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 ―

澤田　貴行（愛知大学地域政策学部）

要旨

人口の減少と空間的な偏りが課題である現在において，人口の正確な把握は政策立案において重要である。しかし，公的統計などを利用して関心のある地域の人口直接的に把握をすることは，収集できるデータの内容などから困難である。そこで，地域の人口を入手が容易なオープンデータを利用し，推定する手法を提案する。具体的には，地域を国土データ基盤として整備された道路，公共施設や土地利用状況などを様々な地域特徴量として表現し，それらと地域人口の関係を機械学習した予測モデルを構築して人口の推定を行う。

キーワード：人口減少，機械学習，サポートベクトル回帰，地域特徴量，地理情報システム

1. はじめに

現在，人口の減少と過疎や過密のような空間的な偏在が認識され，労働力の減少や地域コミュニティの崩壊に繋がるとして問題視されている。このようななか地方政策を行うには，これまでの人口増加を前提とした行政システムから，市町村域より小さな任意範囲の空間（以下，地域という）における人口状況や社会環境などに対応する“きめ細やかさ”を持ったシステムに変化していく必要がある。このために地域の住民やその生活を，集中や偏在といった面的な様相として把握し，状況を踏まえた政策を立案する必要があり，国勢調査を代表とする公的統計を中心に分析が行われている。しかし，公的統計の目的は国全体や市町村域内の総数を知ることであり，国勢調査では調査区域の形状を調査の行いやすい範囲として設定するなどの理由から，その結果は，必ずしも知りたい地域とは一致しない。また，小地域の組み合わせとして，地域を見ようとすると，地域以外の人口がノイズとして混入し，正確な把握は困難となる。なお，このような地域における人口の現状把握の難しさを，小西[i]は，地方公共団体の統計データ活用の状況を分析し，地方計画の策定に小地域統計データが利用できていないと指摘している。

地域の状況を利用して目的とする値を推定した研究には，以下のものがある。澤田ら[ii]は，目的を地域の人口把握とし

て，地域に重なる国勢調査の小地域人口を，地域と重なる部分と重ならない部分に分け，それぞれのテレポイント数の割合によって求め，その人口を足し合わせる手法を提案した。しかしながら，テレポイントは一般的に入手し難いデータであり，また，その精度には疑義があることが課題である。堤ら[iii]は，目的を大都市圏の地価推定として，都市圏の任意地点で網羅できるデータとして，土地利用面積，最寄駅から主要駅までの都心鉄道距離などを説明変数にして，trans-Gaussian krigingと呼ばれる非線形性を考慮した手法を提案した。しかし，郊外部において推計誤差が大きいことが確認され，地域による説明変量の密度について相違があるという課題を指摘しており，公的データに依存した使い方には，工夫が必要であることを示唆している。

そこで，本研究では，地域の人口推計を入手が容易なオープンデータから地域毎に求めた特徴量（以下，地域特徴量という）の集合（以下，地域特徴量ベクトルという）から推計する手法を提案する。具体的には，地域メッシュ[ivv]の基準地域メッシュ（緯度差30秒，経度差45秒で，1辺の長さは約1km，以下，3次メッシュという）を一つの地域と捉え，国土交通省により整備されたオープンデータから得られる道路，公共施設などの存在を地域毎に空間的計測により定量化した地域特徴量ベクトルと地域人口を一組として構築したデータセットを構築する。その上で構築したデータセットを利用して地域の人口と地域特徴量ベクトルの関係として，機械学習することで人口推定を行う。

2. 地域の人口推計

2.1 教師付き機械学習

地域の人口と地域特徴量ベクトルの関係を捉えるために予測モデルは目的変数を，地域の人口とし，説明変数を地域特徴量ベクトルとして，学習する教師つき機械学習手法により構築する。教師付き機械学習を用いた人口推定の概要を図1に示す。図1では，まず，地域の本来の人口と地域特徴量ベクトルからなるデータセットを用意し，訓練とテストのデータセットに分割する。次に訓練フェーズとして，訓練データセットを教師データとして，地域人口と地域特徴量ベクトルの関係を機械学習することで予測モデル

