

札幌市における患者動態調査に関する データ処理ソフトウェアの開発 I

宮 腰 昭 男

Abstract

In this paper we intermediately report the basic idea on the integrated data processing system for the patient behavior research in Sapporo urban area. This system is constructed by the three hierarchical subsystems. The first is the descriptive statistical data processing subsystem, the second the analytical subsystem for the specific purpose, the last the figure processing one for the analytical results or statistical information given by other subsystems. In second subsystem, we developed the Algorithm on the identifying medical regions using the hierarchical clustering. This improvemental algorithm is performed on M685 hardware (SORD 32bits microcomputer) with C Programming language.

In the figure processing subsystem, we mainly explain the processing function and data structure describing the urban area information. Finally, it is shown that the micro computer resources have a high performance in the second and third subsystem in spite of interface handicap.

I. はじめに

本調査および調査結果のデータ処理は、社団法人札幌市医師会の依頼を受け行われたものである。依頼者側との数度に亘る打ち合せの中で本調査の目的は、札幌市を中心とした地域医療システムの確立のための基礎データの収集と解析であることが明確となった。さらに、今回の調査のより具体的目的は、

- (1) 患者がどのような疾病で、どこの医療機関を受診しているかという患者の受診状態を地域的に捉え、把握すること。
- (2) 大域的診療圏の設定のための基礎解析
- (3) 地域医療サービスを前提とした病院における適正病床数の設定のための基礎解析
- (4) その他

等である。日医医療システム研究委員、札幌市医師会の飯塚弘志氏が北海道医報第 506⁽¹⁾号の中で“真に適正な医療システム化へ向けて医師が主体性を持ち、官僚主導型でなく、医師会主導型のものを確立してこそ、地域住民の本当の幸福につながるものである”と述べている様に、医療機関側が地域医療サービスという側面を重視し、対応する医療システムの構築のためにマクロ的な患者の受診動態の状況を把握しようとする意義は大きいと思われる。

今回の調査は医師会所属の医療機関を中心としたほぼ札幌市全域にまたがる全数調査を行っている。従って、測定結果の解析は記述統計的手法に重点が置かれている。我々はデータ処理の過程の中で、イ). 素データを中心とした各種データの汎用的で柔軟な連結およびアクセスが行える

データファイルの確立。ロ)。素データを加工し、多様に分類する統計的ツールの開発。ハ)。処理結果の一般ユーザー向けへの集約的加工（例えば図形的出力等をも含む）。ニ)。ハ) に関してのマイクロ・コンピュータの利用等の必要性を痛感した。イ)、ロ) については、データ構造の決定、データファイルへのアクセス法、およびツールの作成等には莫大なソフトウェア開発の労力を必要とするため、既存のソフトウェアを利用することとし、今後の検討課題とした。

本論文では、ハ)、ニ)を中心としたデータ処理用ソフトウェアの開発およびその考え方について述べる。特にハ) については、一般的に統計処理結果はその書式が複雑で結果のボリュームも多い。従って、一般ユーザはその結果を理解するために多大の労力を必要とし、大半の統計情報を読み取らずに死滅させる場合が多い。ユーザが要求する情報は、厳密な誤差範囲等を記述した生の統計結果ではなく、情報から容易に読みとれる全体的傾向（トレンド）や分類した全体の中での比率、平均値、分散等の基本的情報が大半である。特に、記述的統計の場合にはこの性質が強い。故に、統計処理結果の持つ情報を単純で容易に理解できる様な情報に加工することが重要となる。このためには、図形的処理を行った Visual な情報の提供が有力な手段となる。もう一つのハ)に関連した重要な考え方は、情報提供の過程での Interactive 性である。記述統計的データ処理のプロセスでは、多種多様な分類、整理等の処理が必要となり、ユーザは中間的統計処理結果を観測しながら次の段階での分類対象データの抽出や整理法、出力書式を決定することになる。従って、処理結果の集約的加工の段階で見ても、ユーザとの間の双方向的情報の受けわたし機能を持つソフトウェアの設計が必要となる。ユーザ側からの要求情報は簡易なコマンドや Inquiry 情報のみでなければならず、計算機側からの出力情報は集約的で図形的データが最適と思われる。以上のように、統計処理および出力情報提供の段階での会話的機能は、統計処理の効率を向上させ、比較的少ない労力での確で多様な解析結果を提供する長所を持つと思われる。

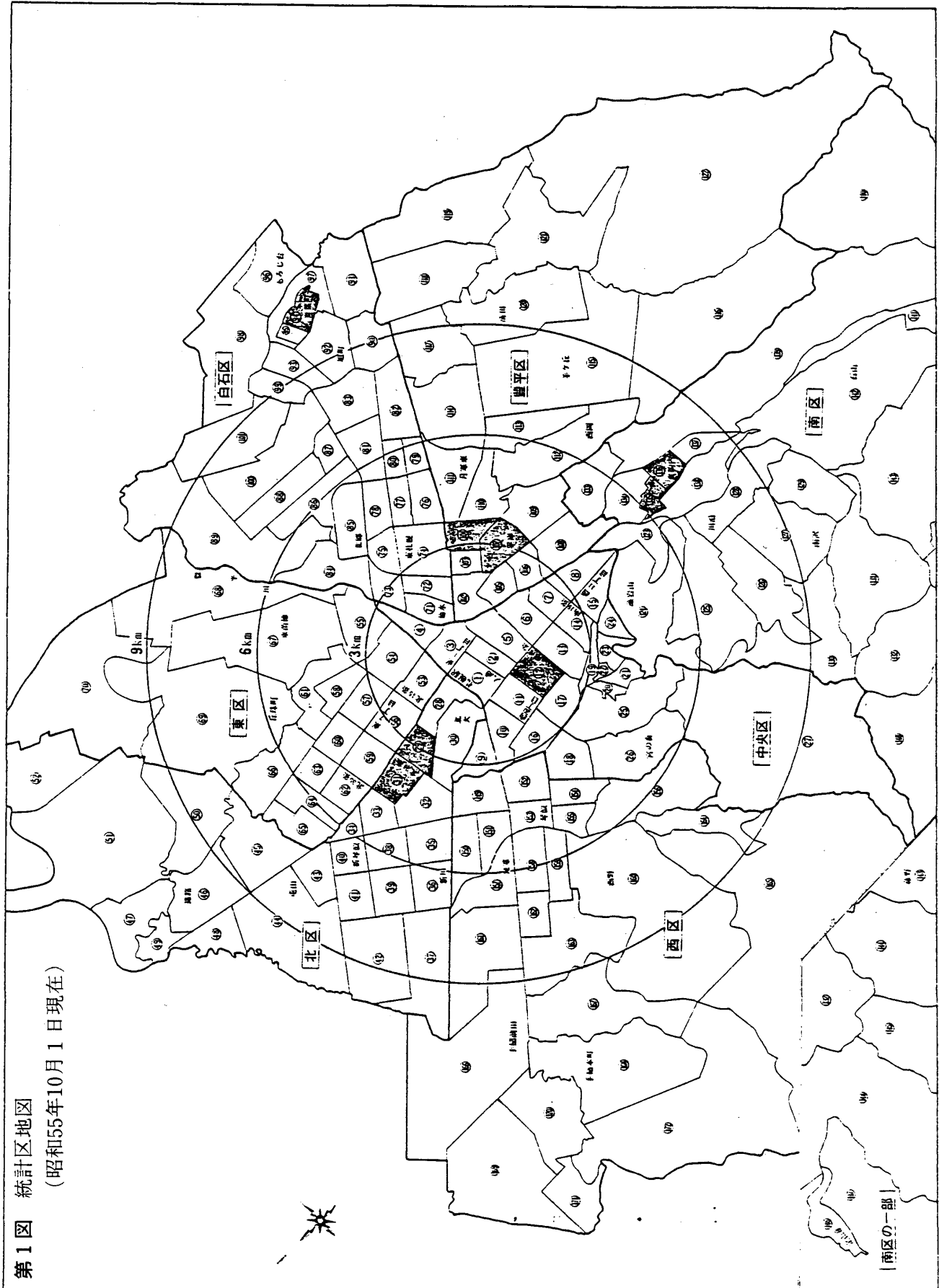
ニ) は、ハ) に関連して統計処理結果の加工と図形的出力をマイクロ・コンピュータで行おうという試みである。素データや統計解析ツール・ソフトウェアはマイクロ・コンピュータのハードウェア的制約によって対象とならないが、中規模程度の処理結果の出力レベルでの加工等には、十分な処理能力を持っている。加えて、大型計算機の TSS のように不用な機能を学習しなければならないという障害もなく、身近で気軽にデータ加工が行えるのは大きな魅力である。

