GPS Coordinates and the Possibility of Micro-based Integration of Statistical Records*

Hiromi MORI**

Summary

This paper discusses, first, the advantages of pinpoint positioning of the surveyed units over area-based location given by area codes such as tract codes and, second, the possibility of integrating records from different sources by using GPS coordinates as the linking key variable. Micro-based integration not only enables the cross-sectional (horizontal) expansion of dimensions but also occasions to create panel datasets in terms of location, including pseudo-panels, even in cases where relevant ID information is not available.

The discussion evidenced that a new type of location information given by GPS technology will open up the possibility to cultivate an untouched frontier in statistics.

1. Background

In contrast to the growing needs for diverse and promised quality data, the production of statistics has been facing increasing difficulties in recent years. The growing sample selection biases due to the decreasing response rate and the retrenchment of human and budgetary resources allocated to statistics are among them. Obtaining new statistical information by conducting new surveys becomes more and more unrealistic now. Under such circumstances, Government statistical bodies in the world are inclined to rely on the more extensive exploitation of existing information, including information obtained through administrative measures.

The Japanese Statistical Act put in force in April, 2009, stipulates that the obtained information should be regarded as a kind of asset with durable attributes. The information should be archived not only to provide data for historical analyses but also to serve as a comprehensive database which enables the creation of new statistical data without conducting another survey. The integration of records from different sources is becoming of outstanding importance in this context for contemporary and

---

* 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

future statistical practices.

The aims of this paper are twofold: first, to review integration of data from the standpoint of regional information, and second, to make an overture toward a possible integration of statistical records with GPS coordinates.

2. Statistical records and the use of statistics

The major form of disseminating survey results has been statistical tables compiled from individual record information captured through surveys. When one appreciates the informational aspect of the tables in relation to the variables that constitute the record, tables can be understood as none other than numerical pictures drawn with a set of variables integrated in a record. Multi-dimensional combination of variables in a dataset of single-sourced records provides a list of survey results. Since these variables are inherently integrated in the respective records, they are also applicable to micro-based analyses, such as multiple regressions.

The activities and behaviors of individuals are generated by numerous endogenous as well as exogenous factors. Individual persons and companies behave or perform their activities in some occasions on one's own accord and being subject to the influence of circumstances in other. Not only coincident events but also various inputs from the past, as well as expectation to the future, govern the activities of the individuals.

If one regards individual records from the time horizontal perspective, the information elements of epoch, aging, and generation effect are immanent in all data that correspond to the relevant variables collected by questionnaire-based surveys, while in the cross-sectional perspective individuals exist, displaying more or less discrepancies among them. These information elements latent in a single record can be brought to light once they are compiled as cross-sectional, repeated cross-sectional, or as longitudinal datasets. The epoch effect is controlled in cross-sectional datasets, while time-series and cohort data control aging and generation effects, respectively. Due to the existing discrepancies among individuals, however, cross-sectional tables are also affected by them, which leads to the over- or under-evaluation of the surveyed results or estimates.

Due to the substantial constraints in terms of surveyed items which carry statistical surveys, it is impossible to portray with a single statistical survey the complete pictures of individuals of multifarious entities and massive phenomena yielded as the result of their activities. Many factors govern the number and kind of variables in designing the survey. Among the planned surveying items there exist some which were exempt from the list for the sake of budgetary or other reasons. The possible overloading to respondents because of the excessive number of surveying issues also accounts for the exclusion. Consequently, numbers of surveying items are limited to reasonable scope. Among the surveying items finally missed in the

questionnaire, there are some which may affect the adopted variables. The two relating factors are occasionally listed as surveying items in different surveys for the identical surveyed units. The presence of variables which surveys are unable to observe should also be noted here, although they significantly affect the performance of the individuals.

3. Patterns of integrating data

(1) Macro- and micro-based integration

(a) Macro-based integration

Cross tables from different sources with the same variables can work as a sort of trigger for data integration. Suppose the two sets of multivariate cross tables by sex and age class, for example, carry m and n units in corresponding cells. They are expected to generate a couple approximate subgroups of the population. Since it seems likely that attributes or activities of these subgroups are comparable, a different set of variables inherent in each subgroup from different sources can be integrated.