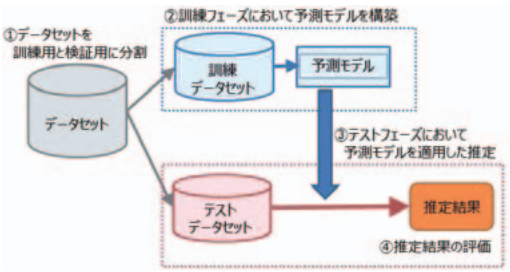


図1　教師付き機械学習による推定

を構築する。さらに，テストフェーズで，テストデータセットの地域特徴量ベクトルに予測モデルを適用して推定人口を算出する。なお，評価は本来の人口と推計人口の差異を評価する。

2.2　サポートベクター回帰

人口を推定する関係を近似する予測モデルの作成には，サポートベクター回帰（以下，SVRという）[vi]を利用する。SVRは分類問題において，近年注目されているサポートベクターマシン（以下，SVMという）[vii]を回帰問題へ拡張したものである。SVMは，教師付き機械学習を利用した識別器であり，分類問題において，入力となる特徴量の高次元空間における最適な分離超平面を見つけるもので，高い汎化能力が示されており，回帰問題への拡張であるSVRも高い汎化能力が期待される。

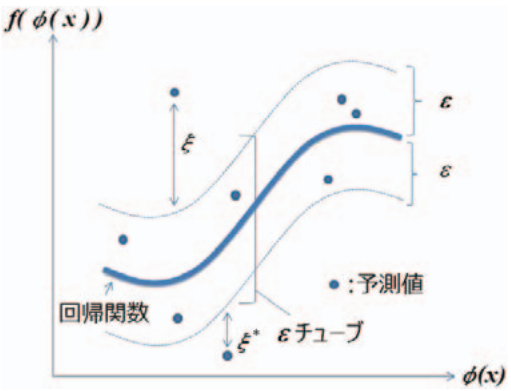


図2　εチューブとスラック変数ξ，ξ^*の関係

SVRは，入力 $x_i \in \mathbb{R}^n, i = 1,2,\ldots,l$ から出力 $y_i \in \mathbb{R}^n, i = 1,2,\ldots,l$ を回帰する非線形回帰の一つである。入力 x_i からなる特徴空間への非線形写像 $\varphi(x)$ を考え，写像後の特徴空間において線形回帰を行う。出力では，εチューブと呼ばれる一定範囲内に入らない場合は外れた分の距離を表すスラック変数ξ，ξ^*に応じたペナルティが与えられるため，ξ，ξ^*が小さくなるような最適な回帰係数を算出する。εチューブとスラック変数ξ，ξ^*の概念を図2に示す。

SVRでは，回帰関数 f を式（2.1）とする。ωはl 次元の重みベクトル，bはバイアス項である。

$$f(x) = \omega^t \Phi(x) + b \qquad (2.1)$$

本稿では，SVRのなかでもε-SVRを利用する。ε-SVRは，予め定めたC>0，ε>0，スラック変数ξ，ξ^*を用いて，式（2.2）として定式化される。

$$\min_{\omega,b,\xi,\xi^*} \quad \frac{1}{2}\omega^t\omega + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^* \qquad (2.2)$$

$$subject\ to \begin{cases} \omega^t\Phi(x_i) + b - y_i \le \varepsilon + \xi_i, \\ y_i - \omega^t\Phi(x_i) - b \le \varepsilon + \xi_i^*, \\ \xi_i, \xi_i^* \ge 0, \quad i = 1,\ldots,l \end{cases}$$

ここで非線形回帰を線形回帰として扱うために，$\omega^t\ \Phi(x_i)$ を $K(x_i, x_j) = \Phi(x_i)^t\ \Phi(x_j)$ と置き換え，ラグランジュ未定乗数法を用い，α_iと α_i^*をラグランジュ乗

数として，式（2.2）は，式（2.3）の最適化問題に帰着する。

$$\min_{\alpha,\alpha^*} \quad \frac{1}{2}(\alpha-\alpha^*)^t K(x_i,x_j)(\alpha-\alpha^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i-\alpha_i^*) \tag{2.3}$$

$$subject\ to \begin{cases} \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0, \\ 0 \le \alpha_i, \alpha_i^* \le C, \quad i=1,\dots,l \end{cases}$$

これを解くと最終的には回帰関数は式（2.4）式となる。

$$f(x)=\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)\,K(x_i,x_j)+b \tag{2.4}$$

