オケージョナル・ペーパー　No.95

階層型ニューラルネットワークモデルによる特異地域の抽出

２０１９年２月

法政大学
日本統計研究所

階層型ニューラルネットワークモデルによる特異地域の抽出

坂本憲昭（法政大学経済学部）

概要

本研究は，数理モデルを構築することが困難な統計データに対してニューラルネットワークモデルを構築し，その学習過程で誤差が大きい学習レコードを特異なレコードとして抽出する手法を提案する。提案手法の基本概念は文献[1]に示しており，本稿は従来手法との比較と新たな統計データに適用して提案手法の有効性を示すものである。

１．はじめに

Table 1 のような統計データの解析や分析をおこなう場合，回帰分析により目的変数の値を予測する関係式を求めたり，説明変数が多い場合には主成分分析や因子分析などにより説明変数の縮約を検討したりすることも考えられる[2]。レコードを分類する目的の場合は，クラスター分析が一般的な手法であろう[2]。しかしながら，「説明変数が多い」「目的変数が複数ある」「説明変数の因果関係が不明で，いわゆる，“どんぐりの背比べ”のような変数の影響があり縮約を避けたい」「レコード数が多い」などの場合は，従来手法の適用は簡単ではない。このような事例としてビッグデータ[1]があり，最近では知見を得るためにAI[2]が盛んに用いられている。AIの基礎はニューラルネットワーク[3]であり，ビッグデータに対して構成が簡単なニューラルネットワークではモデルの構築（学習の成功）ができなかったが，関連する研究成果と計算機の技術進歩により，モデルの構築とその結果にもとづく解析や分析が可能となった。したがって，変数やレコード数が多い統計データに対しても従来手法では所望の結果が得られなかった場合にAIを含む深層学習[3]や機械学習[4]，強化学習[5]などを適用することは容易に考えられる。しかしながら，著者の調査によればこのような研究は少なく，文献[6]を紹介する。この研究は，地域人口を推計する予測モデルを機械学習のひとつであるサポートベクター回帰[4]を用いて構築する。説明変数は，土地利用，標高・傾斜度，道路延長，施設[5]，国土利用計画法など 25 変数，目的変数は地域人口である。一方，本稿の目的は目的変数の推定や予測ではなく，複数の目的変数を有する統計データにおける特異地域（特異なレコード）の抽出であり，サポートベクター回帰では 2 つ以上の目的変数に対応するために工夫が必要となる。本稿は回帰ではなくニューラルネットワークモデルを構築する過程における学習誤差を利用した特異地域の抽出方法を提案する[1]。有効性を確認するため 2 つの適用事例を示す。ひとつは，目的変数が小地域における男女別年齢階層別の流入移動者数のポテンシャルであり，説明変数は 22，学習レコードにあたる地域数は 2300 である[1][7]。この 2300 地域のなかから特異な地域（レコード[6]）を抽出することが目的である（５章参照）。ふたつめは，関東甲信越地方の一都六県における区・市を地域区分とした日常生活に必要な 22 種類の事業所数（店舗数）を説明変数，夜間人口密度と昼間人口密度を目的変数とした統計データである（６章参照）。

---

1 Bigdata，**大容量のデジタルデータ**
2 Artificial Intelligence **人工知能，厳密な定義は異なるがここでは一般的な表現として用いる**
3 Neural Network
4 Support Vector Regression, SVR
5 **公共施設・避難所・バス停・駅**
6 **ニューラルネットワークで学習するレコードを学習レコードといい，本稿の学習レコードはすべて地域である**

本稿の構成は，ニューラルネットワークの概要を２章で説明し，３章は提案手法の原理を示す。４章は２つの基本的な従来手法による検討，５・６章が適用事例である。７章でまとめと課題を述べる。

Table 1 研究対象とする統計データのイメージ

レコード	説明変数 1	説明変数 2	説明変数…	目的変数 1	目的変数 2	目的変数…
A	***	***	…	**	**	…
B	***	***	…	**	**	…
C	***	***	…	**	**	…
…	…	…	…	…	…	…

※イメージであり，***は数値データ，…はデータの継続をあらわす

２．階層型ニューラルネットワークについて

説明変数の各データを入力データ，目的変数の各データを出力データとして，ニューラルネットワークモデルを構築する。ニューラルネットワークをFigure 1に示し（詳細は文献[3][4][5][8]等を参照），ここでは必要な設定を説明する。Figure 1の記号○をユニットと称する。入力層のユニット数は入力データ数であり本稿では説明変数の数[7]に等しく，出力層のユニット数は出力データ数であり目的変数の数となる。中間層は隠れ層とも呼ばれ，中間層の列数をm，本稿は中間層の各列のユニット数を同一としてnとおく。一般的に中間層のユニット数は入力データ数の 2 倍がよいとされているが，中間層のユニット数や列数を多くするほど良い結果になるとは限らないため，m, nの値はシミュレーションにおける学習状況で決定する。また，モデル構築のためのシミュレーション内容を簡単に説明すると，学習レコードごとの入力データ（説明変数）を入力層に入力し，中間層で重み係数の乗算やユニットごとの総和計算，シグモイド関数による処理などをおこない，Figure 1において左から右方向に計算を進める。最終結果が出力層のモデル出力値となり，教師データ[8]との誤差を求める。すべての学習レコードを対象とした平均二乗誤差（MSE：Mean Squared Error）があらかじめ設定した収束値以下になることを目的として，中間層すべてのユニットごとの重み係数の学習（修正）をおこなう。多くの学習方法のなかから，本稿は一般的な誤差逆伝搬法（BP：Backpropagation）を用いる。重み係数の修正後，再び学習レコードごとに計算をおこなう。平均二乗誤差が収束値以下になるまで繰り返し修正計算をおこなう。

従来の一般的なニューラルネットワークは中間層の列がひとつであり，多量のデータを扱う場合には学習成功に至らないことが多いが，計算機の処理能力の向上により中間層を複数・複雑にしても現実的な処理時間で学習が終了[9]するようになった。なお，学習成功に至らない場合，繰り返し計算が無限に実行されるため繰り返し回数の上限を設けておく。以下に学習のプロセスを示す。

学習のプロセス

Step.1 学習レコードの説明変数を入力→計算→モデル出力

Step.2 誤差＝教師データ（目的変数）－ Step.1 で求めたモデル出力

Step.3 すべての学習レコードについて Step1，Step2 の計算をおこない，平均二乗誤差を算出

Step.4 平均二乗誤差が収束値以下なら学習成功で終了，
未達なら中間層のすべてのユニットの重み係数を修正して Step.1 に戻る。ただし，計算の繰り返し回数が上限値以上となれば学習失敗で終了

シミュレーションに用いるプログラムは C++による著者作成である。

---

7 Figure 1 では５章のデータに基づき入力データ数 22，出力データ数 1 と表現した

8 真値または正解。学習レコードごとにモデルの計算結果が一致することが目的の値である

9 学習レコードすべてを対象にした平均二乗誤差が収束値以下になること

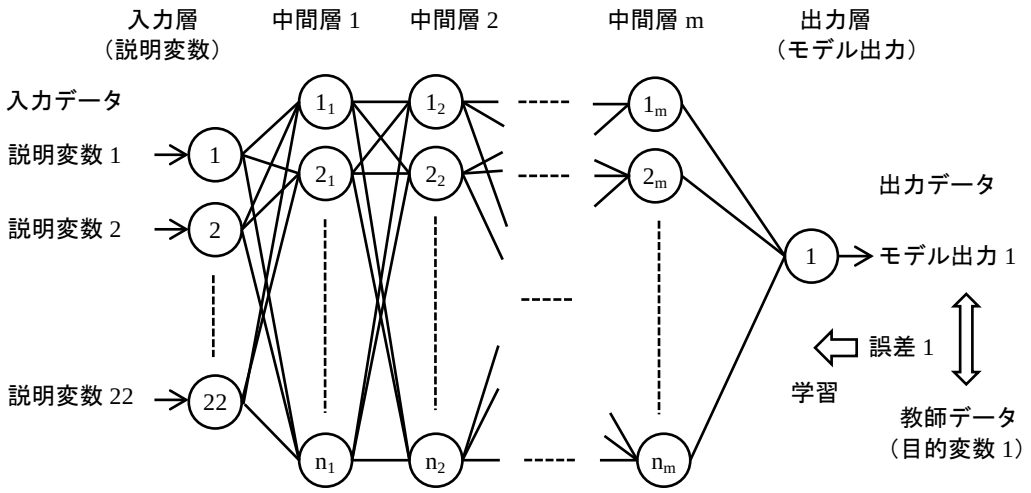


Figure 1 ニューラルネットワーク

３．ニューラルネットワークモデルによる特異地域の抽出原理

ニューラルネットワークの学習過程において，誤差が大きい学習レコードは入力データに何かしらの欠損や異常，特異性があることは容易に考えられる。たとえば，紙に０，１，～，９と書かれた文字をカメラ等で計算機に取り込んでデジタル化した値を入力データ，出力データが０，１，～，９のいずれかの数値に該当するかを判断するニューラルネットワークモデルを考えた場合，紙に書かれた文字が活字であればモデルの学習が容易であり正答率（または認識率という）が 100%近くになるが，手書き文字の場合には下手な字体ほど正答率が落ちる。すなわち，活字ならば学習過程において誤差が小さい学習レコード，手書きの文字は個性があり誤差が大きくなる学習レコードである。このほかにも，たとえば学習レコードとして犬と猫のさまざまな顔写真を多量に準備し，与えられた写真の犬の可能性・猫の可能性を判断するニューラルネットワークモデルの場合，犬のような猫の写真を“犬である確率 30%，猫である確率 70%”と出力する事例を想像していただきたい。本稿はこの状況を利用して，誤差が大きい学習レコードは特異なレコードとみなすことで“統計データのなかから特異性を有するレコードを抽出する手法”を提案する。