以上のように、我々は統計解析の処理結果をユーザの要求するような形態での確で集約的に提供するためにはどのような基本的考え方がソフトウェア開発上必要なのか、および、そのためのマイクロ・コンピュータ・リソースの利用可能性等について具体的なシステムを構築する過程の中の考察を中間的に報告するものである。

なお、現在、患者動態調査データの解析は継続されているが、社団法人札幌市医師会との関係で解析結果の公表は現時点では行えない。

II. 調査およびデータ構造について

本調査は社団法人札幌市医師会に所属する医療機関を中心に他の国公立病院（例えば、札幌市立病院等）の協力を得、昭和 57 年 11 月 18 日（金）に行われた。調査期間は当日のみとし、患者動態に関しては、当日の外来患者、紹介患者についての調査情報をすべての患者について医療機関側が測定することとした。入・退院患者に関しては、当日より 1 ヶ月遡った期間中のすべての対象患者についての測定を行うこととした。他方、医療機関側の測定情報として個々の機関の医療情報（例えば、機関種別、経営主体、診療科目、医師数、その他）を測定している。回収は同年 11 月 25 日を〆切日として郵送で行われ、80%以上の高率となり、外、紹、入院患者数は約 85、



第1図 統計区地図
(昭和55年10月1日現在)

000 ケースである。このうち患者データに関する欠損率は約 5 % 程度である(機関ファイルの欠損率も同様に 5 % 程度である)。

素データ・ファイルは大きく二つに分類され、一つは患者ファイルであり、これらは更に患者属性により、外来患者ファイル、入院患者ファイル、紹介患者ファイルに分割されている。他方は機関ファイルであり、このファイルはケース数は少ないがデータ長が比較的長いファイルであり、患者ファイルとはデータ構造が異なっている。これらの素データ・ファイルはサブファイル化が可能であるが、データ処理の都合上構造化されておらず、各々が独立した単ファイルであり、順編成ファイルである。患者ファイルは約 30 byte のデータ長を持ち、機関ファイルは約 70 byte 程度のレコード長であり、各変数フィールドには数値データのみが記入される。患者ファイルでの変数は約 18 であり、機関ファイルでは 32 変数程度である。各ファイルのデータ構造については詳細に記述できないが、患者ファイルでは、患者 1 人に対して、何如なる居住区よりどこの医療機関へどの様な傷病名で移動しているかが主な情報として測定されている。今回の動態調査では、測定者が医療機関側であり、各患者が応答していないため測定者が大量の測定記述を行わなければならない、データ構造をあまり複雑にできなかった。従って、患者ファイルでは調査目的をを絞り込み患者の移動形態に重点をおくこととしている。

患者の移動情報を測定する場合、この情報を詳細に記述しようとする膨大な長さのコードを必要とする。我々は地理的移動情報の最小精度を札幌市内において国勢調査の人口統計区 172 によって構成した。札幌市以外の流入患者については、特に北海道内の場合、各支庁の市町レベルまでを限界精度としている。172 人口統計区は各国勢調査で小規模の変動が存在するが、比較的固定的であり、国勢調査結果データとの比較検討を行えるという長所を持つ。また、これらの統計区は行政区別や支部別のクラスタ化が行い易く、境界線上に住居区が存在していない。特に統計区を選択した最大の理由は、これ以上の精度によって地域設定を行った場合、データ処理迄に長い時間を必要とすることになり、出力書式や図形的処理が複雑で不能になると思えたからである(図 1 の統計区地図を参照)。医療機関ファイルの各機関の所在地も同様にこの人口統計区で表現している。

III. 流出入患者マトリックス

素データのデータ・クリーニングの段階での考え方については省略するが、基本的には測定されたデータは出来るだけ利用する様に、クリーニング処理レベルで発生した欠損データは先のアンケート調査票に戻り修正し挿入することとした。この Verify 段階でのデータ処理は極めて長い時間を要するため、アンケート票よりの転記およびコード化記述プロセスでの厳重なチェックが必要である。

ここで述べるデータ処理システムは全体システムのサブ・システムである診療圏設定のためのそれであり、I. の (3) については未だ考察されていない。

診療圏設定のためのデータ処理サブシステムは以下の様な処理目的を持つ。

- (a) 各医療機関に対する診療圏を考察するのではなく、大域的に札幌市全域でどの様な動態傾向を持つかを解析する。
- (b) (a) を基礎とし、複数個の統計区をクラスタ化した様な形で、患者の移動行動範囲の閉じている地域が存在するかを検討する。
- (c) (b) に関連して若しその様な Region が存在するならば、なぜ閉じるのかを医療機関側および地域データ(交通網等)よりそのメカニズムを解析する。
- (d) 患者の移動距離を設定し、分類化する。

(e) 診療所および病院に分割し、外来と入院患者の動態を解析する。特に、病院に重点をおく。

(f) 処理結果の統計区レベルでの図形的出力ソフトウェアの開発。

診療圏に関連したデータ解析において本質的に重要なデータ構造が存在し、それらは、統計区別の患者流出データおよび患者流入データである。

患者流入データは以下の様な構造を持つ。

任意の統計区 j (≤ 172) に対して、

j ;	1	2	3		i		i		172	173	174	175
	P_{1j}	P_{2j}	P_{3j}		P_{ij}		P_{jj}				$P_{174,j}$	$P_{175,j}$

となる。上記テーブルにおいて i は 1 より 175 までの値を持つ。1~172 は人口統計区に対応するが、173, 174, 175 はダミー統計区であり、各々、市内不明分と道内および道外を表わすものとする。同テーブルは、 j なる統計区に存在する医療機関に対して各統計区 (1~175) から流入する患者数を P_{ij} によって与えている。各統計区に対して流入患者数を見るのであるから同様のテーブルは 172 個存在する (但し、医療機関の存在しない統計区では $P_{ij} = 0$, $i = 1, 175$ となる)。

