The Expansion of Data Dimensions by the Micro-based Integration of Statistical Records*

Hiromi MORI**

Summary

From the dawn of the history of modern statistical surveys up until the late 20th century, expanding needs for statistics have been chiefly met by launching new surveys. Practices of official statistics, however, seem to have changed its phase drastically in recent decades. National statistical authorities of most countries are now challenged to accommodate the expanding needs for the qualified statistics under the tough budget and human resource constraints together with the respondents' attenuating cooperation to the survey taking.

Due to the continuous budget cuts, statistical authorities are driven to take the alternative policies by partly replacing traditional survey taking with other possible measures. Extensive use of information captured through administrative process for the statistical purposes is one of the fusible options. Another option is to create new statistical information by integrating existing surveyed and administrative records. With regard to this, many national statistical authorities have launched the official statistical data archives where multi-sourced individual records are stored not only for the historical use but also for the data production purposes.

This paper focuses the discussion on the extended possibility of creating statistical information by integrating existing individual records. The argument partly refers to the data fusion which supposed to be a natural extension of the idea. As will be evidenced from the discussion, it not only expands the scope of available information, but also contributes substantially to enhance the quality of statistical cognition that an approach simply through a whole set of existing stand-alone survey results is unable to attain.

1. Background

An era of 19th and 20th centuries is hallmarked by the dominant presence of survey statistics, in which censuses and surveys have been the major channels for national statistical authorities to obtain statistical data. Many new surveys were introduced one after another to satisfy the emerging socio-economic needs for new statistical data. Although they were sporadic in nature in the early days of statistical practices with no meaningful relationships among them, then in a course

---

* Contents of this paper are partly based on the presentation "Exploring Usability of GPSed Records - A data typological approach" made at the workshop "Statistical Innovation: Use of GPS and GSM data and integration" organized by Statistics Netherlands on September 6, 2010 in Heerlen, and further elaborated based on inputs revealed at the workshop.

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

of decades a set of stand alone surveys have been organized to form a system of statistics. The introduction of the concept of population in statistics has marked an epoch to systematize surveys, because under the new system many surveys became able to be related to the population captured by the censuses. Official statistics in the latter half of the 20th century is basically characterized by the system constructed by censuses and surveys of more or less independent nature.

In the high era of the survey statistics, the socio-economic developments prepare in itself a drive to create the coming new stage. It is the metamorphoses in survey conditions that seem to have brought about this turnabout in statistical practices. Due to the peoples' enhanced self-consciousness of the privacy since the 1960s, respondents became increasingly reluctant to cooperate in survey because of the possible threat of privacy disclosure risks and thus respondents tend to evade surveys. Furthermore, the swollen financial deficits during the long lasting depression urged the governments more efficient performance. Budget and the number of personnel allocated to statistical practices were appropriate targets for downsizing policy. On the other hand, for governments which are subjected to work out policies to redistribute the retrenching pieces of pie and for businesses which are obliged to make decisions under augmenting uncertainty, more detailed and high quality official statistics became prerequisites for their successful operations.

Apart from prosperous 1960s, there no longer exist sufficient resources for national statistical authorities to launch new surveys to meet the ever expanding statistical needs. Exploring untouched frontiers of existing captured information including that of administrative records was among possible breakthroughs. It was in such historical context that statistical authorities of many countries directed their attentions to the integration of existing data as an effective countermeasure to meet the growing requirements. The motivation of this paper originates from the understanding that the turn of the century from the 20th to the 21st is also marked as a historical turning point in the development of official statistics.

The aims of this paper are threefold. First, it will shed light on the advantages of questionnaire-based surveys over the so called "table-based surveys" from the standpoint of data integration. Second, it will discuss the patterns of data integration together with attributes of the generated datasets. And finally it articulates statistical meanings of data integration which indicates that the integration is not simply the expansion of dimensions of variables of existing respective individual statistical records but also deeply involved in statistical cognition of the population. The discussion in this paper as a whole will evidence the fact that questionnaire-based surveys not only enable to yield multifarious tabulated results but also involve elements of potential expansion of dimensions by integrating individual statistical records from multiple sources.