ここで$K(x_i, x_j)$は，カーネル関数と呼ばれ，代表的なものに式（2.5）で示す線形カーネル，式（2.6）で示す多項式カーネルや式（2.7）で示すガウシアンカーネル（Radical Basis Function，RBFカーネル）などがある。

・線形カーネル

$$K(x_i,x_j)=x_i^t x_j \tag{2.5}$$

・多項式カーネル

$$K(x_i,x_j)=(\gamma\, x_i^t x_j+\alpha)^p \tag{2.6}$$

・RBFカーネル

$$K(x_i,x_j)=exp(-\gamma\,\|x_i-x_j\|^2) \tag{2.7}$$

ただし、$\gamma=\frac{1}{2\sigma^2}$、$\sigma>0$

p，γ，αはパラメータであり，この値によって回帰特性は大きく変わるが，問題に適したパラメータの推定によって高い汎化性能が得られることが示されている。加えて，ε-SVRでは，ε，Cもパラメータとなる。なお，誤りを許すように制約を緩めることを「ソフトマージン」と呼び，Cにより制御することができる。しかし，「ハードマージン」と呼ばれるCを大きくしたときには過学習をおこしやすくなるため，利用に際しては，汎用性を考慮しなければならない。

2.3　地域特徴量と特徴量ベクトル

地域人口を推計するため，人口と地域を特徴づけた地域特徴量ベクトルからなるデータセットを作成する。地域特徴量は，図3に示すように地域に存在す

図3　GISによる地域の特徴量算出

る地物を地域の外形によって抽出することを地理情報システム（Geographic Information System，以下，GISという）により行い，個数，長さ，面積として計測する。また，地物自身の持つ属性値も利用できる。

さらに本稿では，地物が地域に与える影響を地物までの“近さ”と周辺にある地物の“多さ”として考慮するために，さらに2つの特徴量を求める。“近さ”に関しては，地物からの距離として表現できるので，地物を0と1の2値でラスター化したのちに，施設からの遠さを距離として表した同心円状の形状を算出し，新たなラスターデータのセル値として格納しておき，このラスターデータの地域にあるセル値の平均値を地域特徴量とする。これは地物までの最短距離を表すことから近傍距離値と定義する。また，“多さ”に関しては，地物の存在を0と100の2値でラスター化したのちに，図4に示すように中心セルからの同心円状の近傍セルの平均値を新たなラスターデータのセル値として格納しておき，地域にあるこのラスターデータのセル値の平均値を地域特徴量とする。これは近傍地域の地物の多さを表すことから近傍地物値と定義する。近傍距離値と近傍地物値により地域内における地物の存在だけでなく，近傍地域を含めて地物の影響を考えることができる。

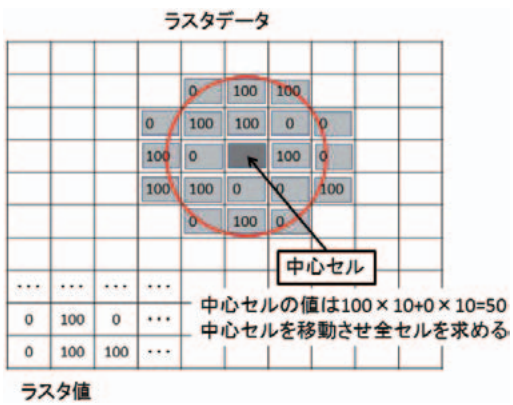


図4　近傍地物ラスターデータの作成

2.4　人口推計モデル

これらを踏まえ，本稿における人口推計モデルの概要は図5のようになる。ここで推定する人口は，政府統計の総合窓口（以下，e-Statという）[viii]より取得できる国勢調査の結果を利用する。地域特