最初にテストデータを作成し，提案手法の有効性を確認する（Table 2 参照）。入力データ x_1, x_2, x_3（説明変数）について，次式により教師データ（目的変数）を求める。係数に根拠はなく，また，本稿の目的から説明変数と目的変数間の数理モデルが存在しない[10]ことを前提条件に，*Sin*関数と乱数（*Rnd*）を含めている。絶対値を取るのは一般的なニューラルネットワークは 0～1 の数値を扱うためである。特異な学習レコードとして No.8,9,10 を与える。

$$y = |0.1x_1 - 0.2x_2 + 0.15x_3 + 0.05Sin(5Rnd)| \quad (1)$$

10 物理的な数理モデルが存在するならば特異なレコードが解析により厳密解として得られると想定しているため，因果関係が成立しないようにする

Table 2 テストデータ

レコード No	x_1	x_2	x_3	y
1	1	1	1	0.0155
2	1	1	0	0.0524
3	1	0	1	0.2001
4	1	0	0	0.1500
5	0	1	1	0.0715
6	0	1	0	0.1504
7	0	0	1	0.2000
8	0.5	0.5	0.5	0.0304
9	0.25	0.25	0.75	0.0393
10	0.75	0.5	0.75	0.1091

ニューラルネットワークの設定値は，中間層ユニット数（n）6，中間層列数（m）1，収束値 0.01 とした。シミュレーションの結果，繰り返し計算が 1512 回目で学習成功となった。その学習過程（シミュレーション過程）において，誤差の絶対値上位 3 レコードを Figure 2 に示す。約 1200 回を過ぎた頃からレコード No.8,9,10 が常に誤差上位 3 レコードであり抽出成功と判断する。５章ではこの原理を用いて特異地域を抽出する。

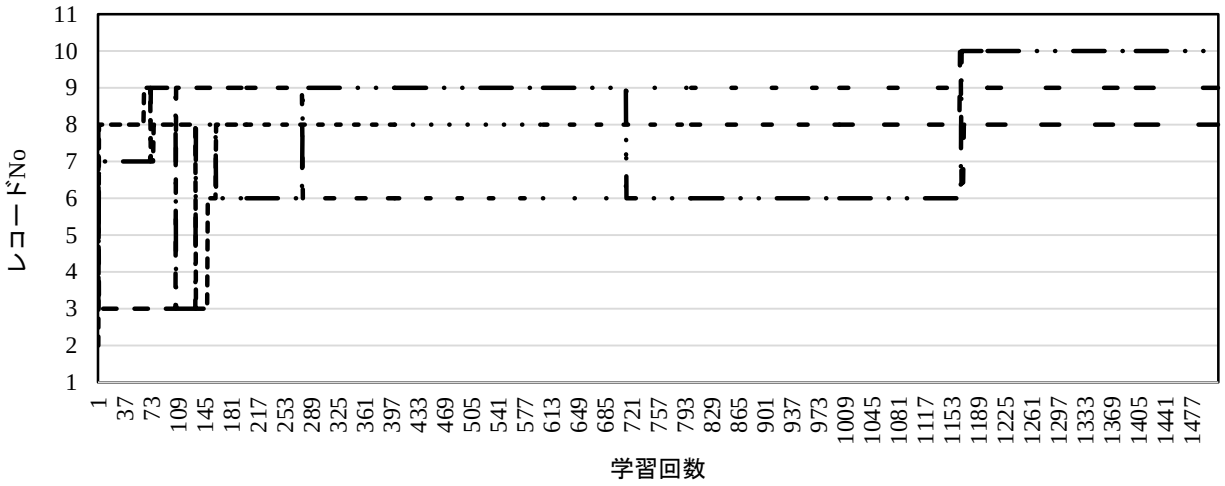


Figure 2 学習過程における誤差（絶対値）の上位 3 レコード

４．従来手法

３章で述べた“対象データのモデルを得るために機械学習を適用する。その学習過程の誤差を利用して特異レコードを抽出する”，従来の抽出手法として，学習はしないがクラスター分析やランダムフォレスト[11]，１章文献[6]で紹介したサポートベクター回帰など多くの手法がある。本稿は最適な手法の追求ではなく，ニューラルネットワークモデルによる抽出手法の提案であり，ここでは比較のために回帰分析とクラスター分析を取り上げる。

誤差が大きいレコードを特異地域とするならば，回帰分析における残差（誤差）に着目することも考えられる。イメージを Figure 3 に示す。近似直線から離れた〇で囲んだデータは外れ値であり，直線で描いた予測モデルとの誤差は大きい。すなわち，誤差が大きいデータは外れ値である。Table 2 のテス

11 Random Forest

トデータに対して Excel の分析ツールを用いて回帰分析をおこなった結果を Table 3 に示す。相関が悪いが各データと回帰式との誤差を Figure 4 に示す。特異なレコードとして No.8,9 を抽出できるが, No.10 を抽出することができない。

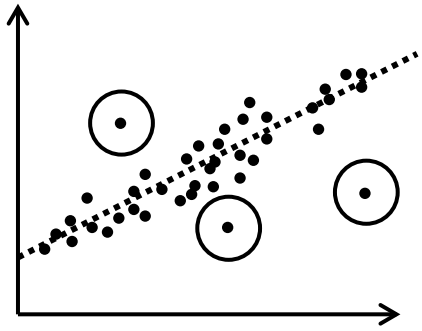


Figure 3 回帰分析における外れ値

Table 3 テストデータの回帰分析結果

データ数	10
重相関 R	0.63
重決定 R2	0.40
補正 R2	0.10
標準誤差	0.066

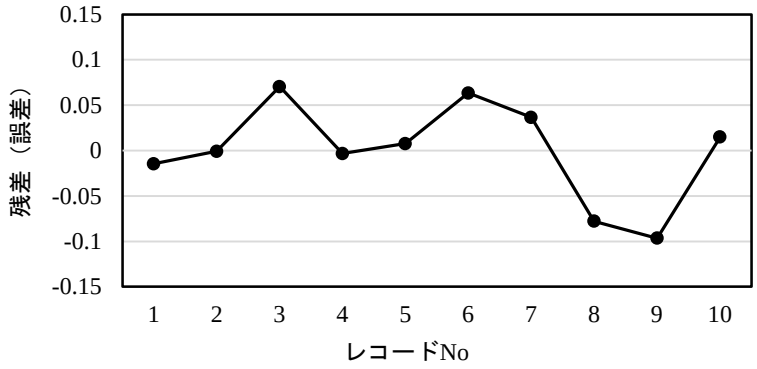


Figure 4 回帰分析における各レコードの残差

次に，特異なレコードの抽出方法としてクラスター分析を適用する。レコードNo.8,9,10 が独立した分析結果になれば抽出できるといえるが，クラスター分析を用いる場合「データの正規化有無」，分析手法として「最近隣法・最遠隣法・メディアン法・重心法・郡平均法・可変法・ウォード法」,「距離・平方距離」と選択肢が多くあり，その選択によって結果が異なるため本稿の目的である特異レコードを抽出したという判断が難しい。結果の一例を付録Figure ①～④ に示す。Figure ②,④であればレコード No.8,9,10 が独立しているとみなせるかもしれないが，その判断は正解が既知だからである。確認のため R[12]（dist関数の標準設定を使用）を用いた場合はレコードNo.8,9,10 が最初にグループ化され（Figure ⑤参照），そのほかのレコードは異なることが理解できるが，特異なレコードとして抽出されているとの判断は難しい。

12 オープンソース・フリーソフトウェアの統計解析向けのプログラミング言語及びその開発実行環境 https://www.r-project.org/ （2019 年 2 月 10 日確認）

５．流入移動者数のポテンシャルデータにおける特異地域の抽出

5.1 適用する統計データ[7]

文献[7]に示された町丁字区分における人口移動の傾向を分析するための統計データに提案手法を適用する。町の流入移動率は男女間および年齢によって異なり，また，たとえば地域Aから転出して地域Bに転入を考えるとAB各地域の常住人口の男女および年齢構成は異なる。一方，人々の移動は性別や年齢だけでなく，ほかのさまざまな作用因によっても影響されるが，作用因は地域間で異なり，作用結果の総体がそれぞれの地域の現実の流入移動者数を作り上げている。逆に言えば，性別・年齢分布が等しい地域間であっても，ほかのさまざまな作用因により流入移動者数が異なる。文献[7]は研究事例として新潟県新潟市の町丁字区分（レコード数：2300 地域）を取り上げ，この性別・年齢以外のさまざまな作用因を考察するために， 2015 年度国勢調査結果 [13]よりTable 4 に示す統計データを得ている（グループ分けは本稿独自のものである）。これが本稿の説明変数に相当する。