同テーブルにおいて、

$$I_j (j \text{ なる統計区へ流入する全患者数}) = \sum_{i=1}^{175} P_{ij} \dots \dots \dots (1)$$

となり、 P_{jj} は自統計区内での発生患者のうちで自統計区内で医療サービスを受ける患者数であるから、

$$GI_j (j \text{ なる統計区への他統計区より流入する全患者数}) \\ = \sum_{i=1}^{175} P_{ij} - P_{jj} = I_j - P_{jj} \dots \dots \dots (2)$$

となる。(1) 式の I_j を受診者数と呼び、自給率を S_j とし、他統計区の j 統計区に対する依存比率を D_j とすると、

$$S_j = P_{jj}/I_j \times 100 \quad (\%) \dots \dots \dots (3)$$

$$D_j = GI_j/P_{jj} \dots \dots \dots (4)$$

となる。ここで、(3), (4) 式において I_j および P_{jj} が零となる統計区については計算をしないこととする。

理想的な状態で考えて、 S_j が極めて高く、 D_j が低い比率を持てば対応する統計区に存在する医療機関は、高い医療サービスを行っていることとなり、一つの閉じた診療圏は設定可能な統計区となる。しかし、現実には単独の統計区ではこの様な状態は存在せず、複数個の統計区の clustering によって閉じる傾向を持つと思われる。

つぎに、流出患者データについて述べる。

流出患者データは流入患者データと同様の構造を持っていて、つぎのようになる。

i ;	1	2	3		i		j		171	172
	Q_{i1}	Q_{i2}	Q_{i3}		Q_{ii}		Q_{ij}		Q_{i171}	Q_{i172}

ここで、 $i = 1, 175$ となり、 $j = 1, 172$ となる。同テーブルは、統計区 i で発生した患者の各統計区 (≤ 172) への分布 (Distribution) を示している。 i 統計区的全患者発生数を O_i とすると

$$O_i = \sum_{j=1}^{172} Q_{ij} \dots \dots \dots (5)$$

となり、 Q_{ii} ($=P_{ii}$) は自統計区に残留する患者数であり残留度 R_i は、

$$R_i = Q_{ii}/O_i = P_{ii}/O_i \times 100 \quad (\%) \dots \dots \dots (6)$$

となる。流出度 T_i および流出比率 E_i は、

$$T_i = (O_i - Q_{ii}) / O_i = \left(1 - \frac{Q_{ii}}{O_i}\right) \times 100 \quad (\%) \quad \dots\dots\dots (7)$$

$$E_i = (O_i - Q_{ii}) / Q_{ii} = \frac{O_i}{Q_{ii}} - 1 \quad \dots\dots\dots (8)$$

となる。(7) 式と (8) 式より

$$T_i = \frac{E_i}{1 + E_i} \times 100 \quad (\%) \quad \dots\dots\dots (9)$$

なる関係が存在する。(6), (7), (8)式において $O_i = 0$, 即ち, 患者発生数の零の統計区は除外する。 R_i が高く, T_i や E_i が低い統計区は医療サービス上自立的な統計区と見ることができる。これらの指標は (3), (4) 式の S_j, P_j と相補性を持っている。

上記の二つのデータ構造において, $i = 1, 172$ および $j = 1, 172$ なるデータ・ボリュームに対してデータは閉じていて且つ保存する。加えて, $i = j$ に対して $P_{ij} = Q_{ij}$ (但し, $i, j \leq 172$) となる。これより K なる流出入患者マトリックスを作成する。 K は 175×172 なるサイズを持ち, 非正方, 非対称マトリックスとなる。

上記のマトリックスを作成するためには, 各統計区に対する流出又は流入数を統計区毎にクロス集計すれば良い。特徴的なことは, 流出のみのクロス集計をすれば (10) 式のマトリックスではデータが保守的であるため自動的に流入患者数も求められることである。(10) 式では, $Q_{ii} = P_{ii}, Q_{jj} = P_{jj}$ ($i, j \leq 172$) であり, 同マトリックス上で i を固定し横 (Row) に見た場合, 流出分布が与えられ, j を固定し縦 (Column) には流入度数が与えられることになる。

流出入マトリックス K は, 診療所および病院に対して各々, 外来患者用, 入院患者用, 紹介を除く全患者用の合計 6 枚が作成される。 K は診療圏解析のための基本データ・マトリックスであり, これらの 6 枚のマトリックスは多様なアクセスと加工処理を可能とする柔軟なデータの構造を持たなければならないと思われる。

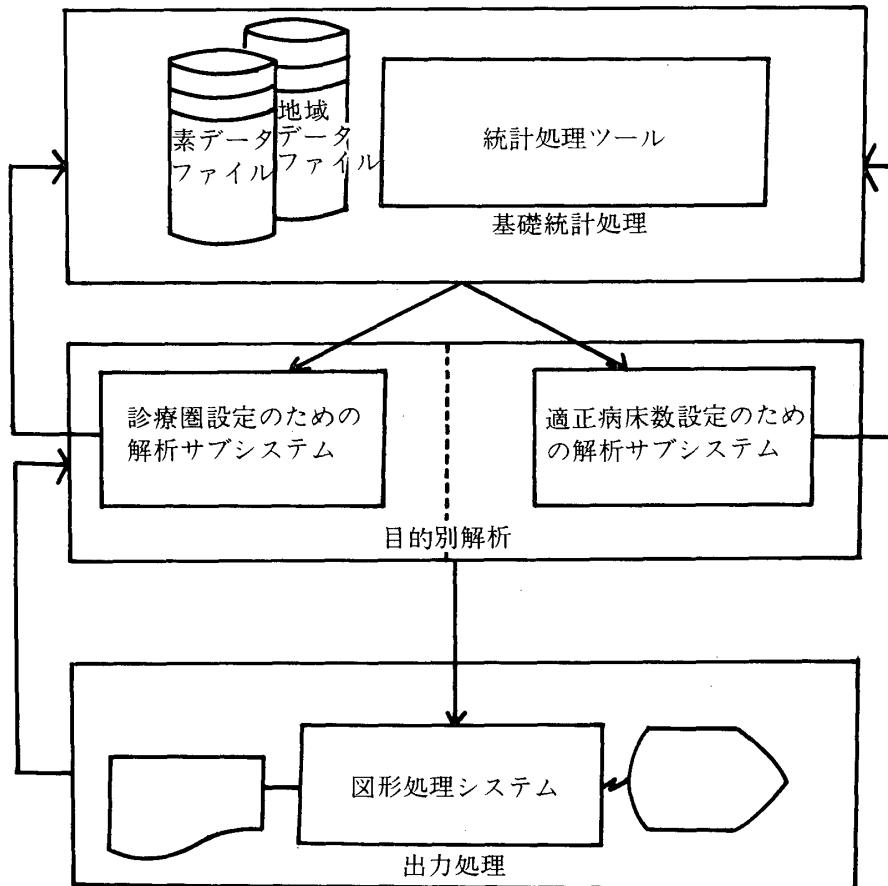
	j																						
$i \backslash$	j	1	2	3	4 i											j 172					
1	Q_{11}	Q_{12}	Q_{13}	Q_{14}												Q_{1j}							
2																							
3																							
⋮																							
i	Q_{i1}	Q_{i2}	Q_{i3}														Q_{ii}						
⋮																							
j																	Q_{jj}						
⋮																							
174																							
175																							

