gibson, James L., "The Political Freedom of African Americans: A Contextual Analysis of Racial Attitudes, Political Tolerance, and Individual Liberty," Political Geography, 14 (1995), 571-599.

Gilbert, Christopher P, "Religion, Neighborhood, Environments, and Partisan Behavior: A Contextual Analysis," Political Geography Quarterly, 10 (1991), 110-131.

Glaeser, Edward L. and Glendon, Spender., "Who Owns Guns? Criminals, Victims, and the Culture of Violence," Paper presented to the American Economics Association, Chicago, 1998.

Goyder, John; Lock, Jean; and McNair, Trish, "Urbanization Effects on Survey Nonresponse: A Test Within and Across Cities," Quality & Quantity, 26 (1992), 39-48.

Groves, Robert M., "Designing Surveys Acknowledging Nonresponse," Unpublished

report, National Academies, 2006.

Groves, Robert M., "Nonresponse Rates and Nonresponse Bias in Household Surveys," Public Opinion Quarterly, 70 (2006), 646-675.

Groves, Robert M., "Research Synthesis: Nonresponse Rates and Nonresponse Error in Household Surveys, Unpublished report, Institute for Social Research, 2005a.

Groves, Robert M., "Total Survey Error," Plenary Presentation to the American Association for Public Opinion Research, Miami Beach, May, 2005b.

Groves, Robert M. and Couper, Mick P., Nonresponse in Household Interview Surveys. New York: John Wiley & Sons, 1998.

Groves, Robert M; Singer, Eleanor; and Corning, Amy, "Leverage-Saliency Theory of Survey Participation: Description and an Illustration," Public Opinion Quarterly, 64 (2000), 299-308.

Grubesic, Tony H. and Matisziw, Timothy C., "On the Use of ZIP Codes and ZIP Code Tabulation Areas (ZCTAs) for the Spatial Analysis of Epidemiological Data," International Journal of Health Geographics, 5 (2006), 58ff.

Harter, Rachel; Eckman, Stephanie; English, Ned; and O'Muircheartaigh, Colm, "Applied Sampling for Large-Scale, Multi-Stage Are Probability Designs," Unpublished NORC report, 2008.

Harvey, Bart J. et al., "Using Publicly Available Directories to Trace Survey Nonresponders and Calculate Adjusted Response Rates," American Journal of Epidemiology, 158 (2003), 1007-1111.

Hill, Mark E. and Rosenwaike, Ira, "The Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages," Social Security Bulletin, 64 (2001/2002), 45-51.

Holmes, Thomas J., "Localization of Industry and Vertical Disintegration," The Review of Economics and Statistics, 81 (1999), 314-325.

House, James S. and Sharon Wolf. 1978. "Effects of Urban Residence on Interpersonal Trust and Helping Behavior," *Journal of Personality and Social Psychology* 36:1029-1043.

HUD (Department of Housing and Urban Development), "A Picture of Subsidized Households – 1998," Washington, DC: HUD, 1998.

Jencks, Christopher and Mayer, Susan E., "The Social Consequences of Growing Up in a Poor Neighborhood," in Inner-City Poverty in the United States, edited by L. E.

Lynn, Washington, D.C.: National Academy Press, 1990.

Jencks, Christopher, "What is the Optimal Level of Inequality?" Unpublished paper, Harvard University, 1999.

Jenkins, Stephen P. et al., "Linking Household Survey and Administrative Record Data: What Should the Matching Variables Be?" Discussion Paper 489. German I Institute for Economic Research, 2005.

Johnson, Timothy and Cho, Young Ik, "Understanding Nonresponse Mechanisms in Telephone Surveys," Paper presented to the American Association for Public Opinion Research, Phoenix, May, 2004.

Johnson, Timothy P.; Cho, Young Ik; Campbell, Richard T.; and Holbrook, Allyson L., "Using Community-Level Correlates to Evaluate Nonresponse Effects in a Telephone Survey," Public Opinion Quarterly, 70 (2006), 704-719.

