統計個票情報の情報特性について

MORI, Hiromi / 森, 博美

(出版者 / Publisher)
法政大学経済学部学会
(雑誌名 / Journal or Publication Title)
経済志林 / The Hosei University Economic Review
(巻 / Volume)
76
(号 / Number)
4
(開始ページ / Start Page)
403
(終了ページ / End Page)
427
(発行年 / Year)
2009-03-09
(URL)
https://doi.org/10.15002/00004033

【研究ノート】

統計個票情報の情報特性について

森　　博美

はじめに

統計あるいは統計調査に関する次のような素朴な疑問が，本稿執筆の直接の動機となっている。その一は，統計が集団を集計量として記述したものであるとする伝統的統計観に対する疑問である。統計＝集計量とする統計観は，多数事例を集計することによって調査等で収集される個体に関する原情報（以下，統計原単位情報）に内在する撹乱要因の作用を相殺し，集団としての安定的な計数を得ることができるとの認識をその存立依拠としている。これまで統計調査の結果が基本的に集計表の形で提供され，利用されてきたのは，主としてこのような事情からである。統計原単位情報を集団として集計量の中に埋没させることではじめて統計という認識資料の価値を見出すのではなく，一つひとつの統計原単位情報それ自体に何らかの情報的価値があり，それから出発することで，これまでとは異なる新たな統計像を描けるのではないかというのが，筆者の疑問の根底にある。

いま一つの疑問は，統計の調査結果の意味づけに関するものである。統計調査は，調査時点における瞬間撮影によく例えられる。仮にそうであるとするなら，調査結果としての統計は，時空間の運動体をある特定の時点で切り取った断層静止画像（snapshot）を数値的に表現したものというこ

とになる。統計とは，本当にこのような静止画像的なものなのであろうか。この機会にこの点についても検証してみたい。

統計調査の実査過程は，調査の対象となる個人，世帯，企業それに事業所といった個体に帰属しあるいは個体に関係した情報を統計原単位情報として調査個票に写し取る作業工程である。そこで得られた質的カテゴリーのコード化情報あるいは量的データは，数値から構成される一連のレコードを構成する。

本稿では，上のような問題意識から，デジタル画像，音声情報，さらにはメッシュやGISといった空間を記述する各種情報と統計個票情報とを比較することで，データ構造から見た統計個票情報の情報特性を明らかにしてみたい。

1．デジタル画像・音声情報とそのデータ構造

（1）デジタル画像情報のデータ構造

（a）静止平面画像

デジタル静止平面画像情報は，画素（picture cell：ピクセル）と呼ばれる個々の微小面の位置を示す座標情報と各画素にリンクされた色調あるは濃淡レベルを示すトーン情報とから構成される。そこでは，画面を構成する画素数が多いほど，すなわち個々の画素単位が微小であるほど解像度の高い画像が得られる。またトーン情報[1)]は，各画素が持つ色調や濃淡のレベルを決定するもので，レベルの区分数が大きいほどグラデーションの変化をより滑らかな移行として反映することができる。

単色画像の各画素のトーン情報は1次元のデータによって与えられ，またカラー画像の場合には，各画素について，R（赤），G（緑），B（青）の３原色のそれぞれがトーンレベルを持つ３次元情報として与えられる。

（b）動画

動画は，短時間に多数の静止画像を連続的に表示することで，残像現象という視覚的錯覚[2]により動態現象のように見せているに過ぎない。従って，このような動画をデジタルデータとして表現するには，各画素に対して，１次元（単色）あるいは３次元（カラー）のトーン情報に，それぞれ時間情報をあらわす１次元のデータを付加することで作られたデジタルデータセットが動画情報を与える。そこでは，時間情報を指定することにより特定時点での静止画像が得られる。

（ｃ）立体画像

立体画像は，３次元の座標情報を持つボクセル（volume cell: voxcel）と呼ばれる微小立体が画素に相当する立体画像の単位となる。ちなみにボクセルによる画像表現の例としては，医学等で使用されるX線CT（Computed Tomography）スキャン画像[3]や核磁気共鳴MRI（Magnetic Resonance Imaging）画像がある。なお，これらの画像におけるトーン情報は静止平面画像の場合と同じく単色画像については１次元，カラー画像は3次元のデータ形式を持つ。

このように，画像データは，ピクセルやボクセルといった画面上の位置情報によって識別される画素を単位とし，それに対するリレーショナル・データとしてトーン情報がリンクされるというデータ構造を持つ。言い換えれば，画素がトーン情報のいわば担い手として機能しているのである。そこで本稿では，被写体の濃度・色彩特性をトーンとして与える情報要素をdata body，またその担い手である画素のような単位情報をdata carrierと呼ぶことにする。ちなみに，単色とカラー静止画像のdata bodyはそれぞれ１次元，３次元で与えられ，動画については，時間要素も含め，それぞれ２次元と４次元の数値データの組が各画素に対応づけられる。なお，ここでのdata carrierである画素は，画像フレーム中の位置情報を示すものに他ならず，それが写す被写体そのものの位置情報を与えるものではない。このように，画像情報の場合，画素というdata carrierの単位情報は，認識の

対象である被写体とではなく，撮影（認識）主体側の装置と関係づけられており，data bodyを構成する各変数値は，あるアングルで被写体を切り取った際のフレーム中の各画素が与える位置情報に照応した被写体の色彩トーン情報を数値化したものである。

