岡山大学環境理工学部研究報告
第8巻，第1号，pp.9-17，2003年3月

インターネットを利用した統計調査分析システムの構築

藤野友和*　　垂水共之†

Development of a statistical survey analysis system using Internet

Tomokazu Fujino, Tomoyuki Tarumi

(Received November 25, 2002)

We constructed the online analysis system of the Okayama Activity Area Survey in place since 1979. This is one way to solve the problem in case of releasing the result of a statistical survey. In this paper, we treat the detail of the structure of the system.

Keywords: microdata, statistical disclosure control, web based system, online database

1　はじめに

近年，マイクロデータ（匿名化標本データ）の公開に対する需要の高まりに伴い，その統計的開示制御（Statistical Disclosure Control）に関する研究が多く行われている．マイクロデータを外部に向けて公開する場合，悪意を持った侵入者により，統計調査に参加した個人のデータが特定され（個体識別），本来侵入者が知り得なかった個人の情報が流出する（情報漏洩）という危険性，すなわちプライバシーの侵害の恐れが生じる．統計的開示制御とは，このようなことが起こる可能性をできる限り小さくするためにマイクロデータに施す統計的な処理のことである．一度でも個体識別や情報漏洩が発生すれば，その統計調査に対する社会的な信用はなくなり，それ以後の統計調査の実施に支障をきたすことにもなる．データの公開者は，このようなことがないよう，マイクロデータに対して統計的開示制御を行うが，分析者が受け取る，統計的開示制御が施されたデータは常に情報の損失を伴う．しかし，分析者の真の目的は元のマイクロデータを手に入れることではなく，それを分析して新たな知見を得たり，企業であれば利益を得たりすることであろう．このように考えると，プライバシーの侵害の可能性をできるだけ小さくしながらも，分析者が真の分析結果を得られるような仕組みがあればよいということになる．近年のネットワーク環境の急速な発展により，ネットワーク上に統計調査結果を分析できるシステムを構築し，外部に向けて公開することが可能になった．このようなシステムの利点は，分析者が真のデータに接することなく真の分析結果を得ることができるということだけでなく，これまで行われてきたような報告書や物理メディアなどによる調査結果の配布方法に比べ，調査結果の配布にかかるコストや労力を削減できることにある．

アメリカの政治・社会調査のための大学協会(ICPSR:Inter-university Consortium for Political and Social Research) は協会に加盟している機関に対して，統計調査データの提供を行っている組織である．ICPSR の公開する Web サイトにおいて，SDA(Survey Documentation and Analysis) というオンライン上での統計調査データの分析システム[1] が公開されている．このシステムは加盟機関外からも利用可能で，約50種類の統計調査データに対して基本的な分析を実行することができる．分析機能は，単純集計，クロス集計や相関行列の出力などであるが，このシステムにおいて興味深いのはカテゴリの併合をしたり，異なる変数から一つの新規変数を生成したりして再集計を行う機能 (Recoding) が実装されていることである．SDA では，複数のデータソースに対して分析機能は共通であるため，個別のデータに特化した分析は実行できない．本稿で提案するのは，特定の統計調査を想定した分析機能を持つオンラインシステムである．

*岡山大学大学院 自然科学研究科 資源管理科学専攻
†岡山大学 環境理工学部 環境数理学科

[1] http://www.icpsr.umich.edu/

表1：各調査年における有効回答者数

回数	調査年	有効回答者数	回数	調査年	有効回答者数
第1回	1979年	6999名	第6回	1991年	7695名
第2回	1981年	8301名	第7回	1994年	7823名
第3回	1983年	8324名	第8回	1997年	7891名
第4回	1985年	7092名	第9回	2000年	7797名
第5回	1988年	7775名			

表2:データ本体の一部

id	address	sex	age	job	job_place	job_trans	job_time	hos_mild	hos_heavy
0001	101	1	2	1	101	4	2	101	101
0002	101	1	2	4	228	5	1	228	228
0003	101	1	3	1	101	3	3	101	101
0004	101	1	3	1	107	3	3	101	101

山本・垂水 (1998) は後述の岡山行動圏調査に関する集計システムを構築した．このシステムは，テキストファイルでサーバー内に保存されているデータを分析者の CGI(Common Gateway Interface) を通した要求に従って，perl スクリプトにより記述されたプログラムにより単純集計やクロス集計を実行するというものであった．これらの集計には，セルの秘匿という統計的開示制御が適用されており，前述の分析者の真の目的は達成できないが，我々は現在に至るまで，様々なシステムの改良や分析機能の実装，追加を行い，2002年10月に岡山行動圏調査分析システム Version2.0 として公開した．本稿では，このシステムの構築に関する報告を行い，その際の注意点や，システムに関する問題点などについて考察する．なお，本システムの URL は

