Possible Expansion of Individual Statistical Records by Loading with Derived Variables

Hiromi MORI*

Summary

In the long-term downsizing phase of the society, statistical authorities become increasingly dependent, in place of the conventional measures, on a wide range of possible substitute procedures for obtaining statistical data under touch resource constraints. Due to the continuous and large-scaled cutback in human and budgetary resources, measures which substitute statistical surveys, such as more extensive use of administrative records for statistical purposes and more exhaustive cultivation of existing data have attracted growing concerns in recent years. It is in such historical context that data integration is regarded as one of the promising breakthroughs from the confronting issues.

This paper brings to light the possible expansion of individual statistical records by loading with a set of derived variables obtained from existing multi-sourced data. The study was originally inspired and encouraged by an intellectual input gained by an idea of derived variables which are practiced in British census micro data (Sample of Anonymised Records: SARs). As the discussion will address, endogenous expansion of record information not only covers data integration based on the surveyed variables but it further extends to involve non statistical information obtained through survey and administrative processes. Among others, one of the major objectives of this paper is to elucidate the outstanding importance of geographical information given as GPS coordinates in exploring information potential of existing individual records. As will be evidenced in the course of discussion, GPS coordinates obtained either by the direct capturing by means of mobile terminals or by converting addresses into coordinates may be expected to tap new frontiers in use of statistical data.

Introduction

Statistical data obtained through censuses and surveys were compiled as a set of tables and have been disseminated mostly as printed copies. In recent decades, internet has acquired ever growing importance in terms of a channel of supplying data

* Faculty of Economics, Hosei University
4342 Aihara, Machida-shi, Tokyo, JAPAN 194-0298
Email: hiromim@hosei.ac.jp

for wide-ranged users. Anonymised individual records which are termed usually as "micro data" for public use (PUMS) and those for licensed users also worked as the effective additional data providing channels together with on-demand data processing services provided either by statistical authorities or outsourced agencies.

It is true that among micro data users there exists some who tried to explore potential usability of obtained information through cross-sectional as well as longitudinal linking of individual statistical records. As for the possible expansion of variables of existing individual statistical records through horizontal (cross-sectional) as well as vertical (longitudinal) record linkage, the author has discussed already in a forerunning essay of this book〔Mori 2011a〕. It would be relevant, therefore, to focus the discussion here on the possible expansion of information of individual statistical records through assimilated imputations based on existing variables. It is expected that existing individual statistical records can amplify their information potentials by subjoining derived variables generated from variables which the existing individual statistical records have with them.

The aim of this paper is twofold: first to address some practical examples of expanding dimensions of individual statistical records through reasonable measures of integrating data that can be practiced in relation to the existing variables and second to give a brief review on the implications of their outcome in terms of usability of the data.

1. A set of information obtained through questionnaire-based survey

Statistical questionnaires carry a set of surveying items to be answered by respondents. Those items which are ascribed to multifarious attributes of the respondents are usually called "face-sheet." The rest of questions which compose major body of the questionnaire constitute the "domain" segment of the survey. In addition to the surveying items which constitute (a) face-sheet and (b) domain, questionnaires usually carry a set of columns which fall in the following categories from (c) through (f).

Columns to fill the township codes and survey tract codes (c) are categorized as the first group. Filled names and telephone numbers of the respondents (d) are also used to make probable *ex post facto* inquiries about the responded answers. Field surveyors or enumerators are also requested to fill columns regarding external issues such as location attributes of the survey tract and the exterior characteristics of the buildings, which are categorized as group (e). Some surveys also carry columns to fill the names, addresses and identification code of the surveyed units such as establishments and companies, which fall in the category (f). The obtained information through (f) is not employed for statistical purposes in the narrow sense but is used primarily to compile the population directory for conducting sampling surveys.

Entries in returns are transcribed electronically into magnetic units which store raw individual statistical records (hereinafter termed as RISRs). RISRs usually carry, together with survey identification code and the year of conducting the survey, variables which fall in categories (a) through (c), (e) and occasionally (f). Since survey returns occasionally give inconsistent entries and missing variables, the RISRs are to be processed subsequently to correct and impute data in the editing steps finally to achieve the edited individual statistical records (hereinafter termed as EISRs).