1. The advantage of questionnaire-based surveys over table-based surveys

Official statistics dates far back to the ancient ages when the sovereigns conducted censuses for the purposes of military mobilization and taxation which give statistical snapshots of the population

at a particular reference date. Under the pre-modern society, besides such static representation of the states, powers such as governments and churches have documented events of the respective persons at relevant stages in their life course. Churches kept records of baptisms, marriages and burials in parishes and notary public offices and tax offices maintained records of transactions of real estates and other goods upon occasion. They documented the relevant events systematically in a whole scope of areas so far as the power exercises its validity and thus the captured records are dynamic in nature and give a number of monthly or annual events occurred during the referenced period of time.

An era that took over the liberal economies was characterized among others by the extensive government intervention into economy. In order to settle upon relevant policy measures and their successful operations, governments became increasingly dependent on the relevant statistical data which traditional sources could no more satisfy in kinds as well as in timeliness. It is worth noting that early statistical surveys in modern era have also carried the policy-oriented nature.

(1) table-based surveys

At the dawn of the history of modern statistical surveys, the so-called "table-based surveys" were major means of collecting statistical information. In early days surveys were poorly organized and sometimes national statistical authorities gave merely the list of the survey items. In the early surveys neither measuring units such as number of pieces, metric tons or currency nor reference date of surveys were clearly instructed. The wishes and ideas of survey designers were conveyed down through hierarchical order to regional levels and finally to the field workers. How to design the survey format was substantially left to each local authority.

From the outset, the surveys had the clear image of the tables to be compiled. In operating the surveys the field workers directly filled respective cells in the table format with aggregate number of survey items for the pertinent region. The collected information at field was reported upward through hierarchical order finally to achieve national totals. The aim of the survey was among others to obtain total sum by survey items. However, due to the absence of uniformed measuring units in earlier surveys, regional subtotals could occasionally achieve no national total.

Although the survey items such as attributes as well as activities of the surveyed units are inherent in individual persons, households, establishments and companies, they are simply counted as a group totals in the table-based surveys. Since the information of the respective surveyed units is merged together from the outset in each cell of the table, it is not possible to disaggregate them subsequently and to reestablish individual records. Inability of establishing individual records in the table-based surveys also slams the door to the micro-based data integration.

(2) questionnaire-based surveys

Units such as individual persons, households, establishments and enterprises have their own attributes and perform their daily activities by establishing varied relations among units. The total entities of these issues comprise the object of statistical cognition. Since individuals compose

ultimate units to be surveyed, the introduction of questionnaire-based surveys has opened up a new arena in the development of survey techniques in terms of obtaining statistical information not in aggregate but in individual manner.

Questionnaire-based survey is distinguished from its predecessor in the manner how original statistical information is collected in survey practices. In the table-based surveys the obtained original statistical information was already aggregated at the time when field workers fill the forms. Since individual survey records have one to one correspondence to the respective surveyed units in the questionnaire-based surveys, questionnaires record individual unit's attributes, behaviors, activities and outcomes thereof in disaggregate manner. Thus, the surveyed datasets are given as N×M matrix, where N and M denote the number of the surveyed units and the number of surveyed variables, respectively. By counting the number of cases using respective variables, obtained information is able to be processed so as to yield statistical tables with any arbitrary combination of surveyed variables. Furthermore, new categorical variables generated through combining several existing variables can expand the scope of variables, which contribute to enrich the statistical outputs.

The evolution of survey methods from the table-based to the questionnaire-based ones has marked an epoch in the sense that it can manage to compile statistics subsequently based on the individual records that allows to expand enormously the scope of statistical outputs. Because of the less developed data processing technologies, however, it was rather recent when national statistical authorities became able to enjoy their potentials in full scale by yielding a wide spectrum of statistical outputs.

2. Attributes of statistical information

This paragraph will identify characteristic feature of statistical information in comparison with other digital information such as pictures and sounds.

(1) characteristics of picture and sound information

Digital still pictures are drawn by tone data which indicate colors and shade carried by respective picture cells (pixcel) which have positional information. The smaller the picture cells are, the clearer becomes the portrayed pictures. When one replaces the data carrier from picture cell to volume cell (voxcel), the dataset can portray three-dimensional cubic pictures.

Digital sounds are composed by three elements: pitch, intensity and tones. The digital compound of sounds resolves itself into intensity carried by microscopic time units termed as "sampling". The fidelity of sounds expressed in digital data depends on the frequency of sampling.