（2）デジタル音声のデータ構造

音声は，声帯や楽器のような振動物が気圧の変化を作り出し，それが直接あるいは電波信号等によって伝えられ，スピーカー等の振動物によって再生されたものが聞き手によって認識されるものである。電波信号も含め，従来は音源での音の発生から受信にいたる音声伝達の全過程がアナログによるものであった。デジタル化された音声データとは，途中の伝達過程をデジタル情報によって代替したものである。最初からデジタル情報として音声が採取されるケースもあるが，一般にはアナログ情報信号として発生された音声が，デジタイザーといわれる機器によってデジタル変換される。このようにして数値化された音声データは，スピーカー等の振動装置を内蔵した機器によって大気の振動というアナログ情報へと逆変換されることで，聞き手側で音声として認識される。

音声は強弱，高さ，それに音色という３種類の要素を持つ。それらは時とともに振幅，波動サイクル，そして波形を変えるアナログ波（アナログ音声）という形で本来的には表現されるものである。

音声の三要素のうちまず音の強弱（音量）については，聞き手側の感知面の空気を押し引きする圧力の強弱として知覚される。それを波動として表現した場合，それは波の振幅の大きさに相当し，より振幅の大きい波動ほどより大きな音声として認識される。このような音の強弱を示す振幅情報は，デジタル音声情報では常態ゼロを中心とする正負のデータの絶対値の大きさによって表現される。また，音声の二つ目の要素である音の高さは，周波数という波動の循環単位であるサイクルによって決まる。単位時間内での波動が描くサイクルの頻度が多い波ほど高い音として認識され

る。ところで，仮に同じ大きさと高さを持つ音声であっても，人あるいは楽器によってその音色は異なる。音声の第三の要素である音色の違いは，波動の減衰パターン，すなわち音声波動の波形の違いに起因する。

このように，音声は三種類の異なる要素を持つが，アナログ音声情報のデジタル情報への変換は，原理的には音声を拾う頻度情報と音声の振幅の大きさを記述する情報という二種類のデータだけでそれを行うことができる。言い換えれば，音声を構成する三要素は，時間単位と音量単位という二種類のデジタル情報に還元される。

このうち時間単位に関する情報は，微小時間単位としてのサンプリングによって与えられる。単位時間（例えば１秒間）における音量情報採取の回数を多くすることで，よりきめ細かな時間単位で音源の音声を拾うことができる。一方，音量情報は反映すべき音量の全範囲をいくつのレベルに段階区分するかというもので，その区分数の多寡はビット数によって決まる[4)]。すなわち，ビット数が大きいほど個々のレベルの刻みを細かくすることができ，音量についての音飛び，いわゆる量子化誤差をより小さくすることができる。従って，サンプリングの頻度と音量区分のビット数を大きくすることで，音源のアナログ波形により近い形で音声のデジタル変換を実現することができる。

サンプリング頻度と音量のビット数を十分大きくすることで，音の高低に関わる振動数（サイクル），振幅，さらには波形も含めて，アナログ音源が持つ音声により近いデジタル音声データを得ることができる。デジタル音声情報がアナログ音源の持つ波動にどの程度接近しうるかは，結局，サンプリングと振幅を表現するビット数，すなわち情報量に依存する。スピーカー等の音声再生装置の精度さらには聞き手が持つ識別能力を超えるデジタル音声情報は，仮に情報技術的に許容されたとしても，実質的には一種の過剰品質となる。

ここで，デジタル音声情報をデータ構造の観点から整理すると，それはサンプリングという時間単位をdata carrierとし，それに関係づけられたリ

レーショナルな１次元の音量情報をdata bodyとしていることがわかる。なお，音源毎にマルチトラック録音されたデジタル音声データのdata bodyは多次元データを維持しており，音源毎に独自の編集や操作が可能である。しかし，それらをひとたびmixingしてしまえば，その音声情報は１次元の，またステレオ用に２トラックに編集された２次元情報として与えられる。このように，デジタル音声情報については，そのdata bodyがそれぞれ独自の変数要素として意味を持つものとmixingにより統合値（集計量）として意味を持つものとがある。

２．空間情報のデータ構造

（１）地域メッシュ統計

メッシュ統計とは，地域をメッシュといわれるグリッド（方形）に区分し，個々の方形内に所在する個体に関する統計情報を，総計や比率といった統合値（集計量）として表示する表章方式である。その表章の基盤となるのが，表１に示したような一連の区画体系を有する地域メッシュである。

表１中の第三次地域区画（基準地域メッシュ）からは，さらにその一辺を２等分，４等分，そして８等分した約500m四方の1/2地域メッシュ，同じく250ｍ四方の1/4地域メッシュ，それに1/8地域メッシュ（同約125m）が作成される[5)]。なお，これら各レベルの地域メッシュにはそれぞれ１桁のコードが付与されており，例えば最小単位である1/8地域メッシュは，全

表１　地域メッシュの区画体系と基準地域メッシュ

区画の呼称	１辺の長さ	区画数	区画コード	備　考
第一次地域区画	約80km	176	４桁	
第二次地域区画	約10km	4,885	４＋２桁	第一次区画の１辺を８等分
第三次地域区画（基準地域メッシュ）	約１km	386,877	４＋２＋２桁	第二次区画の１辺を10等分

体で11桁からなるメッシュ・コードを持つ。

地域メッシュ統計は，電子地図，メッシュ枠情報，国勢調査区及び最小の地域集計単位である基本単位区の境界情報，それに事業所等の所在情報，さらには基本単位区内の個体レコード情報を用いて作成される。その作成は，事業所等の所在地又は基本単位区の位置を電子地図上に反映し，各個体レコードにそれらが所在するメッシュ・コードを対応させるという方法で行われる。なお，同定と呼ばれる地域メッシュへの対応づけには，個体データを地域メッシュに直接対応づける個別同定と調査区等の集計区域単位で対応づける調査区（基本単位区）同定とがある。