Table 4 説明変数内容

説明変数	1	2	3		
統計内容	未婚率	有配偶率	単独世帯率		
説明変数	4	5	6	7	8
統計内容	持ち家率	借家率	一戸建て 世帯率	共同住宅 世帯率	3 階建て以上 住居世帯率
説明変数	9	10	11	12	13
統計内容	雇用者率	自営・家族 従業者率	農林漁業者率	製造業 就業者率	サービス業 就業者率
説明変数	14	15	16	17	18
統計内容	公務員率	管理・専門 従事者率	サービス職業 従事者率	農林漁業 従事者率	生産工程 従事者率
説明変数	19	20	21	22	
統計内容	1 年未満 居住者率	5 年未満 居住者率	10 年以上 居住者率	20 年以上 居住者率	

上記の説明変数とは独立に，常住人口の男女および年齢構成から町丁字区分における流入移動者数ポテンシャル [14]を算出している。この値が本稿の目的変数に相当する。

$$\text{流入移動者数ポテンシャル} \equiv {M_{si}}\Big/{\sum_g (P_{sgi} R_{sg})} \tag{2}$$

s：男女別区分

g：年齢区分別

i：新潟市内 i 地域

M_{si}：新潟市内 i 地域，男女別，流入移動者数

P_{sgi}：新潟市内 i 地域，男女別，年齢区分別，常住人口

R_{sg}：新潟市内平均，男女別，年齢区分別，流入移動者数

13 https://www.stat.go.jp/data/kokusei/2015/index.html （2019 年 2 月 10 日確認）

14 読者の誤解を招かない限りポテンシャルと略す場合がある

Table 4 に示したように説明変数のグループ分けが考えられるがデータ間に因果関係はなく独立しており，(2)式に示したとおり目的変数の値も説明変数を使わずに算出する。すなわち，本稿の目的である数理モデルを求めることが困難な統計データである。Table 5 に一部の地域における説明変数の一部と流入移動者数ポテンシャルの値を示す。

この統計データにおいて特異地域を抽出することにより，その地域においては性別・年齢以外のなんらかの作用因もしくは複合的作用の存在をあきらかにすることが可能となる。その知見は自治体の施策や地域の改善，たとえばインフラ整備や公的機関の統廃合，さまざまな助成金などの検討に役立つ。

Table 5 流入移動者数ポテンシャル（一部のデータ）

地域 新潟県新潟市	説明変数 1	説明変数 14	説明変数 15	説明変数 21	目的変数
	未婚率	公務員率	管理・専門 従事者率	10 年以上 居住者率	流入移動者数 ポテンシャル
北区太郎代	0.2873	0.0119	0.1314	0.6312	0.0468
北区島見町	0.3476	0.0098	0.1518	0.5269	0.1250
北区太夫浜新町１丁目	0.4221	0.0210	0.2370	0.5372	0.1579
北区太夫浜新町２丁目	0.3791	0.0137	0.2104	0.5057	0.1558

5.2 回帰分析による特異地域の抽出

回帰分析の誤差で特異レコードを抽出する方法を考えた場合，４章では満足な結果が得られなかったが，提案手法との比較のために適用する。R の*lm*関数により回帰分析[9]した結果 [15]，

決定係数　0.8962，　**調整済み決定係数** 0.894

$$\text{目的変数の値} = -0.31797x_1 - 0.29946x_2 + \cdots - 0.08052x_{21} - 0.01156x_{22} + 0.16566 \quad x\text{は説明変数} \tag{3}$$

が得られた。決定係数が約 0.9 であることから成立していると判断する。各レコードの説明変数（Table 5 参照）を用いて(3)式で求める目的変数の値と，Table 5 の流入移動者数ポテンシャルとの偏差を Figure 5 に示す。このなかでプラス誤差・マイナス誤差の各上位 3 レコードの地域は，中央区神道寺 2 丁目，北区松浜町，江南区早苗 2 丁目，東区若葉町 2 丁目，江南区早通 6 丁目，中央区西堀前通九番町である。

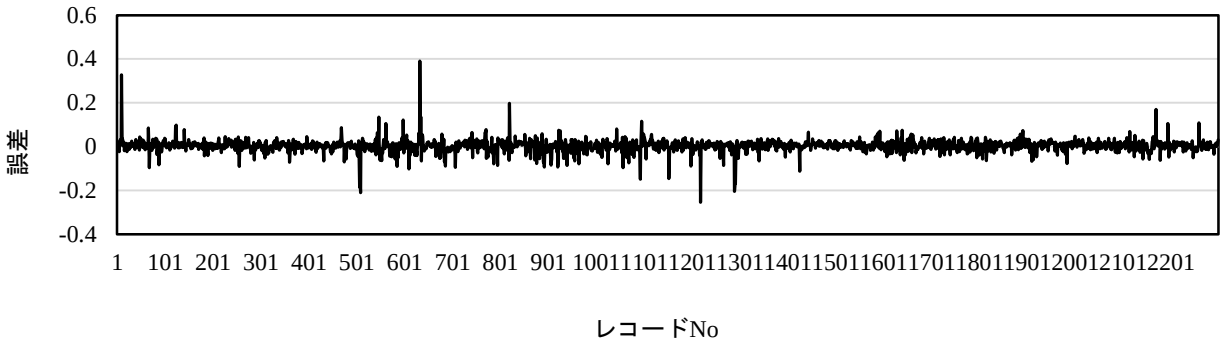


Figure 5 回帰分析による各レコードの誤差

15 Excel 分析ツールの回帰分析は 16 変数までしか対応していないためソルバーで回帰する必要がある。それに比較して R は適用が容易である

5.3 ニューラルネットワークモデルによる特異地域の抽出

5.3.1 データの正規化

一般的なニューラルネットワークは，入力および出力データの範囲が 0〜1 である[10][11][12]。**5.1** 節のデータにおいて，説明変数ごと，および目的変数の各最大値で各データを除算してデータを 0〜1 の範囲にする。

5.3.2 ニューラルネットワークおよび学習シミュレーションにおけるパラメータ設定

入力データ（説明変数）: 22，出力データ（目的変数，流入移動者数ポテンシャル）: 1，学習レコード数（新潟市の町丁区分地域数）: 2300 である。ニューラルネットワークの中間層ユニット数（n），中間層列数（m）は，n = 22, 33, 44，m = 1, 2, 3 としてすべての組み合わせで各 5 回シミュレーションをおこない，学習回数が 1000 の時点における平均二乗誤差を比較して最小値となる値を選択する。その結果は n=44，m=2 である。このほかの設定値として収束値は 0.2，繰り返し回数上限値 30000 とした。

学習成功した 10 ケースを得るまでシミュレーションを繰り返し，学習成功率は約 30%であった。この結果は容易に収束値にはならない妥当な設定と考える。もし収束値を大きくして学習を容易とすればモデルの精度が悪く，4 章に示したような回帰精度が悪いのと同等である。逆に収束値を小さくした場合は学習成功に至らず，逆に学習に成功した場合は各学習レコードの誤差の差が僅差になり，特異なレコードが埋もれてしまう可能性がある。参考までに収束値について一般的な知見はなく，ニューラルネットワーク研究における課題のひとつである。

5.3.3 抽出結果と考察

学習に成功した 10 ケースのうち，プラス誤差が大きい上位 10 地域とマイナス誤差が大きい上位 10 地域を抽出し，このプラス・マイナス誤差上位各 10 地域をあわせて各上位 3 地域を抽出した（Table 6 参照）。

回帰分析による抽出地域と提案手法の抽出地域は江南区早通が近い地域であるが，このほかは異なる。モデルの違いにより一致しないことは自明であるが，正解が不明なため“誤差が大きいレコードの特異性の内容”と“**5.1** 節で述べた研究目的”のために抽出したすべての地域を考察する。比較する地域として同じ区の同町または丁違いを取り上げ，視覚で考察するため区ごとに描いたグラフが Figure 6〜Figure 10（横軸 No.1, No.2, No.3,…: 説明変数，Y: 目的変数）である。その考察を Table 7 に示し，各説明変数の傾向について同様な傾向があると考える地域を 2 ケースにわけて抽出し，さらに，傾向の違いを顕著にするため各説明変数で最大最小間の差が 0.5 未満のデータを削除した結果を Figure 11（Table 7 注①），Figure 12（Table 7 注②）に示す。特に提案手法の結果（Table 7 および付録 Figure A〜Figure F）から，

・流入移動者数ポテンシャルが高い傾向：未婚世帯，借家，共同住宅，5 年未満の居住，
・流入移動者数ポテンシャルが低い傾向：持ち家，ファミリー世帯，5 年以上の戸建て，
・職業については，サービス業が流入移動者数ポテンシャルに若干の影響を与える。