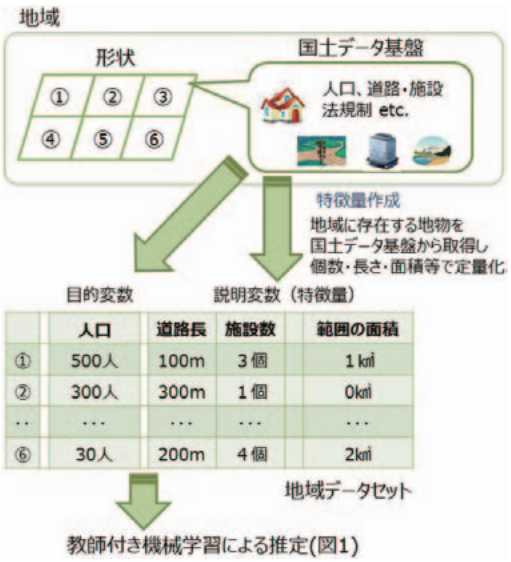


図5　人口推計モデルの概要

徴量の算出では，国土データ基盤として国土交通省国土政策局国土情報課より提供される国土数値情報ダウンロードサービス（以下，国土数値情報という）[ix]から取得できるデータを利用する。国土数値情報では，国土，政策区域，地域，交通という4つのカテゴリに，土地利用，公共施設，バスルート等のさまざまな地物や国土利用計画法[x]等で定められた範囲等のデータが提供されている。取得したデータは，地理空間情報として実在する位置・形状としての空間情報と，それに与えられた名称等の非空間の属性情報から構成されるため，GISにより空間的な計測や必要な属性情報等の取得が可能である。

3. 評価実験

人口推計モデルの評価を行うための対象地域を愛知県，長野県，静岡県とし，それに含まれる地域から作成する地域データセットから人口推計モデルを構築し，推定結果の評価を行った。

3.1 地域データセットの作成

地域データセットにおける1つの地域は，3次メッシュ（1辺の長さ約1km）の1つのメッシュとした。3県に含まれる第1次地域区画（1辺の長さ約80km）は，図6で示す5136，5137，5138，5236，5237，

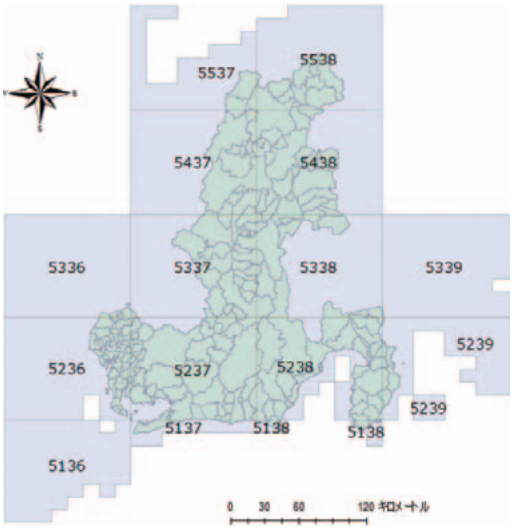


図6　取得した地域データセットの範囲

5238，5239，5336，5337，5338，5339，5437，5438，5537，5538の15コードであり，その中に含まれる3次メッシュは，26,419件であり，これを対象地域とし，地域データセットを作成した。

3.2 地域人口

e-Statから平成22年国勢調査（国勢調査-世界測地系1kmメッシュ）につい

表1　3次メッシュ人口の特徴
（総数26,419）

平均値	507
中央値	2
標準偏差	1,417
最大値	16,332
最小値	0

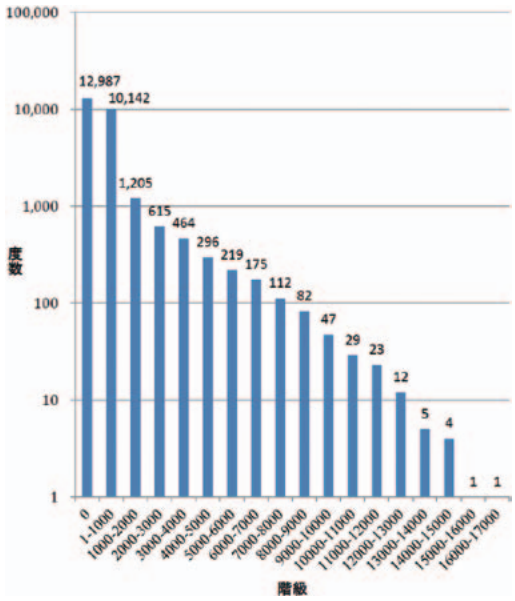


図7　3次メッシュ人口の状況(1)