各メッシュに対応づけられた個体レコード情報の集合は，メッシュ・コード別に集計され，提供される。なお，一つの地域メッシュに表章される人口や世帯数が著しく少ない場合，統計上の秘密保護のために，人口総数など限られた集計結果だけが表章され，それ以外の諸項目については秘匿あるいは隣接するメッシュと統合して表示される。

ところで，地域メッシュは一次統計の集計処理結果の表章形態としてだけでなく，緑地被覆度マップのような環境あるいはハザードマップのような防災分野などでも広く活用されている。これらは，単独のあるいは複数の調査結果，さらには行政情報をメッシュ・コードを介して相互にリンクし，指標化した結果を表示したものである。

地域メッシュ統計の場合，各メッシュが画像情報における画素に相当するdata carrierとして機能し，各メッシュに対応するメッシュ・コードをキーにdata bodyを構成する種々の統計情報が相互に関連づけられる。

（2）統計GISのデータ構造

（ｉ）GISのデータ構造

GISでは様々な情報が測地系という位置情報をキーに相互にリレーショナルなデータとして管理されていることから，それらを電子地図上でデータを検索，加工，編集し，結果を表示することができる。電子地図では，

自然的・人工的オブジェクトに関する情報が，幾何情報と属性情報とに分けてデータ化されている。

このうちまず幾何情報は，河川や海岸線といった自然的オブジェクト，道路や建造物といった人工的オブジェクトそれに行政区の境界等の形状を点や線として表示するデータである。これらに関するデータは，二次元平面をグリッド状の格子に分割した画素（ピクセル）の表現に適したラスタ型データと座標情報を用いて幾何学的にオブジェクトを表示するベクタ型データという二種類の形式のデータによって管理されている。

ラスタ型データには各画素に色情報が格納されていることから，この形式で維持管理されている情報は，土地利用や人口密度，地価などの社会経済データの分布の表現に適している。しかし，このタイプのデータは，情報を画素単位で管理していることから大きな情報サイズを必要とし，またオブジェクトの位置識別ができないという難点を持つ。一方，ベクタ型データは，道路，河川，行政境界等の線情報，土地利用などの面情報というオブジェクトの表示に用いられる。なお，ベクタ型データは座標情報に基づいており，二地点間の距離計算や特定の地点や線から一定距離範囲内の地域を抽出する空間分析を行うことができる。このため，ベクタ型地図情報は，最適な経路探索結果を表示するナビゲーションシステムや圏域居住者の人口属性分析のようなバッファリング分析を活用するエリアマーケッティングなどの分野で広範に使用されている。

他方，オブジェクトの属性情報は，幾何情報とは別にレイヤと呼ばれる層別に整理され，座標値や地域コードといった位置情報と関連づけた表形式のリレーショナル・ファイルとして管理されている。このため，オブジェクトの属性情報については，独自に登録，更新，分類，検索，解析といった情報処理を迅速に行うことができる。属性情報に関する検索あるいは解析結果は，基本単位区や調査区，さらには行政施設の所在地点情報などと関連づけて，あるいは独居高齢者が居住する住宅や交通事故発生地点の分布のように行政行為によって把握された属性情報とともに地点情報に関

係づけて表示することで，避難誘導体制の準備あるいは道路の改善整備といった行政面でも活用されている。現在では，登録等の行政事務は基本的にコンピュータ化されている。そのため，得られた各個別情報に位置情報を付加するだけで，少なくとも情報技術的にはそれをレイヤに編成することができる。また，GPS機器の広範な普及により，市民参加型で収集された各種の生活情報をGISにデータベースとして登録することで，広く情報の共有化がはかられている。

このようにGISでは，電子地図という基盤情報上に道路，鉄道，各種施設，土地利用区分，さらには様々な行政情報といった各種のオブジェクト情報が位置情報をキーとしてレイヤ毎にデータベースとして維持，管理されており，それらのレイヤが地図上に重層配置（オーバーレイ）されている。

（ii）統計GISのデータ構造

統計GISでは，各種の統計データが基本的にポリゴン・コードという位置情報をキーとして電子地図に統合（integrate）されている。それをデータ構造の面から見れば，ポリゴン・コードをキーとして相互にリンクされた多次元のリレーショナル・データである属性情報としてのdata bodyが，ポリゴンというdata carrierによって担われていることになる。例えば，都道府県や市区町村といった行政区分によるデータの場合，２桁あるいは５桁のコードからなるdata carrierが，集計量というデータ形態を持つdata bodyをそれぞれ担っている。各レイヤのdata bodyを構成する諸変数は，同一レイヤ内あるいはポリゴン・コードを解して変数相互を集計量ベースで相互にリンクさせて解析することができ，その解析結果を地図上に表示することができる。

わが国の国勢調査データの場合，調査区を構成する基本単位区[6)]が最小のポリゴンである。基本単位区に該当する個体レコードはそれぞれ共通の9桁からなる基本単位区コードが付与されたレコード形式を持ち，data bodyを構成する集計量としての各変数値が，原理的には統計GISで表示可