これらの特徴と異なる地域が特異地域として抽出されている。さらに，神道寺 2 丁目（Table 7 特記 1）は，居住年数が短いから流入移動者数ポテンシャルが高いのは理解できるが，持ち家率が高い。この地域には一般住宅はなく入所型の介護老人保健施設がある。東区若葉町 2 丁目（Table 7 特記 2）は，神道寺 2 丁目と同様に入所型の介護老人保健施設があるほか，新興住宅地がある。

Table 6 誤差上位による特異地域の抽出

プラス誤差			マイナス誤差		
地域記号	住所	該当ケース数 [16]	地域記号	住所	該当ケース数
A	西区寺尾台 1 丁目	10/10	D	江南区早通	10/10
B	南区小坂	9/10	E	北区村新田	8/10
C	中央区入船町 1 丁目	8/10	F	西区須賀	8/10

Table 7 回帰分析結果を含む抽出地域の考察

Figure.	住所 [17]*	各説明変数の傾向と Google マップ参照結果について	流入移動者数 ポテンシャル について	注
6 中央区	C：入船町 1 丁目 付録 Figure C 参照	10 年以上の居住率が高い	低い	①
	神道寺 2 丁目	持ち家率が高く，居住年数が短い（特記 1）	高い	②
	西堀前通九番町	中央区神道寺 2 丁目と同傾向だが居住年数が長い	中央区神道寺 2 丁目より低い	②
7 北区	松浜町	中央区西堀前通九番町と同傾向だが未婚率が高い	高い	②
	E：村新田 付録 Figure E 参照	中央区入船町 1 丁目と同傾向 ポテンシャルは計算ミスではなく該当者なし 10 年以上および 20 年以上の居住者率が高い	ゼロ	①
8 江南区	早苗 2 丁目	未婚の単独世帯率が高い 借家で共同住宅率が高い，居住年数が短い	低い	①
	D：早通 付録 Figure D 参照	10 年以上の居住率が高い アパート等が多い	低い	
	早通 6 丁目	顕著な相違は見られず	低い	
9 西区	F：須賀 付録 Figure F 参照	顕著な相違は見られず 同等ほかの近隣地域とも顕著な相違は見られず	低い	
	A：寺尾台 1 丁目 付録 Figure A 参照	サービス業従事者およびアパート率が高い 説明変数の No1.〜No.8 の変化が小さい	やや高い	
10 東区 南区	東区若葉町 2 丁目	持ち家率が高く，居住年数が短い（特記 2）	高い	②
	B：南区小坂 付録 Figure B 参照	10 年以上の居住率が高い ほとんどが水田の地域	低い	

16 たとえば 9/10 は学習成功 10 ケースのうち 9 ケースで誤差 10 位以内に含まれている地域
17 住所の A~F は Table 6 の地域記号，記号がない地域は 5.2 節で抽出した地域である