It is remarkable that both digital picture and sound information are common as far as their structures of information are concerned, because one or several dimensional variables which form a so to say data body are carried by the data carriers such as pixcel (voxcel) and sampling. These two set of information elements are united to form a single record which has one-to-one correspondence

to respective pixcel (voxcel) and sampling. Worth noting here is that data carriers such as pixcel and voxcel do not belong to the object to be portrayed but they are given as coordinates of the shooting frame. The same notion is valid also to the sounds. It is not the sound source itself but the quality of operating apparatus such as digitizers that governs the density of information.

(2) characteristics of statistical information

Besides a set of survey items that are filled according to the response offered by each surveyed unit, survey questionnaires cover also issues such as location code, family code, member code etc. Filler's name and phone number are used to implement the possible *ex post* inquiries to the relevant respondents. Some surveys have also columns which local statisticians and field workers document on their own accord such as the relevant attributes of the surveyed tract and the exterior or structure of the residential facilities. As the Japanese Establishment and Enterprise Census, which was recently remodeled into the Economic Census, has aimed also to give population of establishments and enterprises for various surveys, census questionnaires have also columns to fill the name and address as survey items.

The information obtained through questionnaires are read, coded when necessary and stored as a set of individual-based records together with survey identifier information. Diagram 1 illustrates the rough image of individual record of household surveys.

□	date of survey		location codes			ID		survey items																
						family code	member code	attributes							surveyer's remarks									
								household			personal													
survey identification code	year	date	prefectural code	city code	survey tract code	**family sample number**	**seq. number of persons in family**	two or more, one person households	types of family	…	sex	age	education	…	…	item1	item2	…	survey item1	survey item2	survey item3	…	…

Diagram 1 Illustrated example of an individual survey record

Characteristic features of statistical records transcribed from the returned questionnaires can illustrate as follows. The data body with multi-dimensional variables is carried by identifiers which have one-to-one correspondence to the surveyed units such as individual persons, households, enterprises etc. Take the colored picture for example, three dimensional variables that correspond to the three primary colors form the data body which is carried by respective pixcels. They are polymerized or melted together to generate one color. In contrast, statistical records are distinguished from digital sound or picture information by showing off the importance of respective variables.

Table 1 gives a rough sketch on data carriers and the manner how variables are carried by them by type of information.

types of information		data carrier	relation between variables
sound		sampling	polymerization of variables
picture	plane	pixcel	
	cubic	voxcel	
statistical records		unit ID	overlay of variables

Table 1 Data crriers and the carried variables by type of information

As is obvious from the above discussion, in case of statistical records, it is not the units given by the observing apparatus but the surveyed units that carry data body information. In other words, apart from other digital information, one can overlay multidimensional variables over respective unit IDs (statistical ID numbers of establishments and persons). Individual statistical records are also distinct from others by the fact that the overlaid multi-dimensional variables can have meanings not only in combination but also respectively. Such informational characteristics inherent in statistical records suggest the possibility of expanding dimensions by unit-based integration of data which addresses another advantageous element of questionnaire-based surveys over table-based ones.

3. Expansion of dimensions by integrating data

A set of information obtained through surveys usually form an individual record. In case when the records share identification number assigned to respective surveyed units, records from varied sources are easily linkable. Variables such as names, addresses, date of births and telephone numbers also help work as possible linking keys in statistical matching. Matching individual records from different sources can yield a new individual record with expanded number of variables. The extension of individual statistical records' dimensions of variables through linkage is termed here as the "micro-based integration".

By means of micro-based integration records are integrated not in aggregated manner as is the case of macro-based integration but individually. The micro-based integration can effectively substitute the full-scaled survey with no remarkable expense otherwise required. The advantage of the micro-based integration over the macro-based one lays in the fact that it provides the individual statistical records with directly expanded dimension of variables. The datasets based on the expanded individual statistical records are fairly more informative in terms of applicability than the macro-based linked datasets.

(1) Horizontal integration

Irrespective of macro- and micro-based integration, statistical data from different sources are linkable as far as they share relevant variables that can work as matching keys. In case when the time differences between the sources are ignorable, the data can expand their dimensions

cross-sectionally by integrating aggregate data or individual records from different sources. The cross-sectional expansion of information can be termed here as the "horizontal integration".