Johnston, S. Claiborne et al., "Endovascular and Surgical Treatment of Unruptured Cerebral Aneurysms: Comparison of Risks," Annals of Neurology, 48 (2000), 11-19.

Jones, Dale E., Religious Congregations and Membership in the United States, 2000. Nashville: Glenmary Research Center, 2002.

Kalsbeek, William D.; Yang, Juan; and Agans, Robert P., "Predictors of Nonresponse in a Longitudinal Survey of Adolescents," Proceedings of the American Statistical Association, 2002. Washington, DC: American Statistical Association, 2002.

Kalton, Graham and Daniel Kasprzyk. 1986. "The Treatment of Missing Survey Data." *Survey Methodology* 12(1): 1-16.

Kawachi, Ichiro; Kennedy, Bruce P.; and Lochner, Kimberly, "Long Live Community: Social Capital and Public Health," The American Prospect, 35 (1997a), 56-59.

Kawachi, Ichiro; Kennedy, Bruce P.; Lochner, Kimberly; and Prothrow-Stith, D., "Social Capital, Income Inequality, and Morality," American Journal of Public Health, 87 (1997b), 1491-1498.

Kennedy, Courtney; Keeter, Scott; and Dimock, Michael, "A 'Brute Force' Estimation of the Residency Rate for Undetermined Telephone Numbers in an RDD Survey," Public Opinion Quarterly, 72 (2008), 28-39.

Kennickell, Arthur B., "Analysis of Nonresponse Effects in the 1995 Survey of Consumer Finances," Paper presented to the American Statistical Association, Anaheim, August, 1997.

Kennickell, Arthur B., "Darkness Made Visible: Field Management and Nonresponse in the 2004 SCF," Paper presented to the American Statistical Association, Minneapolis, August, 2005.

Key, Wanda L. and Miracle, Stephanie, "What Are 'Bad' Addresses and What Do You Do with Them – Exactly," Paper presented to the National Postal Forum, Washington, DC, May, 2009.

Kim, Jibum; Smith, Tom W.; Kang, Jeong-han; and Sokolowski, John, "Community Context and Cooperation Rate" GSS Methodological Report No. 108, Chicago, NORC, 2006.

Kim, Jibum; Smith, Tom W.; and Sokolowski, John, "Ecological Correlates of Cooperation Rate in the 2002 and 2004 General Social Survey," Paper presented to the American Association for Public Opinion Research, Montreal, May, 2006

Kojetin, Brian A., "Characteristics of Nonrespondents to the Current Population Survey (CPS) and Consumer Expenditure Interview Survey (CSES)," Proceedings of the American Statistical Association, 1994. Washington, DC: American Statistical Association, 1994.

Krieger, Nancy et al., "Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-based Measure and Geographic Level Matter?" American Journal of Epidemiology, 156 (2002a), 471-482.

Krieger, Nancy et al., "Zip Code Caveat: Bias Due to Spatiotemporal Mismatches Between Zip Codes and US Census–Defined Geographic Areas: The Public Health Disparities Geocoding Project," American Journal of Public Health, 92 (2002), 1100-1102.

Latkin, Carl A. and Curry, Aaron D., "Stressful Neighborhoods and Depression: A Prospective Study of the Impact of Neighborhood Disorder," Journal of Health and Social Behavior, 44 (2003), 34-44.

Lee, Barret; Oropesa, R.S.; and Kanan, James W., "Neighborhood Context and Residential Mobility," Demography, 31 (1994), 249-270.

Lepkowski, James M. and Couper, Mick P., "Nonresponse in the Second Wave of Longitudinal Household Surveys," in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002.

Lessler, Judith T. and Kalsbeek, William D., Nonsampling Error in Surveys, New York: John Wiley & Sons, 1992.