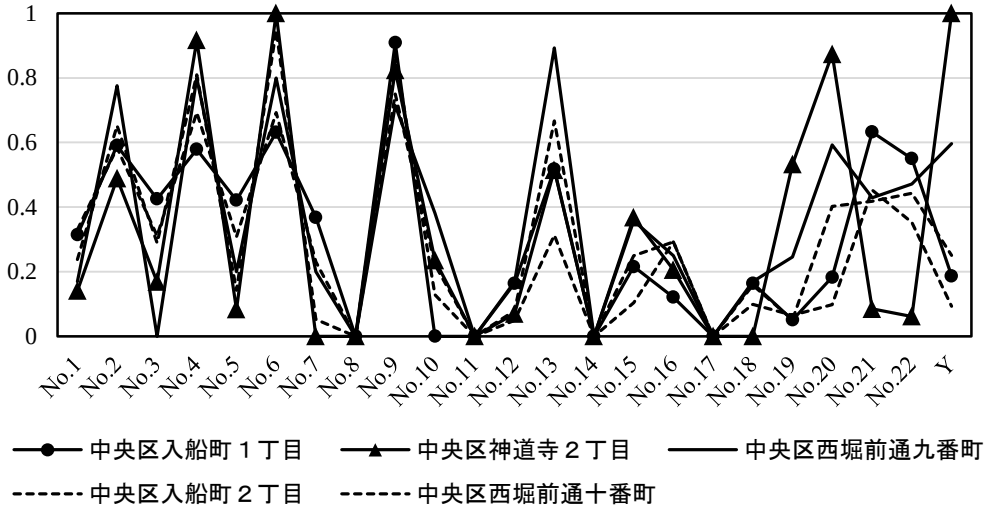


Figure 6 中央区で抽出された地域の考察

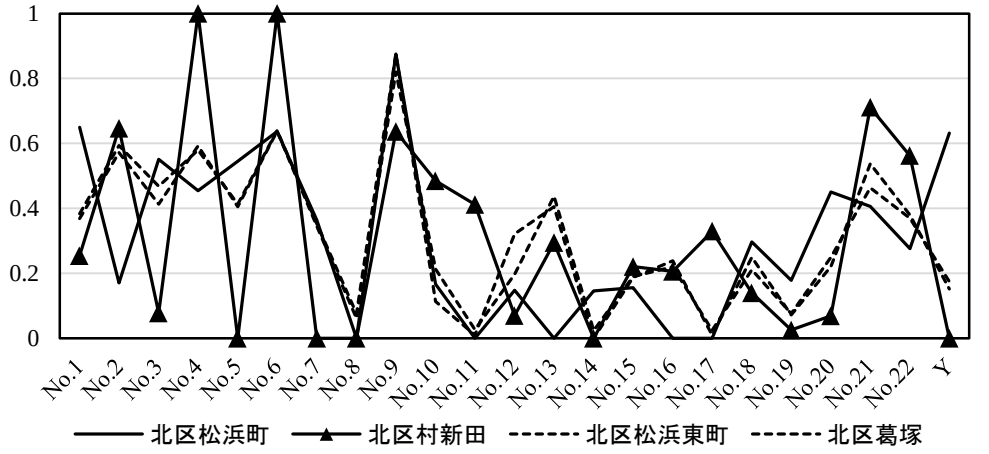


Figure 7 北区で抽出された地域の考察

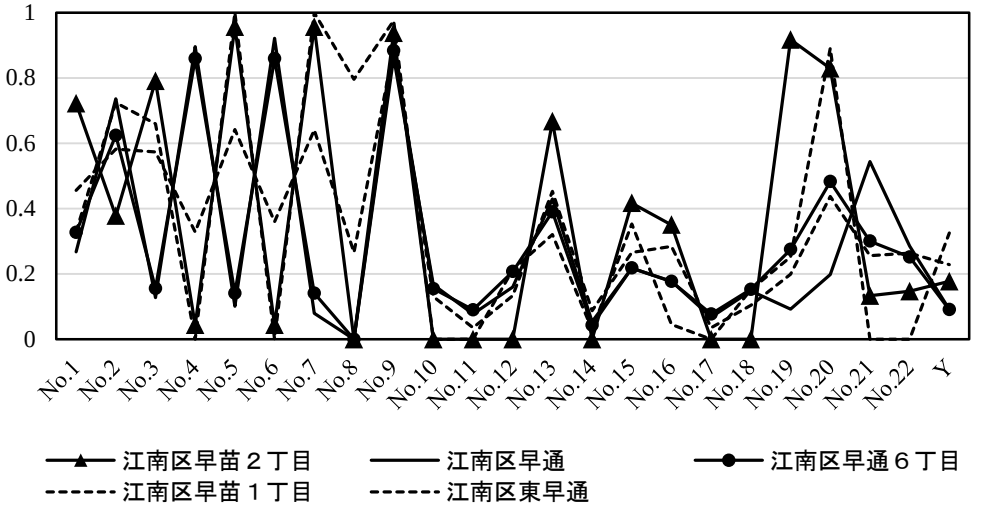


Figure 8 江南区で抽出された地域の考察

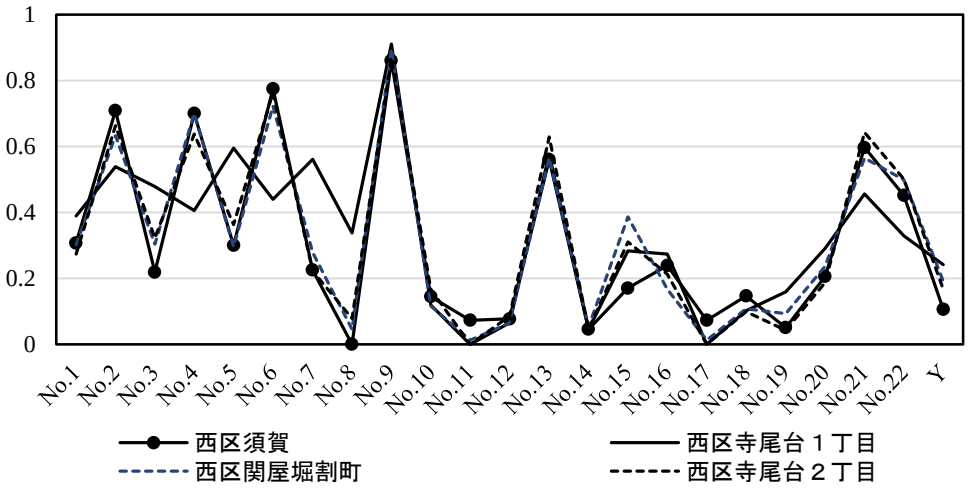


Figure 9 西区で抽出された地域の考察

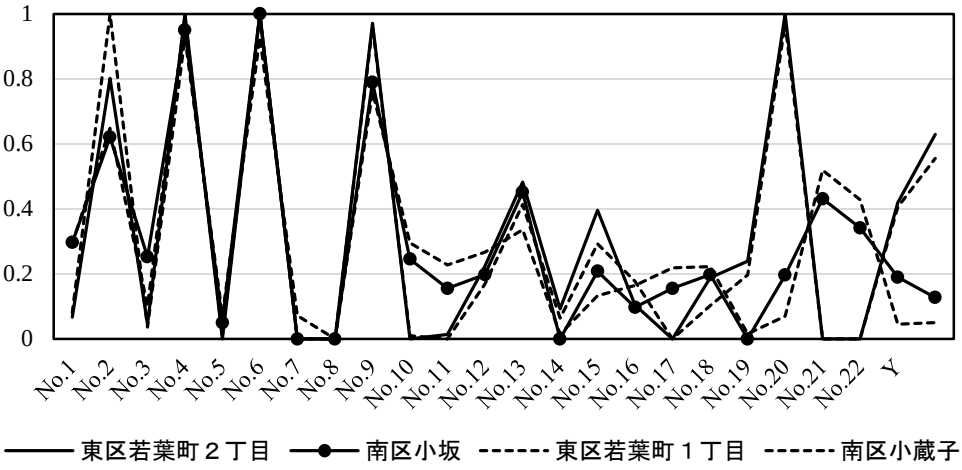


Figure 10 東区・南区で抽出された地域の考察

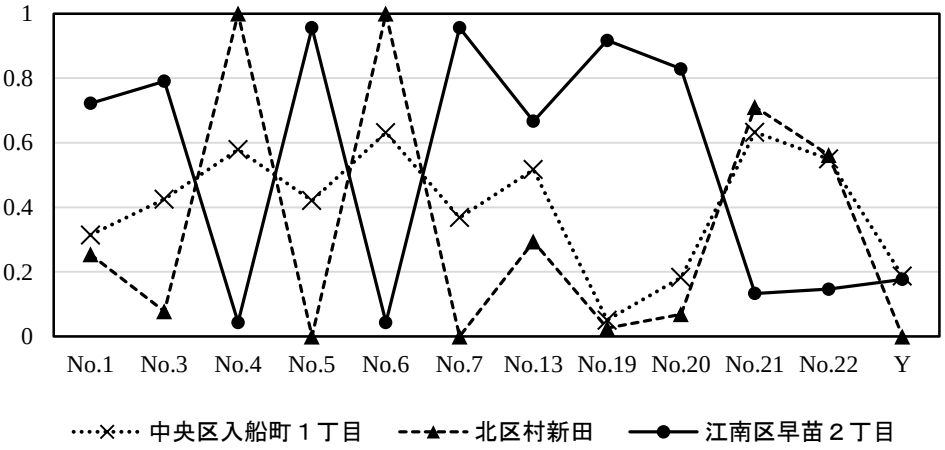


Figure 11 考察結果より詳細な比較（Table 7 ①参照）

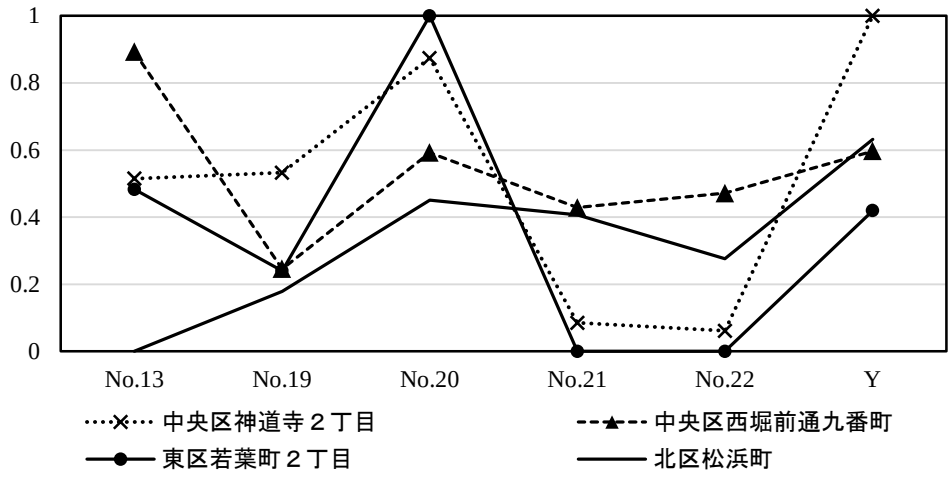


Figure 12 考察結果より詳細な比較（Table 7 ②参照）

６．事業所数と人口密度データ

本章は目的変数が複数の場合を対象として提案手法の有効性を確認する。目的変数が複数の場合，従来手法として，目的変数の数だけ回帰モデルを作る方法や複数の目的変数に対応したモデルを使う方法がある。後者の方法は，たとえば正準相関分析，ランダムフォレスト回帰，ニューラルネットワーク回帰である。すでに４章で述べたように，大きい誤差なら特異なレコードとして抽出する方法が考えられるが，目的変数の数だけ誤差があることに注意しなければならない。たとえば，目的変数が A,B ふたつの場合，A+B の誤差で判断するか，A の誤差が大きく B の誤差が小さい場合，その逆などを考慮する必要がある。この事例では A+B の誤差を扱うこととする。

6.1 事業所数データ

ある地域の事業所数がほかの地域と比較して多いか少ないかは，人口当たり，または人口密度当たりで評価することが多いがそれはランキングであり，ほかの地域と業種ごとの事業所数が同等数あるかを評価することは難しい。著者は面積と人口から事業所数を求める数理モデルを利活用して，対象地域の業種ごとの事業所数を評価する手法を提案している[13][14]。文献[15]は，関東甲信越地方 [18] において島しょ部，および町・村を除外した市区地域における各事業所数を集計し（Table 8 参照），この統計データからその地域の面積・夜間人口・昼間人口を引数として，事業所ごとの数を求める数理モデルを確立している。しかしながら，各事業所について実際の店舗数と数理モデルが算出した標準的な店舗数との差を得ることはできるが，これは事業所ごとの数値（偏差）であり統計データ全体で特異な地域をあきらかにすることはできない。そのため提案手法の適用が有効である。

事前に５章と同じく回帰分析による手法を確認する。目的変数が 2 つのためそれぞれで回帰をおこなった結果を以下に示すが，相関が 0.6 前後であり適用できないと判断する。

目的変数 1（夜間人口密度）：決定係数　0.64，　調整済み決定係数　0.60

目的変数 2（昼間人口密度）：決定係数　0.61，　調整済み決定係数　0.56

18 東京都，神奈川県，千葉県，埼玉県，群馬県，茨城県，栃木県，山梨県，長野県，新潟県

Table 8 統計データ

入力データ（説明変数）21 各事業所数，各店舗数	飲食店，コンビニエンスストア，居酒屋，すし，ラーメン店，在宅介護サービス，デイサービス，歯科，内科，薬局，タクシー，クリーニング，生花店，新聞店，郵便局，市区町村機関，理容店，学習塾，進学塾，保育園，幼稚園
出力（教師）データ（目的変数）：2	夜間人口密度，昼間人口密度（元データは面積・夜間人口・昼間人口であるが考察のために密度とした）
学習データ：208	市区の地域紙

6.2 ニューラルネットワークモデルの構築と抽出結果

５章と同じ方法で構築する。繰り返し回数 1000 の時点での誤差を付録 Table Ⅰ に示す。顕著な違いはないがもっとも結果がよい中間層ユニット数 42，中間層列数 3 とし，また，収束値 0.01 に設定した。学習成功の頻度は 5 回の学習成功を得るために 2 回の学習失敗である。

5 回の学習成功結果の中から誤差が大きい上位 6 地域「常総市，横浜市金沢区，日野市，那須塩原市，沼田市，千葉市稲毛区」が特異地域として抽出した結果である。

6.3 抽出地域の考察

抽出した地域に対して目的変数の夜間人口密度と昼間人口密度の和が最も近い地域と，近隣の市を比較した結果を Figure 13～Figure 17 に示し，その考察を Table 12 で述べる。特定分野の事業所が多い，新幹線停車駅の影響があるなどの特異性を有する地域が抽出できたと判断する。

