# Constraints in Use of the Data Due to the Insufficient Obtaining of Location Information and a Breakthrough in Statistics*

Hiromi MORI†

## Summary

Location information regarding the surveyed units, for example, of households, establishments and enterprises, has been collected in surveys as area information, such as tract codes. The traditional location codes do not support the one-to-one correspondences between the unit and its location information, despite the fact that each unit has inherent unique positional information in terms of its location.

This paper discusses the following issues. Firstly, it portrays a rough historical sketch on how location information of the surveyed units has been captured in surveys. Then it discusses several constraints in using survey results which the ambiguity of the data due to the insufficiently obtained positional information under questionnaire-based surveys gives rise to. Latest developments in information technology have opened up a new arena for the application of GPS also for statistical purposes. Rest of the paragraphs will highlight a breakthrough in regional analyses that will render GPSed records more practical.

Keywords: location, GPS, grid statistics, census tracts, questionnaires

## 1. Introduction

Due to the advanced information technology, together with the wide-spread use of reasonable price handheld PCs, the Geographic Positioning System (GPS), originally introduced as a military invention, brought about wide-ranged revolutionary changes not only in economic activities but also in a wide scope of social lives. Accurate positioning based on the latest GPS technology gave births to numerous new businesses and triggered another burst in the existing industries. Daily lives became more and more involved in this technology.

Statistics seems to be one of the latestcomers in terms of applying this modern technology. Several reasons may account for the fact. Firstly, traditional statistics were based on the notion that statistics were given not as individual records but

* An earlier version of this paper was published in March 2011 on *KEIZAI-SHIRIN* (*The Hosei University Economic Review*), Vol.78, No.4.

† Faculty of Economics, Hosei University
4342 Aihara, Machida-shi, Tokyo, JAPAN 194-0298

primarily as aggregate data. In modern censuses, enumerating activities have been conducted at census tracts, which are arranged to cover exclusively the whole scope of national territory. Census tracts were introduced with an intent of avoiding oversights in enumeration as well as multiple counting. Although some surveys inquire about respondents' addresses, in most cases, location information of the surveyed units is given either by regional or tract codes.

The tract code has been used as a minimal unit to address the location information of the surveyed units within the tract. In other words, the surveyed units at large in a particular tract have shared identical code numbers to represent their locations. Although the surveyed units, such as households, enterprises and establishments exist at the distinct location at the time of survey, the surveyed records only put the diluted information about units' location derived from the tract-based survey operations. For conventional table-based aggregate data users, the ambiguity of data, which is ascribed to the insufficiently obtained location information, caused no serious constraints in their use.

Secondly, there exists an unavoidable trade-off between the in-depth use of statistical data and the confidentiality. When one pursues the more extensive use of obtained statistical data, the more obvious becomes the risk of confidentiality disclosure. In this sense, the table-based aggregate data were an appropriate form of statistics which can stand with requirements of statistical confidentiality. While GPSed individual records are expected to explore new frontiers in employing the data, national statistical authorities have been rather hesitant because of the privacy issues. It was quite recently that statistical authorities of some countries started employing GPS coordinates as a machine to produce better quality data as well as for more intensive use of the existing data.

GPS coordinates are substantially distinct from addresses described in letters. Since they are given in digital form, they are easy to be processed. One can handle the data with a simpler vector algorism for various analytical purposes. Furthermore, they can be employed as an effective key variable to integrate individual records.

Last but not least, there were substantial constraints in terms of the precision of positioning the location with the GPS. Despite many advantages in data handling, the use of GPS coordinates for statistical purposes have been still sporadic up until today. Positioning the location with insufficient precision may partly account for the tardy introduction of GPS technology into statistical practices.

Thanks to the successful launching of the satellite, non-military use of GPS also became able to enjoy the sufficient precision of obtained coordinate information. Besides, a wide diffusion of reasonable price GPS terminals also paved the way to the systematic application of the latest technology for the statistical practices.