(i) Cross-sectional integration of individual records

When individual statistical records from different sources have common ID information on the surveyed units or its substitutes such as derived numbers, names and addresses, one can compile integrated records through one-to-one matching. Such integration of data contributes to expand the information by multiplying the number of variables not only quantitatively but also qualitatively.

First, the integration can expand the information of original datasets quantitatively. The increased number of available variables is able to provide users with a wider scope of variables for cross-tabulation and regression analyses which respective sets of variables from individual records are unable to afford. Thus, the integrated datasets can provide users more informative statistical materials to analyze the reality.

Furthermore, the integration benefits also the quality of analyses. It contributes to enhance the quality of the statistical cognition by producing less biased results. When surveys did not cover whole set of variables that may affect the events, the obtained analytical results entail the possible biases. For example, when a pair of variables has other variables which commonly exerts influence on both of them, the pseudo correlation occurs between the former two variables. If independent variables do not cover whole scope of influencing variables that affect the performance of dependent variables in the regression analyses, the residuals hold systematic biases influenced by the unknown variables. In such case, the estimated parameters are of more or less some biased ones. The expanded cross-over usability of variables enabled by the individual-based integration may contribute to improve the quality of the analysis through liquidating otherwise unavoidable biases.

(ii) Individual-based integration of static and dynamic records

Individual statistical records can also be extended heterogeneously. By using uniform enterprise codes, different types of statistical records i.e. static and dynamic records can be linked together. Take foreign trade statistics for example, if enterprise record are integrated with trade data, the newly created records can document the possible effects of R&D investment to trade. By using these new types of datasets, one can assess, for example, the trade ramification effect induced by the investment promotion policy such as subsidies and remission of taxes.

(iii) Cross-sectional integration of hierarchical datasets

Household datasets usually carry hierarchical attributes. Besides information on households, they also have information on respective family members together with linking key variables. Expansion of dimensions of variables through the micro-based horizontal integration is possible not only for records of identical surveyed units but also among relational units such as family members.

There are two types of integration that fall in this category. First, the expanded dimension of variables through integrating individual statistical records over generations can document the relationships of family members between generations. For example, the possible impact of parents' attributes such as educational attainment and occupation on their children's behavior such as involvement in the labor market can be identified with this type of datasets.

Second, the integrated individual statistical records among family members can bring under light their behavioral relations which are unable to be identified with nonintegrated datasets. By applying integrated datasets of the time budget survey, one can elucidate the manner how family members perform concerted actions such as dining, spending free time together.

(ⅳ) Variable-based fusion of individual statistical records

Among multiple-source records there are some with comparable variables. Even in cases when exact matching among them was not warranted, by employing statistical matching procedures one can achieve extended individual statistical records. Let us term such expansion of statistical information as "data fusion" which composes a segment of data integration. One can regard the data fusion, defined in this way, as one of the effective measures to amplify the potentials of existing data which accounts for the peripheral parts of those generated by the data integration. Since it is quite rare that the identical units are surveyed in small sample surveys, data fusion seems to hold validity as an effective proximate measure to expand the dimension of variables. Moreover, data fusion is also privileged as the confidentiality friendly measure of data expansion.

Closer-distanced individual records in terms of attributes and other variables from different sources can be fused each other to generate the extended records. It is worth noting that they are pseudo in nature, because it is not always identical units' records that are linked together to generate the dataset. Despite the pseudo nature of fused records, they can be applicable to yield some approximate results unless the data integration measures are applicable.

(ⅴ) Location-based fusion of individual statistical records

A single survey result provides a snapshot of the surveyed units at a particular date drawn with a single cross-sectional dataset. GPS coordinates (x, y) give a definite positioning for the surveyed record, while individual statistical record data in traditional format have shared the same geo-codes, such as tract codes, municipality codes and others. In case when the location information represents small areas such as census tracts, whole surveyed units that fall within a respective area should carry an identical code number in terms of location. Individual statistical records loaded with GPS coordinates are generally distinguished from traditional area-coded ones by their disaggregate form of location code.