能な最小単位の基盤情報となる。しかしながら，基本単位区の場合，該当する世帯数が平均しても20～30と限られていることから，統計上の秘密保護の観点から，そのままの形で公表するのには問題がある。このため，現在は，国勢調査データに基づく統計GISのdata bodyは，町丁字（9桁コード）が一般に利用可能な最小のポリゴン単位となっている。ちなみに，米国人口センサスの場合，基本単位区に相当する最小の地域表章単位はブロックと呼ばれ，ブロック・コードを持つセンサス匿名（ミクロ）データがPublic Use Microdata files（PUMs）として提供されている。なお，ブロック・コードについては，商務省センサス局が開発し民間にも供用されている地図情報システム（Topologically Integrated Geographic Encoding and Referencing data base: TIGER）とも連動しており，エリアマーケッティング等の分野で民間事業者の間で広く利用されている。

3．統計個票の情報特性

（1）統計個票と収集情報

表式調査を基本的調査形態とする黎明期の統計調査と異なり，近代統計調査は個票による統計原単位情報の収集をその特徴とする。表式調査に対する個票調査の調査技術面での優位性は，統計作成の出発点となる統計原単位情報の形態にある。表式調査の場合，もともと集計量として収集された計数を積み上げることによって統計が作成される。これに対して，個票調査では調査対象となる個々の調査単位に関する個体情報が統計原単位情報として収集される。統計作成の面でこの違いは決定的である。なぜなら，個票調査では統計的把握の最小単位であり現実存在としての調査単位と1対1で対応づけられた各変数（調査事項）に対応する統計原単位情報が得られることから，理論的には個票が持つ変数の任意の組み合わせについて集計結果表の作成が可能となるからである。

ここで，個票調査が実査過程でどのような個体情報を調査単位から収集しているかを見ておこう。個票調査では，調査単位の名称（氏名）や所在地（住所），調査単位の属性，それに調査単位の活動や行動の記録あるいは意識などを調査事項とする調査票によって，個人，世帯，企業あるいは事業所といった調査単位としての調査客体から統計原単位情報が収集される。この他にも調査によっては，調査員が独自に調査区等の関連情報を収集する調査員記入欄を持つ調査票もある。各調査単位に関わるこれらの統計個票情報は，磁気媒体に転写され，調査単位別のレコード形式を持つデータに転化する。

（2）調査個票におけるcarrierとbody

図１は，世帯調査における世帯員の調査個票から転写した個体情報のデータ・レイアウト･フォームを例示的に示したものである。

図１　調査個票から作成される転写データ・レイアウト・フォーム（例示）

					世帯属性				属　性				統計把握項目							
調査アイデント	調査年月	地域符号	世帯一連番号	世帯員番号	一般・単身の別	家族類型	・	・	年齢	性別	・	・	調査員記入欄				調査項目１	調査項目２	・	・
													項目１	項目２	・	・				

調査個票そのものに記載された情報とそれを転写したレコードとの間には，それぞれが保有する情報に関していくつかの相違が認められる。図１の項目のうち，調査アイデントと調査年月は，調査履歴を示す調査メタ情報であり，世帯一連番号と世帯員番号は世帯と世帯員とのリンクキーとして，位階的（hierarchical）データ構造を持つデータセットを編成するために調査実施者側で事後的に付与されたものである。さらに，地域符号は，調査区に対応するものとしてレコードの空間的位置情報を与える。

その一方で，調査個票に記載されている情報で図１の転写レコードでは

削除されているものもある。調査個票に記載されている氏名（事業体の名称），住所（所在地），連絡電話番号等がそれである。これら個体を識別する情報は，回収後に調査票の審査を行う際に記入内容を調査客体に対して照会するのに用いられる。それらの記入内容は，事業所の名寄せ集計の際の照合キー情報として用いられる場合以外にはその情報が集計に使用されることはない。また，それらは個人の秘密や企業の営業上の秘密保護の観点からも特に慎重な取り扱いを必要とする情報とされている。わが国で個体を識別するこれらの情報が転写された個体レコードから削除されているのは，このような事情によるものと考えられる。

以上の考察を踏まえて個票情報がどのようなデータ構造を持つかを次に検討してみよう。

上述したように，調査個票による統計調査では，各調査単位毎に統計原単位情報が収集される。このような統計個票情報をデータの構造という観点から見るなら，個体の識別情報（あるいはその代理変数としてのID番号等）が，調査個票で収集される一連の情報のdata carrierに相当し，他方，上記のデータレコードに含まれる諸変数がdata body部分を構成する。なお, 行政記録についても現在ではその大半がコンピュータ処理されており，結果的に個体単位で管理されている情報も少なくない。この種の個体行政情報は，基本的に個票イメージのデータ構造を持つ。

４．統計個票情報の情報特性

変数の多寡を別にするなら，data bodyがdata carrier情報によって担われる多次元ベクトルの形で表記されるという意味では，画像・音声・空間情報と統計個票情報とは共通性を持つ。しかしながら，data carrierの性格とdata body部の変数の機能の面で，統計個票情報と他の諸情報との間にはいくつかの相違があるように思われる。

（1）data carrierについて

すでに見たように，静止画像の場合ピクセルあるいはボクセルという画素がdata bodyを担っており，動画については時点情報を持つdata bodyを各画素が担う。また音声情報では，サンプリングという時間単位情報が個々の微小時間における音量を示すdata bodyを担っている。さらに，地理空間のデータ表現としての地域メッシュと統計GISにおいては，メッシュあるいはポリゴンという地理的座標系と対応づけられた画素的空間のコード情報がdata carrierとして当該地域の特性情報であるdata bodyの担い手となっている。