This paper is organized as follows. The first paragraph gives a brief historical outlook on how statistical surveys have obtained the location information. Various

constraints in use of statistical data due to the insufficient obtaining of the location information will be addressed in the ensuing paragraph. The third paragraph shows some examples how and for what purposes the GPS coordinates are obtained in some statistical authorities including Japan. The fourth paragraph will discuss a breakthrough in the use of statistical data with GPS-armed individual records. The final paragraph concludes.

## 2. How statistical surveys have obtained the location information

At the dawn of the history of modern statistical survey, the so-called "table-based surveys" (tabellarische Erhebung) were major way of collecting statistical information. Field workers directly filled respective cells in the table with aggregate number of surveyed items for the pertinent region. Although the surveying items such as attributes as well as activities of the surveyed units are inherent in individual person, household, establishment and company, they are simply counted as a group totals in this type of surveys. Table-based surveys are, among others, distinguished from their successors termed as "questionnaire-based surveys" in terms of survey technique by inseparable mergers of individuals into a group. In surveys that belong to the former category, therefore, respective regions where surveys were conducted represent location information of the surveyed units in question.

Introduction of the questionnaire-based surveys has opened up a new scope in the development of survey techniques in obtaining statistical information not in aggregate but in individual manner, which brought about outstanding progresses in producing a wide spectrum of statistical outputs.

Evolution of survey technique from the table-based to the questionnaire-based ones has accompanied another institutional adjustment for the successful operation of the surveys. The census tracts, which partition whole coverage area into mutually exclusive sub-areas, were introduced to avoid under- as well as over-counting and, hence, to guarantee the quality of the surveyed results. Since the introduction of the questionnaire-based surveys, census tracts have played, up until today, as the basic regional units where field workers operate the surveys.

As stated above, census tracts were originally introduced as an institutional machine for the successful operations of the censuses. They, however, have a particular implication with regard to obtaining the location information of the surveyed units. By conducting the questionnaire-based surveys, statistical authorities can collect individual data about surveying items from the surveyed units such as persons, households, establishments, enterprises and so on. Except those who have no fixed abode, immobile residential units accommodate dwellings for the overwhelming majority of families. Although some business units, such as the owner-driver taxis and mobile shops, carry on their actual business activities in mobile

manner, most of the establishments and enterprises perform their activities at certain business or industrial sites.

Motivation of this paper derives from an idea that the information obtained through the questionnaire-based surveys pertains to or reflects, for example, the attributes and activities of the surveyed units that are located at a particular geographical point, i.e. a dwelling unit or site where business activities are carried on. Put differently, a set of information offered by the surveyed units are related to some particular geographical point. This fact suggests that each respondent, i.e. household, establishment and enterprise in questionnaire-based surveys is connected inherently to the particular geographical point.

Census tracts, introduced as an institutional machine that enables field workers to avoid possible enumeration failure, however, have consequently offered the insufficient location information to the survey results. They did not give the one-to-one correspondence between the surveyed units and their actual location that is potentially allowed by the questionnaire-based surveys but simply the n-to-one correspondence. Because of the insufficient location information, statistical analysts had to put up with "diluted" information in terms of the location of the units. This insufficiency in positioning the units produces a number of constraints on the use of the results.

Figure 1 illustrates examples of traditional household and establishment/ enterprise record layout forms.