.....(10)

IV データ処理システム

本システムを構築する場合, その解析目的である診療圏設定や地域環境を踏まえた適正病床数

の設定に対してデータ解析上のプロトタイプが存在しないことを考慮しなければならない。例えば、診療圏解析においても患者の移動情報を把握しなければならないことは直感的に理解できるが、これらの情報をどのように解析し所望の解析結果を得るかというアルゴリズムが存在しない。加えて、アルゴリズムが存在したとしても、その解析結果が現実の患者動態を踏まえてどの程度の精度で診療圏情報を記述しているかが問題となる。従って解析とデータ処理は試行錯誤的となり、常に素データレベルへ戻り、分類・整理・抽出および統計解析が行われなければならないことになる。適正病床数の算定は厚生省において一定の算定方式が存在し画一的に計算されるが、この方式は札幌市およびその周辺地域の一次、二次医療を含めた地域医療サービスの特殊性を考慮していない。この特殊性や固有性は今回の患者動態調査や地域データの比較等を通して一定の妥当性のもとに抽出されなければならないが、この抽出法は存在していないのである。



第2図 データ処理システムの全体的構成

第2図に本システムの全体的な構造が記述されているが、患者動態調査の素データファイルと地域データファイルが存在し、これらのデータは統計処理用アプリケーション・プログラムによって解析処理し、I.の(2)、(3)のためのデータとして加工する。これらの結果データは診療圏設定や病床数解析のためのアルゴリズムに対して一つの素データ・ファイル的構造を待つ。解析アルゴリズムによって処理された結果情報は更に加工され、図形データに集約される。解析者はこれらの結果を観測して現実的要請にある精度で符合するかを判断し、統計処理レベルへ戻り再び統計処理を行うか、または、アルゴリズムレベルへ戻り入力パラメータの変更やアルゴリズムそのものの改善を行うことになる。

本章では、処理システムの中で診療圏設定のためのサブシステムを中心に述べる。

同サブシステムは、IIIで述べた複数個のKマトリックスを中心に(1)式から(9)式までの各指標データおよび札幌市地域構造データ(例えば、統計区人口、世帯数、年齢構成、性比、面積等)からなる加工データ群を内包している。これらのファイルは、172又は175からなる統計区番号を基本キーとして持っている。サブシステムの中核は、Kマトリックスを加工しCluster化するアルゴリズムであり、他の二つのファイルは参照ファイルとしても利用される

IV-1. 階層的クラスタリングのアルゴリズム⁽³⁾

本アルゴリズムはクラスタリングにグラフ論的考え方を利用した技法でありSlater⁽²⁾の考え方に基づいている。

クラスタリングの基本行列は上記のKマトリックスであり、以下の様に再定義する((10)式ではKの要素は Q_{ij} となっている)。

$$K = \{k_{ij}\}, k_{ij} \geq 0 \quad (i, j = 1, \dots, n) \quad (11)$$

ここで、(10)式の i については175までの値をとっているが、(11)式では $n=172$ を上限値としてKを正方化する。即ち、流出のうち市内不明、道内(札幌市以外)、道外分は考慮しないこととする。

(11)式の行列Kを正規化した行列を

$$P = \{p_{ij}\}, p_{ij} \geq 0 \quad (i, j = 1, \dots, n) \quad (12)$$

とする。正規化とは、Pのすべての行和と列和がある数に等しくなるように各要素をスケールングすることである。

定義1. 行列Pの行(列)のインデックスを有向グラフのノードと考えて、 $p_{ij} \geq t \geq 0$ が満たされた時、その時に限り、ノード i からノード j への有向枝を定義する。この有向グラフを、 t における有向グラフという。

さらに、有向グラフについて次の定義を行う。

定義2. ノード i からノード j の向きで有向枝が存在する時 $a_{ij} = 1$ 、それ以外の時 $a_{ij} = 0$ とした二値列 $A = \{a_{ij}\}$ を、その有向グラフの隣接行列という。

定義3. ノード i からノード j が一連の有向枝で結ばれている時、ノード i からノード j は可到達であるといい、この時 $a^*_{ij} = 1$ 、それ以外の時 $a^*_{ij} = 0$ とした二値行列 $A^* = \{a^*_{ij}\}$ を、その有向グラフの可到達行列という。また、各ノードは自分自身に対して可到達であるとする。

定義4. グラフの有向枝の方向を無視した上で、ノード i とノード j の間が一連の枝で結ばれている時、ノード i とノード j は弱連結な関係にあるという。

定義5. ノード i とノード j が相互に可到達である時、ノード i とノード j とは強連結の関係にあるという。

定義6. ノード k から可到達であるノード集合を k の上位集合、ノード k に可到達であるノード集合を k の下位集合という。

定義7. M 次元正方形行列 $M_1 = \{m^1_{ij}\}$ 、 $M_2 = \{m^2_{ij}\}$ において、すべての i, j に対して、 $m^1_{ij} \leq m^2_{ij}$ となる時、 $M_1 \leq M_2$ と定義する。

上記の定義をもとに以下にアルゴリズムを述べるが、アルゴリズムの概略は第3図に示すように行列Pに対してあるしきい値 t における有向グラフを決定し、グラフ中の強連結成分を捜してそのノード群を一つのノードで代表させる。最後に、残ったノード群から弱連結な関係にあるノード群を捜して一つのクラスタにする。あとは、しきい値 $t (\geq 0)$ を順次変化させて、各々の t に対して上述の処理を行って樹状図を作る。また、強連結成分を一つのクラスタとすることもできる。

更に第3図のアルゴリズムの Stepwise の構成は以下のようになる。

Step.1 しきい値 t の設定

Step.2 有向グラフの決定

(i) 隣接行列 $A_t = \{a_{tij}\}$ の決定

```
for i=1 to n
  for j=1 to n
    if  $p_{ij} \geq t$  then  $a_{tij} = 1$  else  $a_{tij} = 0$ 
```

(ii) A_t の対角要素を 1 とする。

(iii) 可到達行列 $A^*_t = \{a^*_{tij}\}$ の決定

```
for i=1 to n
  for j=1 to n
    if  $a_{tij} = 1$  then  $A_t$  の  $j$  行を  $A_t$  の  $i$  行へたし込む。
```

Step 3. 強連結成分の探索

```
for i=1 to n
  if ノード  $i$  が未探索 then
    for j=i+1 to n
      if  $a^*_{tij} = 1$  かつ  $a^*_{tji} = 1$  then
        ノード  $j$  をノード  $i$  が代表する強連結集合に入れる。
        ノード  $j$  を探索済み集合へ入れる。
```