Worth noting here is that the GPSed cross-sectional records also have an advantage in expanding their information potential by means of micro-based data integration. Among individual statistical records from different sources such as censuses, sets of heterogeneous surveys and administrative records, there may exist some which carry identical coordinate information.

However, such cross-sectional record linkages are "pseudo," because it is not necessarily the relevant business units that were combined with each other as unified records in extended dimensions. Under the budget and human resource constraints, the latest developments in statistical practices of the world have shed light on data integration as one of the possible cultivation of information potential. Records with a multiplied number of variables generated by the coordinate-based cross-sectional data integration among heterogeneous business records may allow intensive analyses that a single set of records could never achieve.

Unlike area-coded records, which share an identical polygon code number among surveyed units, each household record in GPSed datasets, for example, has a unique location code relative to the coordinate information of the dwelling unit. Although the multiple-floor apartment houses may possibly be codified by one and the same pair of coordinates, coordinates may still retain their validity as location indicator. Because GPSed records are able to cope with any buffering zones, irrespective of the one-to-one or one-to-n correspondence to the coordinates.

As the recent developments in GIS tell, the 3D GIS is already put in practice partially. GPS coordinates applied for statistical purposes are also expected to expand their dimension vertically by introducing an additional variable that denotes floor information.

Expanding the informational potentials of existing sample survey data by fusing records horizontally including location-based data fusion is left for further cultivation. Despite the pseudo manner of data linkage, the expansion of dimension of variables achieved through data fusion is expected to bring about new findings that existing stand-alone datasets could never provide.

(2) Vertical integration

(ⅰ) Panel datasets

The surveyed units such as individual persons, households, enterprises and establishments are existent in time and space. In other words, they inherently carry cross-sectional attributes under the time and spatial constraints. They come into being, change statuses from time to time experiencing various events in the course of life and finally cease to exist. Despite such dynamism of units' existence, traditional statistics has portrayed them simply with a series of cross-sectional statistical snapshots.

The same series of aggregate data or the respective individual records obtained from a series of surveys can enrich their information through integrating them over the time dimension. Expanding dimensions of individual statistical records by compiling the aggregate time series datasets and the panel datasets, in which individual records are linked longitudinally, is termed in this paper as the "vertical" expansion of the existing data.

A series of surveys conducted repeatedly will give the repeated statistical snapshots. These snapshots usually comprise repeated cross-sectional datasets. Leaving aside censuses, it is quite rare that the same surveyed units are chosen as samples in sampling surveys. Repeated cross-sectional datasets, therefore, do not portray snapshot observations of the same set of surveyed units. When the same units are surveyed repeatedly, one can compile panel datasets that are given by the $N \times T$ matrix, where N denotes the number of the surveyed units and T the periods. However, the number of the surveyed units in each snapshot is not always the same in the panel dataset because of the attrition of samples. Including the unbalanced datasets with an unequal number of surveyed units in each snapshot, we simply refer to them as panel datasets.

The term panel is defined here in broader sense that also involves pseudo panels. For example, the time series matrix with aggregate statistics as a set of variables for respective groups and a set of time series individual records do not necessarily support the longitudinal attribute of the surveyed

units. The time series aggregate datasets by region, with which economic panel analyses are practiced, for example, in the field of local authorities' public finance, belong to the pseudo panels.

（ii） Longitudinal integration

In case when the same surveyed units are surveyed periodically like in censuses, the individual records can be integrated longitudinally. Among current survey practices carried out by national statistical authorities, there are some panel surveys which are deliberately designed to obtain responses from the same surveyed units repeatedly. When the same units are surveyed repeatedly in a series of surveys, one can compile panel datasets that form a matrix of N surveyed units and T periods for each surveyed variable.

The panel datasets are applicable to wider scope of analyses that neither the repeated cross-sectional nor time series datasets can afford to. Firstly, since reference units' individual statistical records are linked longitudinally, one can compile cross-tables according to the subpopulations which have dynamic nature, such as those who have changed status from employees to the unemployed or *vise versa* during a certain period of time. These datasets are effective to bring under light differences among subpopulations in terms of the changing statuses.

Secondly, the vertically expanded dimensions of variables enable datasets to apply for the panel analyses. The panel datasets enjoy comparative advantage over other traditional types of datasets such as cross-sectional, repeated cross-sectional and time series datasets in providing less biased analytical results and thus contribute to enhance the quality of statistical cognition. The difference in difference (DID) method, for example, can provide results free from possible intrusions of individual effect which other analytical methods such as the before and after analysis is unable to achieve.