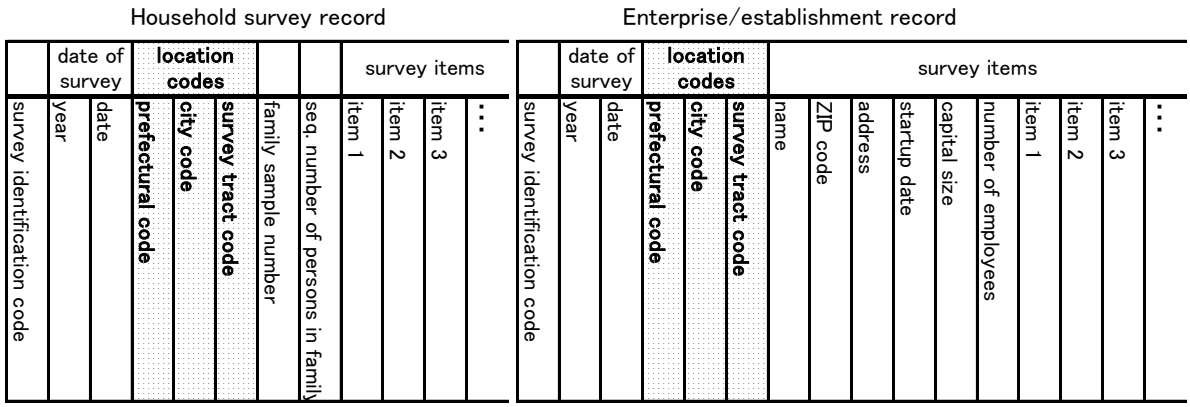



Figure 1. Examples of record layout forms

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

Due to these traditional statistical ideas, together with several technological constraints, statistics remained tolerant of the insufficient use of the location

information inherent in the surveyed records. Although the surveyed units such as households, establishments and enterprises mostly have definite location information regarding the field of their daily activities, survey records documented them not at their particular points, but merely as one of the component units within the tract. Instead of the definite positional codes inherent to respective surveyed units, a tract code number was given to all surveyed units that belonged to the particular tract. Each unit's location information was collected not as a geographical point, but as a parcel of area where the units actually locate.

As administrative districts, such as prefectures, cities, towns and villages, are systematically partitioned into tracts, the ambiguity, which derived from insufficiently collected location information, did never raise serious problems in tabulating the region-based results. The diluted nature of location information in the traditional questionnaire-based surveys, however, reveals a set of problems which the individual survey records had carried in latent manner when one opts to employ the data for different analytical purposes.

## 3. Constraints in use of the data due to the insufficient obtaining of the location information

(1) Problems caused by border rezoning

While Japan had more than 12,000 cities, towns and villages in the 1950s, the number had diminished drastically to 1,750 by the year 2010. The annexation and reorganization of municipalities are real threats to statistical comparability, since they require enormous amounts of clerical work to adjust historical data to the newly annexed or partitioned boundaries. The rezoning of boundaries renders time series regional data less consistent.

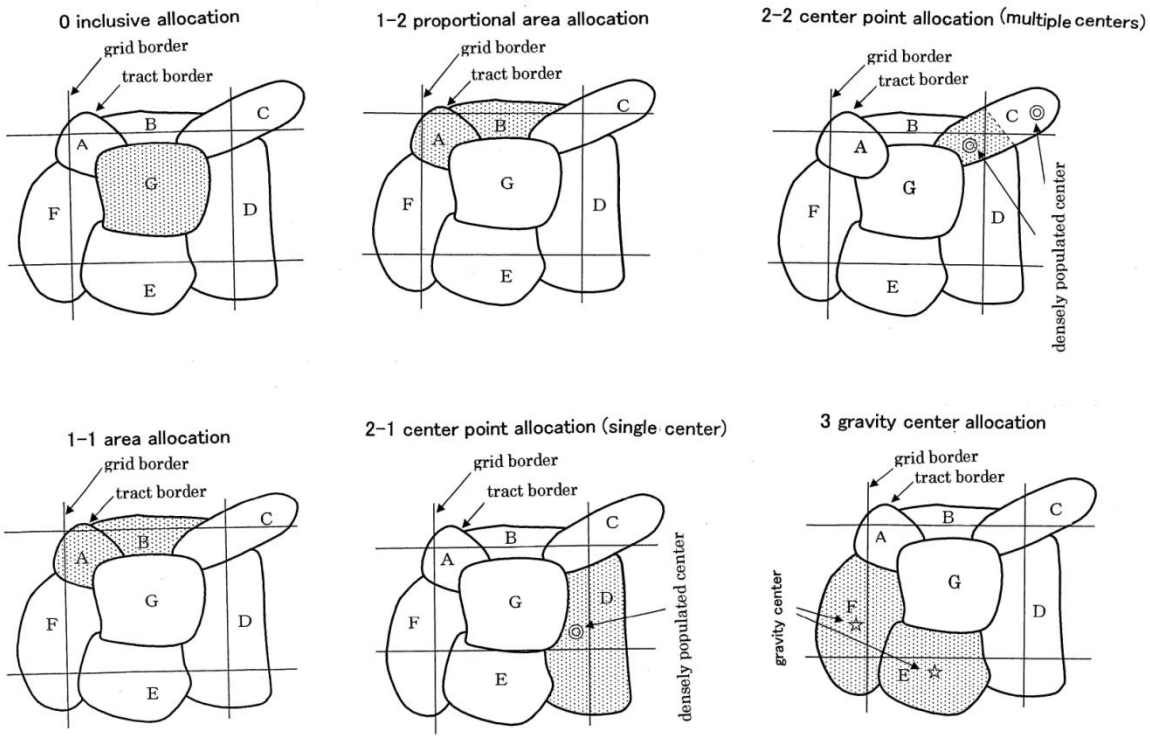
Census tracts are not totally immune from boundary rezoning. The completion of new roads and railways and the development of new residential areas make the existing tract maps obsolete. Some tracts have been partitioned and then annexed to several neighboring tracts, while several others have been totally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

