ISSN 0385-2148

研究所報

No.41

Exploring Potential of Individual Statistical Records

2011年11月

法政大学
日 本 統 計 研 究 所

研究所報

No.41

Exploring Potential of Individual Statistical Records

2011年11月

法政大学
日 本 統 計 研 究 所

Preface

Time and space i.e. the observed time and location information of the surveyed units or statistical groups comprised thereof are of vital importance for statistical data. Many statisticians at the dawn of the modern statistics directed their interest to the relevant treatment of these factors in survey designing and practices.

In modern census, numerous small areas called “census tracts” were organized to cover exclusively the whole scope of the enumerating area so as to defend from possible oversights and overlaps. Since local authorities are in charge of survey practices, whole area of each municipality was subdivided into a number of census tracts. Thanks to such manner of tract settings, tract-based data could adjust themselves to municipality-based survey results. Traditional census and surveys have disseminated results, in principle, according to the respective hierarchical orders of local authorities. Tables compiled from regional coded individual records usually support regional comparisons. However, local authorities are occasionally reorganized to generate new political and administrative entities. Reorganized bordering of regions has given rise to serious constraints in terms of time series comparisons of regional results. Enormous clerical works were required to obtain comparable data out of regional coded records.

Grid area statistics was then introduced to obtain robust results totally free from reorganization of regional areas. Compiling grid-based area statistics out of tract-based records, however, still requires enormous clerical jobs for reallocation of units. Besides the requirements in the process of data compilation, grid area statistics has rather limited applicability due to the incongruity with administrative units, although the introduction of smaller level grids helps to portray phenomena with higher resolution quality.

Thanks to the removal of selective availability, successful launching of the quasi-zenith satellites and the enormous progress in hybrid positioning technologies, GPS now gives fairly good estimates for location coordinates. The promised precisions which GPS can now enjoy gave rise to a wide scope of new businesses and occasioned innovation to traditional industries.

Statistics is rather one of the latecomers in terms of using location information given by GPS. It is not longer than a decade when some statistical authorities of the world started to turn their attention to the potential usability of GPS coordinates for statistical purposes. Location information given by GPS coordinates is expected not only to cultivate new frontiers of statistical use but also to contribute to improve the surveyed data.

The basic concept which governs this book is derived from an idea that statistics thus far has insufficiently exploited the location information immanent in individual statistical records, although the introduction of questionnaire-based surveys had

paved the road for future cultivation.

Part one will discuss GPS information and exploring new arena of statistical data by loading with GPS coordinates. The second part of this book is assigned to describe varied aspects location information with regard to the micro-based integration of statistical data.

The essays carried in this book are part of researches funded by Japan Society for the Promotion of Science: “International Comparative Studies on Archiving System of Official Statistical Data”(#22330070) and “Multi-faceted Studies for Exploring New Frontiers of Official Statistics by Using GPS Information”(#40105854)

November 2011

Japan Statistics Research Institute

Exploring Potential of Individual Statistical Records

CONTENTS

Part one

GPS coordinates and statistical information

Exploring the Usability of GPSed Records: A data typological approach*

Hiromi MORI†

Summary

Up until quite recently, location information on surveyed units, for example, of households, establishments and enterprises, has been collected as area information, such as tract codes and municipality codes, in which one-to-one correspondences between the unit and the location information are not provided, despite the fact that each unit has inherently unique information in terms of its location.

This paper first addresses the issue that ambiguity of data due to insufficiently obtained location information under questionnaire-based surveys gives rise to several constraints in their use. Latest developments in information technologies have opened up new possibilities for the application of GPS for statistical purposes. One can create GPSed records by assigning relevant GPS codes to respective survey results. Compared with non GPSed records, GPSed records appear to yield several benefits. Thus, the remainder of the paper highlights the potential function of GPS codes with respect to the possibilities of cross-sectional as well as longitudinal data fusion, which is expected to explore new frontiers in integrated data production by expanding the dimensions of existing records. The discussion in this paper also implies that GPS coordinates are one of the possible key variables applicable to the integration of data.

Keywords: GPS, data archive, data fusion, data integration

1. Introduction

In modern census, enumerating activities have been conducted at each census

---

* This paper is based on a 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 Central Bureau of Statistics Netherlands on September 6, 2010 in Heerlen. An earlier version of this paper was already published in March 2011 on *Statistics* (The Japan Society of Economic Statistics), No.100.

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

tract, which exclusively covers the whole scope of national territory. Census tracts were introduced with the intent of avoiding counting failures as well as multiple counting. Although some surveys inquire about respondents' addresses, in most cases, location information of surveyed units is given either by region or tract code. The tract code has been used as the minimal unit to represent the location information of surveyed units within the tract. In other words, surveyed units in a particular tract have shared identical code numbers to represent their locations.

Modern information technology has provided a substantial breakthrough in obtaining the location information of surveyed units. Due to advanced information technology, together with the wide-spread use of reasonably-priced handheld PCs, the Global Positioning System (GPS), originally introduced as a military invention, is now widely applied in various fields as a civilian technology.

The aims of this paper are twofold: first to document a set of problems caused by the insufficient collection of the location information of surveyed units, and second to draw a sketch of the potential uses of GPSed records with regard to the typology of data.

2. Statistical surveys and the dual nature of surveyed records

In conducting surveys, information is collected from surveyed units such as persons, households, establishments, enterprises and so on, through the use of questionnaires. The information obtained concerning surveyed units is usually arrayed as a record format, which, however, has a dual nature.

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 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, a set of informational data offered by surveyed units are 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 censuses, 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 when units

change their locations over a period of time. Although the majority of the surveyed units continue to stay 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 when 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 formerly being unsettled dwelling units, while establishments and companies can launch their business activities either at newly developed industrial sites or ones that were formerly 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, 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 small area. Because insufficient location information was obtained, statistics 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.

Household survey record

	date of survey		**location codes**					survey items			
survey identification code	year	date	**prefectural code**	**city code**	**survey tract code**	family sample number	seq. number of ind. in family	item 1	item 2	item 3	・・・

Enterprise/establishment record

	date of survey		**location codes**			survey items									
survey identification code	year	date	**prefectural code**	**city code**	**survey tract code**	name	ZIP code	address	startup date	capital size	number of employees	item 1	item 2	item 3	・・・

Figure 1. Examples of record layout forms

3. Information constraints of tract-based records

(1) Problems caused by border rezoning

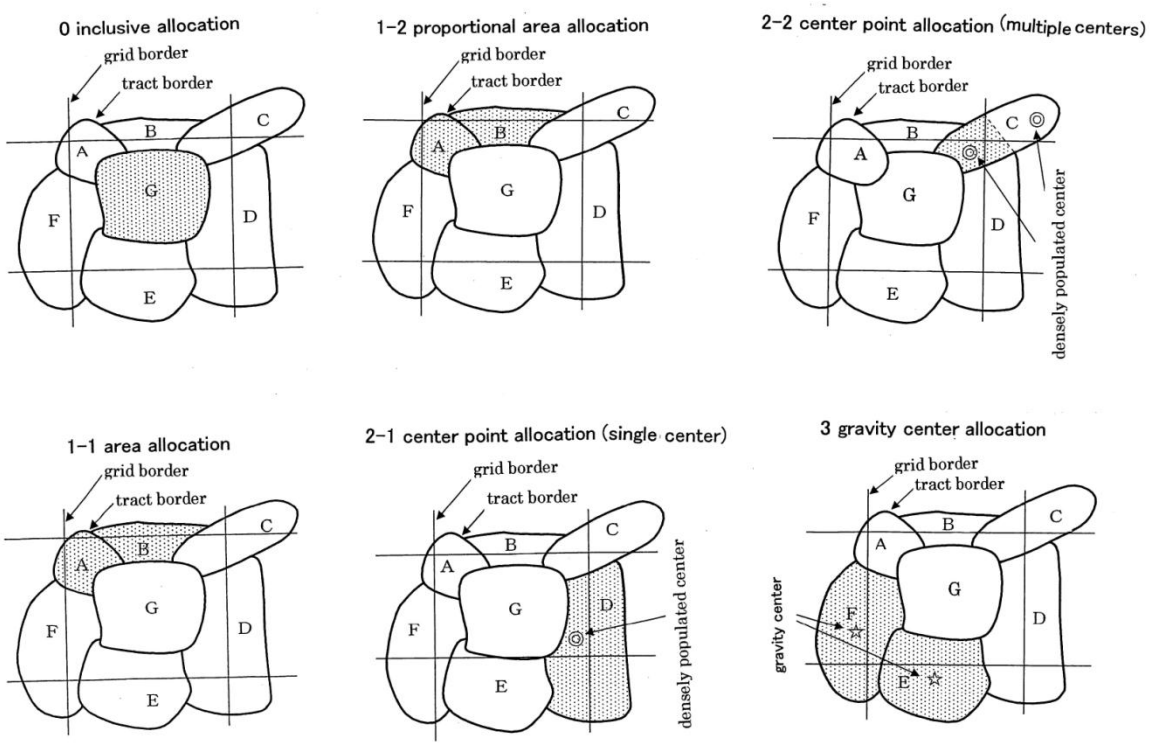
While Japan had more than 12,000 cities, towns and villages in the 1950s, the number had diminished drastically to about 2,200 by the year 2005. The annexation and reorganization of municipalities are real threats to statistical comparability, since they require enormous amounts of clerical work to adjust historical statistics to the newly arrayed 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 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 introduced in Japan in 1990 as the minimal survey tract area of a more or less durable nature. Although these are expected to be more robust than census tracts, they still are not always free from restructuring.

(2) Allocation of surveyed units in tracts

Grid Square Statistics were introduced in Japan based on the 1970 census data. 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.

Since the geodetic line partitions 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, 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 data.



Source: http://www.stat.go.jp/data/mesh/pdf/gaiyo2.pdf (pp.24, 26 and 28)

Figure 2. Allocation of tract units among grid squares

(3) Inadaptability of data for buffering analysis

Buffering analysis is now widely used to identify buffered polygon areas with a fixed distance surrounding 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 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 surveyed units in tracts to 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 improvements in the accuracy of digital map software, together with widespread use of information terminals furnished with various GIS softwares, GPS now enjoys a wider acceptance in daily life as necessary information infrastructure.

Official statistics, however, are relative latecomers in applying GPS for their practices.

Directly obtaining GPS coordinates through mobile terminals and indirect access to them either by means of address-GPS converting software or by applying appropriate calculation methodologies served as a powerful driving force 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. GPS coordinates collected by field workers in address canvassing for the 2010 U.S. Population Census are used to pinpoint the residences of non-responders, and thus to improve the response rate. Statistics Poland is planning to collect GPS coordinates in the 2011 Census.

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 (hereinafter termed GPSed records), it would be convenient to provide here a rough image of GPSed records. The diagrams in Figure 3 document images of a data format for GPSed records.

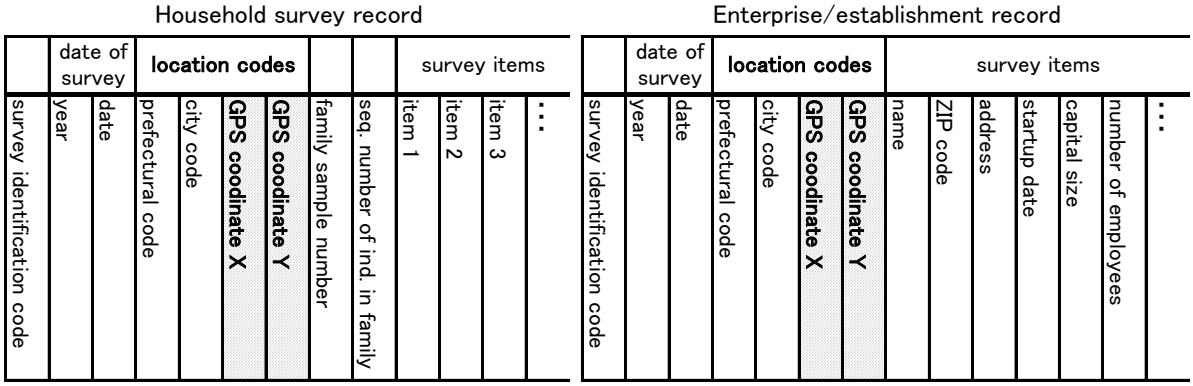


Figure 3. Examples of GPSed records

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 be surveyed intrinsically at the very point of their existence. It was not until the obtaining of coordinate information that statistics 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 to liquidate the ambiguity described above. By doing so, all archived records

will be able to cope with any form of 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 advantages, GPSed records appear to have additional attributes. The following paragraphs will discuss potential uses of GPSed datasets with regard to the typology of data.

5. GPSed records by type of datasets

Datasets can be classified into several subcategories by kinds of surveyed units and forms of datasets. Additional variables that account for the datasets will also be introduced to characterize the specific nature and usability of GPSed datasets.