Step 4. 弱連結成分の探索

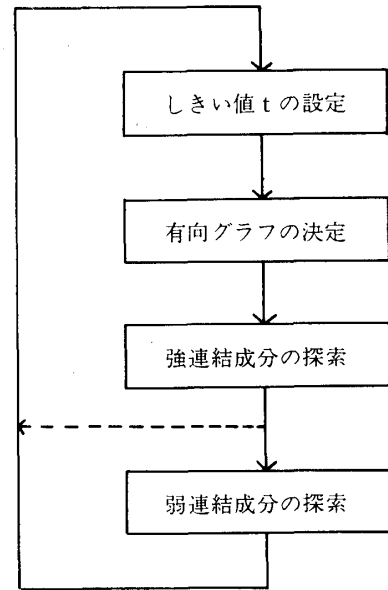
```
for i=1 to n
  if ノード  $i$  が未探索 then 関数 search ( ) を呼ぶ。
```

関数 Search (i)

- (i) ノード i の上位集合に属し、未探索なノードをすべて探し出す。
- (ii) もしそのようなノードが存在するならば、そのノードをノード i の弱連結集合へ入れ、探索済み集合へ入れる。
- (iii) (ii) のノードについて、関数 search () を呼ぶ。
- (iv) ノード i の下位集合に属し、未探索なノードをすべて探し出す。
- (v) (ii), (iii) と同じ処理をする。

しきい値 t における隣接行列 A_t を決定することは、しきい値 t における有向グラフを決定することと同値である。従って、Step (2) (j) のようにして A_t を決定する。Step (2) 終了時に A_t は可到達行列 A^*_t になっている。なお、Step 2 における演算は論理演算である。つぎに、強連結成分は相互に可到達なノードであるから、 A^*_t においては、対角成分に対して対称な成分が共に 1 になっている。Step 3 では、これを利用して強連結成分を捜し出している。Step 4 で search () という関数が再帰的に定義されているのは、あるノードの弱連結成分の中には、その上位集合にも下位集合にも属さないものがあるためである。また、Step 2 (iii) は Warshall のアルゴリズムで、その処理は高速である。さらにつぎのような考え方を導入するとより高速化が可能となる。今、しきい値 t_1, t_2 が、 $t_1 > t_2$ ならば $A_{t_1} \leq A_{t_2}$ が成立する。従って、しきい値をこのように変化させて、 A_{t_1} の変化した要素についてのみ Warshall のアルゴリズムを適用して、 $A^*_{t_2}$ を得る。

クラスタリングの結果は A^*_t マトリックスによって直接見ることができ、しきい値 t ごとにその時のクラスタを統計区地図上に重ね合せ観測した方が即時的に判断しやすい (統計区地図作成のためのソフトウェアについては IV-2. で説明する)。本アルゴリズムでは、一つのしきい値



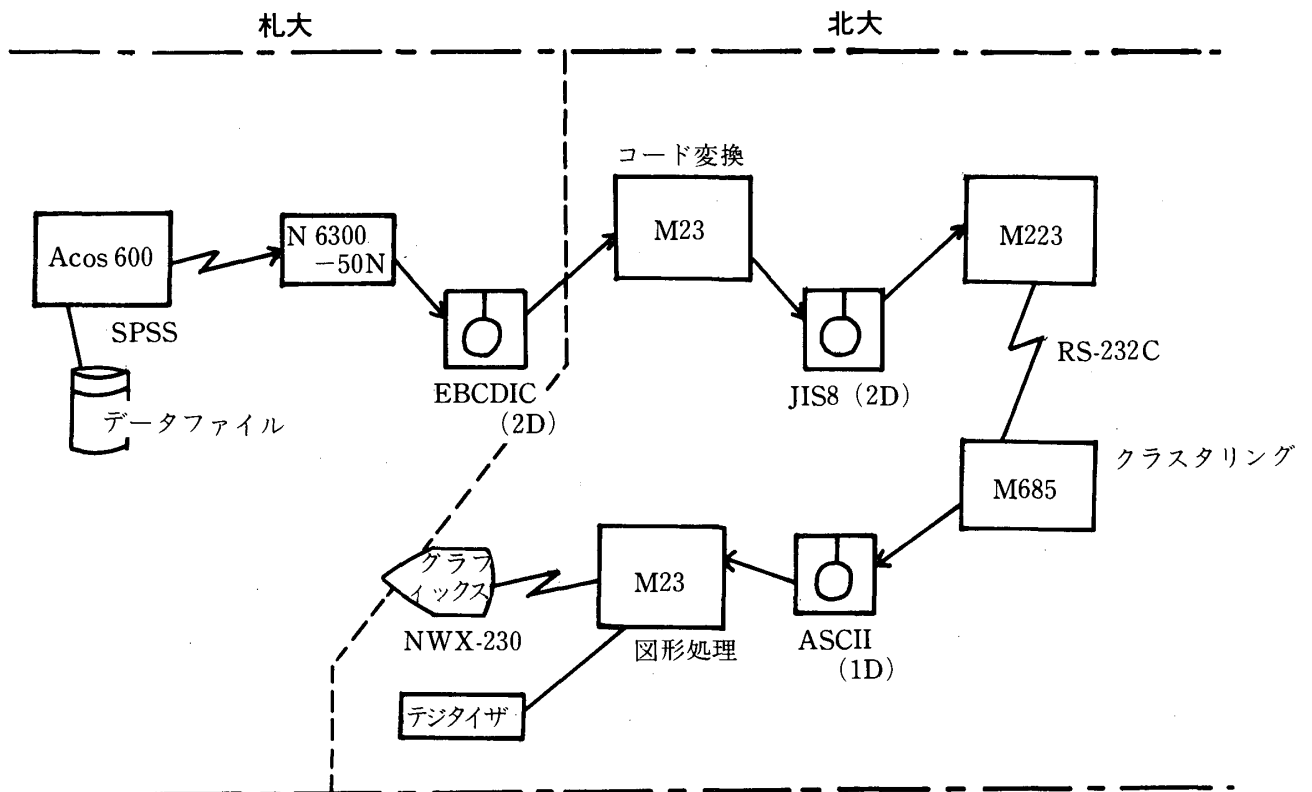
第3図 アルゴリズムの概要

t に対して一つのクラスター群が構成される。従って、ある精度や外的な条件によって診療圏設定のための基準が与えられている場合には、 t を変えることによってこの基準を満たしうるクラスターを会話的に行ない、試行錯誤的に求めることになる。処理結果の可視的出力化はこうした探索の効率を向上させるうえで重要な考え方である。本アルゴリズムの会話的 (Interactive) 機能は、しきい値 t の設定が中心であるが、この他に、いくつかの統計区を併合したり、行政区別に動態状況を見るための問い合わせ機能なども付加することができる。

他のファイルは、クラスター化された統計区内での人口構成や性別別データ、および受診者数、受診率等の値を求めるために参照される。クラスター化された統計区群で自給度が高く、流出度が低くなれば、ある程度一つの診療圏が設定のための基礎データが抽出されたと考えて良い。