The Basic Unit Block (BUB) was newly introduced in Japan in 1990 as the minimal survey tract area of more or less durable nature. Although the BUBs are expected to be more robust than the census tracts, they still are not totally free from restructuring.

(2) Allocation of the surveyed units in tracts

Grid Square Statistics were introduced in Japan based on the 1970 census results. 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 termed as "basic grid squares" and "half grid squares," respectively.

Since the geodetic lines partition areas mechanically into a set of uniform grids, they can be independent of any municipality rezoning and of tract reorganization. As case 0 in Figure 2 illustrates, 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, the surveyed units are allocated more or less by approximation (case 1-1) or by calculation (cases 1-2, 2-1, 2-2 and 3). In either case from 1-1 through 3, an ambiguity occurs in converting tract-based data into grid-based ones.




Source: http://www/stat.go.jp/data/mesh/pdf/gaiyo2.pdf (pp.24,26 and28)

Figure 2. Allocation of tract units among grid squares

(3) Inadaptability of data for buffering analysis

Buffering analysis is now widely used to identify statistical characteristics of the buffered polygon areas with a fixed distance surrounding the specified input features, which can be polygons, lines or points. Since buffer polygon borderlines do not necessarily coincide with those of tracts, borders usually intercross. Similar to the grid estimates, estimates for the buffered polygons, therefore, are usually subject to the ambiguity caused by inconsistent borders. Buffered circles and polygons usually have indented fringes due to the discordance of bordering.

All these difficulties, yielded in the process of allocating the surveyed units in tracts to the relevant areas, derive from insufficiently obtained location information in surveys.

## 4. 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 the remarkable improvements in the accuracy of digital map software, together with the widespread use of information terminals furnished with GIS software, GPS now enjoys a wider acceptance in daily lives as necessary geographical information infrastructure.

Official statistics, however, are relative latecomers in applying GPS for their practices. In the U.S. approximately 143,000 field workers engaged in the so-called "address canvassing operation" during four months 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 through the address canvassing operation were used to pinpoint the residences of non-responders in the 2010 Population Census in the mobile map carried by field workers. The newly adopted latest devises are expected to hike the response rate drastically and thus to improve the quality of the result. Statistics Poland is also planning to capture the GPS coordinates in the 2011 Census.

Japanese Statistics Bureau obtained GPS coordinates of establishments through matching addresses from the Establishment and Enterprise Census data with those in on-the-shelf digital map database provided by a private company. GPSed individual records are used to compile the grid statistics of 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 localizés" (RIL) which carries GPS coordinates as location information. The demographic department of the Institute, which is in charge of updating the RIL, obtains the coordinates in a following way. By purchasing road centerline information from the national geographical authority (Institute Géographique National: IGN), the department calculates coordinates that seem to correspond to each address. Since some residential buildings occasionally share the same address, there may happen that more than 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 respective coordinate information.