（iii） GPS-based longitudinal expansion

One of the characteristic features of the repeated cross-sectional GPSed datasets is the possibility of longitudinal expansion of data dimensions. As for the nature of the surveyed units, we will focus our discussion in the following paragraphs on GPSed records of the surveyed units with a rather stable nature in terms of their geographical locations. Thus, locations, i.e. the inhabited dwelling units and sites where establishments or enterprises engage in their economic activities, are currently our major concerns in discussing GPSed records. Individual statistical records loaded with GPS coordinates involve in themselves a potential moment to separate the dual nature that is latent in the surveyed records. This separation will turn out to be pronounced in the repeated snapshot datasets such as repeated cross-sectional and longitudinal datasets.

(a) Business datsets

When one focuses concern to the location information of the surveyed units given by the GPS coordinates of the sites where establishments or companies currently operate business activities, a new type of dataset, i.e. a pseudo panel dataset of establishments or companies will be generated by fusing the records which share location information by applying the coordinates as linking key variables. The panel dataset compiled in this way is pseudo in nature, because establishments or companies that perform their business activities at the respective sites are not necessarily the

identical units. Businesses being performed at a particular site may alter by the exits of units followed by substitute entries of others during the reference period. However, as an overwhelming majority of business units is expected to carry out their activities at the same sites where they have hitherto operated, we can regard the compiled datasets as a panel in the broader sense. Thus, panel-based analyses would be applicable to these types of business datasets.

(b) Household datasets

Repeated cross-sectional GPSed household datasets give a chain of snapshots focused on the behaviors and activities practiced by households over time. One can bring to light households' various dynamic aspects by each region using GPS loaded datasets.

When one regards the repeated cross-sectional GPSed datasets from the GPS coordinates viewpoint, individual household records can be reorganized as pseudo panel datasets. Similar to the business datasets, those compiled from the repeated cross-sectional GPSed household datasets are still pseudo in terms of longitudinal attributes of the unit, because coordinates are tagged not directly to the respective households, but to the location of the dwelling units. Even in cases when household records maintain the unchanged coordinates in the repeated cross-sectional datasets, they do not always support the continuous settlement by the same family. There may happen an alternation of families in dwelling units under question caused by the moving out of a family followed by another family's moving in.

It is well expected, however, that in the majority of cases, families continue to reside at the same dwelling units. Unless panel datasets in the true sense are available for households, the pseudo panel datasets compiled by means of record linkage using GPS coordinates as matching keys would be applicable as one of the feasible options of a secondary approach to the family's demographic event analyses.

5. Statistical implication of data integration

Unlike digital media records such as sounds and pictures, individual statistical records are generally characterized by tenth and hundreds of variables carried by key variables representing the surveyed units. Multi-dimensional variables which compose data body, however, do not necessarily cover exhaustively the whole factors that govern the existence of the surveyed units. Several constraints, which govern the actual survey designing process, are responsible for it.

The surveyed records only document the results observed in the survey process. As the conventional survey designing process evidences, in addition to the issues adopted as survey items many other factors are also involved in molding the total attributes of the surveyed units. Among such items there are some which were finally given up to take because of the budgetary or survey burden reasons, although the survey designers have realized their importance. Furthermore, it is also probable that, due to the inadequately designed surveys, survey designers miss some of the statistically observable factors to include in the lineup of the survey items. In addition to these items, there still exist some which are unable to observe statistically, although they seem to have

significant effect upon the existence of the surveyed units.

The diagram 2 illustrates the structure of the categories of variables that account for the statistical entity of the surveyed units.