2. Dissemination channels of survey results

EISRs are further processed in several dimensions. Firstly, they are compiled and tabulated for dissemination. Although printed reports used to have been the major medium of releasing tabulated results, webs are now acquiring wider acceptance in the arena of making public the survey results.

Secondly, the data archives of official statistics that stockpile datasets compiled from the EISRs now play in some countries a outstanding role in disseminating the survey results, where users can obtain the required tables according to their respective analytical purposes by means of queries. For the confidentiality reasons, EISRs are usually stored in the form of data cubes and some tables which contain cells with rare cases below the threshold are masked partially.

The third channel of disseminating the results is the on-demand data processing services. EISRs are processed in-house according to the users' requests and the obtained results are examined carefully by the competent committees such as the panel from confidentiality perspective before release. In this channel, users are capable to make indirect access to the EISRs that afford comparatively wider options in processing the data.

Anonymised individual statistical records (hereinafter termed as AISRs) are to be categorized as the fourth channel of disseminating the data. EISRs are anonymised through various measures to liquidate or alleviate the risks of the possible disclosure of confidentiality. Although anonymising measures yield more or less information loss from EISRs, users are qualified to process for themselves AISRs datasets according to their own research purposes.

3. Exploring potential of individual statistical records - horizontal and vertical expansion

It is worth noting that, among existing manifold channels of disseminating the survey results, the micro-based channels which are listed above as the third and fourth ones are less constrained with regard to the application of data compared with the first two table-based channels due mainly to the disaggregated form of the analytical

materials.

In addition to the broader flexibility which analysts can enjoy in data processing, individual statistical records are also distinguished from aggregated data (macro data) in terms of possibility of exploring information potential which they have immanently possessed. By making effective use of the competent linking key variables such as uniform identification numbers assigned to respective surveyed units, one can easily achieve to expand the dimension of variables by integrating individual statistical records cross-sectionally, which the author terms "horizontal integration" of records. Individual records can also be integrated in time horizon perspective to form panel datasets. Longitudinal integration of individual records can be called as "vertical integration." As for the discussion regarding the possibility of exploring information potential of the individual statistical records, see〔Mori 2011a〕.

4. Exploring information potential through endogenous expansion of variables

In the UK a special dissemination model was set up in the 1990s for making access to samples of anonymised record data from the 1991 Population Census. Two different types of micro data sets, which are called SARs (Samples of Anonymised Records): 2 percent individual SAR and 1 percent household SAR, became available in 1993. The Census Microdata Unit of the Centre for Census and Survey Research (CCSR) at University of Manchester is in charge of providing data to academics as well as for business users.

SARs have enjoyed wider acceptance not only among academics but also business users due to the disaggregated nature of the data which provides wider options compared with aggregate data in the data processing operations. However, in addition to particular attribute ascribed to the form of data as disaggregated datasets, SARs seem to enjoy their reputation from another perspective. As the following discussion in this paper will demonstrate, it is worth noting that SARs are especially distinguished from other micro-based datasets such as those released from the Archive at Essex University or Longitudinal Study data on account of a set of derived variables loaded to the respective individual records.

Followings are the list of derived variables adopted for the 2001 SARs.

Table 1　List of derived variables adopted for the 2001 SARs

〔**derived person variables**〕 Education Deprivation; Employment Deprivation; Health and Disability Deprivation; Housing Deprivation; Generation Indicator
〔**derived household variables**〕 Number of Usual Residents in household; Number of Persons in household aged 65 or over; Number of Cars in household; Number of Household Members with poor health; Number in Household with limiting long-term illness; Number of Employed Adults in household; Household with Students away during term time; Multiple Ethnicity Household Indicator; Social Grade of HRP; Number of Families in households; Family Type; Dependent Children in family; Sex of FRP; Economic Position of FRP; NS-SEC of FRP; Persons per Room; Occupancy Rating of Household; ONC imputed person/household; Number of EDIS donors; Indicator marking records that have been imputed; Synthetic indicator of LA

Source: http://www.statistics.gov.uk/census2001/sar_update.asp

A set of derived variables attached to the existing individual statistical records are expected to contribute to enhance the usability of the data by expanding the dimension of variables and thus to provide wider options for analysts in conducting the data processing operations. Put differently, an idea of loading the existing records with derived variables suggests the possibility of expanding information potential of individual statistical records. Some examples by type of data will be discussed in the following sections.