The 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 powerful driving forces for statistics 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.

Besides such applications of GPS coordinates in the survey process, they are expected to provide a wider dimension of inputs to statistical practices. As one of the major aims of this paper is to address the characteristics of individual records with GPS coordinates, it would be convenient to provide here a rough image of GPSed records.

The diagrams in Figure 3 document the images of a data format for GPSed records.

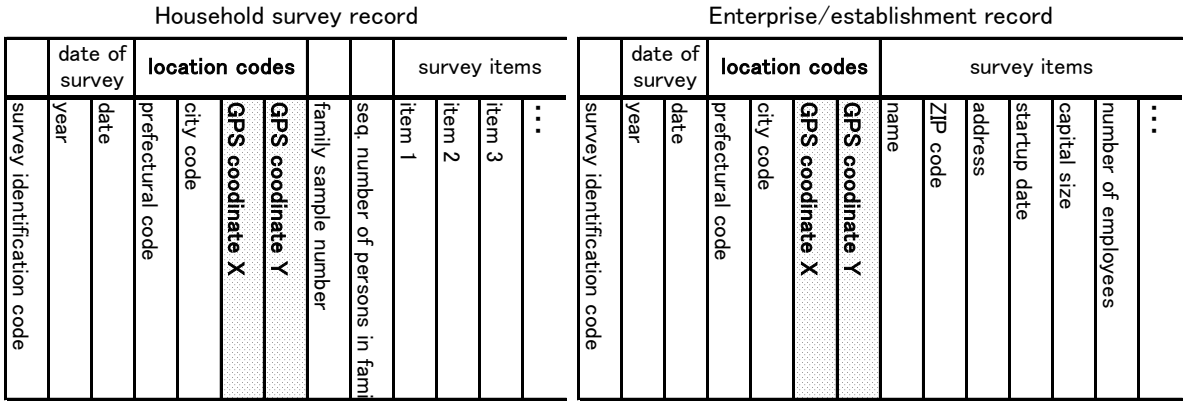



Figure 3. Examples of GPSed records

Unlike tract-coded records, GPSed records provide definite positional information of the surveyed units. As stated above, ambiguity in the use of data derives substantially from the area-based location positioning. GPS coordinates are more appropriate variables than tract codes in terms of identifying the geographical points of surveyed units' actual existence, although they still represent small areal grids.

Once GPS coordinates are tagged to individual records by some measures or other, it becomes possible to allocate the surveyed units not by the estimation but by direct assorting of surveyed units according to the coordinate information. Units such as households, establishments and enterprises have to be surveyed intrinsically at the very point of their presence. It was not until the obtaining of coordinate information that statistics became able to employ location information on an extensive scale.

GPS coordinates tagged to each record as one of the unit's basic attributes will enable to liquidate the ambiguity described above. By doing so, all archived records will be able to cope with every patterns of area zoning. GPSed time series individual records can also enjoy longitudinal comparability in full scale. Furthermore, they are qualified to compile statistics that can meet any buffered zones.

## 5. Breakthrough

This paragraph discusses how GPSed individual records can break up bottlenecks generated by the insufficiently obtained location information by types of surveyed units and datasets.

(A) Cross-sectional GPSed business datasets

As figures 1 and 3 documented, a pair of GPS coordinates (x, y) corresponds to each surveyed record, while the 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. GPSed records are distinguished from non-GPSed ones, among others, by a one-to-one correspondence of surveyed record with its location code. Since GPS coordinates provide an individual record with accurate pinpoint information in terms of each unit's location, GPSed records can be free from ambiguity in allocating units into respective regional areas that non-GPSed records were unable to do.