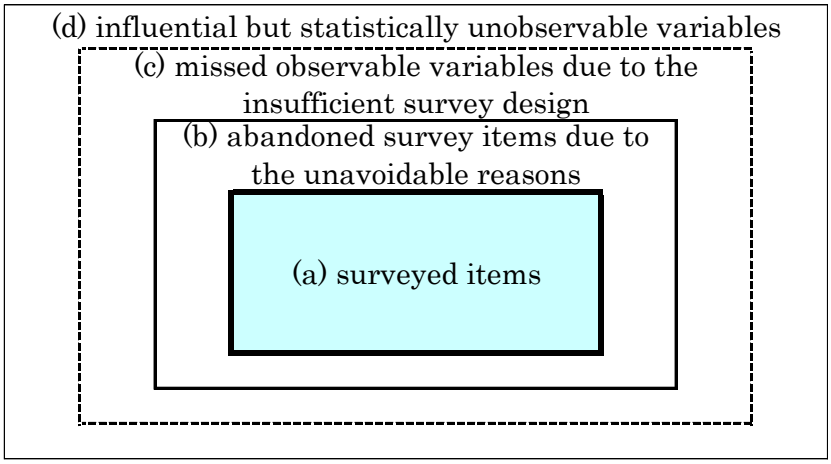

Diagram 2 surveyed variables and other variables that account for the surveyed units

The ultimate entity that statistics ought to document regarding the surveyed units is the composite of these whole categories of variables illustrated in the diagram. An individual statistical record obtained through the survey process only portrays the limited aspects of the object regarding the survey items and squeezes into residuals the affects of whole rest of variables including possible causal affects given rise to by a set of unidentified structural variables.

Japan's government statistical system is known as a typically decentralized one. Each ministry plans, designs and carries out surveys on their own account chiefly for administrative purposes with disposable budgets and resources. The proposed survey plans are examined by statistical coordinating body from the standpoint to relieve survey items, to avoid overlaps among surveys, and thus to avoid excessive reporting burdens. Some of the proposed survey items are often left out due to the excessive survey burden levied to the respondents. Possible duplications among surveys are also to be avoided. Even in cases when some surveying items have crucial importance to portray statistically the surveyed units, they are not accepted for due reasons.

Because each ministry operates survey practices almost independently under the decentralized statistical system, surveys have been conducted in Japan under poorly organized conditions among surveys. The fact that respective surveys are carried out basically in "stand alone" manner negatively affects the quality of statistical portrayals of the surveyed units. Thus, it sometimes occurs that a set of interrelated variables are surveyed in different surveys on the same surveyed units, although each of them affects the surveyed units as structural factors in concerted manner. A single survey results subsequently lead to yield the biased documentation of the surveyed units.

Suppose different source of data were integrated together, a combined set of variables will enable to liquidate possible biases caused by incapability of applying variables that belong to the category (b) illustrated in the diagram. With the exception of systematic biases caused by unobservable variables and white noise, those caused by variables that fall in categories (b) and (c)

should be liquidated as much as possible. It is worth noting here that biases generated from variables which belong to the category (b) are rather of institutional nature rooted in the existing statistical system.

Concluding remarks

Exploring the uncultivated informational potentials in government statistics motivated this paper. As arguments in this paper have evidenced, the archived individual statistical records have rich potentials to create new information by integrating or fusing existing records. Horizontal as well as vertical integration of individual records are expected not only to expand the scope of available variables but also to contribute to achieve less biased results. These facts suggest that statistics still has vast uncultivated frontiers to reclaim.

Discussion here does not confine its scope to information obtained through statistical surveys but it further can be extended to those captured in the process of administrative practices. Launching the relevant institutional structure for the systematic archiving of multi-sourced individual statistical records provides its informational basis. In order to transform information potentials into being by integrating records, which the archived individual statistical records carry in latent manner, a series of institutional setups, such as introduction of uniform enterprise cording system and standardization of geospatial information, seem to be prerequisite for its extensive functioning.

Switching the statistical system from traditional system of "stand-alone" statistical surveys to the integrated system of multi-sourced individual records paves the way to the full-scaled cultivation of existing information potential. The micro-based archiving system of statistical records that make possible the extensive cultivation of existing data by integrating records in varied way is expected to be the due infrastructure of official statistics for the coming era. The basic idea of the system of social and demographic statistics (SSDS) that was brought forward by U.N. Statistical Commission in the 1970s seems not to have been realized up until today. The micro-based system of statistical records should be spelled anew as the system of social, demographic and economic statistics (SSDES).

Acknowledgement

This research has benefited from financial support through the Grant-in-Aid: "International Comparative Studies on Archiving System of Official Statistical Data" (#22330070) of Japan Society for the Promotion of Science.