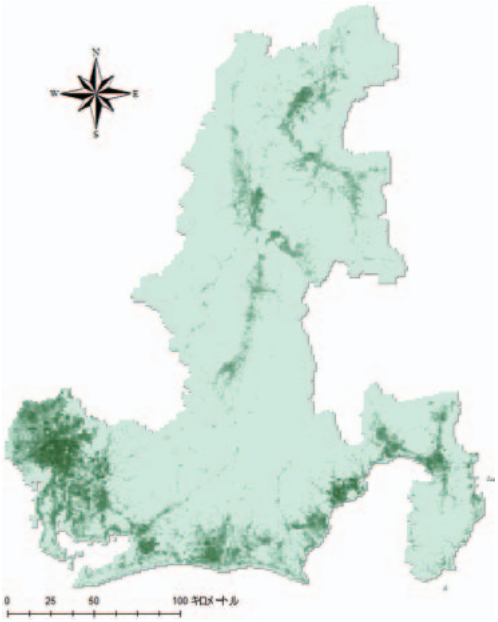


図8　3次メッシュ人口の状況(2)

て，第1次地域区画15コード分を取得し，26,419件の地域データセットに格納した。地域人口に関する基本的な統計量を表1に示す。また，人口を0及び1,000人単位で集計した頻度を図7に，3次メッシュの人口数の多さを色の濃さで表現した地図を図8に示す。これらより，地域人口は大きく偏っていることが分かる。

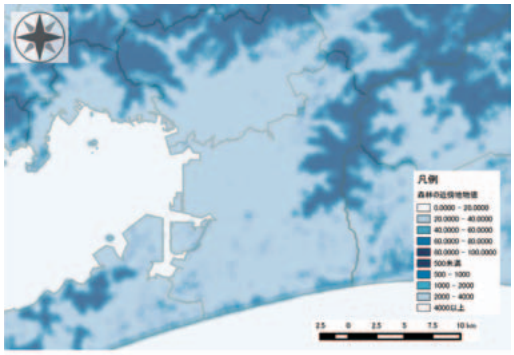
3.3　地域特徴量

国土数値情報を利用した地域特徴量の算出には，2章3節で述べたように様々なものを利用することができるが，3次メッシュにおける取得の容易さを考慮し，3次メッシュと同範囲であるデータとして，土地利用，標高･傾斜度，道路密度・道路延長を取得した。なお，土地利用に関しては，より詳細な細分メッシュ(1辺の長さ約100m）も取得した。また，人口と関連の高いことが予想される地物を収録したデータとして，避難所，公共施設，バス停，駅も取得した。さらに，国土の基本構想を表した国土利用計画法における農業，森林，都市，自然保全，自然公園の地域に関するデータも取得した。以下に取得データ毎に地域特徴量の算出法を示す。

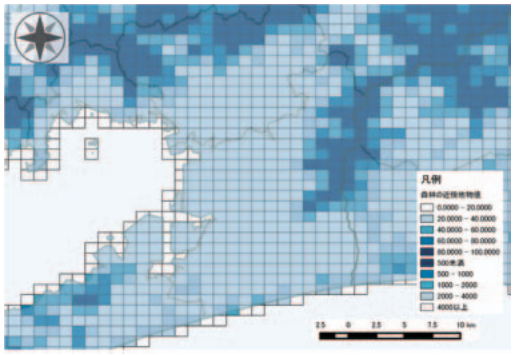
○土地利用

平成21年度の土地利用3次メッシュデータ[xi]と土地利用細分メッシュデータ(1辺の長さ約100m)[xii]について，第1次地域区画15コード分を取得した。3次メッ

シュデータでは，土地利用種別に基づき土地利用種別毎の面積を格納しているため，その面積を地域ごとの土地利用種別毎の総和で正規化したのち地域特徴量とした。また，細分メッシュデータは，細分コード毎に土地利用種別を判別した結果を格納しており，土地利用種から求められるそれぞれの“多さ”を表現する近傍地物値を地域特徴量とした。近傍地物値の算出は，土地利用種別毎の範囲を解像度50mとした範囲ラスターを作成したのち，中心セルからの近傍距離を1kmとして検索したセル値を集計して求めた平均値による近傍地物ラスターを構築し，地域に含まれる近傍地物ラスターのセル値の平均を求めた。図9に森林から求めた近傍地物ラスターと近傍地物値を示す。



森林（種別500）から作成した近傍地物ラスター



近傍地物ラスターから作成した近傍地物値

図9　近傍地物ラスターと近傍地物値（豊橋市周辺）

土地利用種別は12利用区分（田，その他の農用地，森林，荒地，建物用地，道路，鉄道，その他の用地，河川地及び湖沼，海浜，海水域，ゴルフ場）であり，土地利用種別毎に，3次メッシュデータから正規化済み面積の12次元，細分メッシュデータから近傍地物値の12次元を地域特徴量とした。