IV-2. 具体的データ処理システムの構成

患者動態調査結果の記述統計を中心としたデータ処理は、本学 Acos-600 システム (日本電気) におけるアプリケーション・パッケージ SPSS で行われている。診療圏設定のための IV-1. のアルゴリズムでは K マトリックス ((11) 式) が入力データの中心となるため SPSS 処理ではクロス集計が行われる。これらの結果は、Acos-600 のパーマネントファイル上に吸い上げられ、N6300-50N (Acos-600 にオンラインで持続しているインテリジェント端末) を経由し、IBM 形式の 2 D の FDD へ落とす。以後のデータ処理は北大工学部電気工学科系統講座において行われる。IBM 形式の FDD は SORD の M23 により 2 D の JIS 8 単位コードに変換され FDD へ落とす。ついでこのデータ媒体は、データ入力用の SORD, M223 により RS-232 C 経由で M685 (SORD 社の 32 ビット・マシン) のディスク上に移されることになる。M685 では IV-1. で説明したクラスタリング・アルゴリズムが C 言語でインプリメントされているため入力用の K マトリックスが処理され、与えられたしきい値ごとのクラスターがファイルに書き込まれていく。こ



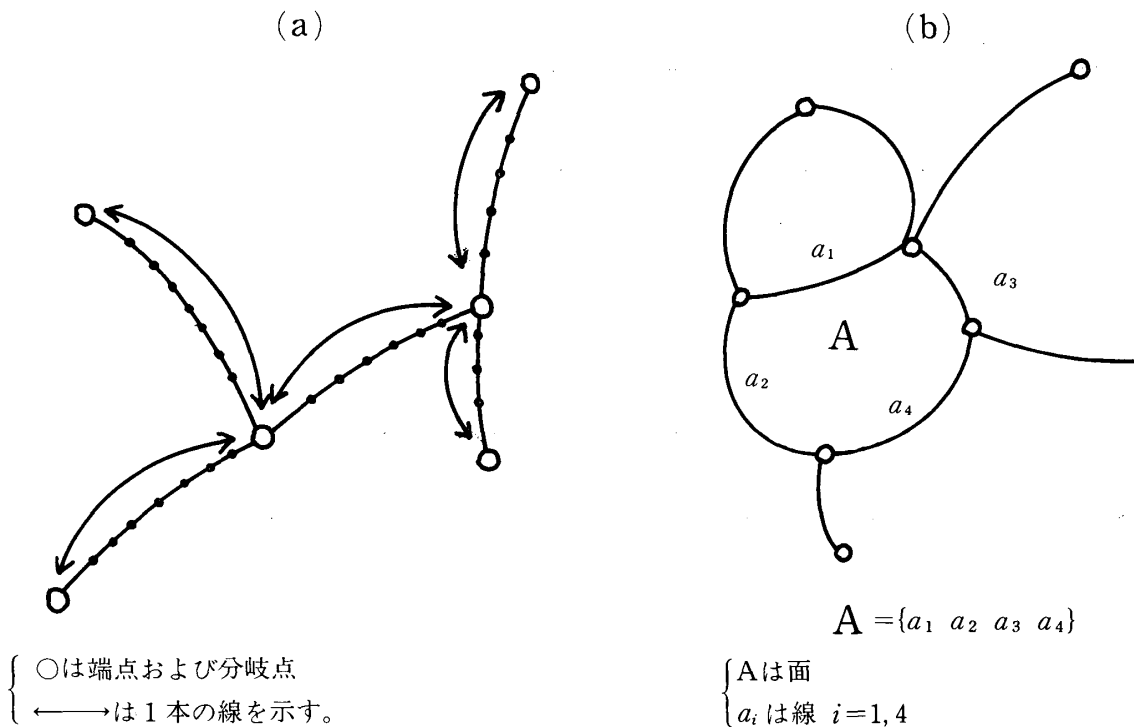
第 4 図 診療圏設定のためのデータ処理サブシステムの概要

のファイルはつぎの図形処理用入力データとなるため CP/M ベースの 1 D の FDD に落とされ、M 23 で処理されることになる。M 23 では統計区地図作成用のデータファイルや簡単な統計処理ツールが用意されていて、地図情報や処理結果は M 23 で制御されるグラフィックス端末 NWX-230 上にディスプレイされることになる (第 4 図参照)。以上の様に、北大側での大きなデータ処理機能は二つ存在し、一つはクラスタリングであり他は図形処理機能である。ここでは、図形処理機能において最も基本となる図形データのデータ構造について説明し、機能についても簡単に触れる⁽⁴⁾⁽⁵⁾。地図情報を計算機処理するには、図面をデジタル化する必要がある、このデジタル化された図面は、正方格子配列の平面上に表現されると考えられる。このデジタル化された平面上での点の連結性には 8 連結と 4 連結の概念が存在するが、本図形処理システムでは、垂直、水平方向のみの 4 連結概念を使用している。地図データでは、点、線、面が基本的構造となるが、4 連結デジタル平面上でつぎの様にこれらを定義する。

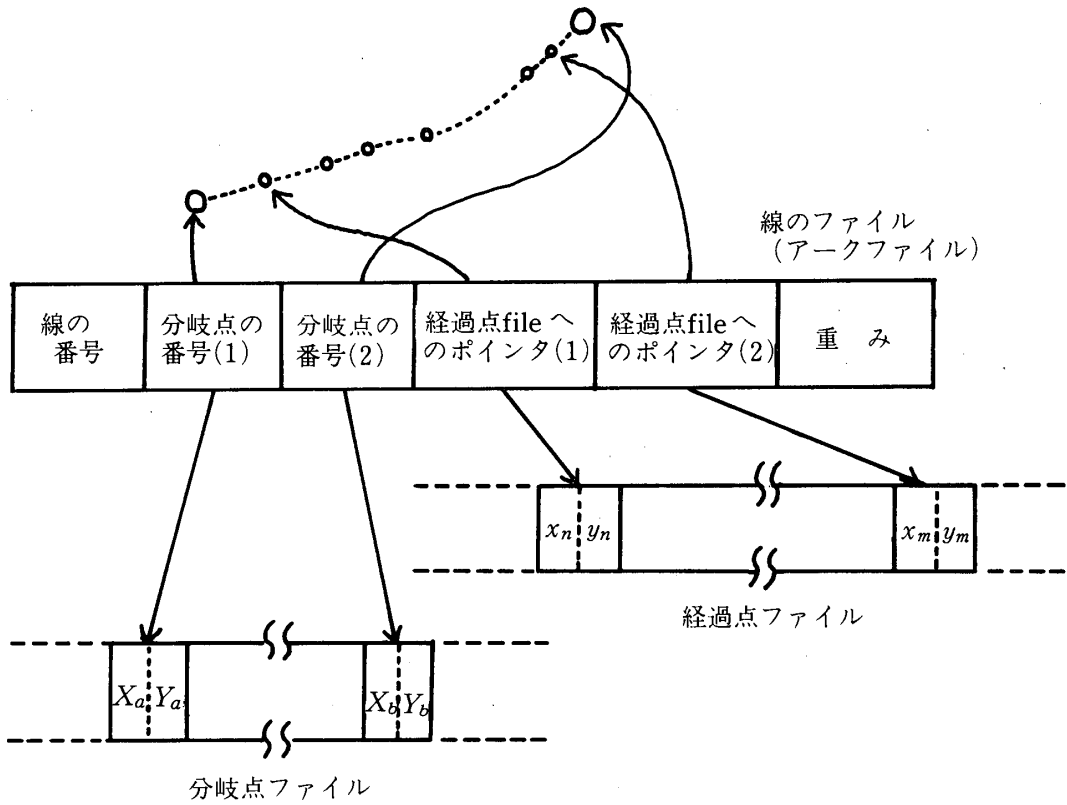
- i) 点：格子点、また点の種類としてはつぎの 4 つを考える。
 - 孤立点：他の格子点と結ばれていない点
 - 端 点：他の 1 点としか結ばれていない点
 - 経過点：他の 2 点としか結ばれていない点
 - 分枝点：他の 3 点もしくは 4 点と結ばれている点
- ii) 線：端点と端点、端点と分岐点、または分岐点と分岐点の間には経過点しかない点の並び (第 5 図 (a) 参照)。
- iii) 面：線に囲まれた領域 (第 5 図 (b) 参照)。