However, in this case, the yielded records with an expanded dimension of variables give only "pseudo" integration, because they generally do not represent the identical group of the population. The greater the distance between data sources, the more fictitious becomes the integration. An array of aggregate data from different sources may constitute a virtual dataset in which respective variables are integrated in aggregate manner. One can establish this type of integration as a "macro-based integration." Annual regional statistics that carry contemporaneous indicators from varied sources by prefecture and municipality are an example of macro-based integration.

It should be noted here that macro-based integration is distinct from micro-based integration by the peculiar manner of the relationship between variables. Indicators are not related to the respective units that compose the group, but rather to the relevant group as a whole. Indicators are integrated as the aggregated variables so far as they characterize attributes or performances of the seemingly identical population. In other words, the aggregated variables from different sources are linkable through the seemingly identical population.

(b) Micro-based integration

A set of information obtained through surveys usually forms an individual record. In cases where the records share identification numbers assigned to respective surveyed units, records from other sources are linkable. Variables such as names, addresses, dates of birth, and telephone numbers also assist in statistical matching. The matching of individual records from different sources can yield a new individual record of the multiplied number of variables. The extension of individual records' dimensions through linkage here is termed as "micro-based integration."

Through micro-based integration, the records are integrated not in an aggregated manner, as is the case with macro-based integration, but individually. The advantage of micro-based integration over macro-based integration lies in the fact that it yields the expanded individual records, which are more informative in terms of application than the aggregated ones.

(2) Horizontal and vertical integration

Irrespective of macro- and micro-based integration, statistical data from different sources are linkable as far as they share relevant variables that can function as a matching key. In cases where the time differences between the sources are ignorable, the dimensions of data can be expanded cross-sectionally by integrating aggregate data or individual records from varied sources. The cross-sectional expansion here can be termed as the "horizontal integration."

Expansion of the dimensions of records through micro-based horizontal integration is possible not only for records of identical surveyed units but also among relational units, such as members of the same family. The latter type of integrated records may document effects that are likely to work over generations, for example, from parents to children and simultaneous actions among family members that are unidentifiable with individual records.

The same series of aggregate data, or the respective individual records obtained from a series of surveys, can have their information enriched by integrating them over the time dimension. Compiling the aggregate time series datasets and the panel datasets of longitudinally-linked individual records produces the "vertical" expansion of the existing data. The term "panel" is defined here in the broader sense that also involves pseudo-panels; for example, the time series matrix with aggregated statistics as a set of variables for respective groups and a set of time series individual records which do not necessarily support the longitudinal attributes of the surveyed units.

4. Area codes and the macro-based integration of multi-source data

In modern census, enumerating activities have been conducted at each census tract, which exclusively covers respective municipalities such as prefectures, cities, towns, and wards in the large cities, and thus the whole scope of national territory. Tracts also serve as sampling frame for most surveys.

Since census and surveys are usually conducted as questionnaire-based surveys, each responder or the surveyed unit are captured as a component of the group of units that are present in the tract. The tract-based survey results provide relevant data in compiling regional statistics, because tracts are organized systematically in conformity with administrative bordering.

Regional areas given by municipalities such as prefectures and cities have also served as a key variable to integrate data from different sources. Let us take *The*

Social Indicators by Prefecture, for example. It carries hundreds of statistical indicators by region obtained from various data sources, including administrative records. One can regard a chain of indicators as a set of aggregated variables overlaid on respective regional codes. Indicators such as regional aggregate data, regional averages, and ratios are characterized as a kind of integrated record. The creation of records in such a way is termed in this paper as "macro-based integration". The respective indicators are usually treated as variables that constitute individual records in "pseudo" sense that provide data for regional regression analyses.

As far as region-based results of Japanese population census are concerned, census tracts were the basic regional units to compile them up until the 1985 census. The Basic Unit Block (BUB) was introduced in Japan in 1990 as the minimal survey tract area which consists of approximately 25 households and, in principle, corresponds to the town block. Thereafter, small area statistics such as subdivision of municipalities by *cho* / *aza* are compiled based on the BUBs.

While Japan had more than 12,000 cities, towns, and villages in the 1950s, the number had diminished drastically to about 2,200 by the year 2005. The annexation and reorganization of municipalities are real threats to statistical comparability, since they require enormous amounts of clerical work to adjust historical statistics to the newly-annexed or partitioned boundaries. The rezoning of boundaries renders time series regional data less consistent.