Loosveldt, Geert and Carton, Ann, "Utilitarian Individualism and Panel Nonresponse,"

International Journal of Public Opinion Research, 12 (2002), 428-438.

Luttmer, Erzo, "Understanding Income Redistribution: The Role of Interpersonal Preference, Information, and Mechanism Design," Unpublished Ph.D. Thesis, Harvard, 1998.

Lynn, Peter, "PEDAKSI: Methodology for Collecting Data about Survey Non-respondents," Working Paper for the Institute for Social and Economic Research, 2002-05, University of Essex, 2002.

Lynn, Peter at al., "The Effect of Extended Interviewer Efforts on Nonresponse Bias," in Survey Nonresponse, edited by Robert M. Groves et al., New York: John Wiley & Sons, 2002.

Marcus, Pamela M et al., "Extended Lung Cancer Incidence Follow-up in the Mayo Lung Project and Overdiagnosis," Journal of the National Cancer Institute," 98 (2006), 748-756.

Marker, David A.; Judkins, David R.; Wingless, Marianne, "Large-scale Imputation for Complex Surveys," in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002.

Martin, Elizabeth, "Unfinished Business," Public Opinion Quarterly, 68 (2004), 439-450.

Massey, Douglas S. and Eggers, Mitchell L., "The Ecology of Inequality: Minorities and the Concentration of Poverty," American Journal of Sociology, 95 (1990), 1153-1188.

Massey, Douglas S.; Gross, Andrew B.; and Shibuya, Kumiko, "Migration, Segregation, and the Geographic Concentration of Poverty," American Sociological Review, 59 (1994), 425-445.

McLeod, Jane and Edwards, Kevan, "Contextual Determination of Children's Responses to Poverty," Social Forces, 73 (1995), 487-516.

Mellor, Jennifer M. and Milyo, Jeffrey, "State Social Capital and Individual Health Status," Unpublished Paper, College of William and Mary, 2004.

Meng, Xiao-Li, "A Congenial Overview and Investigation of Multiple Imputation Inferences under Uncongeniality," in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002.

Merkle, Daniel M. et al., "Unintended Consequences: The Coast of Purging Business Numbers in RDD Surveys," Public Opinion Quarterly, 73 (2009), 484-496.

Minato, Hiroaki and Luo, Lidan, "Towards a Better Estimate of Working Residential

Number (WRN) Rate Among the Undetermined: An Application of Survival Analysis," Paper presented to the American Association for Public Opinion Research, Phoenix, May, 2004.

Montaquila, Jill M. and Brick, J. Michael, "Unit and Item Response Rates, Weighting, and Imputation Procedures in the 1996 National Household Education Survey," Working Paper 97-40, National Center for Education Statistics, 1997.

Moore, Danna L. and Tarnai, John, "Evaluating Nonresponse Error in Mail Surveys," in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002.

Moore, Laura M., "A Multi-Level Analysis of Attitudes Regarding Women in Politics: Why is the United States South Different?" Unpublished Ph.D. Thesis, University of Maryland, 1999.

Moore, Laura M. and Vanneman, Reeve, "Context Matters: Effects of the Proportion of Fundamentalists on Gender Attitudes," Social Forces 82 (2003), 115-139.

Murray, Mary Cay, "Impact of Changes in the Telephone Environment on RDD Telephone Surveys," Paper presented to the American Association for Public Opinion Research, Nashville, May, 2003.

Nolin, Mary Jo et al., National Household Education Survey of 1999: Methodology Report. Washington, DC: National Center for Education Statistics, 2000.

Obenski, Sally, "Models and Applications of Administrative Records Research," Workshop on Data Linkages to Improve Health Outcomes, Washington, DC, September, 2006.

O'Hare, Barbara C.; Ziniel, Sonja; and Groves, Robert M., "Modeling Components of Response Propensity in Centralized Telephone Surveys," Paper presented to the American Association for Public Opinion Research, Miami Beach, May, 2005.