これらのdata carrierには，２つの特徴的な点がある。

画像のdata carrierは被写体に帰属するのではなく，撮影（映像）フレーム内の座標情報によって与えられる。また，音声についても，音源それ自体がサンプリング情報を与えるのではなく，デジタイザーの性能あるいは音声採取の際に使用可能な情報のサイズがそれを決定する。さらに，地域メッシュやGIS統計の場合にも，メッシュ・コードやポリゴン・コードは，域内のオブジェクトの特性情報とではなく，緯度，経度あるいはGPSといった測地系に対応づけられたcarrierキー情報に他ならない。このように，data carrierの位置情報がデータが写しとる被写体や音声さらには地理空間内の実体としてのオブジェクトに依存するのではなく，観測機器あるいは物理的な条件の方に依存するというのがdata carrierの第１の特徴である。

こういった情報におけるdata carrierの第２の特徴は，機器の性能や処理可能な情報のサイズに専ら依存する一種の階層構造を持つ点である。観測機器の性能や処理できる情報量の制約が許せば，より詳細な画素（ピクセル，ボクセル）についての画像表現ができ，また音声についても，サンプリングの頻度を大きくすることで，限りなくアナログに近い波形で音声をデジタル化することができる。識別可能な個体情報の空間表現として行政

等で利用されている一般のGISについては，GPSという測地系上の位置情報がdata carrierとして当該個体の属性や特性に関する情報を担っている。一方，地域メッシュ統計やGIS統計における空間情報のcarrierであるメッシュやポリゴンといった区域情報については，統計上の秘密保護のためにdata bodyが集計量形態をとる場合に，調査区あるいは基本統計区といった統計調査技術上の制約を受けるケースもあるが，どの区画レベルで処理するかは，その処理に投入可能な資源や作業量に依存する。

これらに対して，統計における調査個票情報については，統計が把握の対象とする個人や世帯それに企業や事業所といった個体，すなわち個々の調査単位それ自体がdata bodyの担い手である点が上記の各情報の場合と本質的に異なる。実はこのことが，調査個票に対して集計量とは全く異なる情報特性を付与する。すなわち，data bodyが，氏名（名称）や住所（所在地）といった存在としての調査単位と一体化された識別情報によって担われているというまさにその点が，調査個票情報に対して上述した他の情報と異なる情報特性，すなわち，識別情報あるいはそれから導出された個人識別番号や企業コード等のID情報をキーとした個体単位でのbody情報の潜在的連結可能性という情報特性を付与することになる。

（2）data bodyについて

すでに見たように，静止画像データのdata bodyを構成するトーン情報は，単色画像については1次元，カラー画像の場合は3次元ベクトルとして与えられる。カラー画像の場合，三原色の各トーン情報が与える濃度を重合（mixing）することで当該画素の色調が決定される。そこでは，3次元の各変数値はそれ自体としてではなく，重合の組合せパターンを指示する情報として意味を持つに過ぎない。

音声情報のbodyを構成する振幅情報は基本的に1変数であるが，マルチトラックレコーダーによって複数のトラックに分けて録音された音声のdata bodyは多次元のベクトル構造を持つ。マルチトラックの音源データに

ついては個別の編集が可能であり，サンプリングの時点を調整することで新たな音声を既存の音源情報に追加することができる。なお，それらからマスターレコーダーにミックスダウンされたデータは，モノラル音声の場合には1組の，またステレオ音声の場合には2組の振幅情報に圧縮還元されることになる。このように，音声情報の場合，個々のサンプリング時点情報に照応する並存する複数の振幅情報を重合したデータそのものも，音声情報のdata bodyとして意味を持つ。

地域メッシュや統計GISの場合，各レイヤ上で個々のメッシュやポリゴンに属する各個体レコードは，いずれもメッシュやポリゴン・コードを共有する。このことは，これらのコードをリンクキーとして該当する複数の個体に関するdata bodyを構成する諸変数が相互に接合できることを意味する。ただし，data carrierの識別情報であるメッシュやポリゴン・コードと域内に所在する個体との間には一般に1対複数の照合関係が成立していることから，data bodyの変数の接合は，個体ベースによるものではなく，あくまでも集計量としての接合利用可能性を与えるだけである。

地域メッシュ統計や統計GISの場合，data carrierは画素的性格を持つ。このため，これらについては，画像における動画と同様に，新たに時間要素を情報として取り込むことで，それらを時空間システムとして展開することができる。これは，例えば地域別の人口増減率のように，当該空間について集計量としての統計情報の差分を管理することで，時点間の変化や空間的な波及状況を把握するものである。そこでは，メッシュやポリゴンのdata bodyの特定の変数あるいは複数の変数から導出された各種統計指標が集計量あるいはmixingによるトーン情報として描くいわば動態メッシュや動態GISが，地価の波及や高齢化の進展といった現実における変化の時間的推移についてのいわば動画的な認識情報を提供しうる。なお，ここで，地価のように対象地域の観測単位が固定的である場合と地域的に移動しうるものの場合とで動態の意味が異なる点に留意する必要がある。なぜなら，例えば，動画的表章の対象域内の個体が移動可能な主体である場合，

この動画によって表現されるのは，その時々に当該メッシュやポリゴンという域内に，いわばcross sectional（横断面方向）に存在する個体についての静止画像を連写したものに他ならないからである。それは，同一対象集団をlongitudinal（縦断面方向）に追跡した動画表現とは明らかに異なる。