Allocating units in bordering areas to pertinent areas has been an extremely labor-intensive exercise in compiling grid square statistics. As cross-sectional GPSed datasets can cope with any regional zoning, it may be possible to complete it almost automatically with the help of coordinate information. It is quite reasonable that the Japanese Statistics Bureau converts address data to GPS coordinates in compiling grid square statistics from the Establishment and Enterprise Census data. They can also handle any claims in elaborating polygons required in various buffering analyses.

Cross-sectional GPSed business datasets may be applicable to the following analyses. Firstly, they can provide effective datasets for analysis of various aspects of industrial clusters. The analyses of territorial location of clusters, their economic size and density by region and industry are of major concerns among geographers.

The U.S. Census Bureau was exceptionally quick in assessing the damage caused by the intense Atlantic hurricanes Katrina, Rita and Wilma in 2005 with GPSed establishment records (Jarmin S.Ron and Miranda J., 2009). This case study offers a smart example demonstrating the potential usability of GPSed datasets, for example, in the field of disaster prevention. Central and local governments of most countries have already furnished with various hazard maps. One may easily assess the extent of damage by overlaying GPSed records on hazard maps using coordinate information as linking keys.

(B) Cross-sectional GPSed household datasets

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. Multiple-floor apartment houses may possibly be codified by one and the same pair of coordinates.

As French practices in the RIL shows, it is probable that hundreds of residential units occasionally share the identical coordinate information. Although in neither cases each residential unit and GPS coordinates do hold one-to-one correspondence, coordinates may still retain their validity as a location indicator, because they give a reasonable approximation in terms of the location of the units in question.

Theoretical researches for developing and putting 3 dimensional GIS in practice are now under way. GPS coordinates are also expected to expand their dimensions, for example, by introducing an additional variable that denotes floor information. Some local authorities have already furnished with the floor-based location maps of facilities.

GPSed household datasets are more informative than tract coded ones in analytical usability, because they are qualified to accommodate themselves to a wide spectrum of regional zoning. One can assess the number of casualties from natural disasters such as floods and earthquakes by overlaying GPSed records upon hazard maps. Statistical assessments of governmental services may also be possible by scoring accessibility to public facilities. GPSed household datasets capable of meeting any buffering analyses are also attractive to businesses in mining potential local markets by calculating the size, compositions, density and income distribution of subpopulations in relevant buffering areas.

(C) Repeated cross-sectional GPSed business datasets

Since coordinates are distinct in indicating the location of the units, one can obtain results not by estimation but by the direct counting of units through a vector algorism applicable to any level of polygons. GPSed records can display their advantages over other location codes especially in time series regional comparisons. Once individual records are archived with appropriate coordinate information, the datasets are to be released from every constraint in time series comparisons that was formerly caused by the restructured borders. Allocating units to each pertinent polygon by the help of coordinate information will make possible the prospective as well as retrospective regional comparisons.

Repeated cross-sectional GPSed business datasets obtained by a series of surveys will offer users a periodical chain of snapshots on the activities of business units and their behaviors. They can be applied to analyze, for example, the dynamism of an industrial cluster. With these types of datasets one can draw a series of pictures that illustrate the trend of diffusion or contraction of industrial clusters and can analyze business demographic events such as the entry/exit of units to/from the cluster.

(D) Repeated cross-sectional GPSed household datasets

Repeated cross-sectional GPSed household datasets give a chain of snapshots focused on the activities and behaviors of families over time. Thanks to the

coordinates, the datasets can support any restructuring of the regional zones. One can analyze various dynamic aspects of the population and families by each region using this type of dataset. Comparison of the ageing tempo of the population by region is of importance for policymakers who are keen on the reallocation of the budgets.

(E) Longitudinal GPSed business datasets

By the turn of the 21st century, business statistics in most countries had already become equipped with business registers that now serve as a fundamental survey infrastructure as well as a particular machine to produce relevant statistics. Business registers in many countries have already stepped up to the second generation phase as databases with a longitudinal dimension in order to be able to meet the analytical needs of business demography.