A single census or survey result provides a snapshot of the surveyed units at a particular date that forms a single cross-sectional dataset. A series of censuses or surveys conducted repeatedly during the sequence of time will give repeated snapshots. These snapshots usually comprise repeated cross-sectional datasets. Leaving aside censuses, 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. When the same units are surveyed repeatedly in a series of surveys, one can compile longitudinal 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 longitudinal dataset because of the attrition of the surveyed samples. Including unbalanced datasets with an unequal number of surveyed units in each snapshot, the author simply terms such datasets here as longitudinal.

As for the nature of 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. dwelling units usually inhabited by families and sites where establishments/enterprises perform their economic activities, are currently our major concerns in discussing GPSed records. Individual records loaded with GPS coordinates involve in themselves a potential moment to breakaway the dual nature that seems to be inseparably integrated 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.

Table 1 Business/household datasets by type

surveyed units	observation unit	single snapshot	repeated snapshots	
		cross-sectional	repeated cross-sectional	longitudinal
business (enterprise/ establishment)	unit	(A)	(C)	(E)
	site			
household	unit	(B)	(D)	(F)
	dwelling			

6. Possible uses of GPSed records by type of datasets

Categories of GPSed datasets (A) through (F) in Table 1 appear to have particular attributes regarding each surveyed unit and its location information, which govern the scopes and dimensions of their usability.

(A) Cross-sectional GPSed business datasets

As figures 1 and 4 have documented, 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. 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 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, for example, to the following analyses. Firstly, they can provide effective datasets for analysis of various aspects of industrial clusters. The 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 so quick in assessing damages caused by the hurricanes Katrina, Rita and Wilma with GPSed establishment records (Jarmin S.Ron and Miranda J., 2009), where GPS coordinates had not been obtained neither by direct

positioning the sites by field surveyors nor through address matching , but through calculation measures they loaded individual establishment records with coordinates. This case study offers one smart example demonstrating the potential usability of GPSed datasets, for example, in the field of disaster damage prevention. Central and local governments of most countries have already furnished various hazard maps. One may easily assess the extent of damage by overlaying GPSed records on hazard maps using coordinates as linking keys.

It is worth noting that GPSed cross-sectional records also have an advantage in enlarging the information potential of data by means of expanding dimensions through data fusion. Among individual records from multiple 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. The latest developments in statistics have shed light on data fusion as one of the possible expansions of information potential. Records with a multiplied number of variables generated by the coordinate-based cross-sectional data fusion among heterogeneous business records may allow intensive analyses that a single set of records could never hope to achieve.

(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. 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 indicator. GPS coordinates are also expected to expand their dimensions, for example, by introducing an additional variable that denotes floor information.

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 estimate or assess the number of casualties from natural disasters such as floods and earthquakes by overlaying the 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.

Expanding the potentials 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.

(C) Repeated cross-sectional GPSed business datasets

Since coordinates are distinct in indicating the location of units, one can obtain results not by estimation but by the direct counting of units through a vector algorism applicable to any levels of polygons. GPSed records can display their advantages over other location codes, among others, in time series regional comparisons. Once individual records are archived with appropriate coordinates, they will become able to release the data from every constraint in time series comparisons that was formerly caused by restructured borders. Allocating units to each pertinent polygon by the help of coordinate information will make possible 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 behaviors. They can be applied to the analysis, for example, of 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 or exit of units to or from the cluster.

One of the characteristic features of the repeated cross-sectional GPSed datasets is the possibility of longitudinal expansion of data dimensions. When we focus our interest on the location information of 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 continue to carry out their activities at the same sites they have occupied in the past, we regard the compiled datasets as a panel in the broader sense.

When one changes ones viewpoint to the location where each unit was actually surveyed, however, both panel datasets are “genuine” in nature. Put differently, GPS coordinates are qualified to work as effective key variables to generate panel datasets out of unpaneled repeated cross-sectional datasets. Thus, panel-based analyses would be applicable to these types of datasets.

(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 the course of time. Thanks to the coordinates, the datasets can support any restructuring of regional zones. One

can analyze various dynamic aspects of population and families by each region using this type of dataset. Comparison of the ageing tempo of populations by region, for example, is of importance for policymakers who are keen on reallocating budgets.

When one regards repeated cross-sectional GPSed household datasets from the viewpoint of GPS coordinates, individual household records are reorganized into pseudo panel datasets. Similar to the business datasets, those compiled from repeated cross-sectional GPSed household datasets are still pseudo in terms of longitudinal attributes, because coordinates are linked not to respective families, but only to the dwelling units. Even in cases where household records carry unchanged coordinates in repeated cross-sectional datasets, there may possibly be replacements of families in dwelling units caused by the moving out of a family followed by moving in of another family. 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, 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 family demographic events analyses.

(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 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 and longitudinal 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 launching a business (entry), survival, dormancy (suspension) and quitting (exit).

GPSed longitudinal datasets are far more informative than non GPSed ones. Longitudinal records armed with GPS coordinates are qualified to objectify the dual aspect, i.e. unit and site information which the individual records have carried latently. When one focuses upon 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, it will establish the kinds of functions or potentials of the respective sites.

GPSed longitudinal business datasets can identify the following events. When one focuses on the business unit in the dataset, its coordinates provide information

regarding the unit's relocations in the course of time. Since the unit is identified by the competent ID number, one can easily distinguish redeployment from quitting.

GPS coordinates are more advantageous than descriptive address information in terms of data processing in identifying the redeployments of units. Addresses tend to be mistyped, while coordinates can maintain consistency even in cases when addresses are amended by occasional address recording.

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, 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.

By controlling site information, GPSed longitudinal business 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.

(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 many other household files as satellites.

As business registers have evolved, as a matter of course, from the first generation of the business frame that only reflected a static aspect of the business population to ones with longitudinal attributes, household registers will likely follow 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 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 one side, and provides information on the functioning of respective dwelling units in terms of habitation on the other.

If we direct our concerns to 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.

Longitudinal household 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 / the name of householder or differed ones in a series of snapshots. By overlaying family ID number or the name of 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. The coordinates that became extinct in the GPSed longitudinal datasets compiled of household-based survey results may indicate a 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

Official statistics, which have collected information from 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 direct their concerns toward disaggregate data in the belief that the latter could portray novel images on population that aggregate data approaches were unable to attain.

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

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

The use of GPS coordinates in official statistics is one of the recent remarkable developments in world statistics. Due mainly to technological and partly to social

constraints, it was not until quite recently that statisticians began to turn their attention to the extensive use of location coordinates for statistical purposes.

Although questionnaire-based surveys are 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.

This paper has given evidence for the following issues. First, besides information on the surveyed units, individual records also carry information about the locations of units in latent manner. Second, due to the aggregate nature of coding in terms of the location of units, transferred unit records from returned questionnaires, which are insufficient in giving the location information of units, have placed restrictions on their use. Third, as a typological approach has evidenced, repeated cross-sectional and, among others, longitudinal datasets can highlight location information given by the coordinates with special implications that may represent a sort of potential inherent in the respective spots. Finally, corresponding to each type of datasets, we have explored some new possibilities for GPSed records' usability.

As addressed by U.S. practices, GPS coordinates are also promising in producing better quality data. Dutch attempts may explore new potential uses in obtaining statistics otherwise almost unobtainable by conventional methods. Further seminal elaborations in the application of GPS coordinates in statistics may enrich the discussions being put forward in this paper. It is expected that new types of datasets armed with GPS coordinates will explore new frontiers in the field of statistics.

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.

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.

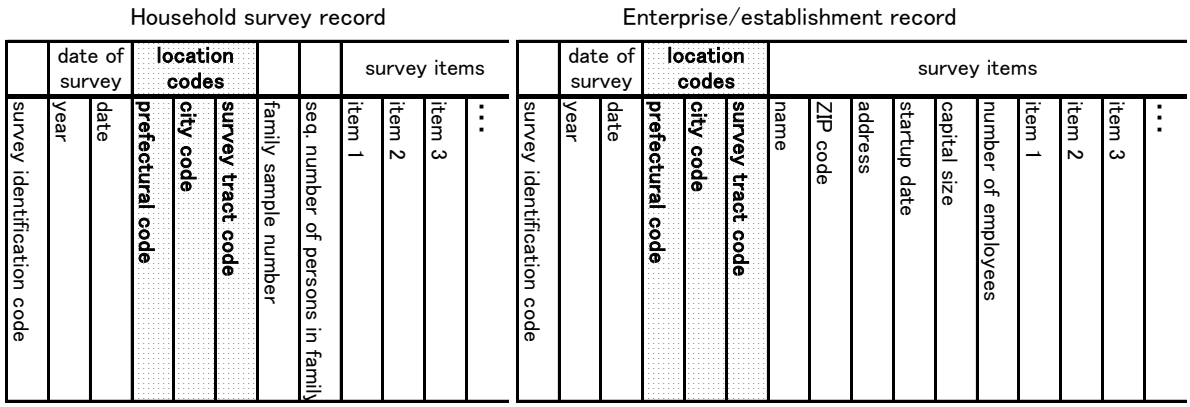


Figure 1. Examples of record layout forms

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.

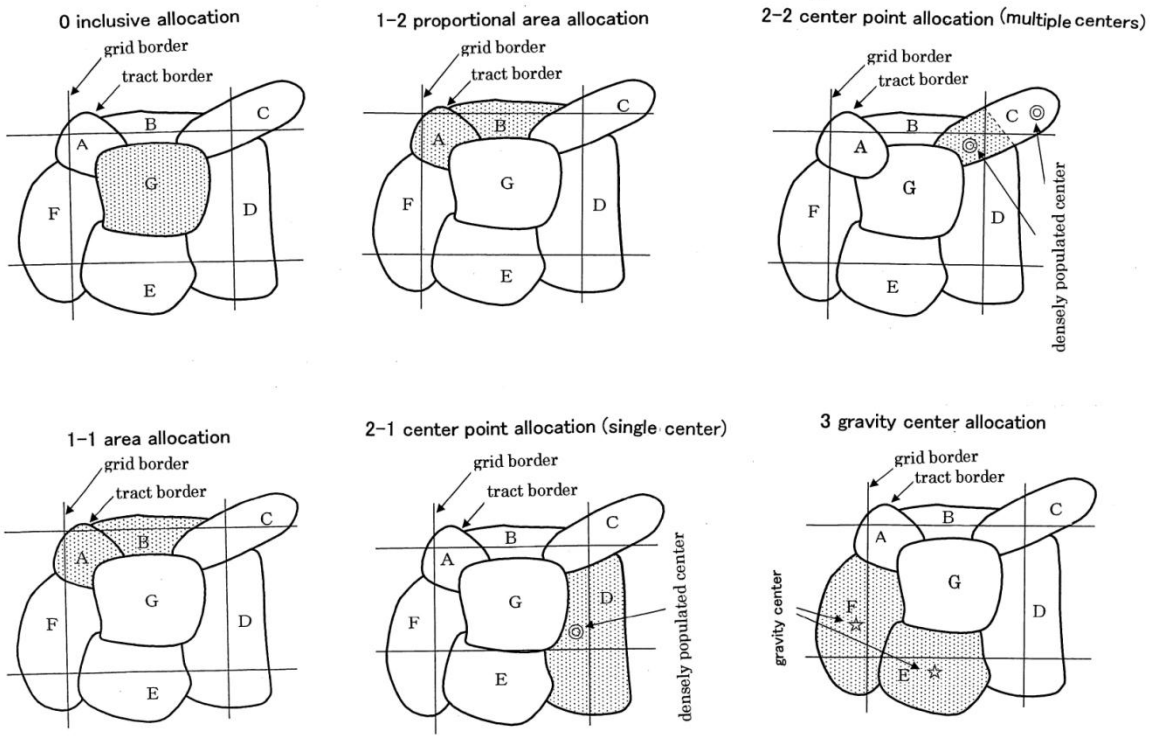
Census tracts are not totally immune from boundary rezoning. The completion of new roads and railways and the development of new residential areas make the existing tract maps obsolete. Some tracts have been partitioned and then annexed to several neighboring tracts, while several others have been totally 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.