Census tracts are not totally exempt from boundary rezoning. The completion of new roads and railways and the development of new residential areas make existing tract maps obsolete. Some tracts have been partitioned and then annexed to several neighboring tracts, while several others have been totally reorganized. Such tract rezoning also disturbs the comparability of small area time series data.

Grid Square Statistics were first introduced in Japan based on the 1970 census results to provide more robust regional units, and thus to compile comparable data in time perspective. Since the geodetic line partitions areas mechanically into a set of uniform grids, the resulting grids can be independent of any municipality rezoning and of tract reorganization. Under this system, the whole national territory is divided into rectangles of about one square kilometer and 500 square meters by longitudinal and latitudinal lines. These grids are called "basic grid squares" and "half grid squares," respectively.

For tracts that are totally included in a particular grid, the whole of their elements are properly allocated to that grid. In the case where the grid borders cross the tracts, however, tract elements, i.e. the surveyed unit records, should be processed in such a way as to cope with the problems of how to allocate them among grids in an appropriate manner. In all remaining cases, surveyed units are allocated more or less by approximation. Although the newly-introduced BUBs still require a certain amount of clerical work to compile grid statistics, by affording more detailed regional

information they could serve in improving the quality of estimates.

This paper is motivated from the idea that geographical codes can operate as linking key variables to integrate data effectively from different sources. When one reviews the above discussions from this perspective, worth examining is the manner in which they integrate the relevant variables as a consistent set of information that constitutes one individual record.

Irrespective of the hierarchical level of zoned areas including grid squares, categories of areas are common so far as it is the aggregated sums or averages / ratios derived thereof that correspond to each area code. The aggregated data share the identical area codes. Put differently, the set of area codes discussed above can work as a platform to overlay macro-based variables from varied sources.

Table 1 illustrates the integration pattern by levels of areas.

Table 1 Patterns of macro-based integration

geographical areas		area codes	attributes of relational key variables	pattern of data integration
administrative districts	prefectures	prefecture code	pixel of raster type	macro-based integration
	cities,wards,towns,villages	city etc.code		
	cho /aza			
census tracts		tract code		
basic unit block		BUB code		
grid square		grid code		

The annual reports of *The Social Indicators by Prefecture* carry a set of annual indicators, i.e. aggregated sums or derived averages / ratios, by prefecture from various sources which can be termed as “macro-based horizontal integration.” These sets of indicators are also reorganized to form a sequence of time series data by region which can be called “macro-based vertical integration.” These datasets provide data for macro-based analyses.

As I have discussed (Mori 2010), these datasets have various constraints due mainly to the insufficient obtaining of location information inherent in the units which compose the elements of each region. Another setup is required before there can be a breakthrough in the utility of such datasets.

5. Location positioning and the possibility of micro-based integration

(1) Dual nature of the surveyed records

The surveyed units such as persons, households, establishments, and enterprises usually exist in time and space. A set of information regarding their attributes, activities, and their results can be captured through questionnaires and

administrative processes and arrayed as a record format. The discussion here is to highlight the dual nature of the surveyed records.

It is obvious that the obtained data, i.e. the various attributes, activities, and results, are ascribed to each surveyed unit. That is, individual records have been regarded as statistical copies of the surveyed unit. Another aspect of the data is less obvious compared with the first one. The surveyed information belongs to or relates to the units that are located at a particular geographical point, i.e. a dwelling unit or site where business activities are carried out. Put differently, the set of informational data offered by surveyed units is related to some particular geographical point. One may term the former "unit information" and the latter "spot information."

Spot information obtained from observations in a single survey is less obvious than unit information, because spot information refers not to the unit itself but to its locational existence. Repeated observations, however, may more clearly address the dual nature of the records. When the same unit has been repeatedly observed in a series of surveys or census, the obtained records may reflect longitudinal change in the relevant unit. When the same spot has been observed in repeated surveys, it will document the kind of activities of one fixed point at different moments.

As these two aspects which the surveyed records inherently possess in a latent manner are substantially dynamic in nature, they may split off in cases where the locations of units change over a period of time. Although the majority of the surveyed units remain at the same spots, the replacement of units may possibly take place in surveys conducted at certain intervals. Different units may be observed in ensuing surveys at the same spot due to the replacement of units, i.e. by a former unit moving out followed by a substitute moving in. The observed spots in the previous survey can disappear, whether or not the dwelling units are existent, in cases where no succeeding tenants accommodate that dwelling unit. It may also be possible that new entrants are surveyed at new spots. Families can be occupants either of newly-constructed or unsettled dwelling units, while establishments and companies can launch their business activities either at newly-developed industrial sites or ones that were unoccupied when the previous survey was conducted.