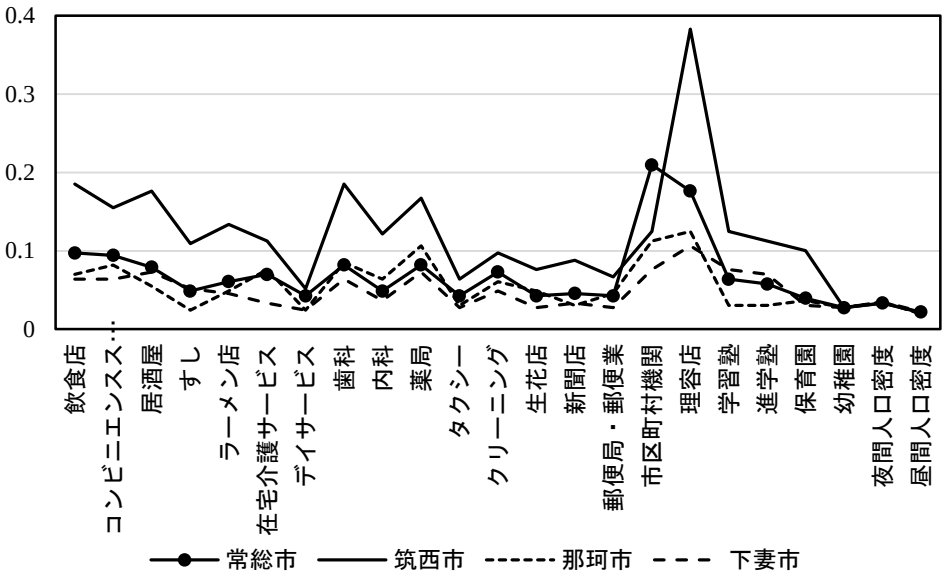


Figure 13 常総市の特異性

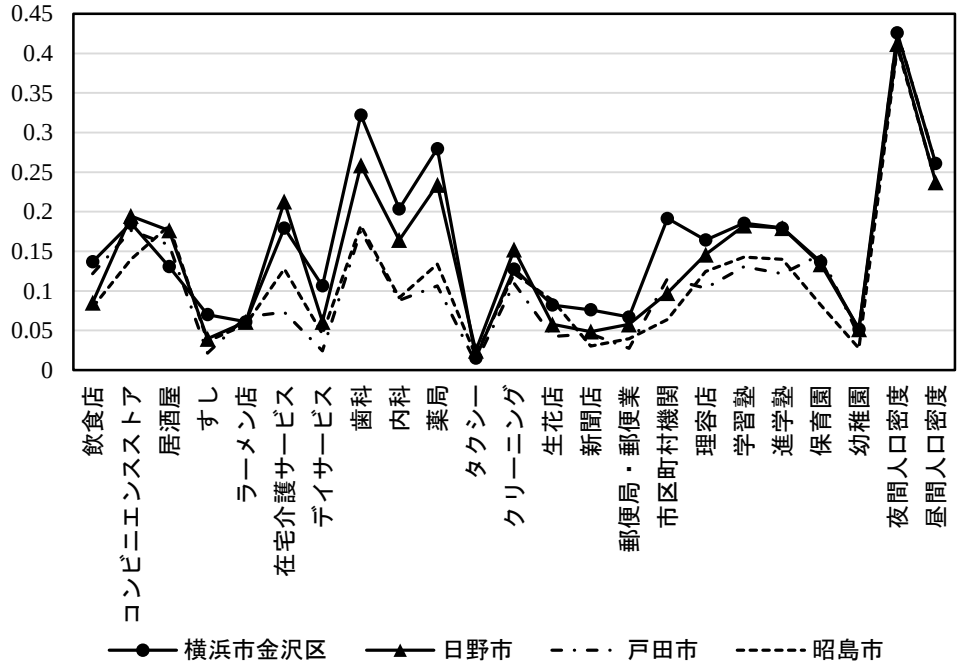


Figure 14 横浜市金沢区と日野市の特異性

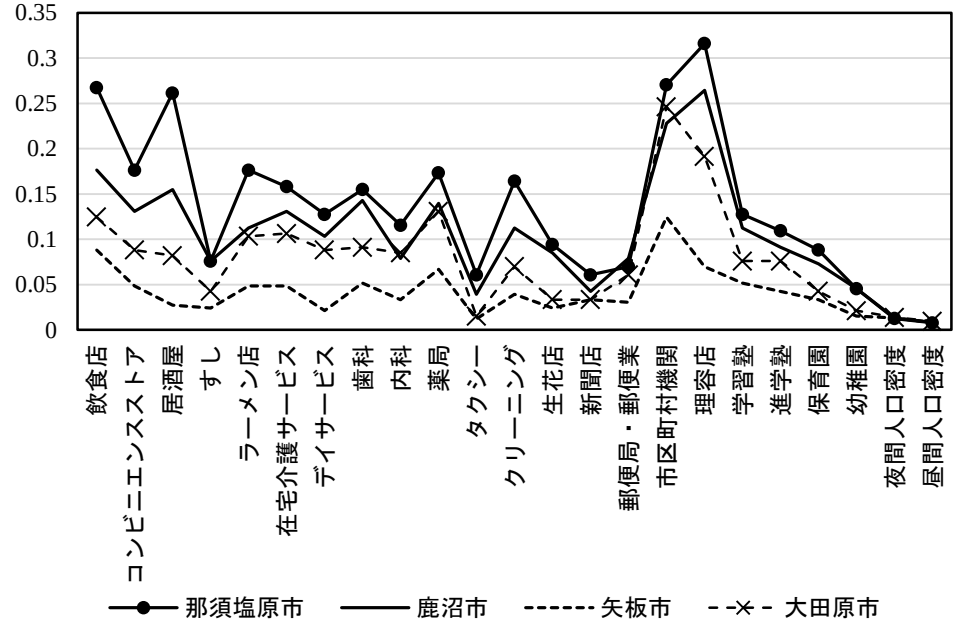


Figure 15 那須塩原市の特異性

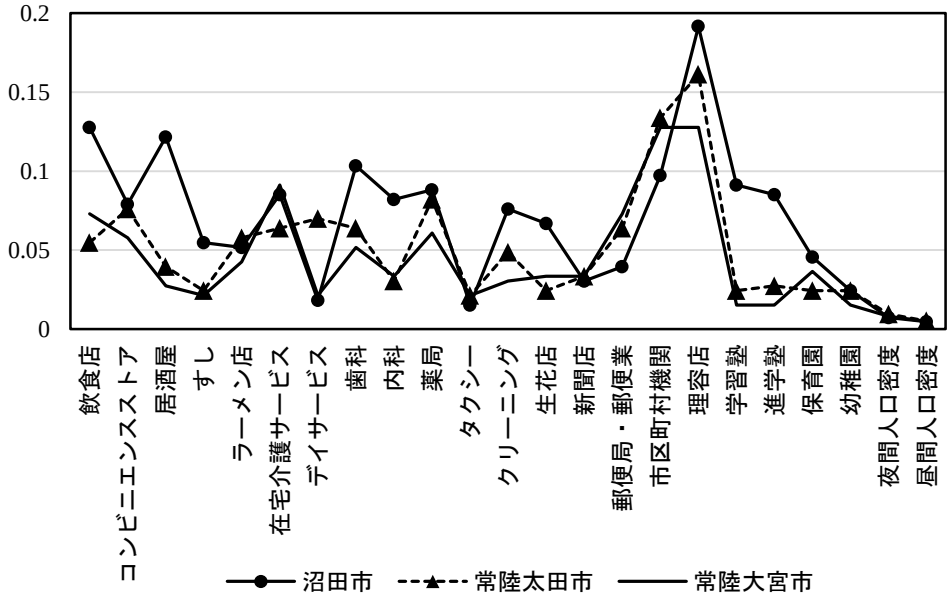


Figure 16 沼田市の特異性

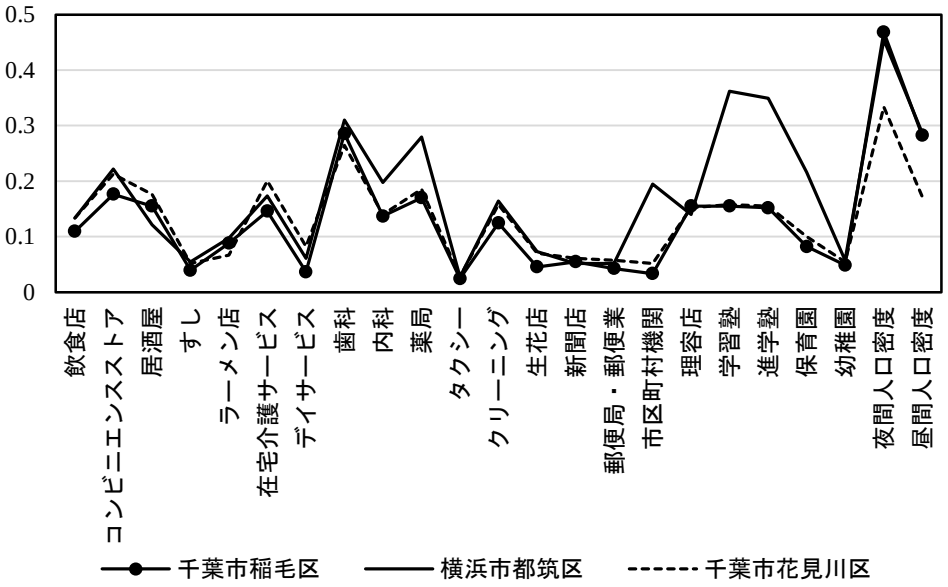


Figure 17 千葉市稲毛区の特異性

Table 12 抽出した特異地域の考察

抽出した特異地域	考察
常総市 （Figure 13）	筑西市を除外すれば那珂市と下妻市と比較して飲食店が多く，特に市区町村機関と理容店の数値が高い。筑西市が抽出されていないことは課題である。
横浜市金沢区 日野市 （Figure 14）	横浜市金沢区と日野市の両人口密度が同等なので両市をあわせている。戸田市と昭島市よりも多くの値が高い。特に両市とも在宅介護サービス，歯科・内科・薬局の医療機関が多い特徴を有する。
那須塩原市 （Figure 15）	両人口密度が 3 地域とも同等であるが，那須塩原市はほぼすべての事業所数が多い。新幹線の駅が 2 つあり温泉地がある。大田原市も観光資源を多く抱えるが新幹線の駅がない。
沼田市 （Figure 16）	常陸太田市と常陸大宮市と比較して事業所があきらかに多い。