A business register, as the core segment of a relational database, forms a backbone for the integration of a wide spectrum of business statistical records both in cross-sectional (horizontal) and longitudinal (vertical) dimensions. A systematic coding of the ID numbers of business units is a prerequisite for the effective functioning of the database. Longitudinal records in themselves contain elements of business demography, such as a birth of the business (entry), survival (continuance), dormancy (suspension) and quitting (exit).

(F) Longitudinal GPSed household datasets

Building longitudinal household databases may currently remain a far-reaching project issue for most countries. However, Nordic countries have already switched over their statistical systems to register-based ones. Central Bureau of Statistics (CBS) of the Netherlands has constructed a modern version of the System of Social and Demographic Statistics (SSDS) as the Social Statistical Database (SSD), which is realized as a relational database with population register at its core segment and integrates multi-sourced household files, including administrative record files, as satellites.

As business registers have evolved from the first generation of the business frame that only reflected a static aspect of the business population to the second generation with longitudinal attributes, household registers will likely follow the similar steps in the future. In this sense, the current status of statistical practices regarding household registers may be rather premature for the following discussion on the potential usability of GPSed longitudinal datasets.

Longitudinal household datasets can be compiled through matching the records by family ID number. In case when the relevant ID number is not available, householders' names will substitute for it. 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 one side, and it provides information on the functioning of respective dwelling units in terms of habitation on the other.

If we direct our concerns to the units, i.e. families 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.

## 6. Concluding remarks

Official statistics, which have collected information from the surveyed units primarily to compile statistical tables, have experienced several historic turnabouts during the second half of the 20th century. Instead of macro-based datasets, the component of which are substantially aggregate statistics, users increasingly directed their concerns toward disaggregate data under the belief that the latter could portray novel and more correct images on the universe that the aggregate-data-based approaches were unable to attain.

Transition of the system of statistics from that made up substantially of stand-alone surveys to the micro-based integration of the surveyed and administrative records is another remarkable development. Collected information, which was formerly of temporary value simply for tabulating purposes, is more and more regarded as a sort of information asset of a durable nature that can meet the long-standing and varied integrated uses.

It is quite reasonable that contemporary needs for statistics require the archiving of obtained data which can withstand long term comparability and enable horizontal as well as vertical expansions of dimensions of the archived records. The focus on GPS coordinates themselves in this paper derives from the anticipation that arming surveyed records with GPS coordinates might be one of the possible options in designing the future data archives.

The use of GPS coordinates in official statistics is expected to be one of the remarkable developments in recent world statistics. Due to the technological as well as social constraints, it was not until quite recently that statisticians began to direct their attention to the extensive use of coordinates for the statistical purposes.

Although questionnaire-based surveys were qualified potentially to collect information that belongs or relates to each surveyed unit, traditional surveys have offered location information in an aggregate manner as tract codes. Because of the ambiguity of available location information that arises from tract-coded records, survey results were subject to several constraints in use, especially in time series regional analyses.

Discussions in this paper are focused to evidence the following issues. The questionnaire-based surveys missed to collect the distinct pinpoint information regarding the location of the surveyed units. Due to the insufficient obtaining of location information, traditional survey results have undergone various constraints in their effective use. By arming individual records with GPS coordinate information, users can allocate the surveyed units to any areas according to their relevant analytical purposes. They can release the surveyed results from various constraints in data processing. Data can be processed free of many constraints which derived from the tract-based survey taking. By doing so, individual records became able to explore informational potentials which the returned questionnaires have inherently carried.

This paper sets aside a lot of issues regarding the statistical use of GPS coordinates. When one discusses GPSed records from the standpoint of intensive as well as extensive employment of existing data, many further possibilities seem to be left for cultivation. Surveys of mobile units, such as person trip, are totally out of scope of this paper. They remain as the subjects for later studies.

REFERENCE

Jarmin S.Ron and Miranda J., (2009) "The Impact of Hurricanes Katrina, Rita and Wilma on Business Establishments: A GIS Approach" *Journal of Business Valuation and Economic Loss Analysis*. Vol.4

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.