Statistics has long been regarded as a science dealing primarily with massive phenomena. In traditional statistics, therefore, surveyed units used to be regarded simply as elements that mold a population or subpopulation. It was only in the latter half of the 20th century that statisticians began to shed light on individual survey records.

Due to these traditional statistical ideas, together with several technological constraints, those working in the field of statistics remained tolerant of the insufficient use of the location information inherent in survey records. Although surveyed units such as households, establishments, and enterprises mostly have definite location information regarding their existence, survey records documented

them not at their particular points, but only as one of the component units of the tract. Instead of specified location codes inherent to respective surveyed units, a tract code number was given to all surveyed units that belonged to a particular tract. Each unit's location information was collected not as a geographical point, but as a small area. Because insufficient location information was obtained, practitioners statistical science had to put up with "diluted" information in terms of the location of units that resulted in a number of constraints on its use. Figure 1 illustrates examples of traditional household and establishment/enterprise record layout forms.

(2) Obtaining GPS coordinates

Developments in information technologies have opened up a new scope in obtaining location information from each surveyed unit. Similar to the Internet, GPS was originally invented and has been utilized primarily for military purposes. Thanks to improvements in the accuracy of digital map software, together with widespread use of information terminals furnished with various GIS software, GPS now enjoys a wider acceptance in daily life as a form of necessary informational infrastructure. Official statistics, however, are rather behind compared with other fields in applying GPS for their practices.

In the U.S., approximately 143,000 field workers engaged in the so-called "address canvassing" operation over four months beginning from April, 2009. Canvassers verified the nation's residential addresses and captured GPS coordinate information for each of these addresses using a personal digital assistant (PDA) equipped with ArcPad software. GPS coordinates collected in the address canvassing operation were used to pinpoint on the mobile map carried by field workers the residences of non-responders in the 2010 Population Census. The newly-adopted latest device is expected to improve the response rate and thus the quality of the result. Statistics Poland is also planning to collect GPS coordinates in the 2011 Census.

The Japanese Statistics Bureau obtained GPS coordinates of establishments and enterprises through matching addresses from the Establishment and Enterprise Census data with those in an on-the-shelf digital map database provided by a private company. The GPSed individual records are used to compile the grid statistics for the establishments.

The French Statistics Bureau (Institute National de la Statistique et des Études Économique: INSEE) maintains a housing unit register termed as "répertoire d'immeubles localizes" (RIL) which carries GPS coordinates as location information. The demographic department that is in charge of updating the RIL obtains the coordinates in the following way. By purchasing road centerline information from the national geographical authority (Institute Géographique National: IGN), the department calculates coordinates that correspond to each address. Since some residential buildings occasionally share the same address, it may happen that more

than one hundred residential units carry the same GPS coordinates in the RIL. In the RIL, therefore, it is not a residential unit but an address that corresponds to the coordinate information.

Directly obtained GPS coordinates through mobile terminals and indirect access to them either by means of address-GPS converting software or by applying appropriate calculation methodologies can serve as a powerful driving force for statisticians to explore the wider dimensions of the applicability of coordinates, not only for the use of data but also for the production of data of improved quality.

(3) Advantages of the GPS coordinates over other regional codes

The GPS coordinates (x,y) are intrinsically a pair of infinite decimals that illustrate an intercept that is changeable depending on the number of figures. The coordinates calculated down to the 6th decimal place correspond to a micro area of one square meter. Thus, they do not necessarily provide any pinpoint information. Moreover, multiple-floor apartment houses or business buildings may possibly be codified by one and the same pair of coordinates. As French practices in the RIL show, it is probable that tens and hundreds of residential units occasionally share the identical coordinate information. Although in either case the GPS coordinates do not support one-to-one correspondence with respective residential units, shops, or offices.

The development of 3D GIS technology is now under way. The introduction of an additional variable may work as far as statistical identification of the surveyed units. Even in cases where a one-to-n correspondence between the coordinates and the units governs, coordinates may still retain their validity as a location indicator, because they provide a fairly good approximation in terms of the location of the units in question.