O'Muircheartaigh, Colm, "There and Back Again: Demographic Survey Sampling in the 21st Century," 2003 at www.fcsm.gov/events,papers2003.html

Oreopoulos, Philip, "The Long-Run Consequences of Living in a Poor Neighborhood," Quarterly Journal of Economics, 118 (2003), 1533-1575.

Pappas, Gregory et al., "Potentially Avoidable Hospitalizations: Inequalities in Rates between US Socioeconomic Groups," American Journal of Public Health, 87 (1997), 811-816.

Peel, Jennifer L. et al., "Ambient Air Pollution and Respiratory Emergency Department Visits," Epidemiology, 16 (2005), 164-174.

Peeples, F. and Loeber, R., "Do Individual Factors and Neighborhood Context Explain Ethnic Differences in Juvenile Delinquency," Journal of Quantitative Criminology, 10 (1994), 141.

Powell, Lisa M. et al., "Food Store Availability and Neighborhood Characteristics in the United States," Preventative Medicine, 44 (2006), 189-195.

Raghunathan, Trivellore and Van Hoewyk, John, "Disclosure Risk Assessment for Survey Microdata," Unpublished report, ISR, University of Michigan, circa 2005.

Rahn, Wendy M. and Rudolph, Thomas J., "A Tale of Political Trust in American Cities," Public Opinion Quarterly, 69 (2005), 530-560.

Raudenbush, Stephen Wand Bryk, Anthony S., Hierarchical Linear Models: Applications and Data Analysis Methods. Thousand Oaks, CA: Sage, 2002.

Regnerus, Mark D., "Moral Communities and Adolescent Delinquency: Religious Contexts and Community Social Control," The Sociological Quarterly, 44 (2003), 523-554.

Rosenfeld, Richard and Messner, Steven F., "Beyond the Criminal Justice System: Anomie, Institutional Vitality, and Crime in the United States," Paper presented to the American Sociological Association, San Francisco, 1998.

Rosenfeld, Richard; Bray, Timothy M.; and Egley, Arlen, "Facilitating Violence: A Comparison of Gang-Motivated, Gang-Affiliated, and Nongang Youth Homicides," Journal of Quantitative Criminology, 15 (1999), 495-516.

Rosenfeld, Richard; Messner, Steven F.; and Baumer, Eric P., "Social Capital and Homicide," Social Forces, 80 (2001), 283-309.

Safir, Adam et al., "Effects on Survey Estimates from Reducing Nonresponse," Proceedings of the American Statistical Association, 2001. Washington, DC: ASA, 2001.

Salvo, Joseph and Lobo, Arun Peter, "The Effect of Neighborhood Characteristics on Nonresponse in the Bronx Test Site of the American Community Survey," Proceedings of the American Statistical Association, 2003. Washington, DC: ASA, 2003.

Sampson, Robert J.; Morenoff, Jeffrey D.; and Earls, Felton," Beyond Social Capital: Spatial Dynamics of Collective Efficacy for Children," American Sociological Review, 64 (1999), 633-660.

Sampson, Robert J., Stephen W. Raudenbush, and Felton Earls, "Neighborhoods and Violent Crime: A Multilevel Study of Collective Efficacy," Science, 277 (1997), 918ff.

Scheuren, Fritz, "Macro and Micro Paradata for Survey Assessment," unpublished report, Urban Institute, 2000.

Shaw, Clifford R. and Henry D. McKay. 1969. *Juvenile Delinquency and Urban Areas*. Chicago: The University of Chicago Press.

Silver, Brian D.; Anderson, Barbara A.; and Abramson, Paul R., "Who Overreports Voting," American Political Science Review, 80 (1986), 613-624.

Slvo, Joseph J. and Lobo, Arun Peter, "The Effect of Neighborhood Characteristics on Nonresponse in the Bronx Test Site of the American Community Survey," Paper presented to the American Statistical Association, San Francisco, August, 2003.