表２は，各種デジタル情報のdata carrierとdata bodyの特徴を要約的に整理してみたものである。

表２　data carrierとdata body

<table>
<tr><td colspan="2"></td><td>data carrier</td><td colspan="2">body変数の機能</td></tr>
<tr><td colspan="2">音　声</td><td>時点情報</td><td colspan="2" rowspan="2">重　合</td></tr>
<tr><td colspan="2">画　像</td><td rowspan="3">画素的情報</td></tr>
<tr><td rowspan="2">空　間</td><td>地域メッシュ</td><td rowspan="3">接　合</td><td>集計量</td></tr>
<tr><td>統計GIS</td><td></td></tr>
<tr><td colspan="2">統　計</td><td>調査単位識別情報</td><td>個　体</td></tr>
</table>

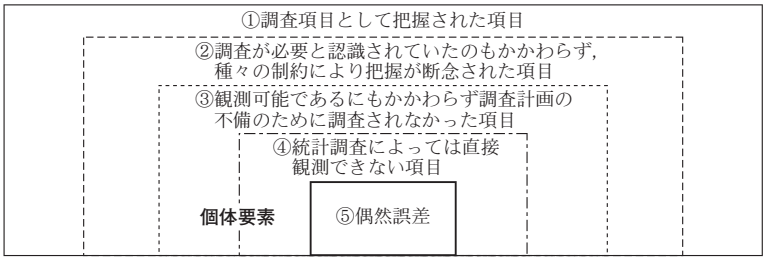
５．統計個票情報の潜在的リレーショナル特性

（１）個体の変数要素と統計調査

画像や音声といったマルチメディア情報と異なり，統計におけるdata bodyは，統計調査によって把握される数十次元あるいは数百次元の超多次元情報から構成される。しかし，data bodyを構成するこれらの変数は，調査単位である個体を特徴づける情報要素である全ての変数を網羅しているわけでは必ずしもない。なぜなら，個体に関する情報要素は，以下のようなものから構成されているからである。①調査票に記載された調査事項，②調査計画段階で調査事項の候補となったものの，予算制約，報告負担あるいは他の調査との重複のために調査が断念された項目，③本来調査すべき項目であるにもかかわらず調査計画の不備により調査事項にならなかった項目，④調査による把握が困難な事項，⑤偶然的要素により調査結果に

反映される撹乱的要素，がそれである。図２は，これらの変数要因の関係を概略的に図示したものである。

図２　個体を規定する変数要因



分散型統計制度であるわが国では，各府省が特定の調査目的の下に自ら投入可能な予算並びに要員といった一定の資源制約の中でそれぞれの調査を企画し，実施している。各府省の調査企画については制度官庁が，調査の実施体制，調査項目の妥当性，回答者の報告負担の量，他の調査との調査事項の重複の有無などに関して統計調整の観点から審査を行い，調査計画が最終的に承認される。統計調査の審査過程で，予算の制約，他の調査との重複，それに回答者の報告負担の軽減といった理由で，本来，調査項目として採用されるべき事項が，やむなく調査項目から除外される場合もある。センサスが与える母集団分布情報を介してセンサスと標本調査，さらには標本調査が相互に体系性をなしているとはいえ，これまで統計調査は基本的にそれぞれ独立の単体（stand alone）的調査として企画，実施されてきた。その結果，本来的にはある変数に対して系統的に作用を及ぼす一群の説明変数を構成しているにもかかわらず，その一部が別の統計調査の調査項目として調べられるケースも起こりうる。

それぞれの統計調査における個票情報のdata bodyをこのような調査の企画，実施と関連づけて見た場合，それを実際に構成している変数要素群とは，統計の企画，審査過程において当初は実施が検討されていた②に該

当する事項が調査項目から削除され，結果的に①だけが調査事項として採用され，その限りで得られた個体に関する情報要素に他ならない。②～⑤に該当する情報要素が調査結果に反映されていないことから，それらは，分析結果に偏りをもたらす要因としてデータに内在することになる。

このような分析結果における偏りをできる限り排除するためにも，それに有意に影響していると考えられる諸変数については，個々の調査において必らず調査される必要がある。そこでは，調査事項の重複とデータに内在する偏りとが，単体的調査において相互にトレードオフの関係にある。調査企画の過程で調査項目として意識にさえのぼらなかった③は論外として，本来，一体のものとして調査項目に加えられるべき②に該当する諸項目が予算的，技術的，あるいは制度的な制約により調査事項から結果的に除外されたケースについては，さもない場合には回避できたはずの偏りを生起させることになる。②に該当する調査項目の中には，同一個体を対象とした別の調査で①として調査されているケースもありうる。

（2）統計個票情報の潜在的リンク可能性

統計個票情報は，調査単位と一体化している。このことは，data bodyに対してその担い手である個体識別情報をキーとした横断面（cross sectional）あるいは縦断面（longitudinal）での個体ベースでの１対１リンク，すなわち完全照合の潜在的可能性を付与する。

（ｉ）data bodyの横断面接合による変数次元の拡張可能性

異種の調査における同一個体に関するレコードを氏名等の直接的個体識別情報あるいはそれから導出されたID番号等をリンクキーとして相互に１対１でリンクさせることで新たなレコードを編成することができる。これは，データの構造としては，同一のdata carrierが担っている異種のdata bodyの変数ベクトルの次元の拡張を意味する。data bodyの変数次元を拡張することで得られた新たなデータセットは，まず統計の作成面では，例えば静態データと動態データのリンクによるこれまでにないタイプの統計情