Unlike tract-coded records, GPSed records provide definite location information of surveyed units. As stated above, ambiguity in the use of data has sprung substantially from area-based locating. GPS coordinates are more appropriate variables than tract codes in terms of identifying the geographical points of surveyed units' existence. Once GPS coordinates are tacked to individual records by some measure or other, it becomes possible to allocate surveyed units not by estimation but by direct assorting of surveyed units according to the coordinate information. Units such as families, establishments, and enterprises will have been surveyed intrinsically at the very point of their presence. It was not until the obtaining of coordinate information that statisticians became able to employ location information on an extensive scale.

GPS coordinates tacked to each record as one of the unit's basic attributes will enable the liquidation of the ambiguity described above. By doing so, all archived records will be able to withstand any form of re-zoning. GPSed time series records can enjoy longitudinal comparability in full scale. Furthermore, they are qualified to compile statistics that can meet any buffered zones.

Besides these, the GPSed records seem to have additional advantages with regard to the micro-based integration of records from different sources.

6. Cross-sectional records and GPS-based data integration

A single survey result provides a snapshot of the surveyed units at a particular date drawn with a single cross-sectional dataset. A pair of GPS coordinates (x, y) corresponds to each surveyed record, while surveyed units with a traditional record format share the same geo-codes, such as tract and other area codes. In the latter case, whole units that fall within a respective area should carry an identical location code number, such as a tract code. The GPSed records are distinguished from non-GPSed ones generally by a one-to-one correspondence of the surveyed record with its location code.

It is worth noting that the GPSed cross-sectional records also have an advantage in enlarging the information potential of the data by means of expanding dimensions through data integration. Among individual records from multiple sources such as census data, 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. The latest developments in statistics have shed light on data integration as one of the possible means of expanding 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 hope to achieve.

Unlike tract coded records, which share an identical polygon code number among surveyed units, each household record in GPSed datasets usually has a unique location code relative to the coordinate information of the dwelling unit. Although multiple-floor apartment houses may possibly be codified by one and the same pair of coordinates, coordinates may still retain their validity as location indicators. GPS coordinates are also expected to undergo an expansion of their dimensions, for example, by introducing an additional variable that denotes floor information.

Expanding the potential of existing data by data fusing records is also valid for household records. Despite the pseudo manner of data linkage, the compiled datasets with multiplied dimensions of variables will enable intensive analyses that may bring about new findings.

7. GPS-based pseudo and genuine panel datasets

A series of surveys conducted repeatedly will give repeated snapshots. These

snapshots usually comprise repeated cross-sectional datasets. Leaving aside census data, we can see that a series of survey results do not necessarily cover the same surveyed unit. Repeated cross-sectional datasets, therefore, do not portray snapshot observations of the same set of surveyed units, yet while 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. However, the number of surveyed units in each snapshot is not always the same in the panel dataset because of the attrition of the samples. Including the unbalanced datasets with an unequal number of surveyed units in each snapshot, in this paper we simply refer to such datasets as panel datasets.

As for the nature of the surveyed units, we will focus our discussion on the GPSed records of 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 / enterprises perform their economic activities, are currently our major concerns in discussing GPSed records. Individual records loaded with GPS coordinates present 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.

(1) Repeated cross-sectional data and pseudo-panel datasets

One of the characteristic features of the repeated cross-sectional GPSed datasets is the possibility of longitudinal expansion of data dimensions. When one focuses one's interest on the location information of the surveyed units given by the coordinates of sites where establishments or companies currently perform their activities, a new type of dataset, i.e. a pseudo-panel dataset of establishments or companies will be compiled by fusing records by means of coordinates. The dataset is "pseudo" in the sense that establishments or companies that perform their business activities at the respective sites are not necessarily identical units. Business being performed at a particular site may alter by the exits of units followed by substitute entries during the period of time in question. However, as it is expected that an overwhelming majority of business units will continue to carry out their activities at the same sites which they have occupied in the past, we regard the compiled datasets as a panel in the broader sense. Thus, panel-based analyses would be applicable to these types of business datasets.

Repeated cross-sectional GPSed household datasets give a chain of snapshots focused on the behaviors and activities practiced by the household over time. One can analyze various dynamic aspects of the household by each region using this type of dataset.

When one consider the repeated cross-sectional GPSed datasets with regard to the GPS coordinates, one can see that individual household records are reorganized into 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 related not directly to the respective households, but only to the dwelling units. Even in cases where household records maintain unchanged coordinates in repeated cross-sectional datasets, there may occur the replacement of households in dwelling units under analysis 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 will 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.