○標高・傾斜度

平成23年度の標高・傾斜度3次メッシュデータ[xiii]について，第1次地域区画15コード分を取得した。3次メッシュデータでは，傾斜度等の値を格納しているため，最大傾斜角度，最小傾斜角度，平均傾斜角度の3次元を地域特徴量とした。

○道路延長

平成22年度の道路密度・道路延長メッシュ[xiv]（3次メッシュ）について，第1次地域区画15コード分を取得した。3次メッシュデータでは，幅員毎に道路延長や1km^{2}当りの換算値を格納しているが，幅員毎に細分化すると値が小さくなり，

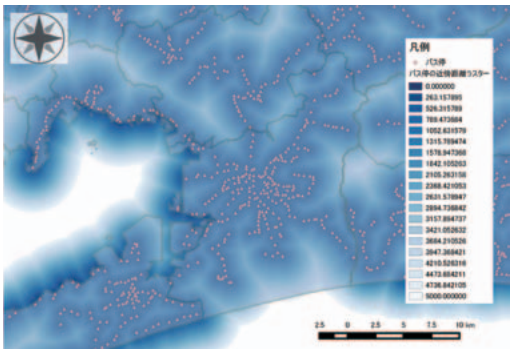
特定幅員のデータ値が欠け，地域の網羅性を欠くことから，幅員合計の道路延長1km^{2}当り換算値のみを地域特徴量とした。

○施設（公共施設・避難所・バス停・駅）

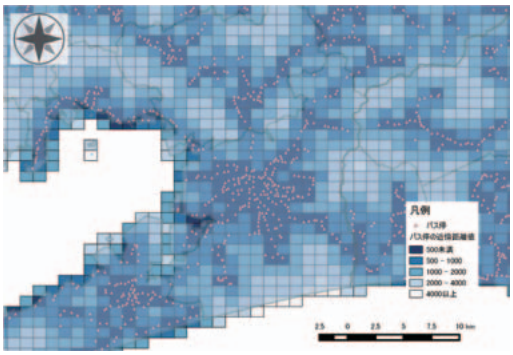
公共施設データ[xv]は平成18年度，避難施設データ[xvi]は平成24年度，バス停留所データ[xvii]は平成22年度，駅データ[xviii]は平成25年度について，それぞれ愛知県，長野県，静岡県の3県分を取得した。各種施設に対して，大小分類等の識別を行う属性値はあるものの，細分化すると値が小さくなり，特定識別のデータ値が欠け，地域の網羅性を欠くことから，地域に含まれる総数（公共施設，避難施設，バス停留所）や長さ（駅）を地域特徴量とした。また，施設の“近さ”を表現する近傍距離値を地域特徴量とした。近傍距離値の算出は，施設の有無を解像度50mとした施設ラスターを作成したのち，施設からの遠さを距離として表す同心円状の形状を算出した近傍距離ラスターを構築し，地域に含まれる近傍距離ラスターのセル値の平均を求めた。また，近傍距離値の距離による影響を緩和するために対数化した値も地域特徴量とした。図10にバス停留所から求めた最短距離ラスターと最短距離値を示す。

○国土利用計画法

国土利用計画法により定められた平成23年度の農業地域xix，森林地域[xx]，自然保全地域[xxi]，自然公園地域[xxii]，都市地域[xxiii]について，それぞれ愛知県，長野県，静岡県の3県分を取得した。地域毎のそれぞれの指定地域を切り出した面積を地域自身の面積で正規化したのち地域特徴量とした。また，地域周辺の指定地域毎の“多さ”を表現する近傍地物値を地域特徴量とした。近傍地物値の算出は，それぞれの範囲を解像度50mとした範囲ラスターを作成したのち，中心セル



バス停留所から作成した近傍距離ラスター



近傍距離ラスターから作成した近傍距離値

図10　近傍距離ラスターと近傍距離値（豊橋市周辺）

からの近傍距離を1kmとして検索したセル値を集計して求めた平均値による近傍範囲ラスターを構築し，地域に含まれる近傍範囲ラスターのセル値の平均を求めた。

3.4　地域特徴量ベクトル

3章3節により求めた地域特徴量を利用し，表2で示す組み合わせにより地域特徴量ベクトルを3種類作成した。地域特徴量ベクトルはすべて25次元であり，最も単純なNo.1をベースラインとし，No.2は考案した近傍距離値と近傍面積値を取り入れたもの，No.3では，さらに近傍距離値を対数化したものである。