千葉市稲毛区 （Figure 17）	人口密度が同じ横浜市都筑区と比較すると教育関係が顕著に少ない。千葉市花見川区と比較すると多くの事業所数が同等にもかかわらず量人口密度が高い。

7．おわりに

数理モデルの構築が困難な多量の統計データにおいて特異なレコードを抽出したい場合，回帰分析，正準相関分析，ランダムフォレスト回帰，ニューラルネットワーク回帰などを用いてモデルを求め，データとモデル計算値との誤差が大きい，もしくは正答率が低いデータを「標準値ではない」「特異値である」「外れ値」として判断することができる。モデルやツールはいくつもあるが，目的変数（出力データ）が複数の場合，ニューラルネットワークは有効な手法である。

従来は多量データに対して中間層の列数がひとつなどの基本的なニューラルネットワークモデルでは学習成功に至らなかったが，最近の計算機の処理能力の向上により大規模なデータに対して中間層のユニット数や列数を増やしても実時間で学習成功が可能となった。

以上の背景から，本稿はニューラルネットワークを用いて多量の統計データから特異なレコードを抽出する方法を提案し, 2 種類の統計データに適用して提案手法の有効性を確認した。提案手法の課題は，具体的に Figure 13 の筑西市が偶然にも特異な地域と考えられたが抽出されていないこと，回帰分析との抽出地域の違い，これらのことから，多量の統計データにおいて抽出すべき地域が既知なケースに対して適用することである。また，収束値をゼロに近い値に設定して学習成功となった場合，特異地域のレコードが埋もれてしまうことが考えられ，この点も課題である。

参考文献

[1] 坂本，ニューラルネットワークモデルによる特異地域の抽出，電気学会全国大会 G401-B1，2019
[2] 山口，高橋，竹内，多変量解析の基本と仕組み，秀和システム，2004
[3] （社）人工知能学会（監），深層学習，近代科学社，2015
[4] 小高，機械学習と深層学習，オーム社，2016
[5] 小高，強化学習と深層学習，オーム社，2017
[6] 澤田，サポートベクトル回帰を用いた地域人口の推定－国土データ基盤から算出した地域特徴量の考察－，愛知大学情報メディアセンター紀要，Vol.26, No.1, 2016
[7] 森，流入移動ポテンシャル指標による移動面での特異地域の検出—新潟市を事例とした小地域統計による分析—，法政大学日本統計研究所，オケージョナル・ペーパーNo.94，2018
[8] 平野，C でつくるニューラルネットワーク，パーソナルメディア，1991
[9] 金，R によるデータサイエンス，森北出版，2007
[10] 白井，岩田，久間，浅川，基礎と実践ニューラルネットワーク，コロナ社，1995
[11] 熊沢，学習とニューラルネットワーク，森北出版，1998
[12] （社）人工知能学会（監），What’s AI，http://www.ai-gakkai.or.jp/whatsai/ （2019 年 2 月 10 日確認）
[13] N.Sakamoto: A Method to Evaluate an Urban Area by Using the Model That Calculates a Number of Facilities from an Area and a Population, Current Urban Studies, 2016, DOI:10.4236/cus.2016.44028
[14] 坂本：事業所充実度の評価に関する検討（タウンページと経済センサス），日本 OR 学会，2017 年秋季研究発表会
[15] 坂本：事業所数評価モデル（政府統計使用による新潟県コンパクトシティの評価），経済志林（法政大学経済学部学会），Vol.85, No.2, pp.147-165, 2018

付録

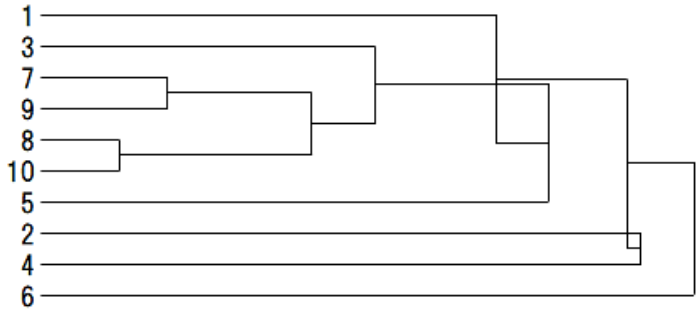


Figure ① 重心法，非正規化，平方距離

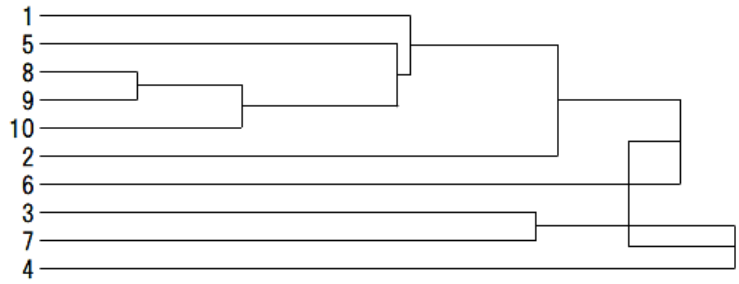


Figure ② 重心法，正規化，平方距離

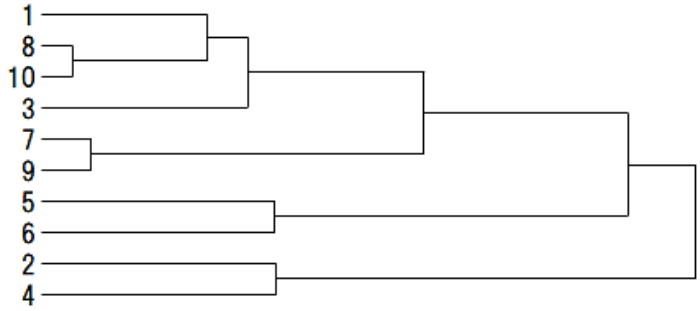


Figure ③ ウォード法，非正規化，平方距離

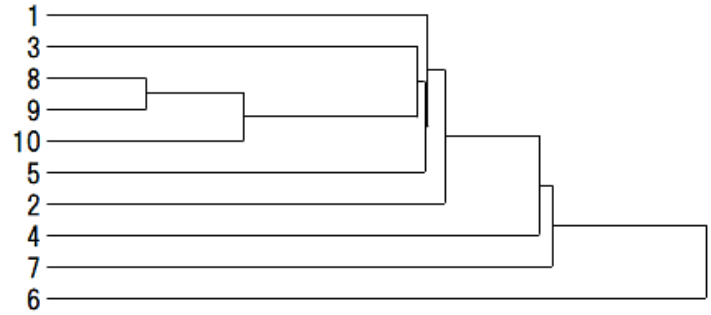


Figure ④ 最近隣法，正規化，平方距離

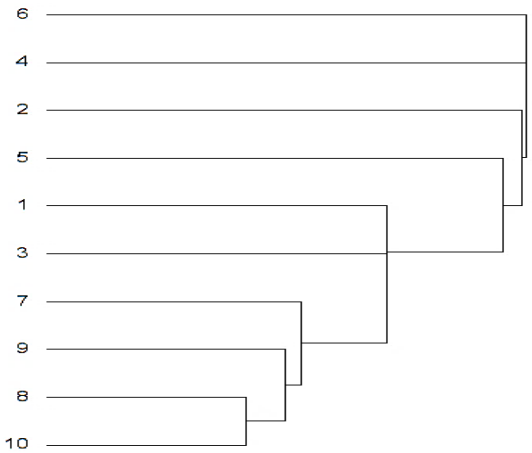
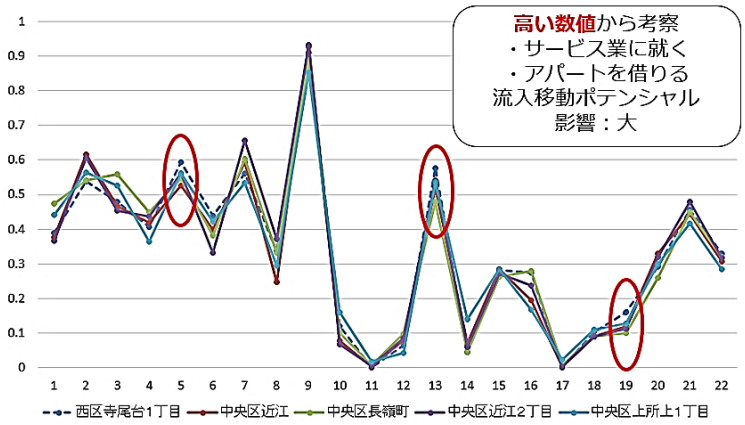