`http://www.f7.ems.okayama-u.ac.jp/shoken/`

であり，分析機能を利用するためには著者への電子メールによる申し込みでユーザー ID を取得する必要がある．

2 岡山行動圏調査について

岡山行動圏調査は，岡山経済研究所が岡山商科大学と岡山大学の指導のもとで1979年から実施されている調査で，1985年度までは2年おきに，以後は3年おきに現在までに9回行われている．調査の主な目的は，岡山県を含む周辺地域の交通環境の変化（本州四国連絡橋，岡山自動車道，山陽自動車道，中国自動車道の開通，岡山空港の開港など）に伴う岡山県民の行動圏の変化を調べることである．質問項目は，性別，年齢や職業に関する基本情報をはじめ，医療圏，交際圏，商圏，観光圏，各年により異なるトピックにより構成されている．医療圏は重症時，軽症時に利用する病院の場所，交際圏は友人や親戚を訪問する場所，商圏は各商品ごとにそれらを購入する場所，観光圏は各観光地ごとの訪問回数についての質問項目からなっている．

表1に各調査年におけるサンプル数を示す．各市町村ごとの住民の行動圏の特性を把握するという観点から，各市町村から最低50サンプルを抽出するという条件の下，市町村を2大市（岡山市，倉敷市），その他の市，町村の3グループに分類し，各グループ内では人口比で標本数を決定するというサンプリング方法をとっている．

3 システム構成

本節では，分析システムの構成について述べる．データベースに対する集計の要求，その他の分析のためのプログラムの実行や分析結果の表示は，スクリプト言語である Ruby[2] により記述された CGI プログラムが行う．データの分析者は，Web ブラウザを利用して分析システムのサイトにアクセスし，フォー

[2] `http://www.ruby-lang.org/`

表 3:項目に関する情報の一部

id	var_e	var_j	item	category	catid
1	id	整理番号			
2	address	住所	tableA	基本情報	1
3	sex	性別	sex	基本情報	1
4	age	年齢	age	基本情報	1
5	job	職業	job	基本情報	1
6	job_place	勤務・通学地	tableAB	基本情報	1
7	job_trans	利用交通機関	trans	基本情報	1
8	job_time	通勤時間	duration	基本情報	1
9	hos_mild	軽症時の病院	tableAB	医療圏	2
10	hos_heavy	重症時の病院	tableAB	医療圏	2

表 4:項目ごとの選択肢を含むテーブル（一例）

code	job
1	つとめ人
2	自営業
3	農林漁業
4	学生
5	専業主婦
6	無職
7	その他
9999	無回答・誤記入

表 5:秘匿処理の例

(a) 秘匿処理実行前

	B_1	B_2	B_3	B_4	合計
A_1	103	128	4	112	347
A_2	120	150	87	63	420
A_3	1	85	31	90	207
合計	224	363	122	265	974

(b) 秘匿処理実行後

	B_1	B_2	B_3	B_4	合計
A_1	×	128	×	112	347
A_2	120	150	87	63	420
A_3	×	85	×	90	207
合計	224	363	122	265	974

ムを通して分析に必要な集計年度，集計項目やオプションと共に分析要求をCGIプログラムに渡す．CGIプログラムは受け取ったパラメータに基づいて各種の分析を実行後に実行結果を表示するためのHTMLを作成し，分析者のWebブラウザに送る．

各年の調査により得られた個票データはフリーのRDBMSであるPostgreSQL[3]により管理されている．各年ごとに調査項目や選択肢が異なるため，データは各年ごとに格納されている．各年のデータは，データ本体のテーブルと基本情報のテーブル，そして各質問項目に対応する選択肢を含む複数のテーブルから構成されている（表2〜表4）．調査結果の全てをデータベースに格納することにより，オンライン分析システムの構築は，このデータベースに対するWeb上のインターフェースを構築することと等価になる．また，RDBMSはデータの更新作業を容易に実行できるため，リアルタイムで変化するデータ，具体的にはWeb上でのアンケート調査に対しても，本稿で提案するようなシステムを構築することが可能となるであろう．

CGIプログラムからのデータベースへのアクセスは，RubyのPostgreSQLにアクセスするための拡張モジュールである"Postgres"[4]を利用することで容易に実現される．これによって呼び出されたデータに対する統計処理にはフリーの統計エンジンであるR[5]を利用するがこれに対するRubyとのインターフェースは開発されておらず，我々はCGIプログラムが子プロセスとしてRを起動するという方法を採用した．具体的には，CGIプログラムがファイルとしてRのコマンド生成し，これを入力として起動したRから出力された分析結果をCGIプログラムが読み込むという仕組みになっている．また，地図の描画を行うために利用するPGPLOT[6]はC言語とFortranのためのグラフィックスに関するサブルーチンライブラリであるが，これに対するRubyのインターフェースとしてRuby/PGPLOT[7]が開発されており，これを利用することにより動的に地図を生成することが可能となっている．

