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.

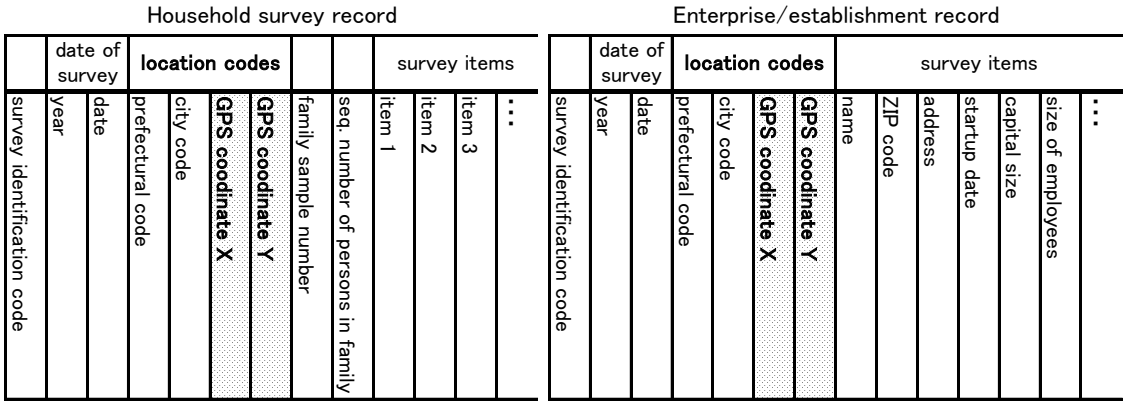

Figure 1. Examples of GPSed records

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.