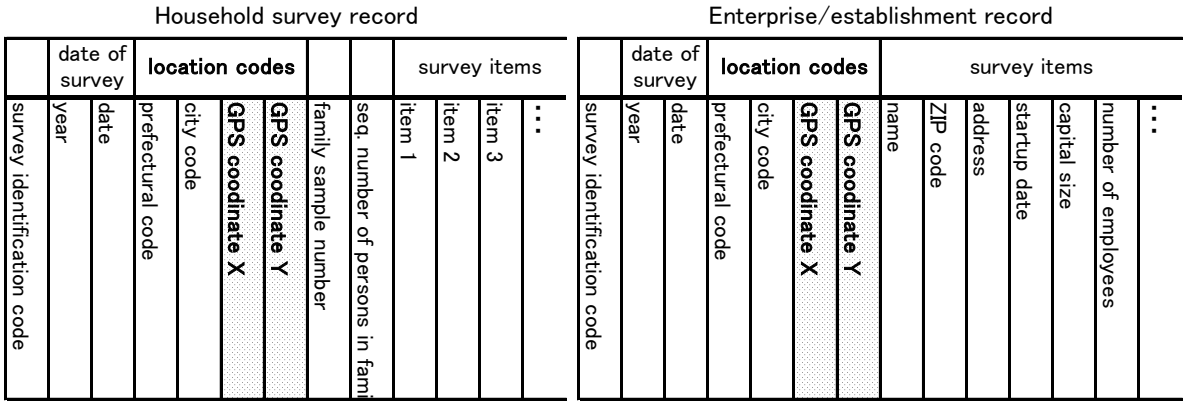


Figure 3. Examples of GPSed records

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.

GPSed Datasets and the Possibility of Exploring the Micro-based Concept of Regional Potentiality*

Hiromi MORI**

Summary

Up until today, statistics have treated location information of the respective survey units quite improperly due mainly to the insufficient obtaining of the relevant information. In the traditional statistical records location information captured through questionnaire-based surveys has been given in principle as the regional codes such as tract code, where the whole surveyed units in the tract in question share one unique code number. Consequently, the n-to-one correspondence between the units and location, which was prevalent in the traditional datasets, has impeded a full-scaled exploitation of statistical data for the regional analyses.

Latest developments in information technology, however, enabled to allocate distinct positional information as the Global Positioning System (GPS) coordinates to the respective surveyed records. Inspired by such technical input, the author brought forward in this paper a concept of regional potentiality that can be explored with GPSed individual records as one of the possible expansions of information collected through questionnaire-based surveys or administrative records. The discussion of the usability of GPSed records by type of datasets evidenced that a wide scope of issues are still left for the future cultivation.

Keywords: GPS, regional potentiality, questionnaire-based surveys, tract code

1. Introduction

In modern censuses, enumerating activities have been operated by each census tract, a set of which exclusively covers the whole scope of national territory. Up until quite recently, location information on the surveyed units, such as households, establishments and enterprises, has been captured basically as area information, such as tract code, which only provides the n-to-one correspondence between the units and the location indicators, despite the fact that each unit has inherent positional

* An earlier version of this paper was published in March 2011 on KEIZAI- SHIRIN (The Hosei University Economic Review), Vol.79, No.1.

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

information. In other words, the regional codes such as tract code attenuate the information value in terms of unit's location that each surveyed unit intrinsically bestowed in the process of field survey operation.

Due to the advanced information technology, together with the wide-spread use of reasonable price handheld PCs, the GPS, originally introduced as a military invention, is now widely applied in various fields as a civilian technology. It has also opened up new possibilities for the application of this positioning technology for statistical purposes. One can compile the GPSed datasets by allocating relevant GPS codes to the respective surveyed records. The diagrams in Figure 1 document images of a data format for GPSed household and business records.

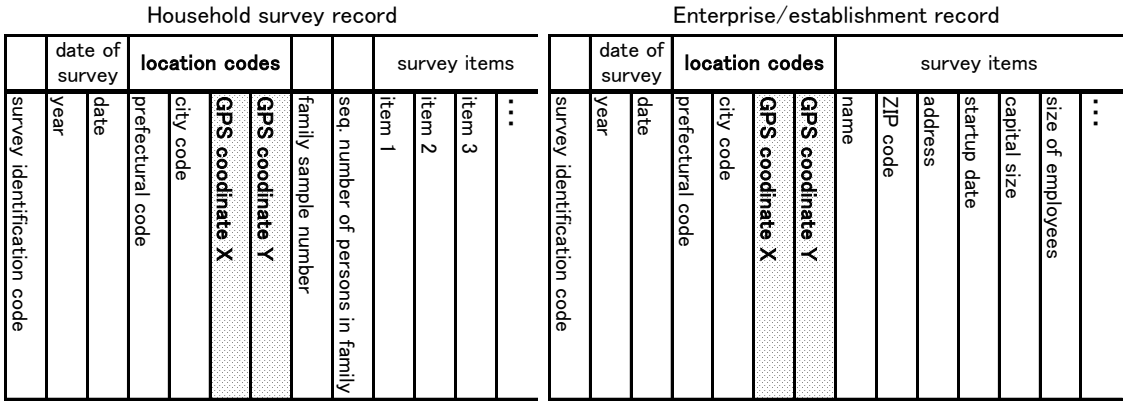


Figure 1. Examples of GPSed records

Among key variables applied for tabulating the surveyed results, those variables such as prefectures and cities were remarkable in providing users with meaningful statistical materials. Developments in the information technology made possible for national statistical authorities to disseminate tables with a wider spectrum of regional demarcation. Growing involvement of a variety of small area statistics in the list of publications evidences the outcome of the developments.

Regional results that constitute quite a few segments of disseminated tables have either one of the hierarchical regional demarcation as tabulating key variable. However, it is worth noting that there are some tables processed in a way as to hold the distinctive regional characteristics or attributes in common among regions. Tables by the size of population, those with densely inhabited districts (DID) tables and those by the level of municipality may fall in this category. These tables carry results processed according to the attributes of population such as its size and density or other institutional classification, regardless of actual geographical locations of respective regions.

Micro-based statistical analyses now enjoy a wider acceptance among analysts. Because of the absence of the relevant micro-based datasets with distinct location identifiers, prevalent model analyses are likely to disregard a variable of substantial

importance i.e. a variable that indicates various geographical attributes of the respective locations where the individual unit actually exists.

The aims of this paper are, first to elucidate the dual nature of the surveyed records, second to discuss characteristics of GPSed datasets, third to give a sketch of regional potentiality that can be drawn with GPSed individual records, and finally to suggest the possibility of measuring the regional potentiality by type of datasets.

2. Statistical surveys and the dual nature of the surveyed records

In the process of survey operation, information is collected from the surveyed units such as persons, households, establishments and enterprises through questionnaires. The information obtained from the surveyed units is usually arrayed as a record format for processing. Up until today, however, statistics have overlooked the fact that the recorded information has a dual nature.

It is obvious that the obtained data, i.e. the documented records of various attributes, activities of the units and their outcomes, are ascribed to the respective surveyed units. The individual records are nothing other than statistical copies of the surveyed units.

Another aspect of the record is less obvious compared with the first one. Since the surveyed units are actual beings in the real world, the surveyed information belongs or relates to the units that are located at a particular geographical point, i.e. a dwelling unit or site where business activities are operated. Put differently, a set of information offered by the surveyed units are related to some particular geographical point. One may term the former aspect of the surveyed records as “unit information”, while the latter as “spot information.”

Observations given by a simple survey are unlikely to address explicitly the spot information of the surveyed unit, because it is inseparably integrated into the unit information. The repeated surveys, however, may more clearly throw light on the dual nature of the records. When the same unit was repeatedly observed in a series of surveys, the obtained records may document the longitudinal changes of the relevant unit. Whether or not identical surveyed units that accommodate the particular spot, when the repeated surveys succeed in obtaining reports, it will turn out to address the activities or performances of the respective fixed points at different moments.

Although these two aspects which seem to be integrated inseparably to generate one record in the single survey, they may split off each other in cases when units change their locations at the subsequent surveys. It is probable that some of the surveyed units are subjected to the redeployment of their location over time. Different units may possibly be observed in the ensuing surveys at the same spot due to the replacement of the units, i.e. by a former unit’s moving out followed by a subsequent moving in. The observed spots in the previous survey might disappear, whether or not

the dwelling units are existent, in cases when no subsequent tenants accommodate that dwelling unit. It may also be possible that new entrants are surveyed at spots which were not documented previously. Families can be occupants either of newly constructed or so far unsettled dwelling units, while establishments and companies can launch their business activities either at newly developed industrial sites or at rental facilities that were unoccupied at the previous survey.

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 a population or its subgroups. It was only in the latter half of the 20th century that statistical analysts began to shed light on the individual survey records.

Due to such traditional notions on statistics, together with several technological constraints, statistics was subjected to be tolerant of the insufficient use of the positional information inherent in the survey records. Although the surveyed units such as households, establishments and enterprises mostly have definite positional information regarding their existence, the surveyed records documented them not at their particular points, but simply as one of the component units of the tract. Instead of the specified positional codes inherent in the respective surveyed units, an unified tract code number was allocated to all surveyed units that belonged to the tract in question. Each unit's location information was linked not to the geographical point, but to the small area. Because of the insufficient capturing of the positional information, statistics had to put up with so far the "diluted" information in positioning the units. Being incapable of identifying the relevant information which the surveyed records had carried in latent manner, statistics have suffered from a number of constraints on exploiting the data.

3. GPSed records by type of datasets

As figures 1 has illustrates, a pair of GPS coordinates (x, y) are tagged to each surveyed record, while a certain number of 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 are to carry an identical location code number. The GPS coordinates provide an individual record with a distinct pinpoint information in terms of each unit's location.

By the way, the datasets can be classified into several subcategories by kinds of the surveyed units and forms of datasets. Additional variables that characterize the types of datasets will also be introduced to account for the specific nature and usability of GPSed datasets.

As for the nature of the surveyed units, this paper focuses the discussion on GPSed records of the surveyed units with comparably stable nature in terms of their

geographical locations. Thus, the geographical locations of the dwelling units usually inhabited by families and sites where establishments/enterprises operate their business activity, are currently our major concerns in discussing the usability of GPSed records. Individual records armed with GPS coordinates involve in themselves a intrinsic moment to split their dual nature that seems to be inseparably integrated 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.

Table 1 illustrates categories of datasets by type of the surveyed units and datasets.

Table 1 Business/household datasets by type

surveyed unit	observation unit	single snapshot	repeated snapshots	
		cross-sectional	repeated cross-sectioanl	longitudinal
business/ household	unit	(A)	(B)	(C)
	site/ dwelling unit			

Categories of GPSed datasets in Table 1 appear to have particular attributes regarding each surveyed unit and its location information, which govern the scopes and dimensions of their usability.

(A) GPSed cross-sectional datasets

A single survey result provides a snapshot picture drawn by the surveyed units at a particular moment. It forms a single cross-sectional dataset, in which a set of information obtained through surveys are associated with a pair of coordinates in the GPSed datasets. The coordinates, which give the positional information of the particular surveyed units, are integrated into the surveyed records.

Households usually lead their lives in a certain residential unit and the units such as establishments, companies and other organizational entities mostly perform their business or other activities at distinct sites. As illustrated in the table 1, in the respective surveyed records which belong to the dataset in category (A), the location where the surveyed units are observed is definitely identical with the place of their daily activities. In this case, each surveyed unit enjoys benefits provided by the facilities i.e. residential units, production equipments together with a set of so-called "external effects" that possibly derive from the environmental conditions of the area.

(B) GPSed repeated cross-sectional datasets

A series of surveys conducted repeatedly over time offer a set of repeated statistical snapshots drawn with the repeated cross-sectional datasets. Population subgroups given by the unpaneled census results and a series of survey results, for example, a group of the unemployed or enterprises that operate business activities in certain industrial sectors, do not necessarily cover the same surveyed units in a chain of surveys. The repeated cross-sectional datasets, therefore, do not portray snapshot observations of the same set of the surveyed units. Nevertheless, since the repeated snapshots are still qualified to bring under light the gross changes of the matters, the

aggregate data given by a series of survey results are applicable to the macroscopic analyses of time sequential changes.

Additional information potentials that are expected to yield many uncultivated benefits in statistical analyses seem to inhere in the GPS-tagged repeated cross-sectional datasets. By using GPS coordinates as the matching key variables, the existing records are capable of expanding their dimensions cross-sectionally (horizontally) as well as longitudinally (vertically).

The horizontal extension is achieved by linking together the individual records from different sources including those from administrative sources as far as they hold coordinates in common. One should note that the extension through integrating records in this way are "pseudo", because records from different souses, which have common coordinate information, do not necessarily mean that they are identical units. Even in cases when the same unit is horizontally linked together, the more distanced in terms of time intervals between the sources of data, the less efficient becomes the integration.

GPSed records can be expanded also vertically by integrating them from the same series of surveys. Similar to the horizontal expansion, the datasets compiled in this way are also "pseudo" in nature, because coordinates are linked not directly to the respective households, but only to the dwelling units. Even in cases where household records carry unchanged coordinates in the repeated cross-sectional datasets, there may possibly occur the replacements of families in the dwelling unit caused by the moving out of a family followed by another family's moving in.

The same cases are also applied to the business units. The datasets are pseudo in the sense that establishments or companies that perform their business activities at the respective sites are not necessarily the identical units. The moving out of one unit followed by another subsequent entry may give rise to the replacement of the businesses.