[3] http://www.jp.postgresql.org/

[4] http://www.postgresql.jp/interfaces/ruby/index-ja.html

[5] http://www.r-project.org/

[6] http://astro.caltech.edu/~tjp/pgplot/

[7] http://www.ir.isas.ac.jp/~masa/ruby/pgplot/index.html

4 秘匿処理について

この分析システムにおける分析機能のほとんどにおいて，特定の調査項目に関する度数表が出力される．第2節で述べたように，各市町村から最低50票の標本が抽出されており，特に人口の少ない町村において標本抽出率が高くなるという状況が生じている．このため，調査に参加した人の特定が，一般の無作為抽出に比べて発生しやすい．このことは，度数表において少数の標本しか含まないセルに属する個体が特定される可能性を高めている．これを防ぐための最も一般的な方法は，一定数未満の標本しか含まないセルを秘匿することであり，この分析システムにおいては，5人未満のセルに対してこれを適用している．これを一次秘匿と呼ぶが，周辺和が既知の状況においては，秘匿されたセルの標本数が再計算可能となってしまう．これを防ぐために，秘匿セルを追加する処理が度数表に適用される．この処理を関連秘匿と呼ぶ．表5は，5人未満のセルに対して一次秘匿を適用し，さらに関連秘匿を行った例である．

5 分析機能

5.1 単年度単純集計

単年度単純集計は，指定した年度における一つの項目に関する度数分布表を出力する．指定された項目の回答が空白であったり，選択肢にないようなコードであったりするような個体に関しては全て「無回答・誤記入」のカテゴリにカウントされる．すなわち，単純集計の最終的な結果は，表4のような選択肢のテーブルと結合した形で出力されることになる．実際の実行画面を図1に示す．図1(a)の選択項目のリストは指定された集計年度の調査項目をCGIプログラムがデータベースから読み込むことにより動的に生成している．ここで選択された項目に関する出力結果が図1(b)のように表示される．

5.2 単年度クロス集計

単年度クロス集計は，指定した年度における二つの項目に関するクロス集計表を出力する．本システムのクロス集計機能で特徴的なのは，秘匿セルに対する補完機能を備えていることである．この機能は，対数線形モデルを仮定することで，クロス表の周辺和から各セルの期待頻度の最尤推定量が容易に計算できるという事実を利用して実現している．なお，具体的な計算方法はAppendixに示す．図2(a)は秘匿セルを含むクロス集計の結果表示画面である．このウインドウ内の補完処理ボタンをクリックすることにより，図2(b)の補完処理のメニューが表示される．このメニューでは，補完処理に用いるモデルを4つの対数線形モデルから選択する．図2(c)が補完処理を行ったクロス集計表の表示画面である．ここでは，クロス集計表と同時に，モデルの当てはまりの良さを示すAICと，周辺和の制約条件を満たすための修正量を示す値が表示される．利用者はそれらのどちらを優先するかによって，適切なモデルを選択することができる．

5.3 相関行列

相関行列出力機能は，10個までの項目に関する相関行列を表示するものである．分析者はこれらの項目と同時に，相関行列を計算する際の「無回答・誤記入」カテゴリの処理方法と連関測度を選択する．「無回答・誤記入」カテゴリの処理方法は