5. Endogenous expansion of questionnaire return information

For the convenience of discussion, let us begin the discussion with giving a definition of endogenous expansion of information inherent in the individual statistical records. The term "endogenous" denotes in this context the generation of new derived variables from the existing ones collected either for statistical or non-statistical purposes in the course of survey operation with reasonable calculation procedures by one to one or one to n matching to the variables obtained from other sources. As the following discussion will evidence, not only statistical variables but also even non numerical data immanent in the existing records can afford to provide key information to generate relevant derived variables which contribute to expand the information potential of the existing individual statistical records.

Following paragraphs will address some examples by type of record units.

(1) Personal individual statistical records

Personal individual statistical records usually carry a variable that addresses age.

It is often the case that the dates of birth obtained from returned questionnaires are converted into age data. It works as one of the variables which compose the face-sheet segment of the questionnaire to compile cross tables by age (or by age class). The importance of this variable is not confined to such usage. Especially from the standpoint of data integration, ages can and should be regarded as the "primary information" from which many variables can be derived by conducting acceptable linking with variables from varied sources.

It is probable that some age groups of population can be arranged so as to mold a peculiar subpopulation whose ways of thinking and behaviors are remarkably distinct from others. Such subpopulation groups are usually called "generation." Wartime and post-war generations, baby boomers, their second generations and the so-called "lost generation" are widely known examples.

It is well known that statistical results are more or less affected by three types of effects: era, age and generation. Some statistical variables may be more strongly influenced than others by generation effect. Although most survey results carry cross a set of tables by 5-year age class, generations are often grouped with irregularity in terms of age, possibly longer or shorter than 5 year of age. Five year age classes, therefore, do not fully meet the analytical needs to portray the probable generation effect. By loading the individual statistical records with newly derived variables generated through re-grouping the age, they will remarkably enhance the information potential of existing records.

Age data can also be expanded in different way. By making reference to one's educational attainment, it will be possible to generate from age data a new derived variable that addresses a fairly good proxy of the year of one's first involvement in the labor market. As many empirical studies have already evidenced, the actual condition of the concurrent labor market significantly affects the subsequent labor involvement behaviors of the new entrants. Economic indicators, such as annual growth rate, annual average unemployment ratio or active job openings-to-applicants ratio, which can be added to the existing records as derived variables by the aid of age and educational attainment data, seem to help provide a set of meaningful additional information to practice micro-based analyses on the employment behavior.

(2) Household individual statistical records

Questionnaires or entry books of household surveys such as the Family Income and Expenditure Survey or the National Survey of Family Income and Expenditure conducted by Japanese Bureau of Statistics carry columns for address data of residential units to be filled by respondents. Information obtained through these columns has been used exclusively for inquiry purposes to edit the improperly responded answers. The address information, therefore, has not been employed for statistical purposes e.g. to compile statistical tables up until today.

It would be well supposed that the geographical points where the residential units are located might more or less affect economic behaviors of the households or respective family members. Despite the existence of probable location-led affects, traditional survey results could not describe their effects due mainly to the inadequate treatment of location information. Because of the fatal absence of disposable data, empirical model-based approaches were obliged to disregard their possible affects. It is apprehended that estimated parameters in traditional manner should more or less carry biases.