Smith, Tom W., Anti-Semitism in Contemporary America, New York: American Jewish Committee, 1994a.

Smith, Tom W., "Developing Nonresponse Standards," in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002a.

Smith, Tom W., "Estimating Nonresponse Bias with Temporary Refusals," Sociological Perspectives, 27 (1984), 473-489.

Smith, Tom W., "Estimating the Status of Cases with Unknown Eligibility in Telephone Surveys," Paper presented to the Second International Conference on Telephone Survey Methodology, Miami, January, 2006a.

Smith, Tom W., "A Generation of Data: The General Social Survey, 1972-2002," GSS Project Report No. 24. Chicago: NORC, 2002b.

Smith, Tom W., "The Hidden 25%: An Analysis of Nonresponse on the General Social Survey," Public Opinion Quarterly, 47 (1983), 386-404.

Smith, Tom W., "A Methodological Analysis of HIV Risk Behavior from the 1988-1998 General Social Survey," GSS Methodological Report No. 92. Chicago: NORC, 1999.

Smith, Tom W., "Survey Non-response in Cross-national Perspective: The 2005 ISSP Non-response Survey," Survey Research Methods, 1 (2007), 45-54.

Smith, Tom W., "Total Survey Error," in Encyclopedia of Social Measurement, edited by Kimberly Kempf-Leonard. New York: Academic Press, 2005.

Smith, Tom W. and Sokolowski, John, "Using Audio-Visuals in Surveys," in The Handbook of Emergent Technologies in Social Research,, edited by Sharlene Hesse-Biber. Oxford: Oxford University Press, forthcoming

Snedker, Karen A.; Herting, Jerald R.; and Walton, Emily C., "Neighborhood Contextual Effects and Adolescent Substance Use: Exploring the Moderating Role of Neighborhoods," Paper presented to the American Sociological Association, Montreal, 2006.

South, Scott and Baumer, Eric P., "Community Effects on the Resolution of Adolescent Premarital Pregnancy," Journal of Family Issues, 22 (2001), 1025-1042.

South, Scott; Baumer, Eric P.; and Lutz, Amy, "Interpreting Community Effects on Youth Educational Attainment," Youth and Society, 35 (2003), 3-36.

Steeh, Charlotte et al., "Are They Really as Bad as They Seem? Nonresponse Rates at the End of the Twentieth Century," Journal of Official Statistics, 17 (2001), 227-247.

Stoop, Ineke A.L., The Hunt for the Last Respondent: Nonresponse in Sample Surveys. The Hague: Social and Cultural Planning Office, 2005.

Stoop, Ineke A.L., "Survey Nonrespondents," Field Methods, 16 (2004), 23-54.

Taylor, Marylee C., "How White Attitudes Vary with the Racial Composition of Local Populations: Numbers Count," American Sociological Review, 63 (1998), 512-535.

Taylor, Marylee C., "Fraternal Deprivation, Collective Threat, and Racial Resentment: Perspectives on White Racism," in Relative Deprivation: Specification, Development, and Integration, edited by Walker and Smith, New York: Cambridge University Press, 2002.

Thomas, Avis J. et al., "ZIP-code-based versus Tract-based Income Measures as Long-Term Risk-adjusted Mortality Predictors," American Journal of Epidemiology, 164 (2006), 586-590.

Tolbert, Charles M.; Lyson, Thomas A.; and Irwin, Michael D., "Local Capitalism, Civic Engagement, and Socioeconomic Well-Being," Social Forces, 77 (1998), 401-427.

Traub, Jane; Pilhuj, Kathy; and Mallett, Daniel T., "'You Don't Have to Accept Low Survey Response Rates' – How We Achieved the Highest Survey Cooperation Rates in Company History," Paper presented to the American Association for Public Opinion Research, Miami Beach, May, 2005.

Turrell, Gavin et al. "The Socio-economic Patterning of Survey Participation and Non-

response Error in a Multilevel Study of Food Purchasing Behaviour: Area- and Individual-level Characteristics," Public Health Nutrition, 6 (2003), 181-189.