Either of these two types of datasets compiled through horizontal or vertical matching by using coordinate information as linking key variables are pseudo in terms of panel datasets from the surveyed units' standpoint. Once one turns the viewpoint to the location where each unit was actually surveyed, however, both panel datasets are "genuine" in nature. Put differently, the GPS coordinates are qualified to work as the effective key variables to generate the panel datasets out of unpaneled repeated cross-sectional datasets. Thus, the information given by the datasets that fall in category (B) can imply a kind of performance of the respective locations.

(C) GPSed longitudinal datasets

When the same units are surveyed repeatedly in a series of surveys, one can compile the panel dataset that can be described by a matrix of $N \times T$ for each surveyed variable where N and T denote the number of the surveyed units and that of snapshots, respectively. However, the number of observations in each snapshot is not always the

same in the panel dataset because of the probable attrition of the surveyed samples. Including the unbalanced datasets with an unequal number of observations in each snapshot, the author simply terms here them as panel datasets, due to the longitudinal nature of the surveyed units.

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 the backbone for the horizontal as well as vertical integration of a wide spectrum of business statistical records. A systematic coding of the ID numbers of business units is prerequisites for the effective functioning of the database. Longitudinal records in themselves contain elements of business demography, such as launching a business (entry), survival (continuation), dormancy (suspension) and quitting (exit).

The GPSed longitudinal datasets are far more informative compared with the non GPSed ones. Longitudinal records armed with the GPS coordinates are definitely qualified to objectify the dual aspect, which the individual records have carried latently. When viewed from another angle, i.e. the standpoint of the units in the GPSed longitudinal datasets, the unchanged coordinates indicate the survival of the unit, while the changed ones suggest its redeployment. If one switches the view to the perspective of location, records illustrate the activities of the units operated at the particular location specified by the coordinates. In other words, it will establish the functions or potentials of the respective geographical points governed by surrounding conditions.

Similar to the longitudinal business records, household records also 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 provides information on the functioning of respective dwelling units in terms of habitation on the other.

4. How does regional potentiality reveal in statistics

(A) GPSed cross-sectional datasets

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 on the mobile map carried by the field workers the residences of non-responders in the 2010 Population Census.

Japanese Statistics Bureau obtained GPS coordinates of establishments and enterprises 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 square 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 localisés" (RIL) which carries GPS coordinates as location information. The demographic department of the Institute which is in charge of maintaining 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 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. Therefore, it is not a residential unit but an address that corresponds to a set of coordinates in the RIL.

The directly captured GPS coordinates through mobile terminals and indirect access to them either by means of address-GPS converting software or by applying appropriate calculation methodologies, however, do not always support the one-to-one correspondence between the coordinates and the respective surveyed units, but a unique set of coordinates often represent several mailing addresses. Despite the appearance of possible one-to-n correspondence, the GPS coordinates give the distinct location information where the relevant units such as households, establishments and enterprises actually exist. Thus, the GPSed single cross-sectional datasets are able to handle any claims laid by various regional analyses including buffering analyses. They are qualified to accommodate the records to a wide spectrum of regional zoning.

GPS coordinates are more advantageous than descriptive address information in terms of data processing in identifying the redeployments of units. Addresses tend to be mistyped, while coordinates can maintain consistency even in cases when addresses are amended by occasional address recording.

GPSed records 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. The distribution of regional population by age or income given by the GPSed cross-sectional datasets of households may provide commercial businesses with information on the anticipated market size for the planned new shops or factory owners with information on the size of mobilizable commuters to the planned new production firms.

It is well supposed that the accessibility to the public transportation might affect people's involvement in job or the occupancy rates of rental offices, residential units and business sites by the tenants. By using this type of datasets, one can even assess the

magnitude of intangible value inherent in the respective sites by comparing the yields produced by sites with comparable conditions including transportation accessibility.

Analysts can organize the diversified regional analyses including the multi-dimensional potentials which the respective sites inherently possess by generating areas at their disposal using GPS coordinates as a key variable.

(B) GPSed repeated cross-sectional datasets

The repeated cross-sectional datasets can also explore their information potentials by arming records with GPS coordinates.

As already described in the paragraph 3, the GPSed repeated cross-sectional datasets carry a genuine panel nature when individual records are linked together through the positional information. They document a set of observations at the fixed point. Since they give the repeated snapshot pictures of the particular point, one can identify the dynamism of potentials over time by comparing them at the point in question estimated by the single cross-sectional datasets. By comparing the trends of job involvement in several regions with comparable accessibility to the public transportation, one can identify regional discrepancies in terms of potentials of attracting residents for jobs. Each parcel of land does not change its prices unanimously but usually in diversified manner even in case where they fall in the same category in terms of land use. Among similarly inhabited commercial zones, some lands reveal upward trend in land price while others are not. Potentials inherent in respective areas may account for the discrepancies. Exogenous effects caused, for example, by the completion of the new roads or by the opening of new railroad stations are assessed by comparing the potentials with sites of comparable regional attributes.

(C) GPSed longitudinal datasets

The GPSed longitudinal datasets can identify the following events. When one focuses, for example, on the business unit in the dataset, changes of its coordinates document the unit's relocations over time. Since the unit is identified by the competent ID number, one can easily distinguish redeployment from quitting of businesses.

By controlling the site information, the GPSed longitudinal business datasets would be applicable to establish the redeployment ratio of business units by size and industry and to compare the ratios between the single and multiple establishment businesses and those between the grouped and single enterprises.

Business units go through a set of demographic events throughout the period of their activities. When one focuses on the coordinates, the surveyed unit records being identifiable by the unit ID number may log the demographic events which the business units experience over time, such as survivals, entries, exits which are performed at a particular location. Thanks to the unit ID number, it is possible to distinguish new entries from the moving in of existing units by redeployment and also exits from the moving out of the units.

By using the GPSed longitudinal datasets of households focused on the dwelling

unit, one can draw a new picture of the habitation behavior of residents that the repeated cross-sectional ones are unable to achieve. Household records reported from residents of residential units with unchanged coordinates may either give the same or different family ID number or the name of householders in a series of snapshots. By overlaying the family ID number or the name of householders on respective coordinates, one can compile a dataset that helps to shed light on the occupancy status of residential units. The unchanged ID numbers suggest that the same families or individuals continue to reside at the same residential units, while the changed numbers indicate replacement of families or individuals. The coordinates that became extinct in the GPSed longitudinal datasets compiled of household-based survey results may indicate a vacancy or a halt of operation as residential units, while the newly emerged coordinates will suggest the new engagements as residential units. The datasets will also be applicable to measure the residential mobility, for example, by region and tenure.

Since the individual records in the GPSed longitudinal datasets carry two matching keys i.e. ID numbers and the coordinate information, to integrate the data, one can compile the genuine panel datasets in terms both of unit and location point by setting aside records only linkable by either one of variables.

The panel datasets achieved through the dual way integration are applicable to a series of analyses mentioned in this paragraph (B) and are expected to yield results with improved precisions, because they can avoid disturbances possibly caused by the insufficiency in integrating records.

5. Concluding remarks

The following ideas have motivated this paper. First, statistical information obtained through the returned questionnaires may possibly mirror a sort of activities of respective location performed by the surveyed units. Second, the location point or area may have distinct potentials inherently which the traditional statistical analyses have overlooked. And finally, the extensive use of the GPSed datasets may be helpful to identify their magnitudes and their changes.

By arming individual statistical records with the GPS coordinates, it seems likely that datasets can acquire additional advantages in their applicability compared with the non GPSed ones. This paper focused the discussion on the analytical value of the location information in statistics that has been treated rather improperly until recently. The author tried to elucidate various attributes by type of datasets and to cultivate a new frontier in the use of statistics.

Worth remarking is that a simple substitution of location code by GPS coordinates expands dramatically the value of information that each individual record has inherently carried. Mori (2010) has already discussed some aspects of its advantages. This paper highlighted a net contribution of GPS coding in exploring the so-to-say

"potentiality" that respective areas inherently carry and its possible changes over time. The fact that GPS coordinates emerged as an effective key variable to carry out the micro-based integration of records is one of the findings of this study. As the discussion evidenced, the genuine panel datasets can be compiled out of non-longitudinal repeated cross-sectional datasets so far as the location is concerned.

This paper has accommodated discussions only from the methodological standpoint, and whole list of issues to be identified through detailed analytical studies are left aside for the future tasks. The analytical studies based on the GPSed datasets are expected to yield numerous new findings and they, in turn, may throw back issues that occasion the reexamination of the typology of datasets put forward in this paper. It is expected that these interactive process between methodologies and analytical researches might provide an initial steps to formulate a new systematic categorization of statistical datasets.

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: "International Comparative Studies on Archiving System of Official Statistical Data" (#22330070) of Japan Society for the Promotion of Science.

Comparison of Precision of GPS Coordinate Data by Obtaining Measure

Noriaki SAKAMOTO*

Introduction

Due to the remarkable improvement in the quality of the coordinate information together with widespread diffusion of reasonable-price electronic devices, GPS（Global Positioning System）has acquired in recent years wider acceptance among social lives. Disaster prevention is known as one of the fields that provide a wide spectrum of concerned issues which extend from drawing ups of hazard maps to prior and *ex post facto* assessments of the damages caused by various disasters.

U.S. Census Bureau has practiced address canvassing operations in 2009 where field canvassers have collected GPS coordinates of respective residential units by clicking handheld PC devices armed with ArcPad software. Together with i-Phone, digital cameras are now enjoying wider popularity as easy access measures to capture location information. Although such disaster-related researches have mostly relied on digital maps provided by private firms and compact digital cameras loaded with GPS software, information regarding the positioning rules adopted in maps and precision of censing technology given by the camera used for observations seem not to be obvious to the public users.

The aims of this paper are twofold: first, to identify some remarks when one uses digital maps and mobile devices in order to capture coordinate information, and second, to clarify and assess by empirical study the possible observation errors in data obtained through digital camera.

1. Obtaining the latitudinal and longitudinal information

This paper will discuss two measures to obtain latitudinal and longitudinal information: one is to obtain coordinate information through addresses-based conversion and another is to capture them by using GPS terminals. Following paragraphs will portray a quick sketch on the obtaining process and some pending problems regarding these two measures will be raised.

(1) Obtaining latitudinal and longitudinal data through address-based conversion

Direct capturing of GPS coordinates for huge number of dwelling units, business sites and social facilities are quite labor intensive and time consuming operation. In

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

order to obtain tens of thousands of location information through these mobile terminals, however, we need huge number. Relevant data required, for example, for compiling hazard maps are usually converted from existing address files. In order to obtain latitudinal and longitudinal information, web-based matching services are provided by Google Maps[5], Yahoo! Maps[6], CSV address matching services[13] and Portal Cyber Japan[14]. These services are provided free of charge and user can make access to the coordinate information with relatively unsophisticated programming steps. As these digital maps are mostly provided by private firms, contents are sometimes amended due mostly to their own accounts with no particular explanation. It is also probable that providing services are occasionally suspended abruptly. Even in cases when services are provided by the government or by other public authorities, services sometimes are stopped due to the budget cuts and other reasons. Central and local governments are major users of hazard maps. Although stable and reliable positioning of the features is required especially for the hazard maps used for the administrative purposes, the above stated issues may yield difficulties. The second subchapter will address some examples for these difficulties.

(2) Capturing coordinate data by means of GPS terminals

In recent years compact digital cameras furnished by GPS capturing software (hereinafter simply referred to as GPS cameras) are supplied to the market at reasonable prices. Mobile smart phones with varied GPS-based services are now enjoying growing popularity among public. These mobile terminals are now widely applied as convenient tools to capture latitudinal and longitudinal information [2]. PC application software which can display photo snapshots from GPS cameras and smart phones are also on sale.

In case of smart phones, a special user contract should be concluded to make the GPS function available. Reasonable price mobile terminals with an appropriate function are required for public use including analyses of disasters and for other academic purposes.

High-quality apparatuses for professional users, of course, give fairly good estimates with observation error of less than some cm in terms of precision. One of the blocking walls, however, which now disturbs wider use of reasonable price GPS terminal, is that they usually suffer from unavoidable observation errors which is termed “final 10 meters problem”[4].

In case of smart phones, positioning data received from satellites is usually adjusted with the location information given by local telecommunication stations, which is regarded to improve significantly the direct observation results. In case of disastrous incidents like earthquakes and floods when radio waves from

telecommunication stations are unavailable due to the widespread suspension of electric power, observation results by smart phones will lessen the accuracy to the levels almost comparable to GPS cameras.

As described above, due to the "final 10 meters problem" from which electronic tips for positioning GPS carried by GPS cameras and smart phones now suffer, latitudinal and longitudinal coordinates of the real features captured by these tools may more or less differ from those displayed on the digital maps. Thus, among well organized academic papers on disaster analyses there are some which carry results adjusted by particular PC software[1][3]. The third subchapter will discuss the assessment of the observation errors.