報を含めた追加的な統計情報の創出可能性を持つ。他方，利用面では，元の各個票情報がそれぞれ保有していたその利用可能性を二重の意味で拡張する。

その第 1 は，利用可能性の外延的拡張である。data bodyの変数次元の拡張は，より広範な変数（調査項目）相互の交差的利用，すなわち，新たな変数を組み合わせた多重クロス表の作表や回帰分析における変数選択の範囲の拡張を可能にする。それは，既存の独立な単体的調査だけからはでは得られなかったより広範な調査結果を提供することで，時空を貫き相互に連関する社会的集団現象の諸側面（その総体を以下では統計把握空間と呼ぶ）に対するより豊富な認識資料を提供する。新たなデータセットが与える利用可能性の外延的拡張は，統計把握空間に関してそれまでは得られなかった追加的な統計的認識資料を提供することで，現実に対する認識の向上に寄与する。

第 2 の利用可能性の拡張は，既存の統計的認識の質の改善への寄与である。統計把握空間に対してより偏りの少ない認識を得るには，その構成要素である個体に関する構造的変数要因を可能な限り統御（control）する必要がある。種々の現実的制約から個々の調査がそれらを網羅的に調査項目として採用することができない場合，得られる結果数字は自と偏りを持ったものとなる。例えば，複数の変数に共通に作用する第三の変数要因が存在する場合，それの作用を統御していないことでしばしばみかけの相関が発生する。また，回帰分析において，説明変数が従属変数に対して系統的に作用を及ぼす変数を尽くしていない場合，結果的に残差項の中には系統的変動要因が残存することになる。残差が系統的な偏りを含む場合，得られた回帰推定値は偏りを持ったものとなる。個体ベースでのリンケージによって得られた拡張レコードが可能にする変数相互の新たな交差的利用可能性は，この種の偏りの回避という意味で，分析結果の質の改善に貢献する。

（ii）data body情報の縦断的接合可能性

統計の把握対象である個人，世帯，企業あるいは事業所といった個体は，もともと横断面と同時に時間軸という二つの方向性を持つ。それは，様々なイベントを経験しつつ縦断的に変貌を遂げる存在でもある。そこでは個体は，誕生しその時々において様々な社会的集団を構成しつつ，時代という共通の場をくぐりながら，個々には時間の経過の中で着実に年齢（継続時間）を重ね，そして最終的には消滅（死亡）に至る。それらは，決して孤立した無機的存在ではなく，時空間の中でその時々に個体間で様々な関係を取り結びつつ思考，行動し，その結果として社会的集団現象の諸側面である統計把握空間を作り上げるそのような主体として存在する。

統計把握空間の時間的側面について，これまで統計は，それを一定期間中に生起したイベントを集計した動態統計あるいは横断面データの差分による比較という形で取り扱ってきた。これに対して，1960年代半ばにアメリカで，同一個体を時間の経過の中で継続的に追跡調査するという新たな調査方式による縦断面データ（longitudinal data）が初めて導入された。このような個体ベースで縦断面にリンクされたデータから構成されるデータセットは，パネルデータと呼ばれる。

国勢調査や事業所・企業統計調査のようなセンサスについては，ほぼ同一の調査事項に関する調査が周期的に繰り返し実施されてきた。統計個票が調査単位である個体をdata carrierとしており，センサスが全個体を対象とする悉皆調査として実施されることから，このように反復実施される横断面調査の調査単位を直接識別する情報あるいはそれから導出されたID番号等をリンクキーとして，それぞれのdata bodyを構成する同一事項に関する変数値を時間軸方向に相互にリンクすることによって縦断面のデータを編成することができる。このようにして編成されたdata bodyの各変数の変数値は，時点という新たな次元の情報要素が付加されたパネルデータというデータ構造を持つ[7)]。

data bodyを構成する各変数が時点要素を持つことから，それは，一回限

りの静態調査による横断面データや反復的横断面データさらには集計量としての時系列データによっては実現できない新たな利用可能性を持つ。

上記のデータ形態に対するパネルデータの分析面での優位性の一つは，時点間の位相要素を集団構成に組み込んだ動態集団類型の編成が可能な点である。時点間の就業異動に基づく就業動態類型や住戸の所有形態の変化を類型化した居住動態類型による各集団の属性分析や行動等に関する特性分析は，同一個体について，複数時点の情報を持つパネルデータあるいはパネル的調査項目を持つ横断面調査データを用いることで初めて可能となるものである。これは，横断面データのパネル化によるデータの分析可能性のいわば外延的拡張といえよう。

パネルデータのもう一つの優位性は，分析結果の質に関係するものである。統計把握空間を構成する個体あるいはグループの中には，個体やグループに固有の特徴を持つ要因も存在すると考えられる。横断面データや時系列データによる分析の場合，分析結果の中にはこのような個体差やグループ差も分離不能な形で混在する。そのために，得られた推計値はそれらの作用の結果として偏りを持ったものとなる。パネルデータの場合，同一個体が継続的に調査されることから，統計調査によって観測不可能な変数要素（図２の④）のうち個体に関して時間に不変な要因を統御，除去することで，純粋に時間的変化の寄与分だけを抽出することができる。このような個体差に起因する偏りの排除という意味で，パネルデータは，これらのデータ形態に対する優位性を持つ。

むすびにかえて

冒頭にも述べたように，本稿における分析は，二つの素朴な疑問から出発している。一つは，統計あるいは統計学がこれまで暗黙の前提としてきた集計量という観念の再検討に関係するものであり，もう一つは，静態量の把握を目的とする統計調査が，調査時点現在で現実を写し取った一種の

瞬間撮影であるとする理解である。前者は統計個票という統計の本源的形態の意味を探ることに，また後者は統計以外のデータ表現と統計個票のデータ構造とを比較考察することによって統計個票情報の情報特性を明らかにするという本稿の分析方法に関係している。