以上の定義を用い、地図 (統計区) データを記述するが、基本的な単位は ii) の線であり、これらが複数個結合され iii) の面を構成することになる。

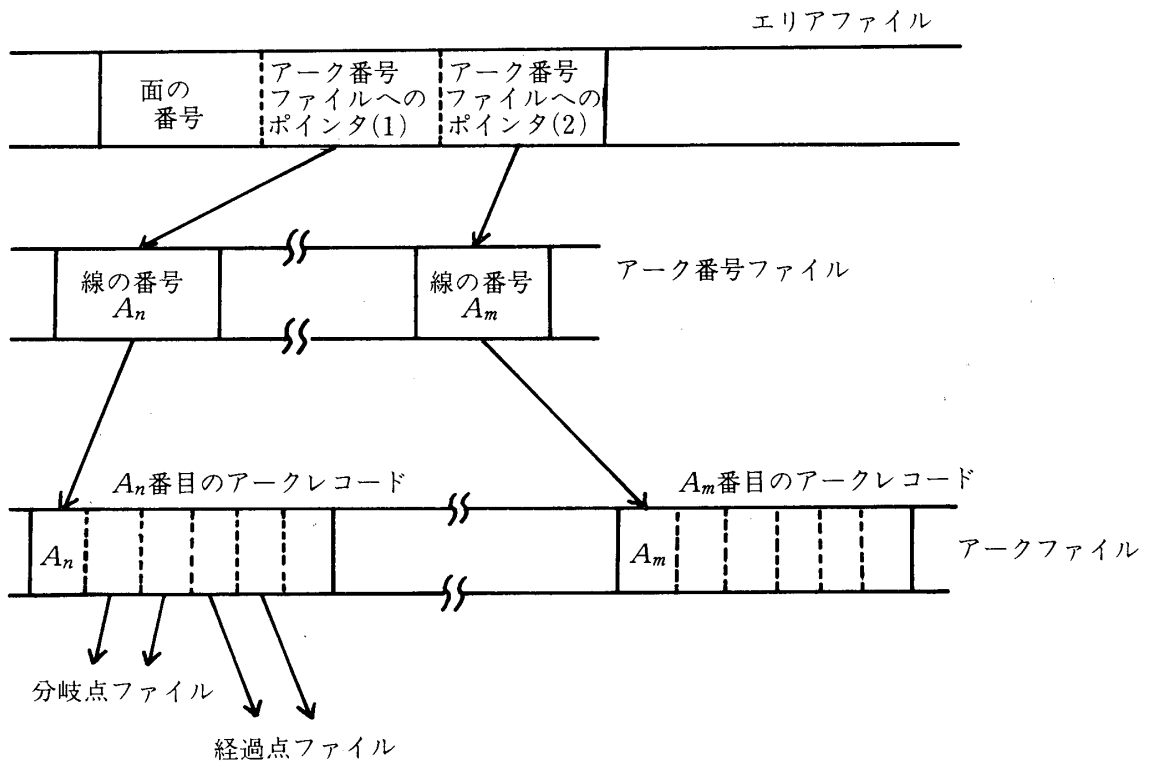
データ構造は、基本的には 3 つのファイルによって構成される。アークファイル (線のファイ



第 5 図 線および面の構造



第6図 線を構成するデータ・ファイルについて



第7図 面を構成するデータ・ファイルについて

ル)、分岐点ファイル、経過点ファイルがこれらであり、アークファイルは他の二つのファイルを管理している。第6図において、アークファイルにおける分岐点の番号はその線の分岐点分岐点ファイルの何番目のレコードであるかを示している。また経過点ファイルへのポインタ(1)、(2)は経過点ファイルの何番目から何番目のレコードかこの線の経過点であるかを示す。重みのフィールドは、線に何らかの重みづけをする時に用い、画像とは無関係なものである。以上の3つのファイルより、任意の線を抽出することはできる。しかし、この3つのファイルだけで、面の抽出や合成を行うことは困難であるので、更に面を表現するためのファイルを用意することとした。ここで1つの面を構成する線の数、面によって違うので第7図に示すようなデータ構造にした。ここで、アーク番号ファイルには面を構成する線の番号が順に並んでいる。そしてエリアファイル中のポインタ(1)、(2)がアーク番号ファイルにおいて、何番目から何番目までがその面を構成する線の番号であることを示す。更にアーク番号のファイルの各々のレコードがアークファイル中のそれぞれのレコードを指している。また、各ファイルの第0レコードの初めの欄にはそのファイルが所有するレコードの個数を書き込んでおく。

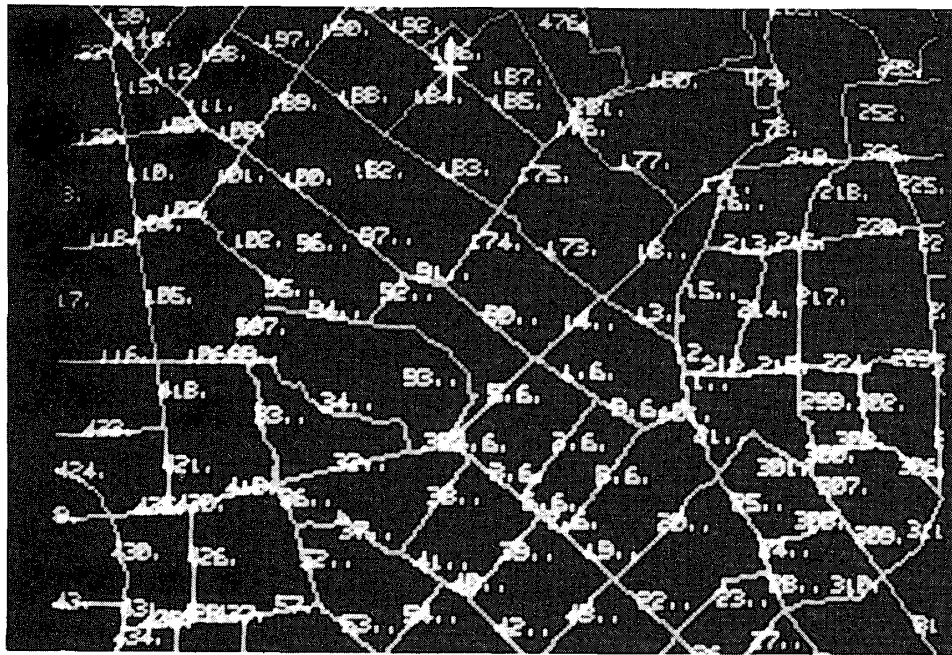
これまで述べてきたように、エリアファイルとアーク番号ファイルによって、面を構成する線を表わし、アークファイルと分岐点ファイル、経過点ファイルによって線を表わしている。これによって、線に関する操作に加えて、面単位の操作が可能となる。