2. Obtaining latitudinal and longitudinal data from digital maps

This subchapter will discuss two web-based obtaining of latitudinal and longitudinal data based on address matching: piecewise obtaining and program-based one.

(1) Piecewise obtaining by direct input of addresses on digital map

Web services provided by digital map businesses respond latitudinal and longitudinal data in varied manners. Some softwares return positional data in response to the given addresses, others display the data in URLs which correspond to addresses. There often occurs the different positioning on maps or different latitudinal and longitudinal data are displayed depending on the employed software. Maps in Figure 1 illustrate the searched results by three groups of web services. Google Maps[5], Yahoo! Maps[6], Mapion[7] and goo Maps[8] compose the first group, which are based on maps offered by ZENRIN Co. Ltd.[15]. Chizumaru[9] falls in the second category which is based on maps offered by SHOBUNSYA Co. Ltd.[11]. MapFan[10] provides matching services based on maps from Increment P Corporation[12].

Figure 1 (a) Google Maps (ZENRIN)

Figure 1 (b) Yahoo! Maps (ZENRIN)

Figure 1 (c) Mapion (ZENRIN)

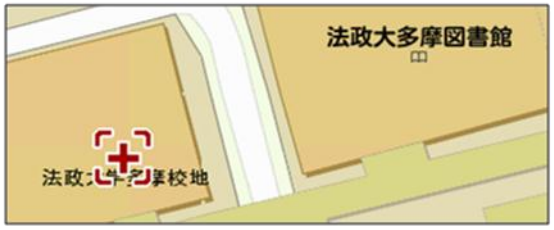


Figure 1 (d) goo Maps (ZENRIN)

Figure 1 (e) CHIZUMARU (Shobunsha)

Figure 1 (f) MapFan (Increment P)

A. Comparison among the same digital maps

Six maps in Figure 1 shows the searched results for the address "4342 Aihara, Machida-shi, Tokyo" where Hosei University Tama Campus is located. As Figure 1(a)-(d) illustrate, so far as the identical digital maps are employed, the same address indicates the identical position on the map even in cases when service providers are different. Reproducibility of pointing on maps is also ascertained in other cases as well. Figure 1(e) and (f) give different positioning from those shown as Figure 1(a)-(d). These facts suggest that possible difference is derived from digital maps on which respective service providers are dependent.

B. Difference in notation of angle and geographical coordinate systems

Table 1 compares the searched results given by web service provider.

Table 1 Obtained latitudinal and longitudinal information by web service provider

No.	Web services	map	Notation of angle	Geo. Co. System	latitude	longitude
1	Google Maps	ZENRIN	degree	World GCS	35.615093	139.297398
2	Yahoo! Map	ZENRIN	degree	World GCS	35.61506244	139.29739242
3	Mapion	ZENRIN	degree	World	35.61182778	139.30058333

				GCS		
4	goo Map	ZENRIN	d,m,s	Not mentioned	35ºN36m42.580s 35.611828	139ºE18m2.100s 139.300583
5	CHIZUMARU	SYOBUNSHA	d,m,s	Tokyo Datum	35 ºN36m41s 35.611389	139ºE18m8s 139.302222
6	MapFan	IncrementP	d,m,s	Not mentioned	35 ºN36m43.5s 35.612083	139ºE17m55.3s 139.298694

referred address: 4342 Aihara Machida-shi Tokyo

As for notation of angle, degree is described by decimal, while degree, minute and second (abbreviated in the table as "d,m,s") by sexagesimal measurement. For reference, although SI units (the International System of Units) given for angle is given by radian, this unit notation is not used in the geographical coordinate system.

Japan now has two geographical coordinate systems: world and Japanese geographical referencing systems. The former is known as international standard of geographical referencing system whose Japanese version is now termed "Japanese Geodetic Datum 2000." As for the transition of the systems from former Japanese system which was called "Tokyo Datum", to the world system see note (1) at the tail of this paper.

As notation of angle and geographical coordinate system are convertible automatically by means of relevant software, we can disregard the difference of the applied geographical referencing systems. The notation of angle and geographical coordinate system are convertible automatically by means of relevant software.

C. Inconsistency of latitudinal and longitudinal data supplied by the identical digital map

As maps in Figure 1 illustrate, the searched address is plotted on maps. Their latitudinal and longitudinal data are also displayed as concomitant information. The first two columns from the right in Table 1 show the obtained coordinate data. As observations 4, 5 and 6 in Table 1 are originally given in degree, minute and second as for the location information, for the sake of reference, the relevant cells for these 3 cases carry degree notation also in the second line.

As observations from 1 through 4 in Table 1 show the searched results based on ZENRIN map, the displayed maps indicate the same spot. However, the obtained latitudinal and longitudinal figures can enjoy coincidence only to the second decimal place. The range of discrepancies in distance between observations 1 and 2 and that between observations 1 and 3 are 3.427 and 463.182 meters, respectively. As is obvious from these observations, the digital map-based searched results of addresses will indicate the same spot. The same spot in maps, however, do not necessarily offer

identical latitudinal and longitudinal figures. As for the significant decimal place, the above example displays maximum distance of 463.182 meters. This topic will be further discussed in the third subchapter.

(2) Program-based obtaining information through Google Geocoding

Users also can obtain latitudinal and longitudinal data by submitting programs compiled of functions provided by web map services. Users, for example, prepare CSV files in advance which give a list of addresses and by feeding them in sequential order they can obtain location information on the map i.e. plots and their latitudinal and longitudinal figures.

As explained above, when users wish to obtain latitudinal and longitudinal data by direct input of addresses on digital map, addresses are processed one by one. The largest advantage of programming-based procedure is that a considerably huge number of addresses can be processed en bloc. It should be noted here, however, that Google Geocoding has a setting of maximum number of coding that can be processed by day. See note (2).

As an example of geocoding, this paper examines Google Geocoding on which most digital camera businesses in Japan are dependent. Table 2 shows the obtained results from two samples of addresses "4342 Aihara Machida-shi, Tokyo" and "1-8-6 Shinkawa, Chuo-ku, Tokyo."

Table 2 Latitudinal and longitudinal data obtained through Google Geocoding

addresses	Obtained date	latitude[degree]	longitude[degree]
4342 Aihara, Machida-shi, Tokyo	July, 2010	35.6152469	139.2953124
	February, 2011	35.6150636	139.2973928
1-8-6 Shinkawa, Chuo-ku, Tokyo	July, 2010	35.6779361	139.7831772
	February, 2011	35.6779361	139.7831772

The observed figures seem to tell two facts. Firstly, as for the first address, the obtained latitudinal and longitudinal figures are different from those in Table 1. This fact suggests that the mixed procedures for obtaining latitudinal and longitudinal information for a set of addresses may produce inconsistency in terms of the captured coordinate data.

Secondly, as latitudinal and longitudinal figures in Table 2 show, the obtained results may differ by date of data capturing operations, although the same tool was employed in the operations. Figure 2 illustrates approximately 190m of discrepancy in terms of distance between the indicated points for the same address by differed date of operation. Although the detailed examination still to be done on the factors which

generate possible discrepancies, it would be worth recording the date and time of capturing data together with coordinate information.

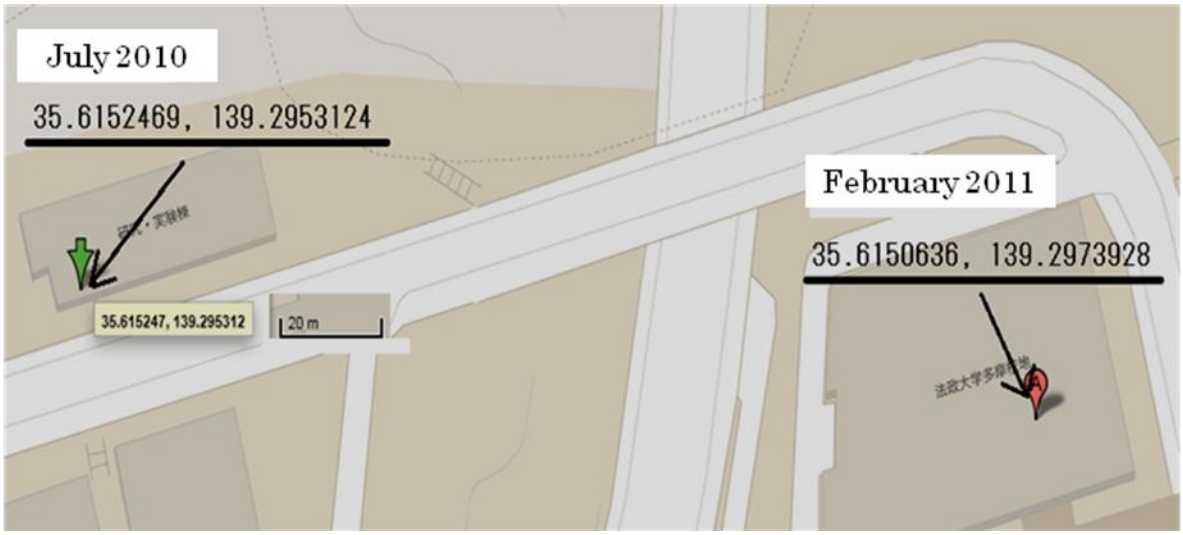


Figure 2　Discrepancy of distance by repeated operations（moved by 189.567m）

3. Possible errors in location information captured through reasonable-price digital compact cameras

Due to the convenience in recording itinerary and keeping memories of spots experienced in the sight-seeing tours, GPS cameras now enjoy expanding market recently. While only two camera manufacturers offered tools loaded with GPS software in March 2010, six businesses (Panasonic, CASIO, SONY, Canon, HOYA(PENTAX), Fuji Film) offer them to market one year later. As stated in 2(2), Together with smartphones, GPS cameras now became one of the handiest and most convenient tools for capturing location information.

This subchapter will discuss possible observation errors and raise some remarks on the use of GPS cameras. The information on the observation operation is as follows:

Employed tool: Panasonic LUMIX DMC-TZ10 released March 2010

Geographical coordinate system: world geographical referencing system (WGS84 same with Google)

Date and time of observation: 13:00 March, 4th, 2011

Weather condition: clear

As referred to above, ordinary electronic geodetic GPS chips suffer from “final 10m problem” in terms of data precision for the time being. User manuals for the GPS

cameras, however, carry only a limited description on the possible observation errors. While those for two brands suggest the hundreds of meters of possible errors, others have no mentions at all. Captured coordinate data are significantly affected by the place and conditions where the cameras took snapshots. It often happens that one cannot capture the relevant information due to the shadowing effect by the neighboring lofty buildings. Figure 3 illustrates observation failure because of the neighboring 5-floor apartment house.

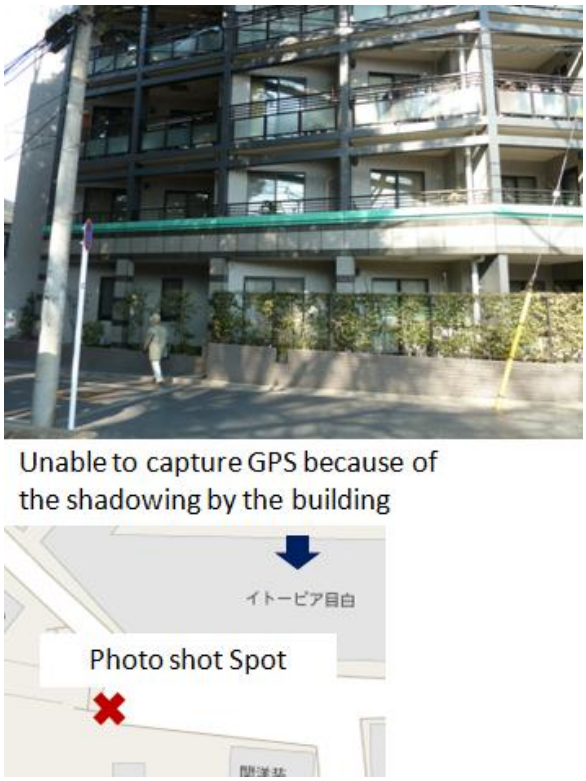


Figure 3 Example of observation failure due to the shadowing by the buildings

At regions with latitudinal condition almost comparable to Japan, differences of 0.0001 in latitudinal and longitudinal degree correspond to about 11 and 9 meters, respectively. These figures suggest that observation errors of about ±10 meters are almost equivalent to the difference of ±1 degree at the forth decimal place.