Address information immanently holds a wide spectrum of possibility in generating various derived variables which enable to amplify enormously the applicability of the existing individual statistical records. By using the address matching procedures, one can load individual household records with GPS coordinates for most addresses in urban areas. Thanks to the widespread modern remote sensing technology, field surveyors became able to collect the relevant coordinate information with sufficient accuracy by simply clicking the handheld terminals.

It is probable that several residential units are linked to the identical coordinates due to the existing address giving system. Hence, a pair of coordinates (x,y) corresponds to one unit or occasionally to a certain number of units. Even in the latter cases, the obtained coordinates definitely give actual geographical location information which a set of relevant units commonly share. As far as the characteristics of the entities which fall in a particular polygon are concerned, the fact that a certain number of the individual statistical records obtained from surveys share identical coordinates will bring about no serious issues for analytical purposes.

Coordinates information can easily be processed to assess the accessibility to public transportation (distance from railroad stations and bus stops), commercial and public facilities, such as stores, banks, schools and medical care centers. Individual statistical records from the Housing and Land Survey of Japan, for example, carry access information from such facilities as categorical variables. Vector data given by the coordinates are privileged to assess them numerically which enable to provide wider options in analyses. The derived variables in this way can represent location-related factors which also govern the performance or the behaviors of the surveyed units.

As was already described in (1) above, these newly added variables which can be derived from address of residential units also effectively expand the scope of usability of individual statistical records.

(3) Business individual statistical records—establishments and enterprises

Establishments and enterprises usually operate their business activities at specified sites conditioned by various environmental factors. It is well supposed that influencing factors may differ among industrial sectors. Business activities of

commercial stores, such as super markets and convenience stores, are more or less affected by the density of regional population, the presence of neighboring rivals and so on. Manufacturing firms are more likely to benefit from accessibility to the public transport, i.e. an easier access to the motorway interchanges, airports, and seaports. Accessibility to water and electricity supplies and among others the presence of massive well-educated working population are of central concerns for them. Agglomeration of multifarious firms also helps promote prosperous business activities through various benefits supplied by neighboring businesses. As for the service sector, density of regional population in terms of human and business may also affect their prosperous performance.

National Survey of Prices conducted by Japanese Bureau of Statistics collects price data for a number of designated commodities and services each five years for commercial establishments such as shops, department stores, supermarkets, convenience stores together with the information on the presence or absence of neighboring rivals. The survey results evidence the cut down effects of commodity and service prices by type of commercial shops caused by the presence of neighboring rivals.

In case when price data collected from stores were loaded with coordinates, varied competitive patterns by neighboring rivals can be identified not by the survey process but simply by the subsequent calculation. It is expected that coordinate information will provide materials with wider perspective of applicability for analyzing possible effects on price creation which is ascribed to the multifarious location attributes of respective commercial stores.

(4) Individual statistical records on residential units

Housing and Land Survey of Japan is a large-sized survey with about 8 percent of sampling ratio. It provides comprehensive data on the actual conditions and tenure of housing units and lands each five years. Variables compiled into a huge set of tables are mostly of return questionnaires origin. Besides these variables, individual survey records, however, carry also additional variables obtained by local staffs and field workers.

Local staffs who are in charge of survey operation are requested to prepare in advance an itemized list characterizing the survey tracts which covers issues such as use-defined land under the Article 8 of the City Planning Law, the building-to-land ratio, a floor-area ratio, sewerage and distance to the nearest railroad stations or bus stops, city parks, public halls/meeting facilities, emergency refuge sites, day service centers for the aged, medical care facilities, post offices/banks, convenience stores, nurseries, elementary schools and junior high schools. In the course of the survey operation, however, field workers are also requested to examine issues, such as type of housing units, housing units by construction material, and road 6 meters in width or

wider by radius of designated range of distance. Added information captured by local and field staffs that constitutes a part of individual statistical records seems to affect the quality of living well-beings.

In case when the surveyed units' records were loaded with the GPS coordinates, the fairly more accurate results could be achieved about the accessibility simply by an electronic *ex post facto* calculation. One can obtain the results with any buffering radiuses to replace inflexible categorical zoning e.g. less than 500m, 500-1km, etc. The introduction of electronic calculation measure not only helps relieve responding burdens and field surveyors' workloads and finally contribute to save budgets but also brings about the notable improvement in terms of accuracy of survey results because of the possible avoidance of inappropriate estimation caused by local and field staffs.