1. 1つのカテゴリとして処理
2. リストワイズに削除
3. ペアワイズに削除

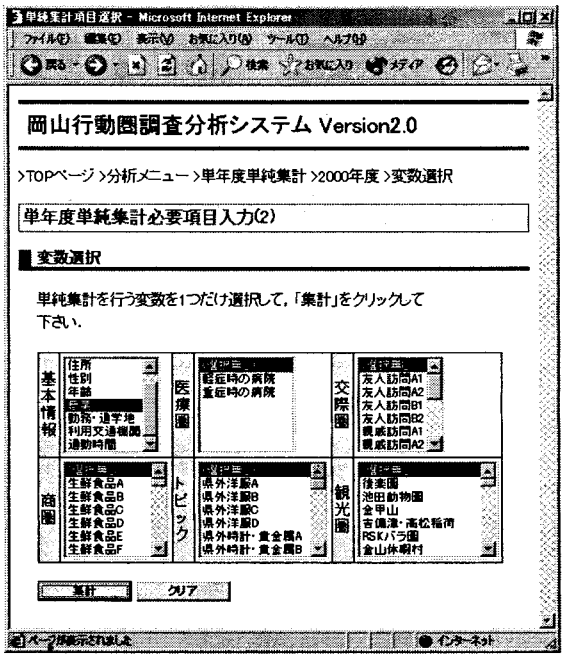
の3通りがある．2と3の違いは，選択された変数の中で1つでも「無回答・誤記入」があるサンプルは削除するのが2であるのに対して，3は相関を計算する2変数のペアごとに少なくとも一方に「無回答・誤記入」があれば，その相関の計算にだけ，そのサンプルを使わないというものである．岡山行動圏調査においては，年齢，性別や住所という個人の基本情報以外の項目については，その質問に関する行動を取った場合についてのみ回答するという形式であるため，無回答は単なる記入漏れということではなく，その行動を取っていないという意味を含んでいるケースが多いと考えられる．分析者はこのことを考慮して，処理方法を選択しなければならない．連関測度はクラメールの連関係数とピアソンの相関係数の2通りが選択可能である．クラメールの連関係数は，$k \times l$ の分割表において2変数間の連関性の度合いを示す測度で