According to the Japanese Construction Law, every building sites must face the road at least 2 meters in widths (Sub-section 1 of Section 43). In order to distinguish the location of the buildings applicable for drawing hazard maps, allowable observation errors are recommended to be less than 1 meter. Most academic researches including those in the field of science dealing with the prevention of disasters are based on the coordinate figures captured by reasonable price mobile terminals furnished with GPS geodetic chips. Taking the actual size of the buildings into consideration, further examination is to be done to judge if observation results

with 10m of possible errors are actually applicable to identify features.

As for the required number of ciphers at decimal place, as have already discussed in 3(1)C, one unit difference in the third decimal place gives 463.183 meters. If one considers the possible size of error of about 10m, a maximum number of ciphers should be fourth decimal place.

Figure 4 illustrates another coordinate capturing practice at a shrine in Tokyo. In this case, field operator failed to capture figures at façade of the feature, he could rather succeed at sides. The first map gives geodetic codes captured at left side of the façade (35.722139 in latitude, 139.713447 in longitude), while the second displays the results captured at its right side (35.722186 in latitude, 139.713475 in longitude). By inputting coordinate data onto Google Maps, operator could get relevant spots which are marked by arrow on maps. Actual photoshot spots marked in "×" and the pointed head of each arrow are somewhat distances. Displayed maps in Figure 4 seem to suggest two facts. Firstly, the shifting of photoshot spots has accompanied corresponding shifts in position of arrows, although the generated shifts are not the same scale. Secondly, latitude and longitude give different results in terms of discrepancies. Discrepancies in latitude are much larger than those in longitude.

(a) left side of the façade (b) right side of the façade

Figure 4 Results of geodetic survey practices

Figure 5 shows the result of another capturing practice at a multi-floor apartment house also in Tokyo.

Figure 5　Geodetic survey of multi-floor apartment house

The arrowed head point in Figure 6 and figures in Table 4 give a set of portrayed results obtained by inputting the captured coordinates (35.71346 in latitude, 139.73006 in longitude).

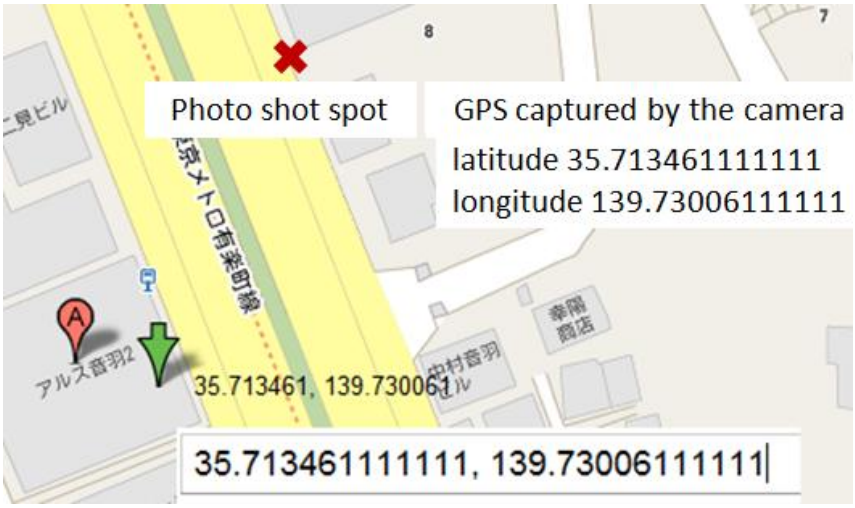


Figure 6　Google mapping of the surveyed results

Table 4　Comparisons of figures by geodetic survey and Google Geocoding

Addresses	Data sources	latitude[degree]	longitude[degree]
1-8-3 Otoba, Bunkyo ward, Tokyo	GPS camera	35.713461	139.73006
	Google Geocoding *	35.714159	139.730285
1-22-18 Otoba, Bunkyo ward, Tokyo	Google Geocoding *	35.713584	139.72995

Google Geocoding *　: coordinate figures obtained through address-based Google Geocoding

The right arrow in Figure 6 shows the result of Google Geocoding. The Google Maps points to another apartment house at opposite side of the road. Due to the difference in latitude, map seems to have chosen other apartment house marked by the arrow (A).

Although GPS cameras and Google Maps adopt the same geographical referencing system, they indicate the location point more or less differently. In order to assess probable size of gaps in positioning, two stages of examinations were practiced.

In the first place, the survey precision of GPS chips in cameras was examined. Japanese latitudinal and longitudinal standard (Figure 7) was chosen as an observation point to assess the errors which may arise in geodetic operation. Table 5 gives standard figures.

Figure 7 Japanese latitudinal and longitudinal standard

Table 5 Coordinate information of Japan latitudinal and longitudinal standard

latitude	35°N39m29.1572s	35.65809922 [degree]
longitude	139°E44m28.8759s	139.74135442 [degree]

Survey operation was repeated ten times at this particular point. The captured figures are listed in Table 6.

Table 6 Captured figures in repeated practices [in degree]

No.	latitude	longitude
1	35.65805278	139.74145556
2	35.65805556	139.74140278
3	35.65806111	139.74146389
4	35.65807222	139.74143333
5	35.65807500	139.7414667

No.	latitude	longitude
6	35.65807778	139.74137500
7	35.65808889	139.74140833
8	35.65810833	139.74133889
9	35.65811667	139.74136667
10	35.65815000	139.74125833

As figures in Table 6 shows, discrepancies are within one unit at the fourth decimal place for latitude as well as longitude. Table 7 carries means, maximum values and standard deviations in meters calculated out of disparities from the true values for 10 respective cases.

Table 7 Observation errors [in meters]

	mean	Maximum value	Standard deviation
Latitude	1.5	5.7	3.4
Longitude	-3.4	9.0	5.5

According to the Table 7, although standard deviations of observation errors are not ignorable, their maximum divergence is shorter than 10 meters. When one considers the size of errors which GPS geodetic chips furnished in reasonable price terminals can possibly generate, the observation result can be judged to be within expected scope.

Concluding remarks

This paper discussed measures to capture latitudinal and longitudinal information widely used in various researches, among others, the performance of GPS cameras and Google Maps in obtaining coordinate data and indication on maps together with a quick review on the type and size of concomitant errors. Followings are the obtained remarks.

(1)Web geocoding by address matching

When one obtains coordinate information through web-based address matching, it is of vital importance to record together with coordinate information the following items: (a) name of the web service, (b) applied digital maps, (c) notations, (d) geographic coordinate system and (e) date and time of capturing data.

(2) Capturing coordinate data through GPS cameras

GPS camera applied for this study is loaded with GPS chip with maximum observation error of ±10 meters which stands ±1 unit at the fourth decimal place in degree. Accuracy of the captured data depends substantially on property of the furnished electronic chips. Significant number of figures for terminals applicable to research purposes is required to be more than 4 at the decimal place. Comparative studies among GPS cameras and other handy terminals and by time and climatic conditions are still to be practiced.

(3) Errors caused by camera-based data capturing and Google Maps

As observations in tables 6 and 7 tell, maximum size of errors is less than 10 meters. Since actual size of the buildings are considered to be large enough, one can well expect that in most cases locations indicated by captured GPS coordinates may correspond to features on digital maps. However, capturing practices in this study identified an interesting fact that captured latitudinal and longitudinal figures are more or less different from those on the digital maps despite the identical coordinate system applied. This issue is also left for future examination.

(Notes)

(1) Geographical coordinate referencing systems

"Standard of the Survey" stipulated in Japanese Surveying Act (Sub-section 2 of Section 11) was amended from Japanese to world geographical coordinate referencing system which was put in forth on 1st April, 2002. Between the two systems there exist some divergences. Latitudes, for example Tokyo and its environs, expressed in Japanese

system will increase its figure by about 12 seconds in the world system, while longitudes by -12 seconds. These discrepancies in angles give approximately 450 meters of divergence to the Northwest direction. The free converting software is already available.

(sources)

http://www.gsi.go.jp/LAW/jgd2000-AboutJGD2000.htm（6 March, 2011）

http://www.gsi.go.jp/LAW/G2000-g2000-h3.htm（6 March, 2011）

(2) Geocoding

Geocoding denotes data processing procedure to convert addresses into geographic coordinates (latitude, longitude). With the assistance of coordinate information, one can plot marker on the map. By means of Google Geocoding API(#), users can make direct access to geocoder via HTTP requests. One can also convert coordinate information to addresses through "reverse geocoding" procedure.

Google Geocoding API provides geocoding services up until 2,500 per day. In case of Google Maps API Premier, users can process 100,000 requests per day. These limitations may be altered without any notices in advance. In addition, a maximum frequency of requests are also stipulated to defend the services from dishonest users. For continuous requests beyond 24 hours and detected illegal users, Geocoding API services are to be suspended temporary. For the repeated violation of the limits set for request, further access to the Geocoding API are blocked.

(#)Geocoding API

Geocoding API are only applicable in combination with Google Maps. Use of Geocoding without displaying on the map is not allowed.

For further information on the use of Geocoding API：http://code.google.com/ intl/ja/ apis/maps/documentation/ Geocoding/#Geocoding

(source)Google Maps API Web site

http://code.google.com/intl/ja/apis/maps/documentation/javascript/v2/services.html#Geocoding （6 March, 2011）

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.

References

[1] Takashi FURUTO, Naofumi SASAKI, Hiroyuki YAMADA, Shigeru KAKUMOTO :「Study of Advanced Collection and Registration of Disaster Information –Municipality Support Using Spatial Temporal Information System in Chuetsu Earthquake(7)-」, Papers and proceedings of the Geographic Information Systems Association, Vol.14, pp.157-160, 2005 (in Japanese)

[2] Katsushige HIRATA and Satoshi KATO :「About the new field work technique that used a digital terminal」, Journal of Applied Survey Technology, Vol.18, pp.101-107, 2007 (in Japanese)
[3] Takashi FURUTA, Mitsuaki SASAKI, Mahito USHI, Kaoru FUKUYAMA and Shigeru KAKUMOTO : 「Collection and Classification of Disaster Information - Spatial Temporal Information handling for Risk Management(6) - 」, Papers and proceedings of the Geographic Information Systems Association, Vol.17, pp173-176, 2008 (in Japanese)
[4] Special feature「Kimetewa Itijyoho (The decisive factor is positional information)」, NIKKEI ELECTRONICS, March 7 issue, pp.43-71, 2011 (in Japanese)
[5] Google Maps：http://maps.google.co.jp/
[6] Yahoo! Maps：http://map.yahoo.co.jp/
[7] Mapion：http://www.mapion.co.jp/
[8] gooMaps：http://map.goo.ne.jp/
[9] Chizumaru：http://www.chizumaru.com/
[10] Map Fan Web：http://www.mapfan.com/
[11] Shobunsha Publications, Inc.：http://www.mapple.co.jp/
[12] INCREMENT P CORPORATION：http://www.incrementp.co.jp/
[13] CSV address matching service（Center for Spatial Information Science, The University of Tokyo）: http://newspat.csis.u-tokyo.ac.jp/geocode/
[14] Denshi Kokudo Portal：http://portal.cyberjapan.jp/index.html
[15] ZENRIN：http://www.zenrin.co.jp/
※ We confirm all URLs on March 10, 2011.

Geographical Information System and Spatial Micro Data: An Introductory Socio-Technological Perspective

Akio KONDO*

1. Introduction

The geographic information indicates information with the relation between the position on the earth and what locates. Because the position on the earth is generally described in latitude and longitude, the geographic information might have a spatial characteristic of extending, and be called spatial data. The one that the geographic information was expressed visible is a map. The history about the map is old, and making concerning the geographic information of Japan has advanced in the Edo era when a real measurement was begun. The method of making a map has developed as cartography for a long period of time. Geographical maps are generally divided into two types; general cartography and thematic cartography. Many of certain maps are called the thematic map, and made a map for a specific theme or subject. For instance, various geographic information concerning regions and countries have been described in the atlas used by many educational institutions. It is included in a map that the description concerning the natural environment such as geographical features and climates, or the social characteristics such as population, trade, industry, and so on.

The theme of such a thematic map is called an attribute, and the spatial data related with these themes is called an attribute data. Of course, various attribute data in a map are visualized and they are also rank-ordered. It becomes possible to treat the mass data easily by the development of information technology and computing technology though it is difficult to individually collect the attribute data not so long ago because it included many things. Digitalization of mapping technology has advanced, and it becomes easy to collect, to manage, and to analyze it by the individual computer as for collecting attribute data or mapping. Therefore, in these contexts, geographic information system (GIS) can effectively aggregate many attribute data and integrate spatial micro data and variant layers of map[1].

2. Social concerns of special dimensions: A history of mapping

Geographical information system is mainly based on and was originally born from geography. Geography has a long history. The original meaning of geography is to describe (graph) land (geo). Land (geo) is a stage where we live, and includes not only natural environment such as geographical features, climate, and ocean but also the

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

social features of economy and culture. It can be said that geo area and place are terms that indicate the extension of the concrete space and show a place where city, village, farm, and other districts are partly located. In other words, it is included in the definition of geography that locality or territoriality represents concrete space, which the technical term as space is high the abstraction level, and, therefore, understood by various meanings. The space is an aspect where we see things while extension concrete land (geo) is directed, and it can be said the concept of related to the ideal way of our recognition for spatial dimensions.