The questionnaire of this survey carry question on monthly rent for rented housing units. Unfortunately, however, as for the owned residential units, neither questions regarding the cost spent for their acquisition (booked price) nor their current value estimates. Absence of the relevant price data regarding the owned residences renders comprehensive economic analysis of residential units quite limited.

Once individual records obtained through this survey were loaded with GPS coordinates, it is expected that the existing records will be able to enjoy possible expansion of inherent information by means of data fusion with those from other sources. There are many sort of land price data available in Japan obtained from varied sources of periodical surveys and administrative records which provide widely covered land price data. Since individual observation data carry location information as addresses, by using GPS coordinates as key link variable, residential and land unit records collected through the Housing and Land Survey may be able to acquire approximate of land price as derived variable by applying those obtained from neighboring spots, although many research works yet to be done until the new variable became actually able to enjoy relevance.

6. Analytical implications of expanding dimension of individual statistical records

As surveys and administrative records merely document some limited aspects of multifarious entity of the surveyed units, the obtained data do not always portray a comprehensive picture of the actual state of their real existence. As paragraph 5 of the forerunning essay of this book〔Mori 2011b〕 has already discussed in detail, among a set of variables which account for the phenomenon, there are not few that appear to be documented in other surveys or administrative records. In such cases, the obtained estimates from one particular set of survey results turn out to be of biased nature.

This paper is motivated originally from the expectation that analysts can enjoy wider option in data processing by integrating or fusing the existing individual records.

A set of newly loaded variables obtained from different sources through matching or converting procedures based on existing statistical or non-statistical variables are expected to touch the new frontiers in terms of analyzing data which even the highly sophisticated measures were unable to have achieved so far as the empirical studies are based on the datasets from surveys of substantially stand-alone nature.

A focal motive of this paper, among others, was to bring under light the GPS coordinates as one of the most effective key variables to conduct data integration which is one of the top concerns in contemporary official statistical practices. It is worth noting that model analyses thus far have generally overlooked the possible affects influenced by location-related factors. Although they seem to have meaningful effects on the dependent variable, a core framework of traditional models was build up of a selected set of independent variables which are supposed to embody location-based factors within them. Consequently, possible affects caused by dependent variables so far untouched in terms of location-related factors are likely to have given rise to more or less biases in estimated parameters resulting from improper treatment of residuals. It is thus that involvement of such variables in model building processes is expected not only to cultivate new arena in regional analyses but also to claim some possible modification of the already established academic achievements.

Concluding remarks

As for the expansion of dimensions of variables of existing individual statistical records through exact matching using relevant identifiers, we already have a lot of achievements in actual statistical practices. Individual statistical records, which are archived in relational manner, keep potentiality of cultivating immanent information by micro-based integration. Even in case when records are not necessarily linked by exact matching, they can also expand their information potential by statistical matching. Under the contracting human and budgetary resources allotted to the statistical production, national statistical authorities of many countries are increasingly keen on compiling new statistical data by more comprehensive use of existing information. This paper discussed some possibilities of fusing the data from different sources as a broader category of data integration.

As described above, this paper is indebted to the methodological input suggested by the concept of derived variables which address the distinctive value added to the UK census micro data (SARs). It is worth noting that individual statistical records have in themselves some endogenous elements to expand their information. Some examples of possible expansion were already illustrated in the discussion.

Discussion in this paper also highlighted the fact that not only statistical variables and descriptive information directly collected from the surveyed units but also information captured by field surveyors in the process of survey operation can help

generate additional information attached to the existing individual statistical records. A net contribution of this paper is, among other things, the full-scaled evaluation of location information which most of return questionnaires usually carry. Addresses filled by respondents in the questionnaires are substantially of non statistical nature in terms of the type of information. By converting addresses into GPS coordinates, the surveyed records can acquire powerful linking key variables to enhance remarkably the information potentials which the individual statistical records have originally carried in latent manner.