$$V = \sqrt{\frac{\chi^2}{N[\min(k,l)-1]}}$$

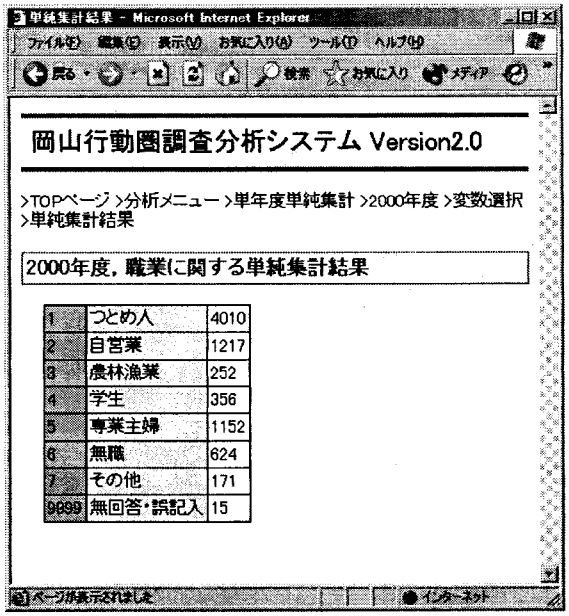
で定義される．ここで，N は観測総数で，χ^2 は

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$$

である．ただし，f_{ij} は各セルの頻度，$\hat{f}_{ij}$ は2つの項目間が独立である場合の各セルの期待頻度，すな

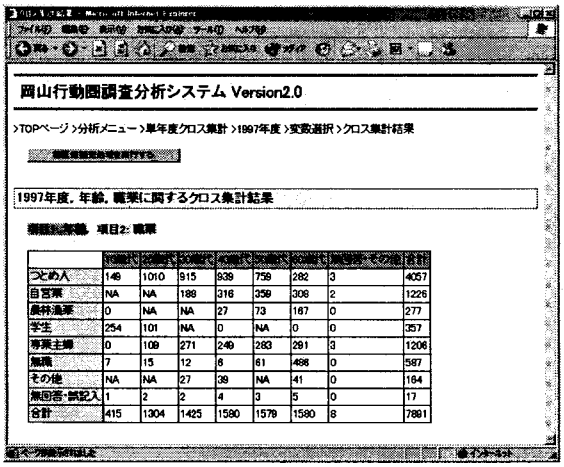


(a) 集計項目入力画面

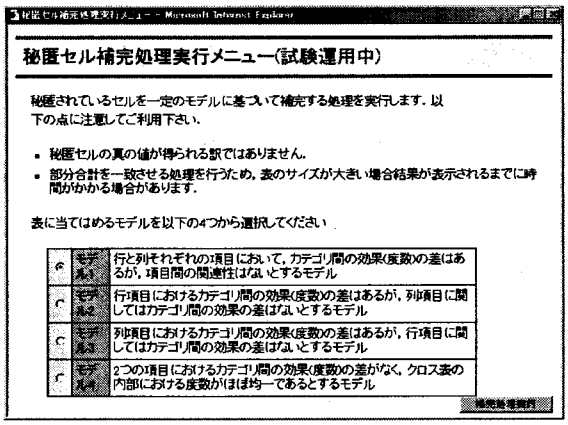


(b) 単純集計結果出力画面

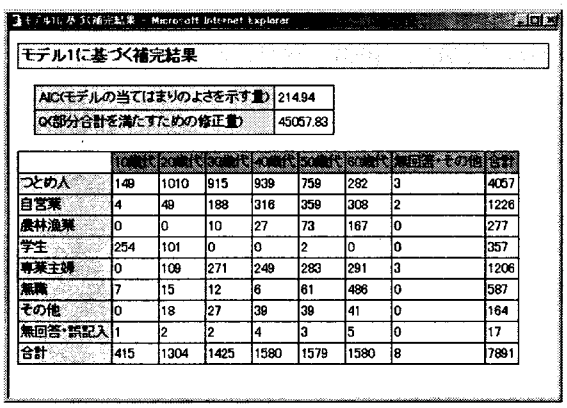
図 1: 単年度単純集計実行例



(a) クロス集計結果出力画面



(b) 補完処理モデル選択



(c) 補完結果出力画面

図 2: 単年度クロス集計実行例

わち

$$\hat{f}_{ij} = \frac{1}{N}\left(\sum_{i=1}^{k} f_{ij}\right)\left(\sum_{j=1}^{l} f_{ij}\right)$$

を表す．分析画面を図3に示す．

5.4 商圏解析・行動圏解析

岡山行動圏調査の調査結果を分析するにあたり，最も必要で有益であると思われる機能が商圏解析と行動圏解析機能である．商圏解析は，調査によって得られた15品目についての買い物場所の情報を使って，ある市町村が特定の商品の買い物に関して，岡山県内の各市町村からどれくらい利用されているかや，ある市町村の住民は，どの市町村を主に利用しているかということを調べるための機能である．行動圏解析は，通勤，利用する病院や交際に関して，同様の分析を行うものである．この2つが異なるメニューに分かれているのは，調査票において，買い物場所に関する選択肢のみ商店街単位になっているため，集計の際にそれらが市町村に変換されなければならないというシステム的な理由からである．

商圏・行動圏解析を行う際に利用する指標は「依存度」と「流出率」である．今，n_{ab} をA市の住民で，B市で行動した（通勤・通学先，商品の購入等）人数とする．このとき，$n_{ab}/\sum_b n_{ab}$ は，A市の住民である行動を取った人数に対する，B市で行動した人数の割合である．ある市町村Aを固定して，県内各市町村Bについてそれぞれこの割合を計算する場合には，これらを総称してA市からの県内各市町村への「流出率」と呼ぶ．これに対して，ある市町村Bを固定して，県内各市町村Aについてそれぞれこの割合を計算する場合には，これらを総称してB市への県内各市町村からの「依存度」と呼ぶ．ある市町村の流出率を調べると，その市町村の住民の行動傾向を知ることができ，依存度をを調べると，その市町村の持つ商圏や行動圏を知ることができる．図4(a)は，指標選択画面であり，依存度と流出率のいずれかを選択する．図4(b)は，対象とする市町村と行動を選択するフォームであり，選択内容を送信すると図4(c)のような結果が出力される．ここでは，1997年度における洋服の購入についての津山市への依存度が出力されている．この結果を視覚的に捉えるための機能として，依存度や流出率の塗り分け地図を出力するものが実装されている．これは，Rubyスクリプトにより記述されたCGIプログラムからC言語のサブルーチンライブラリであるPGPLOTをRuby/PGPLOTインターフェースを利用して呼び出すことにより実現されており，市町村コードとそれに対応した数値をプログラムに渡すと塗り分け地図を生成するという単独で動作する一つのプログラムである．地図の描画には，国土交通省が提供している国土数値情報[8]の行政界データを利用している．図4(d)がこれにより出力された塗り分け地図であり，津山市は，洋服の買い物に関して岡山県北東部の市町村から多く利用されていることが見て取れる．

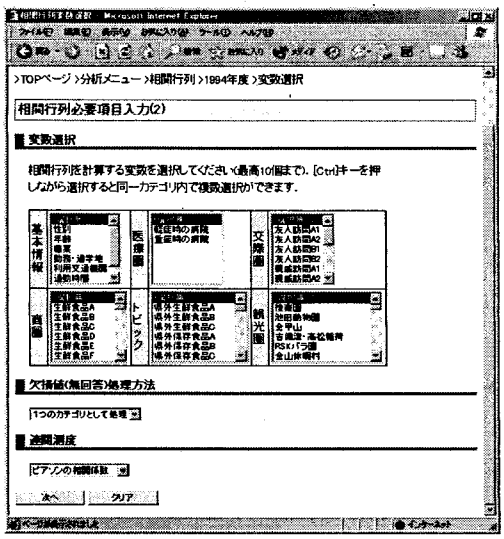
6 考察

本稿では，統計調査データの公開方法をめぐる問題の一つの解決策として，分析者の手元に真のデータがなくても真の分析結果が得られるというオンラインシステムの提案と紹介を行った．一般の民間による統計調査に対して，秘匿処理や特別な分析に関しては個別に検討しなければならない問題があるものの，基本的な分析については比較的容易に本システムと同様のシステムが実現可能であると考えられる（官庁統計に関しては法律上の問題もあり，このような形での公開は難しいかもしれないが，技術的には可能である）．