Spatial dimensions are a way in which we recognize our surrounding, and another way is time-series recognition. It becomes easy to understand the feature and the attribute of space from comparing it with the feature of time-series though these both recognition processes are indivisible for something like the both sides of the coin. Time-series recognition is one of the most fundamental ways when we acknowledge the world, and it may say that the recognition method enables us to think the relation of one thing to another as a causal relation. Moreover, a certain kind of universality is sensitive in the flow of time because all lives of mankind and artificial materials are limited and in another way they have expiration date. Our historical drama sometimes shaken our mind because all things are in flux and nothing is permanent, it is believed that lives are destined to die out eventually. Therefore, time-series recognition processes represent a kind of universality. On the other hand, our understanding of spatial dimensions is more individual or specific in general compared with universality of time-series recognition. This is deeply related with the thing that we locally recognize our surroundings with bounded rationality the famous social scientist Herbert Simon conceptualized. For instance, there is a map in what described by the aspect of the space. When we trace the history of map back, we know the first world atlas was a world atlas of Babylonia in around the seventh century B.C. and followed by Ptolemy's maps around the second century and famous TO map around tenth century. It was characteristics for the early-time mapping that the central area was concrete but the peripheral area was not. These maps depended on the editor's recognition of their surrounding until Age of Geographical Discovery came, and the experiences of many parts of the world came to be collected, and a description different in each place was performed. As remarked, it can be said that the space has an individual and specific feature compared with time-series.

Philosopher Immanuel Kant expected that geography had a central role to collect information around world and then to unite its knowledge[2]. In the 18th century when Kant played active roles, there were a lot of world maps based on the accurate measurement, and the position of geo was expressed numerically by the graticule that centered on two poles and the equator. It connects with catching the aspect of space with generality or more advance by accurate mapping. Development of geographical information system and geoscience has been alongside the history of geographic

measurements in the Kant's philosophy how to generalize spatial features.

Geographic information system is a platform system designed to capture, store, manipulate, analyze, manage, and present all types of spatial data referred geographically. In the simplest terms, GIS is the complex of cartography, statistical analysis, and database technology. Therefore, Figure 1 shows GIS is developed through various academic disciplines and fields such as cartography, surveying, informatics, and computer science, which also expands in application to agriculture, environmental science, engineering, applied economics, and business management.

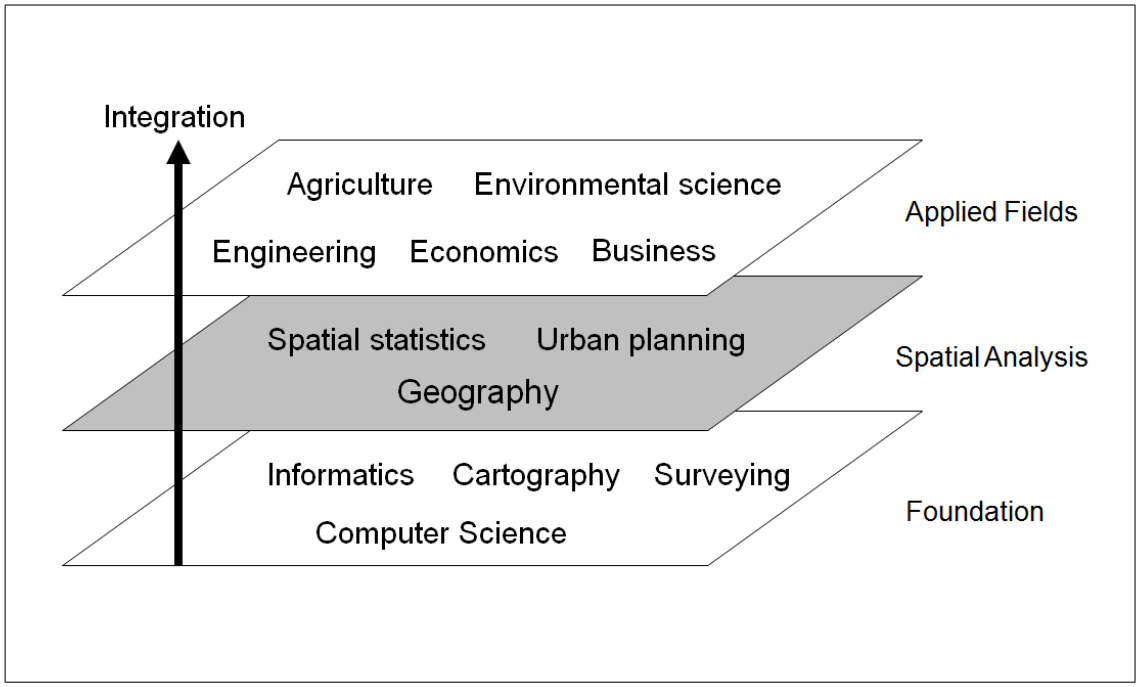


Figure 1. Geographical Information System based on academic fields

This platform system excels in integrity and functionality[3][4]. Therefore, in a general sense, it enables to integrate, store, edit, analyze, share and display geographic information for informing decision making. Also, GIS can be thought of as a system—it digitally creates and manipulates spatial areas that may be jurisdictional, purpose or application-oriented for which a specific GIS is developed. For example, the application of the Internet to GIS has been advanced, and it has a lot of web-GIS sites and clearing houses[5][6]. Thus, it can be said that the geographic information system is a strong tool to enable individual, concrete geographic information to be treated totally in the background of digitalization based on computer technology and informatics. It is important to point out that the development of geographical information system/science is consistent with the social infrastructure in using spatial micro data.

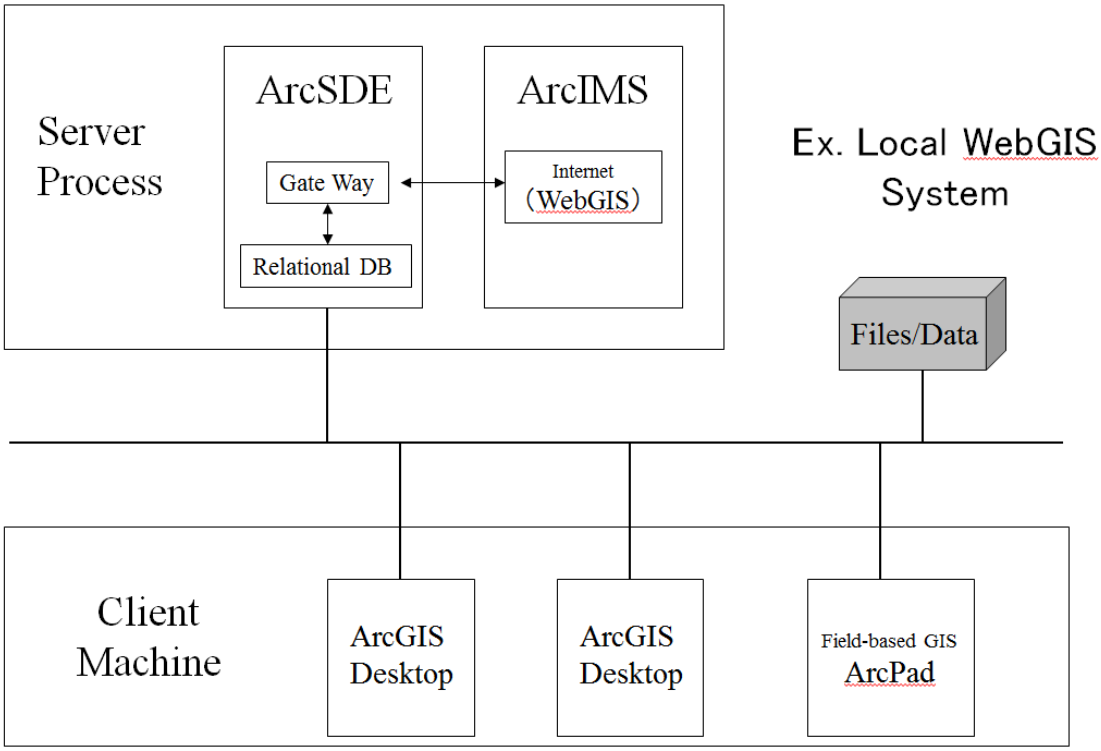


Figure 2. Local web-based GIS: An example

3. Geotechnology matters: technological development of GIS

Along with nanotechnology and biotechnology, geotechnology is one of the top three most important emerging fields according to the U.S. Department of Labor. In recent years, the importance of Geographical Information System in socioeconomic development has become widely recognized. Today's GIS research devotes attention not only to research and development in universities but also to the importance of constructing a national platform system of GIS that effectively connects comprehensive systems incorporating market and social needs, public sectors and private sectors. At the same time, however, various issues have arisen, such as the latent market and social needs, and the emergence of more sophisticated and complex technology.

The history of R & D in geotechnology is not that short. Its inception dates from 1960s. Subsequently, during the 1980's, some geographers expanded the classic method. In this way, the concept of geotechnology has been around for a long time, but it was not until year 2000 that the recent geoscience boom came about. Geotechnology started to draw attention when Bill Clinton, the then President of the United States, launched the "21st century Science and Technology Initiative." The United States regarded geotechnology as a technology that would trigger the next industrial revolution and decided to promote intensive investment in it.

Given these turns of events, geotechnology and related fields was given high priority in the R & D policy of the U.S. Federal Government, and it has invested more than $1 billion during the past ten years. Turning to the European Union, they thus followed the U.S. strategy and geotechnology became a major theme in each of the EU countries. Since R & D on mapping technology have progressed in parallel with the development of information technology in Japan, expectations for geotechnology were initially high in this country.

Recent times have seen the links between mapping and spatial data become less pronounced. According to recent studies of geoscience, research can be categorized into three types. These are pure basic research (mapping technology) by computer science and informatics, pure applied research by engineering and needs-inspired pure basic research by applied fields. The latter type research is recently gaining in importance, because the motives for mapping have become increasingly needs-oriented in recent years. In other words, the most advanced mapping technology cannot be realized without technology-based knowledge, and consequently, this fact affects research in universities and public research institutes. Even national platform for geographical information cannot be performed without considering social needs.

The broad aim of geotechnology is to identify geographically referred data likely to yield the greatest natural, economic, and social benefits. During the 1990s, geotechnology became much more widespread. Since 2000, most advanced geotechnologies including global positioning system have been becoming more sophisticated and complex day by day, and as such, the time and efforts inherent in the realization of geotechnologies have also increased. Many institutions and universities which had previously conducted R&D internally are now partly outsourcing their R&D tasks and the associated efforts, except for their applied field. Development of geotechnologies has reduced the distance between mapping and analyzing spatial data. In universities and national research institutes, for example, those public R&D strategies for geotechnology which were previously considered as precompetitive research have been affected by social needs and public policies. In such a circumstance, a wide range of geotechnological methods are available, some are specifically designed for future work while others are developed from management and planning. Some may not be specifically related to the social needs but are used to provide the basis for useful information sources. Some methods, like spatial multivariate analysis, which were developed by quantitate geographers, have since been borrowed by others. From the range available it is important that the chosen methods are selected as suitable for the purpose for which they are to be used. Exploring possible, probable and preferable spatial futures relies on assumptions about the situation and how we relate to it, which in turn will influence the choice of methods. However, mapping technology is the most common and available processes. Layer setting and collecting spatial data collecting are more like qualitative methods rather than quantitative.

4. Spatial data and mapping technology

It is a need for GIS to use two types of data; GIS data (an abstraction of real object) and spatial data (statistical data on society and economy). GIS data represents real objects such as roads, land use, trees, rivers, mountains and so on with digitalization. Generally, there are two methods used to store data in GIS for both kinds of abstractions mapping references: raster images and vector. A raster data type is any type of digital image represented by reducible and enlargeable grids. It is familiar with digital photographic technology where a picture is consist of the raster graphics pixels as the smallest individual grid unit building block of an image. On the other hand, vector type data are expressed by considering those features as geometrical shapes; points, lines or polylines, and polygons. Spatial data contains geographically referred data. Also, geographically referred data is expressed as an attribute matrix. As for spatial units such as area or regions, geographical matrix is a combination of attribute matrix and interactive matrix which Brian berry designed (Figure 2).

Digital mapping technology is to visualize spatial features to be abstracted along with a spatial axis. The origin of digital mapping is derived from a public tool in governmental institutions. The development of digital mapping in public sector two important aspects in technology foresight. One is that we can apparently confirm the increase of its importance as spatial mapping system and social future vision. We can see the development of digital mapping from a public or private institution's method to an integrated system on a national level. The other is that more and more specific knowledge is needed to build a digital mapping. This originates from the advancement of technologies and diversification of social needs. The digital mapping needs to be considered as a collection of spatial data including various elements of society.