本システムの図形処理上の機能は未だ整備されていないが、つぎの様なものがある。

- 地図表現の精度を変える機能：診療圏設定においては中間段階でそれ程高い精度を必要としない地図上にクラスタを描くことがある。この場合、分岐点同士を線形近似し、地図を表示する。この精度は前記の線形近似レベルから、デジタイザからの座標入力による最も統計区地図に近い精度まで存在し切り替え可能となっている。
- 特定の面の抽出と面の合併：任意の統計区を考察する場合、面の抽出が必要となる。行政区単位での表示の場合には複数個の面の合併が行われることになる。
- 面内へのキャラクタの記入：統計区番号や統計データ等の非画像データを同地図に重ね合せる場合に必要であり、グラフィック画面とキャラクタ画面を使用することになり、画像データファイルと対応する統計区数値データ・ファイルとがポインタにより結合されている。

MWX-230では面のPaintは非常に困難である。これは、グラフィック機能がビットイメージ的なVRAMのハードウェア構成を持っていないことによる(特に地図のような曲線的な線によって囲まれる面の場合にはソフトウェア的に面の境界を識別するのが難しいことによる)。

以上の図形処理システムはPASCAL/MT+によってM23上にインプリメントされている。本システムによる統計区地図の表示例が第8図に示されている。



第8図 (a)
統計区地図の表示例



第8図 (b)
統計区地図の表示例

第8図 統計区地図の出力

V. む す び

本論文では、社団法人札幌市医師会の依頼によって行われた患者動態調査のデータ処理システムについて中間報告した。本システムは未だ開発中の過程にあり、I章の(2)、(3)を含めた統合データ処理システムとはなっていない。これは依頼者との関係で動態調査の記述統計結果等のデータ処理が優先したためである。これらの集計結果は、SPSSの解析結果として直接依頼者側に手渡されている。(2)および(3)の目的のためには単純集計のみならず、より複雑なデータ加工が必要となる。SPSSの中にはクラスタ分析も存在するが、診療圏設定のためのより効率的でコンパクトな形でソフトウェア作成を独自に行っている。このため、我々はクラスタリングのためのアルゴリズムそのものを考察の対象として、グラフ理論を基礎としたアルゴリズムを開発した。このアルゴリズムは32ビット・マイクロ・コンピュータ上において充分実用的レベルでインプリ

メントできることがわかった。ソフトウェア開発上の労力は、マイクロ・コンピュータでの OS や開発ツールが充分安定であれば汎用計算機を使用するより小さなものですむと思われる。

図形処理ソフトウェアについては、図形データの作成等に多くの労力を必要とするが、診療圏設定データ処理システムでは強力なソフトウェアである。特にクラスタ解析の結果の図形出力化は、診療圏設定の様子をわかりやすくユーザに理解させ、その妥当性を既時的に判断させることになる。これは情報が集約され全体的な傾向等が明確となるためである。更に、一つの診療圏の設定が行われた場合、この診療圏を構成するメカニズム等についてより多様な解析が行える可能性を与えてくれる。例えば、設定された診療圏図形上に道路情報データを重ね合せれば、これらのデータの相関の強さや構成要因の大きさが可視的に判断できる。診療圏設定のための解析のように、一定の解析方法が存在しない場合、こうした処理結果の図形出力化は、より精度のよいデータ解析への手掛りを与えてくれることになる。

現時点までのデータ処理システムでは、Interactive 機能はそれ程大規模に取り込まれていない状態である。これは、Interactive 機能を考慮するとソフトウェア開発の労力が急激に増加すること、および、会話的処理の対象アクティビティや個々の機能の関係が明確でなく、全体的なフレーム・ワークがつくりづらいことによる。いくつかの処理モジュールがクラスタ解析や、図形処理ソフトウェアに組み込まれているが、これらは個別的であり、対象とするデータ・ファイル等に強い関連性を持っているため他のプログラムへ簡単に移植できない欠点を持っている。従って、我々は当面必要とする Interactive な機能をすべて調べ上げ、これらをいくつかに分類して標準化するように考えている。各モジュールは機能分担を明確にし、重複化させないようにしなければならない。アルゴリズム上での Interactive 機能は、データ入力時、各種パラメータの設定と変更時、出力形式の変更時に必要である。猶アルゴリズムそのものの変更には、アルゴリズムそのものの構造に関係し LIST 処理的考え方が必要なため今後の課題である。もう一つの機能の対象はファイルであるが、これはファイルの構造に依存してしまう。データ構造に依存させないで各モジュールを標準化させようとする、各種の素ファイルへの加工バンドラが必要となる。ファイルを対象とする会話的機能としては、一つは検索機能であり、他は、ファイル間結合等を含むファイルそのものの取り扱い、レコードの追加、削除、項目変更や追加、ソート機能等である。既に開発した会話的機能に限定して考察した場合、データ処理効率が向上したように思える。

最後に、現時点までの診療圏設定データ処理サブシステムの今後の検討課題について述べる。このデータ処理システムは三つのレベルに分離され個々のレベルでは効率の良いデータ処理が行っている。問題となるのは各レベル間でのデータの受け渡しである。下位レベルで処理結果が入力データとして利用されるが、各レベルでの処理システムが異なるため出力データに互換性が存在しない。マイクロ・コンピュータの入力用データ・ファイルとして最も標準的で利用しやすい FDD は、フォーマットやディレクトリの違いでそのままファイル・アクセスできないためハンドリングしなければならないという欠点を持っている。バンドラの作成と加工のための労力は無視できないものである。従って、我々は札大と北大との間で公衆回線を利用したマイクロ・コンピュータ間のデータ転送方式を開発中である。札大側のシステムは PC 9801 又は PC 8801 を対象とし、モデム付 1200 ボーの手順付とする予定である。

最後に、本報告を作成するにあたり、貴重な御助言を頂いた北海道大学工学部電気工学科系統講座加地郁夫教授、患者動態調査の共同研究者として多大な御指導を頂いた同講座大内東助教授、社団法人札幌市医師会理事小野修氏、日医医療システム研究委員飯塚弘志氏に深く感謝いたします。また、図形処理システムの作成と研究を担当して頂いた同講座修士 2 年の丸井敏郎氏、同 4 年松尾英普氏ならびにクラスタリング・アルゴリズムを担当して頂いた同修士 1 年栗原浩一氏に

場をかりてお礼申し上げます。

昭和 59 年 3 月

参考文献

- (1) 飯塚弘志：北海道医報，第 560 号，No.6 (1983)
- (2) Slater, Harner : "Identifying medical regions using hierarchical clustering", Soc. Sci. & Med., Vol. 14D, pp. 3~10 (1980)
- (3) 栗原浩一，大内 東，加地郁夫，宮腰昭男："可到達行列を用いた階層的クラスタリングのアルゴリズム"，SICE 北海道支部学術講演会，No.5，(1984)
- (4) 丸井敏郎："CP/M マシンによる地域情報処理システムの研究と試作"，修士論文，(1984)
- (5) 松尾英普："デジタル平面における線と面に関する処理システムの試作"，卒業論文，(1984)