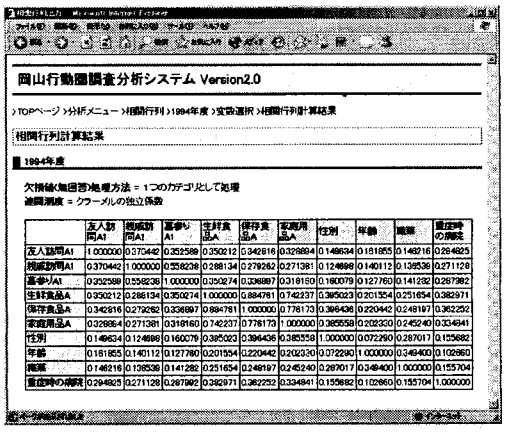
本システムは始めに述べた目的を実現するための一つの完成したシステムとなっているが，今後のシステムの改善・拡張や，他の統計調査データに対する同様のシステムの構築を行う場合に考慮しなければならない点について述べる．本システムでは，集計機能としては単純集計とクロス集計を実装しており，一部地域において標本抽出率が高くなるため，少数の個体しか含まないセルについては秘匿処理を行った．仮に調査対象地域が限定されてなく，標本抽出率が低い場合であっても，実装するクロス集計の次元が高ければ高いほど，個票を公開する状況に近くなり，集計によって得られた標本数1のセルに属する個体が母集団でも一意的な存在になっている可能性が生じる．このような場合には，集計機能に加えて秘匿処理機能も実装しなければならないだろう．このように，新たな機能を実装する場合には，それに伴って生じる個体識別や情報漏洩の可能性を考慮して，必要であれば秘匿処理機能の実装など，必要な措置を講じなければならない．

もう一つは，処理速度に関するシステム上の問題である．岡山行動圏調査における各年の標本数は7000～8000程度であるが，官庁統計をはじめとする大規模な統計調査では標本数が数万～数十万となること