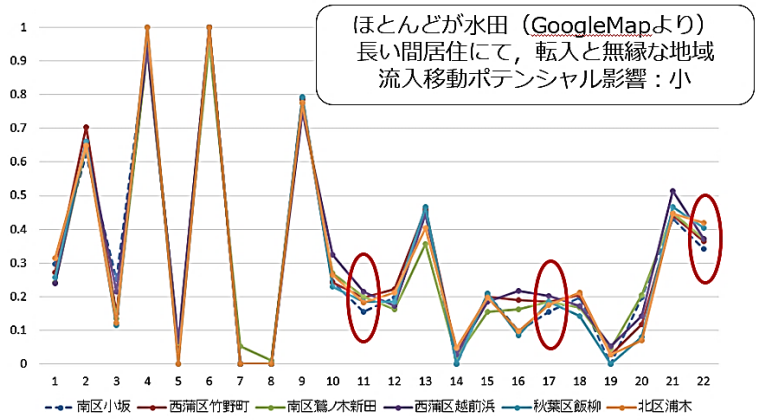
Figure ⑤ R によるクラスター分析



西区寺尾台１丁目

高い数値　5:借家率，13:サービス業就業者率，19:1年未満居住者率

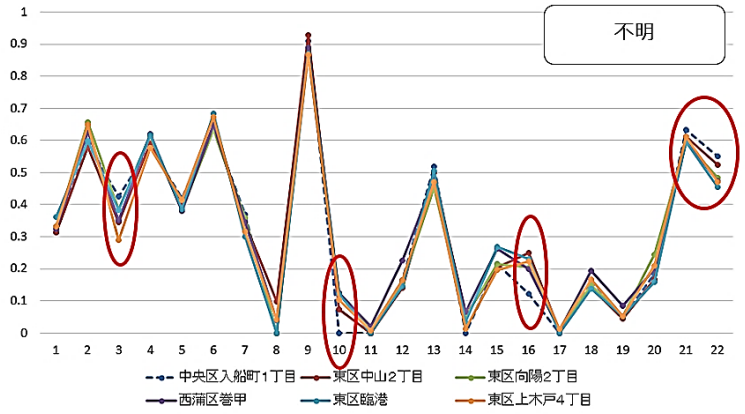
Figure A **西区寺尾台 1 丁目**



南区小坂

低い数値　11:農林漁業者率，17:農林漁業従事者率，22:20年以上居住者率

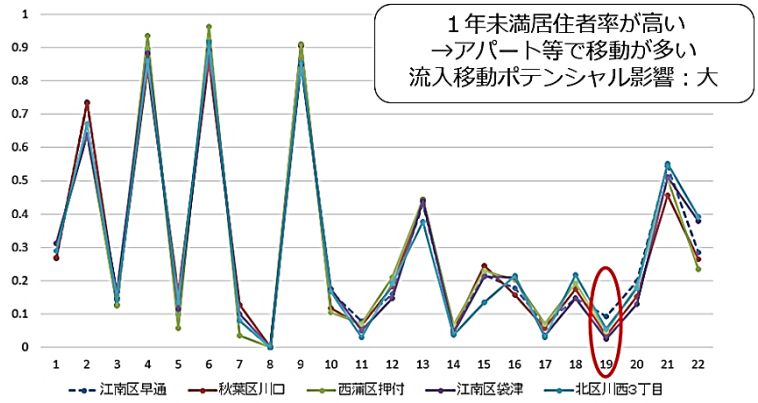
Figure B **南区小坂**



中央区入船町１丁目

高い数値　3:単独世帯率，21:10年以上居住者率，22:20年以上居住者率
低い数値　10:自営・家族従業者率，16:サービス職業従事者率

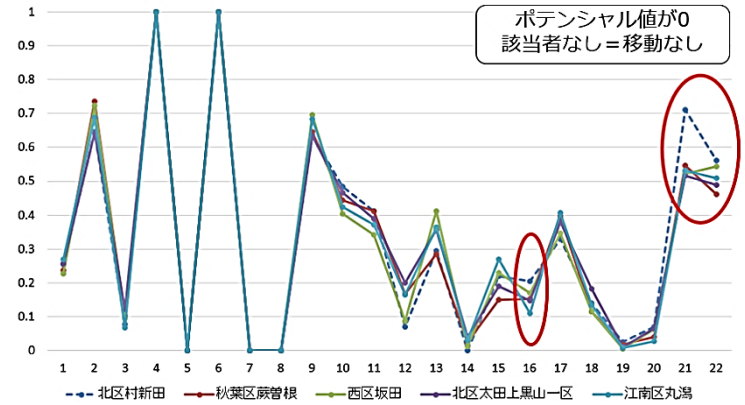
Figure C　**中央区入船町 1 丁目**



江南区早通

高い数値　19:1年未満居住者率

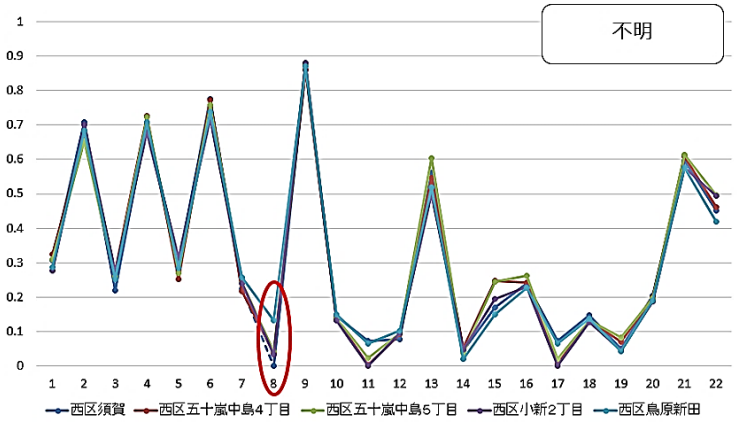
Figure D　**江南区早通**



北区村新田

高い数値　16:サービス職業従事者率，
21:10年以上居住者率，22:20年以上居住者率

Figure E　**北区村新田**



西区須賀

低い数値　8:3階建て以上住居世帯率

Figure F　西区須賀

Table Ⅰ　中間層ユニット数と中間層列数の検討

中間層ユニット数	中間層列数	5 回平均	標準偏差
21	1	2.90	0.05
	2	2.80	0.25
	3	2.74	0.28
	4	2.88	0.37
42	1	3.41	0.32
	2	2.86	0.23
	3	2.43	0.26
	4	2.65	0.23
	5	3.08	0.61
63	1	3.41	0.20
	2	2.80	0.19
	3	2.75	0.17
	4	2.89	0.37
	5	2.90	0.22

日本統計研究所

オケージョナル・ペーパー（既刊一覧）

号	タ イ ト ル	刊行年月
75	鉄道開業前・後期における鉄道沿線域内人口移動について－つくば EX 沿線域内 18 市・区間の移動を事例として－	2017.02
76	首都圏南西翼地域における距離帯間・距離帯内移動について	2017.02
77	首都 60 キロ圏における移動ホットスポットの検出	2017.03
78	地域間移動における転出・転入移動圏とその特徴―首都 60 キロ圏を対象地域として―	2017.04
79	首都 60 キロ圏における 20 歳代移動者の移動圏について	2017.04
80	1880 年ドイツ帝国営業調査構想について―エンゲルの「建白書」を中心にして―	2017.04
81	転出入移動圏から見た地域人口移動の方向的特性について	2017.05
82	ビスマルク政権とプロイセン統計局 1862-82 年―エンゲルのプロイセン統計局退陣をめぐって―	2017.05
83	角度情報を用いた東京 40 キロ圏の子育期世代の移動分析	2017.06
84	移動選好度による居住移動圏の検出－住民基本台帳人口移動報告「参考表」(2012-16 年)による分析―	2017.10
85	九州・沖縄地方の域内移動から見た移動圏とその構造	2018.01
86	QGIS による西武国分寺線沿線の産業構造分析	2018.02
87	The Simulation Results of Expenditure Patterns of Virtual Marriage Households Consisting of Working Couples Synthesized by Statistical Matching Method	2018.03
88	ロジャーズ-ウィルキンス・モデルの東京都の人口への応用	2018.03
89	わが国の三大都市圏における移動圏とその構造	2018.04
90	居住地移動者数の将来動向に関する一考察―2016-20 年期～2046-50 年期の都道府県間比較―	2018.04
91	男女別移動率を用いた移動者数の都道府県別将来推計	2018.05
92	ぐるなびデータを用いた店舗数に関する考察	2018.09
93	表式調査と業務統計における統計原情報の表式的集約について	2018.09
94	流入移動ポテンシャル指標による移動面での特異地域の検出―新潟市を事例とした小地域統計による分析―	2018.09

オケージョナル･ペーパー　No.95
2019 年 2 月 15 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、2326
Fax　042-783-2332
jsri@adm.hosei.ac.jp
発行人　菅　幹雄