Van Goor, Henk; Jansma, Folbert; and Veenstra, Rene, "Differences in Undercoverage and Nonresponse between City Neighborhoods in a Telephone Survey," Psychological Reports, 96 (2005), 867-878.

Voogt, Robert J.J. and van Kempen, Hetty, "Nonresponse Bias and Stimulus Effects in the Dutch National Election Study," Quality & Quantity, 36 (2002), 325-345.

Weakliem, David L., "Race Versus Class? Racial Composition and Class Voting, 1936-1992," Social Forces, 75 (1997), 939-956.

Weakliem, David L. and Biggert, Robert, "Region and Political Opinion in the Contemporary United States," Social Forces, 77 (1999), 863-886.

Williams, Jason et al., "Enhancing the Validity of Foster Care Follow-up Studies through Multiple Alumni Locations Strategies," Child Welfare, 85 (2006), 499-521.

Wirth, Louis. 1938. "Urbanism as a Way of Life." *American Journal of Sociology* 44:3-24.

Zanutto, Elaine and Zaslavsky, Alan, "Using Administrative Records to Impute Nonresponse," in Survey Nonresponse, edited by Robert M. Groves et al. New York: John Wiley & Sons, 2002.

Table 1

Aggregate-Level and Geographic Sources

Source	Geographic Unit	Variables
Energy Information Administration	GIS/Distance	Location of power plants
EPA/Superfund Sites	GIS/Distance	Location/type of Superfund sites; 26 variables in table plus follow-up site narratives
Federal Bureau of Prisons	GIS/Distance	Location of federal prisons
HUD/Subsidized Households	GIS/Distance	Location and characteristics of government subsidized HUs and tenants; 64 variables
StreetPro/MapInfo Professional	GIS/Distance	Location of hospitals, schools/universities, places of worship, government facilities, cemeteries, golf courses, recreational facilities (e.g. zoos, museums), major retail centers, transportation hubs, airports, etc.
Trauma Centers	GIS/Distance	Location and type of trauma centers
PrisonerLife.com	Address	Location of 1507 correctional facilities
FUNDRACE (huffingtonpost)	Address/Zip code	Amount and recipient of donations; 4 variables on amount and party of campaign contributions by zip code

Table 1 (continued)

Geographic Source	Unit	Variables
National Center for Charitable Statistics/IRS	Address/Zip code	Not-for-profits except churches, about 42 variables covering types of organization, assets, and various administrative matters
Claritas	Block Group to County	Demographics, Housing/Property, Automobiles, Financial, Telephone, Purchases, Outdoors, Insurance, Audio/Video, Contributions, Medical, Interest, High-Tech/Computers, misc. About 1,000 aggregate variables
US Census, Decennial Census	Block Group to County	Variables vary by geo level. For Census tract hundreds of demographic combinations by such variables as age, race, ethnicity, gender, marital status, household size, income, labor force status, education, etc.
Dunn & Bradstreet Businesses	Zip code	40 mostly economic variables on 10 million businesses
EPA	Zip code	Various databases: Envirofacts, Toxics Release Inventory, Facility Registry System, Enforcement and Compliance History
Internal Revenue Service	Zip code	38 variables on individual taxes and income
US Economic Census	Zip code, County	Number, type, size of employers
Association of Religion Data Archives	County	466 variables mostly about number of adherents in specific denominations

Table 1 (continued)

Geographic Source	Unit	Variables
Audit Bureau of Circulations	County	Circulation levels for hundreds of periodicals
County Characteristics, 2000-2007 (ICPSR)	County	470 variables, mostly Census and governmental variables
FBI Arrest and Offense Figures	County	63 measures of arrest and crimes reported to police
Presidential Election Returns	County Community/Precinct	Votes for presidential candidates, other offices Sub-county votes are not centrally available