With regard to the possible expansion of the dimensions of variables of the existing individual statistical records, it is noteworthy to refer here to one forerunning trial practiced by the Ministry of Economy, Trade and Industry (MITI) in the Census of Commerce record data.

A survey report on the characteristics of commerce businesses by location provided by the Census of Commerce carries a list of tables by various types of regional areas. The report carries survey results on establishments engaged in wholesale and retail trades based on the definitions of classification of characteristics of regional areas in accordance with the "Large-Scale Retail Store Location Law."

Characteristics of respective regional areas are classified into commerce-integrated areas, office building areas, residential areas, industrial areas, and other areas according to the "use-defined land" stipulated by Article 8 of the City Planning Law. As table 2 shows, MITI further divides commerce-integrated areas into 5 subcategories: areas around stations, city-area-type, residential-background-type, residential-type, and other types.

If individual business records were loaded with these variables derived from location information through GPS coordinates, the newly created datasets with expanded dimensions of variables are expected to cultivate the untouched scope of analyzing business activities. This practice seems to suggest one of the future possibilities of individual-based data integration.

Table 2 Location Characteristics of Areas of Commercial Stores-classification and Defintion

No / Classification	Sub-classification of commerce-integrated areas	Definition
10 Commerce-integrated areas		<Areas which constitute a shopping district in near-commercial areas or commercial areas of "use-defined land" under Article 8 of the City Planning Law> One connerce-integrated area usually forms one shopping district which refers to an area that has 30 or more retailing shops, restaurants and service industries. A shopping center or multi-purpose building, such as a station building or co-operative department store building, that falls under the definition of "one shopping district" is, as a general rule, regarded as one commerce-integrated area.
11	Commerce-integrated areas around stations	<Commerce-integrated areas located around JR or private railway stations> They, however, as a general rule, do not include areas located around streetcar or subway stations.
	Establishments in station wickets	
12	City-area-type commerce-integrated areas	<Commerce-integrated areas located in a busy shopping or office building district in the center (except areas around a station) of a city>
13	Residential-background-type commerce-integrated areas	<Commerce-integrated areas having a residential or housing complex district as past of their background>
14	Residential-type commerce-integrated areas	<Commerce-integrated areas (except those in the center of a city) located mainly along a national route or major road>
15	Other types of commerce-integrated areas	<Commerce-integrated areas that cannot be classified into any of the above four categories, such as shopping districts in tourist resorts and those shrines and temples>
20 Office building areas		<Areas that do not come under the categories of near-commercial areas or commercial areas of "use-defined land" under Article 8 of the City Planing Law>
30 Residential areas		<First-class or second-class low-rise residential building areas, first-class or second-class medium- or high-rise residential building areas, or first-class or second-class residential areas or quasi-residential areas of the "use-defined land" under Article 8 of the City Planning Law>
40 Industrial areas		<Quasi-industrial areas, industrial areas or exclusive industrial areas of "use-defined land" under Article 8 of the City Planning Law>
50 Other areas		<Areas that do not come under any of the above four categories>
	Establishments in toll roads	

〔source〕Census of Commerce-report by characteristics of location (retail trade) , p.25 (partly revised)

Reference

(1)Ministry of Economy, Trade and Industry (2009), *Census of Commerce- report by characteristics of location* (retail trade).

(2) MORI, Hiromi (2011a), The Expansion of Data Dimensions by the Micro-based

Integration of Statistical Records, *Bulletin of Japan Statistics Research Institute*, Vol.41
(3)MORI, Hiromi (2011b), GPS Coordinates and the Possibility of Micro-based Integration of Statistical Records, *Bulletin of Japan Statistics Research Institute*, Vol.41

Acknowledgement
This research has benefited from financial support through the Grant-in-Aid: "Multi-faceted Studies for Exploring New Frontiers of Official Statistics by Using GPS Information" (#40105854) of Japan Society for the Promotion of Science.