[8] http://nlftp.mlit.go.jp/ksj/

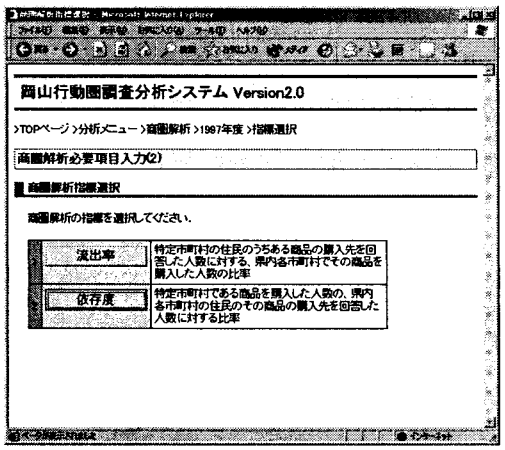


(a) 項目，オプション選択

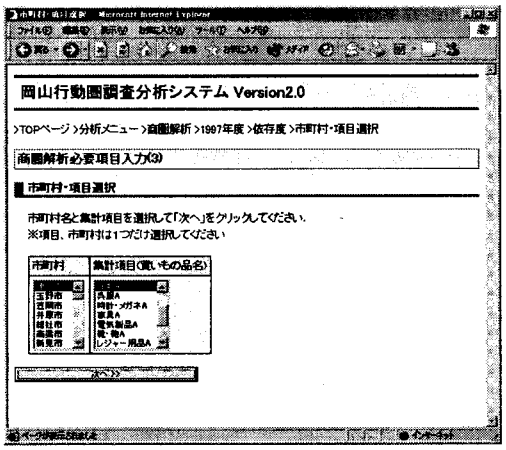


(b) 相関行列出力

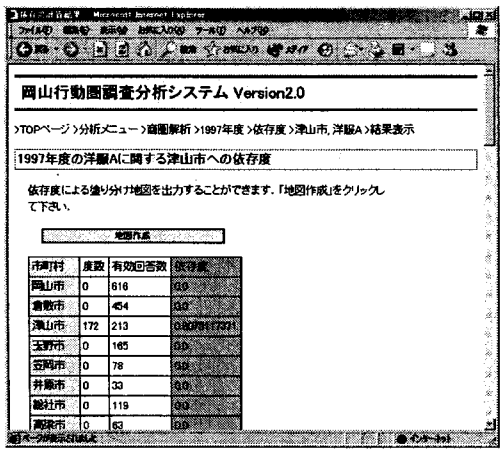
図 3: 相関行列実行例



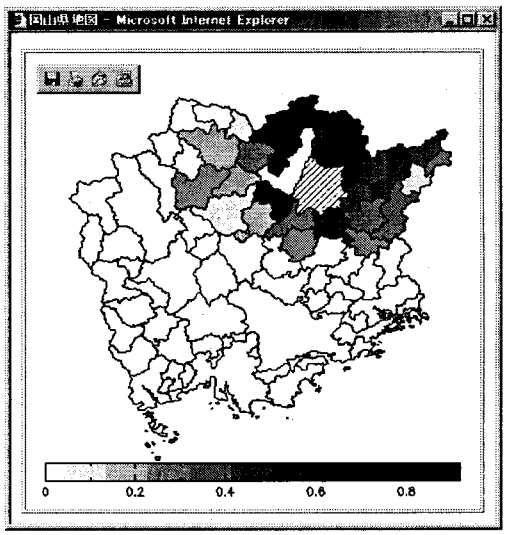
(a) 指標選択



(b) 項目選択



(c) 結果出力画面



(d) 塗り分け地図の出力

図 4: 商圏解析実行例

もある．クロス集計などの分析に必要な集計は，分析者の要求に従って動的に生成されるが，標本数が多ければ多いほどシステムにかかる負荷が大きくなり，処理速度も低下する．このような問題については，集計部分のみを単独の実行プログラムとして作成したり，分散処理を導入したり，ハードウェア的な解決を図るなどして解決されなければならない．また，集計処理だけでなく，複雑な分析機能を実装する場合にも同様のことが言えるだろう．

岡山行動圏調査分析システムの今後の課題は，数量化II類を実行する機能と経年変化を把握できる機能を実装することである．経年変化を把握できる機能は，ある項目の複数年にわたる集計を同時に出力して，その変化を調べるためのものであるが，岡山行動圏調査は各調査年度で質問項目や選択項目が異なる場合があり，この処理をどのように行うかが検討課題である．また，データベースやシステムの効率化は常に考えなければならない問題であろう．

参考文献

[1] Bishop,Y.M.M., Fienberg,S.E. & Holland,P.W. (1975), Discrete Multivariate Analysis: Theory and Practice, Cambridge, Massachusetts: MIT Press.

[2] Deming,W.E & Stephan,F.F. (1940), On a least squares adjustment of a sampled frequency table when expected marginal totals are known, Annals of Mathematical Statistics, **11**, 427–444.

[3] Stephan,F.F. (1942), An iterative method of adjusting sample frequency tables when expected marginal totals are known, Annals of Mathematical Statistics, **13**, 166–178.

[4] 稲葉由之・岩崎　学 (1997), 統計表における秘匿の補完法, 日本統計学会誌, **27**, 3, 263–280.

[5] 山本義郎・垂水共之 (1998), Web上の統計解析システムの構築－ CGIによる統計処理とグラフ描画の実装－, 計算機統計学, **11**, 1, 45–50.

[6] 藤野友和・飯塚誠也・垂水共之 (2001), アンケート調査データの収集と分析～岡山行動圏調査について，日本行動計量学会第29回大会発表論文抄録集，72–75.

[7] 藤野友和・垂水共之 (2002), 統計調査データのオンラインでの有効利用，2002年度統計関連学会連合大会講演報告集，142–143.

Appendix A 分割表における秘匿セルの補完処理

今，Iカテゴリを持つ項目AとJカテゴリを持つ項目Bに関する周辺和を伴う観測頻度を表す$(I+1)\times(J+1)$行列Yを$Y=(y_{ij})$とする．ただし，

$$
\begin{aligned}
y_{i+} &= y_{i(J+1)} = \sum_{j=1}^{J} y_{ij}, \qquad i = 1,\ldots,I \\
y_{+j} &= y_{(I+1)j} = \sum_{i=1}^{I} y_{ij}, \qquad j = 1,\ldots,J \\
n &= y_{(I+1)(J+1)} = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}
\end{aligned}
$$