個体に関わる調査結果そのものは，多次元の数値レコードとして表現される。そこで本稿では，統計の本源的形態である統計個票情報の特質に関して，data carrierとdata bodyというデータ構造の側面から統計以外のデジタルデータと統計個票のデータ構造を比較しつつ，その特徴を考察してきた。その結果，統計個票情報のdata bodyが調査単位に関する個体識別情報というdata carrierによって担われていること，またその帰結としてdata bodyが潜在的に横断面，縦断面のリンク可能性（potential “relationality”）を持つこと，さらには，地域メッシュ統計や統計GISが画像や音声といった他のデジタル情報と統計との中間的性格を持っていることなどが明らかになった。

標本数が限られている標本調査の場合，個票情報のリンク可能性は，あくまでも潜在的可能性にとどまる。なぜなら，特に小規模標本調査の場合，同一の個体が異なる調査において調査客体として標本に選ばれる確率は殆どゼロに等しいからである。その点で，本稿でいうリンク可能性は，むしろセンサス同士，あるいはセンサスと標本調査において現実的意味を持つ。なお，このようなリンク可能性という調査個票データの情報特性は，単に統計調査だけでなく，行政記録から得られるデータとのリンク可能性にも拡張できるものである。

統計個票が持つ個体ベースでの潜在的なリンク可能性という統計情報に特有な情報特性は，統計を個体すなわち調査単位のレベルで捉えることによって初めて析出できるものである。なぜなら，統計を集計量として捉える限り，この種の情報特性は集計値の中に埋没してしまっているからである。このような個体視点によって初めて横断面，そして縦断面方向にdata bodyを拡張したデータセットの編成が可能となる。それは，統計作成面で

は新たなタイプの統計整備の可能性を持つと同時に，利用面では，一回限りの横断面調査データ，反復的横断面データ，そして時系列データといった従来のデータにはない独自の利用可能性を持つ。その意味で，個体から出発した場合の現実に対する認識像と集計量から出発した場合のそれとは明らかに異なる。その点についての具体的な検討は，今後の課題としたい。

注

1）デジタル画像データでは，トーン情報はビット数で表現される。例えばグラデーション表現を必要としない単色（モノクローム）テキスト文書などは，最小の１ビット（2^1= 2 レベル）で各画素のトーンをデジタル化することができる。通常の写真の場合，8 ビット（2^8=256レベル）程度で実用に足るが，業務用のフル規格の画像データでは，一般に24ビット（2^{24}=16777216レベル）が使用されている。

2）１秒間に30こま（/30fps : frame per second）以上であれば，比較的スムーズな動画が得られるとされている。

3）CTスキャン画像は，線源から照射されたX線の線量が検出装置によって受信されるまでの減衰率を測定することでその構造を把握し，その測定結果を画像化したものである。そこでは，骨格，筋肉，空気，血液等について所定の減衰率が±1,000の範囲内のトーン情報として与えられていることから，トーン情報としては，全体で2000以上のトーンレベルが求められる。従って，医療用のX線撮影画像の場合，実用に足る解像度のレベルを得るには，最低限でも10ビット（1024レベル）が必要となる。

4）音声の強弱を示す振幅を何段階で数値化するかはビット数で決まる。例えば，２ビットであればあらゆる強弱の音が2^2= 4 段階で表現される。ちなみに，専門家仕様の機材の場合には，一般に24ビット（2^{24}=16777216レベル）が使用されている。

5）現在，大都市圏域の一部について1/4地域メッシュが提供されている。

6）基本単位区は，学校区，町丁・字などの小地域についての調査結果を提供するために平成２（1990）年国勢調査で初めて導入された最小の地域表章単位である。街区表示方式による住宅表示地区では，基本単位区は原則として一つの街区に対応しており，それ以外の地域では，道路，河川，鉄道，水路といった恒久的なオブジェクトによって境界が区分されている。平成２年国勢調査以降，調査区の設定も基本単位区に基づいて行われており，一般には２つ以上の基本単位区を合わせて一つの調査区が設定される。なお，各基本単位区には９桁からなるコード番号が付けられているが，先頭の６桁が町，丁あるいは字を示すコード（町丁・字コード）に対応している。なお，平成７年国勢調査で初めて導入された町丁・字別集計は，それらを共有する基本調査区コードを持つレコードを集計することによって得られる。（「統計表で用いられる地域区分の解説」http://www.stat.go.jp/data/kokusei/1995/04-02.htm参照）

7）ここでのdata bodyは，静止画像の連写による動画情報が持つ時間要素というdata bodyの変数次元の拡張とは異なる。なぜなら，パネルデータの場合，調査客体はその空間的位置とは無関係に一個の個体として，いわば定点観測的に追跡，記録されるからである。ちなみに，風景を定点観測的に連写する場合，被写体の時間的変化が各画素におけるトーンの変化，すなわちdata bodyの変化として対応する。しかしこの対応は，被写体がたまたま静物であることによるものである。被写体自体が動体である場合，静止画像の連写は，被写体という客体そのものの動きの追跡記録ではなく，データ構造的にはdata carrierとしての各画素が各瞬間において担うdata bodyのトーン情報が結果的に描く瞬間撮影の軌跡に他ならない。

〔付記〕本稿は，2008年11月11日に京都大学学術情報メディアセンターのセミナーにおける報告「政府統計情報のSOCIAL ASSET的性格と統計データベース」の一部に加筆したものである。