There seem to be three major reasons why digital mapping system is needed nowadays by both governments and private sectors. First, because of the rapid increase of social and economic complexity and diversification, it becomes necessary to grasp total conditions strategically based on geotechnology. In other words, more efficient and strategic management is needed to understand the global or local situations. Moreover, because of the increase of collecting digital data by high-tech tools, more selective and concentrated information system is needed. Therefore, digital mapping system increases in importance as a part of GIS.

Secondly, due to the recent digital information overloaded, it becomes more difficult to select or search useful sources individually. This results in the necessity of strategic selection and concentration of spatial data and more focusing on spatial scales, especially in socio-economic phenomenon where the rapid changes is seen. This is the reason why digital mapping and geographical information system has become more important.

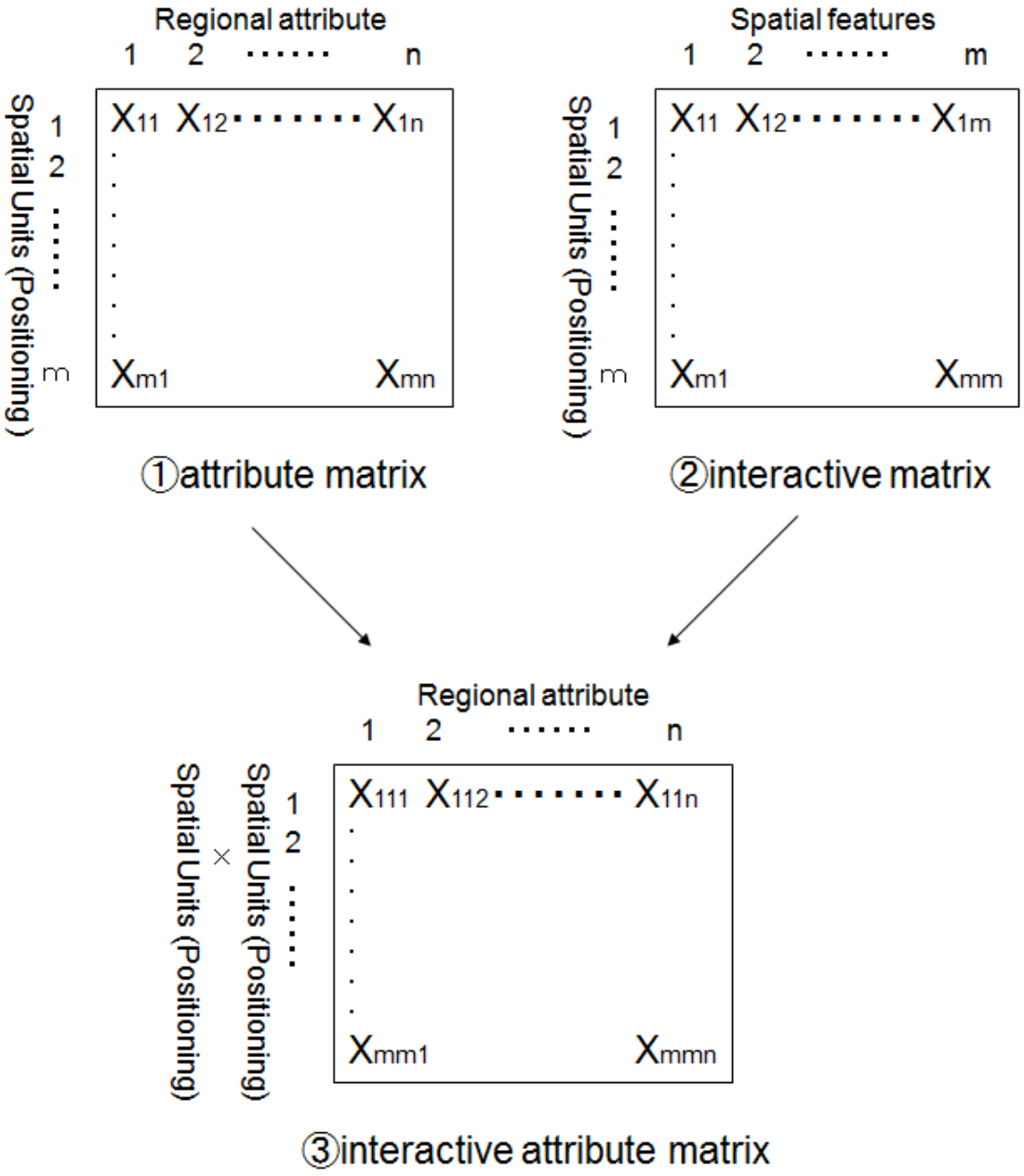


Figure 3. Structure of Geographical matrix

Geographically mapping is based on specific technological knowledge. As technological complexity increases, the search range of technology seeds becomes wider, which means it is necessary to gather not only open information sources such as academic or public research institution, but also closed information such as know-how and implicit knowledge in the related fields. There would be, however, still some amount of difficulties to construct a huge GIS platform. According to the factors of the

digital mapping above, the formula, shapes and components deeply depend on the purposes of mapping. National governments, consortia of academic institutions could be the builders of a GIS platform, and they have various GIS platform (see Figure 4). If a government or public community build a GIS platform, they would have to consider national conditions and reasonableness of public services, and eventually it should be a long-term digital mapping rather than short-term[7][8].

Figure 4. Public GIS platform: An example
Sources: http://portal.cyberjapan.jp/index.html

5. Concluding Remarks

It is now required for governmental and private contributions in geotechnology to see whether expectations for the geoscience field create matching social value and needs upon spatial integration. Sufficiently social effective results are now demanded from GIS platform. The field of geotechnology has been in the limelight in recent years, because the possibility has emerged that it will bring about not only innovations in geography and engineering but also radically remake ordinary information tools. A pioneering example is fullerene, which is used widely in fields from car navigation system to mobile GPS tools. As far as the future potential of geotechnology in

industrial technology is concerned, geotechnology has created high expectations as a technological seed that brings about explosive innovation. However, huge investments and strategic inputs of resources are required to realize it.

Taking the field of geographical information system as an example, this paper points out the importance of a viewpoint that takes a series of processes from mapping to social platform as a single information system. Positioned at the center of this system is Geographical Information System. As a tool that helps develop a combination among spatial features such as nature, environment, society, economy, and so on, and prioritized allocation of efforts and human resources, a geotechnology is also being developed for GIS also in the USA and EU countries. Although a GIS brings about a comprehensive vision, however, there are some problems with technology characterized by quantitative methods such as collecting spatial data. Research and Development of GIS in the amounts of cost is required to reach integration of spatial micro data in the field of geotechnology. The path to integration is extremely unclear, and, furthermore, technological establishment does not necessarily match social needs. Under these circumstances, a geographical information system that functions sufficiently for developing a collection of spatial micro data, be it for the government or a private sector. In this sense, one can also say that the field of geotechnology is an experimental field for various purposes and extremely cutting-edge cases.

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.

References

[1] Yoshio SUGIURA eds.: *Geographical Spatial Analysis*. Asakura-Shoten. 2003. (in Japanese)

[2] Atsuyuki OKABE: *A Challenge for Spatial Information Science*. Iwanami-Shoten. 2001. (in Japanese)

[3] Hiroyuki KOHSAKA: *A Handbook of Geographical Information Technology*. Asakura-shoten. 2002. (in Japanese)

[4] Yuji MURAYAMA and Ryosuke SHIBASAKI eds.: *A Theory for Geographical Information System*. Asakura-shoten. 2008. (in Japanese)

[5] Akio KONDO, Yoshinori NAGANO and Koichi TAKANO: *Building Geographical Information Archive System and Its Advanced Application*. Nihon University Research Institute eds. *Constructing Digital Archive System*. Nihon Univ. Press. 2005. (in Japanese)

[6] Akio KONDO, Yoshinori NAGANO and Hideo OYAGI: Geographical Information System and Spatial Data Analysis: Beyond Digital Archiving, Nihon University Research Institute eds. *Utilization of Digital Archive System and An Agenda*. Nihon Univ. Press. 2006. (in Japanese)

[7] Ryosuke SHIBASAKI and Yuji MURAYAMA eds.: *A Technology for Geographical Information System*. Asakura-shoten. 2009. (in Japanese)

[8] Ryosuke SHIBASAKI and Yuji MURAYAMA eds.: *Building Geographical Information System for Social Platform and Environmental Issues* Asakura-shoten. 2009. (in Japanese)

Part two

Micro-based integration of statistical data

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.

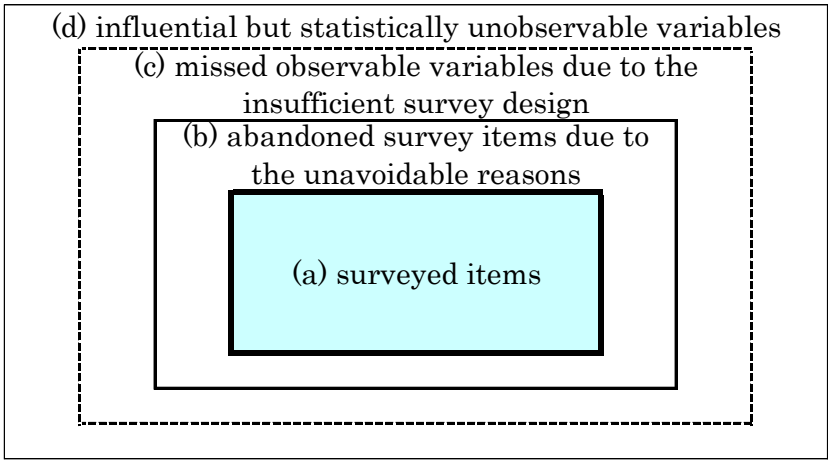


Diagram 2 surveyed variables and other variables that account for the surveyed units

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.

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.

Possible Expansion of Individual Statistical Records by Loading with Derived Variables

Hiromi MORI*

Summary

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

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

Introduction

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

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

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

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

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

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

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

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

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

2. Dissemination channels of survey results

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

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

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

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

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

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

materials.

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

4. Exploring information potential through endogenous expansion of variables

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

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

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

Table 1　List of derived variables adopted for the 2001 SARs

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

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

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

5. Endogenous expansion of questionnaire return information

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

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

(1) Personal individual statistical records

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

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

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

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

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

(2) Household individual statistical records

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

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

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

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

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

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

(3) Business individual statistical records—establishments and enterprises

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

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

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

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

(4) Individual statistical records on residential units

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

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

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

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

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

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

6. Analytical implications of expanding dimension of individual statistical records

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

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

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

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

Concluding remarks

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

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

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

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

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

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

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

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

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

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

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

Reference

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

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

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

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

研究所報(最近刊行分)

号数	タイトル	刊行年月日
17	地方統計	1990.11.30
18	厚生統計	1992.03.31
19	人口移動統計	1993.03.31
20	わが国における外国人労働者	1994.01.31
21	統計調査環境の変容と現状:1994 年	1995.07.31
22	サービス業統計の現状と課題	1996.02.29
23	民間統計	1997.01.31
24	統計環境実態調査	1998.01.31
25	ミクロ統計データの現状と展望	1999.01.31
26	The2000-01 World Population Census and the Related Issues	2000.01.31
27	統計と人権および開発－IAOS 2000 をめぐって	2001.03.15
28	第 4 回日本・中国経済統計学国際会議	2002.03.15
29	職安求職者にみる失業の実態	2002.12.20
30	国連ミレニアム開発目標と統計	2003.10.20
31	Workshops on "the Population Censuses" and "the Use of Census Micro Data"	2003.12.20
32	ミクロデータとその利用	2004.04.20
33	International Symposia on Population Census and Micro Data Archives	2005.01.10
34	政府統計の二次的利用	2005.04.20
35	ジェンダー（男女共同参画）統計	2007.02.20
36	人口センサスの現状と新展開	2007.04.01
37	統計における官学連携	2007.04.20
38	ジェンダー（男女共同参画）統計　Ⅱ	2009.02.10
39	社会生活基本調査とその利用	2010.01.15
40	地方統計の現状と課題	2011.09.15

研　究　所　報　No.41
2011 年 11 月 5 日

発行所　法政大学　日本統計研究所
〒194-0298 東京都町田市相原 4342
Tel 042-783-2325,6
Fax 042-783-2332
jsri@adm.hosei.ac.jp
発行人　森　博美

B U L L E T I N

OF

JAPAN STATISTICS RESEARCH INSTITUTE

No.41 November 2011

Exploring Potential of Individual Statistical Records

CONTENTS

Edited by
JAPAN STATISTICS RESEARCH INSTITUTE
HOSEI UNIVERSITY
TOKYO, JAPAN