である．同様に期待頻度を表す$(I+1)\times(J+1)$行列Mを$M=(m_{ij})$とする．また，セルが秘匿されているかどうかを示す指標として，次のような集合を定義する．

$$S = \{(i,j)|\ y_{ij}\text{が秘匿されている}\}$$

観測されている分割表$Y_{(I\times J)}$は，個体数nの多項母集団からの標本と仮定し，各セルの期待値$m_{ij}, i=1,\ldots,I, j=1,\ldots,J$に対して次の対数線形モデルを考える．

$$
\begin{aligned}
\text{model 1} &: \quad \log m_{ij} = \mu + \alpha_i + \beta_j \\
\text{model 2} &: \quad \log m_{ij} = \mu + \alpha_i \\
\text{model 3} &: \quad \log m_{ij} = \mu + \beta_j \\
\text{model 4} &: \quad \log m_{ij} = \mu
\end{aligned}
$$

ただし，μは固定効果，α_i, β_jはそれぞれ行i，列jの効果を表す．各モデルにおいて周辺和の期待度数の最尤推定値と周辺和の観測度数との間に次のような関係がある．

$$
\begin{aligned}
\text{model 1} &: \quad \hat{m}_{i+} = y_{i+}, \quad \hat{m}_{+j} = y_{+j} \\
\text{model 2} &: \quad \hat{m}_{i+} = y_{i+} \\
\text{model 3} &: \quad \hat{m}_{+j} = y_{+j}
\end{aligned}
$$

これにより，周辺和の観測度数に対して秘匿処理がされていないと仮定すると，各モデルにおいて各セルの期待度数の最尤推定値は次のように計算される．

$$
\begin{aligned}
\text{model 1} &: \quad \hat{m}_{ij} = \hat{m}_{i+}\hat{m}_{+j}/n \\
\text{model 2} &: \quad \hat{m}_{ij} = \hat{m}_{i+}/J \\
\text{model 3} &: \quad \hat{m}_{ij} = \hat{m}_{+j}/I \\
\text{model 4} &: \quad \hat{m}_{ij} = n/(IJ)
\end{aligned}
$$

ところで，秘匿されていないセルの度数が与えられたとき，秘匿セルの条件付期待度数は多項分布の性質により次のように与えられる．

$$\frac{m_{ij}}{\sum_{(i,j)\in S} m_{ij}} \left(n - \sum_{(i,j)\notin S} y_{ij} \right)$$

そこで，この式において秘匿セルの期待度数をその最尤推定値で置き換えた

$$\hat{y}_{ij} \equiv \frac{\hat{m}_{ij}}{\sum_{(i,j)\in S} \hat{m}_{ij}} \left(n - \sum_{(i,j)\notin S} y_{ij} \right), (i,j) \in S$$

は秘匿セルの度数の推定値のひとつとして考えられる．しかしながら，この推定値は一般には周辺和の観測度数に関する制約条件を満たさない．そこで，$\hat{y}_{ij}$ に対して，目的関数

$$Q = \sum_{(i,j)\in S} \frac{(y^*_{ij} - \hat{y}_{ij})^2}{\hat{y}_{ij}}$$

を次の等式制約条件と不等式制約条件

$$\begin{aligned} \sum_{j\in S_i} y^*_{ij} &= y_{i+} - \sum_{j\notin S_i} y_{ij}, \quad i = 1,\ldots,I \\ \sum_{i\in S_j} y^*_{ij} &= y_{+j} - \sum_{i\notin S_j} y_{ij}, \quad j = 1,\ldots,J \\ y^*_{ij} &\geq 0, \quad (i,j) \in S \end{aligned}$$

の下でで最小化するような修正値 y^*_{ij} を逐次２次計画法による最適化計算で求める．ここに，

$$\begin{aligned} S_i &= \{j|(i,j) \in S\}, \quad i = 1,\ldots,I \\ S_j &= \{i|(i,j) \in S\}, \quad j = 1,\ldots,J \end{aligned}$$

であり，それぞれ各行，各列における秘匿セルのインデックスを表す．

以上のような手順で，各モデルに基づいた周辺和の制約条件を満たす秘匿セルの推定値が得られることになるが，一般的には，仮定するモデルが異なると補完されるセルの値も異なる．これらの評価基準としてはAICと，秘匿セルの推定値を周辺和の制約条件に適合させる際に最小化したカイ２乗統計量 Q を考えることができる．ここで考えるAICは，秘匿されていない観測セルに対するモデルの当てはまりの良さを示すものであり，次式により計算される．

$$AIC = -\sum_{(i,j)\notin S} y_{ij} \log \hat{m}_{ij} + 2(\text{パラメータ数})$$

Q については，モデルに基づく秘匿セルの推定値と周辺和に関する制約条件との乖離を示す量と考えることができるので，これもモデル選択の一つの基準となる．これらの評価基準が同一のモデルを選択した場合には，そのモデルにより得られた秘匿セルの推定値を最良の補完値として採用する．しかしながら，これらの基準が異なるモデルを選択した場合には，どちらの基準を優先させるかは一概には定めることはできない．