(2) Longitudinal data and genuine panel datasets

The GPSed panel datasets are far more informative than non GPSed ones. Longitudinal records armed with GPS coordinates are qualified to objectify the dual aspect of the questionnaire information. This means that, besides the unit information, location information which was already existent in the individual records in latent manner is brought to light through repeated surveys. When one focuses upon the surveyed units, unchanged coordinates indicate their survival, while the changed ones suggest the redeployment of the unit. If one switches the viewpoint to sites, records illustrate the activities of the units operated at the particular site specified by the coordinates. Put differently, this process will work to establish a kind of function or potential of the respective sites.

The GPSed panel business datasets can identify the following events. When one focuses on the business units in the dataset, their coordinates provide information on the units’ relocations over time. Since the unit is identified by the competent ID number, one can easily distinguish redeployment from quitting.

Business units go through a set of demographic events throughout the period of their activities. When one focuses on the coordinates, surveyed unit records being identifiable by unit code number may denote the demographic events of the business unit, such as survivals, entries, and exits which come about at a particular site. Thanks to the unit ID number, it is possible to distinguish new entries from the moving in of existing units due to redeployment and also exits from the moving out of units. It is expected that GPSed records can partly substitute for the profiling work of business units, which is actually quite labor-intensive clerical work, through automatic data processing.

Household panel datasets can be compiled through matching records by family ID number. If no ID number is available, householders’ names will substitute for the ID. Similar to the longitudinal business records, household records carry a dual implication. The record tells a story about the units themselves, i.e. families or

individuals who share the dwelling unit on the one hand, and provides information on the functioning of respective dwelling units in terms of habitation on the other.

If one changes one's concern to the units, i.e. households or individuals, a changed set of coordinates will trace the family or personal history of residential moves. This type of dataset is expected to provide relevant materials for analyzing the geographical residential moves of families or individuals in each stage of a family's or an individual's life cycle.

By controlling the site information, GPSed business panel datasets would be applicable to establish, for example, the business unit redeployment ratio by size and industry, and compare the ratios between single and multiple establishment businesses or grouped or single enterprises. Household panel datasets focused on dwelling units can draw another picture of the habitation behavior of residents. Household records reported from residents with unchanged coordinates may give either the same family ID number or the name of the householder, or different ones in a series of snapshots. By overlaying the family ID number or the name of the householder on respective coordinates, one can compile a dataset that helps to shed light on the occupancy status of dwelling units. Unchanged ID numbers suggest that the same families or individuals continue to reside at the same dwelling units, while changed ID numbers indicate replacement of families or individuals. Longitudinal records with vanished coordinates may indicate vacancy or a halt of operation as residential dwelling units, while newly-emerged coordinates suggest new engagements as residences. The datasets will be applicable to the identification of residential mobility, for example, by region and tenure.

7. Concluding remarks

Apart from other space-coded datasets that carry location codes such as municipality codes and tract code, the GPSed datasets can enjoy a wider scope of advantages. This paper focused the discussion on exploring the potentiality of the data inherent in the questionnaire by micro-based integration of records. A novel idea to regard the GPS coordinates as key variables to integrate the data is the cornerstone of the discussion.

GPS coordinates can work as an effective linking key variable in cases where traditional linking mechanisms, such as ID numbers, are unavailable. As discussions in this paper have evidenced, positional information captured through GPS terminals not only integrates individual records horizontally as well as vertically to compile panel datasets including those with a "pseudo" nature, but also yield datasets that, in combination with relevant ID information, enable analysis of the dynamism of the surveyed units.

Although the micro-based positioning of the records involves potential elements to

cultivate an untouched frontier of statistics, the positional variable given by the GPS coordinates has been unreasonably undervalued up until today in Japan. The GPSed datasets created at vast outlays are only used for quite limited purposes.

In the current phase of the development of world statistics, some countries have already steered helm to create the GPS-supported statistical infrastructure applicable not only to the production but also to the more intensive use of the statistics. Its successful completion may substantially affect the redesigning of the official statistics.

References

Mori, Hiromi (2010), "Constraints in Use of the Data Due to the Insufficient Obtaining of Location Information and a Breakthrough in Statistics", *Hosei Economic Review (keizai-shirin)*, Vol.78-4

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.