3.5　実験と評価の方法

作成した地域データセットを利用し，SVRによる人口推定モデルを作成し，地域の人口予測を行った結果を評価した。ここで推定モデルの学習とテストは，図11に示す5分割交差検定により行った。5分割交差検定とは，データセットを5分割し，1つをテストデータセットに，残り4つを訓練データセットとして，すべての分割したデータセットに対する推定を5通りで繰り返すものであり，結果，すべての地域が訓練データセットとテストデータセットに用いられる。

評価指標は，予測人口と本来人口との差の平方を求め，その平均の平方根をとして，式（3.1）で表す平均二乗誤差（Root Mean Squared Error，RMSE）と式（3.2）で表す予測モデルがどの程度，本来人口に当てはまるかを残差平方和と偏差平方和の割合で評価する決定係数（Coefficient of Determination, R^2）を利用した。なお，交差検定のRMSE及びR^2は，交差毎に求められるが，本稿では全地域データセットのそれぞれの予測

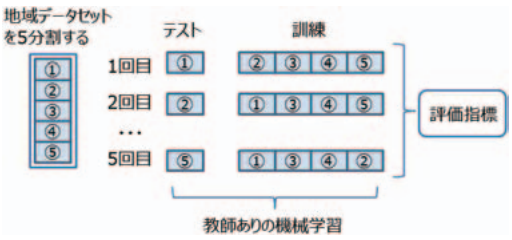


図11　交差検定と推定評価指標

表2　地域特徴量の組み合わせと地域特徴量ベクトル

No.	概　要（カッコ内は次元数）				
	土地利用（12）	傾斜（3）	道路（1）	施設（4）	国土計画（5）
1	正規化面積	最大傾斜角，最小傾斜角，平均傾斜角	幅員合計の道路延長の1km^{2}換算値	施設数，長さ	正規化面積
2	近傍地物値			近傍距離値	近傍地物値
3				近傍距離値（対数）	

結果から求める推定評価指標により評価をした。

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{N}} \quad (3.1)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y_i})^2}{\sum_{i=1}^{N}(y_i - \overline{y_i})^2} \quad (3.2)$$

N：予測対象数

y_i；実測値，$\overline{y_i}$：実測値の平均，

$\hat{y}_i$：予測値

3．6　システム実装とSVRパラメータ

実験システムの構築は，SVMとSVRを比較的に容易に実装できるように開発されたLibSVM[xxiv]を利用し，プログラミング言語C#において動作するlibsvm-NET[xxv]を用いて構築した。なお，SVRに関するカーネル関数などのパラメータは，任意数のデータセットを利用した5交差検定法によりRMSE及びR^2により評価し決定する。

4．実験結果

実験は2つ行う。1つは任意数のデータセットを利用してSVRに関するカーネル関数やそれに必要なパラメータの選定をする実験，もう1つは，決定内容を利用した3種類の地域データセットによる人口の推定をする実験である。

表3　実験で利用したパラメータ

Linearカーネルについて

ε	0.1（デフォルト値）
C	1（デフォルト値），2，4，8

RBFカーネルについて

ε	0.1（デフォルト値）
C	1（デフォルト値），2，4，8
γ	0.01，0.04（デフォルト値），0.1

Polyについて

ε	0.1（デフォルト値）
C	1（デフォルト値），2，4，8
γ	0.01，0.04(デフォルト値)，0.10
α	0（デフォルト値），1
p	3（デフォルト値），4，5

4.1　カーネル関数とパラメータ推定

カーネル関数には，式（2.5）で表す線形カーネル，式（2.6）で表す多項式カーネル，式（2.7）で表すRBFカーネルについて，No.1の地域データセットから無作為に抽出した5,284件(地域データセットの1/5相当)を利用し，表3に示すパラメータ値により評価をした。ただし，εは式（2.2）よりカーネル関数には依存しないため，どのカーネル関数にも同様の効果があることが想定され，デフォルト値$\varepsilon = 0.1$により固定して行うこととした。

●線形カーネルについて

$\varepsilon = 0.1$とし，Cに関して値を変化させた結果を図12に示す。

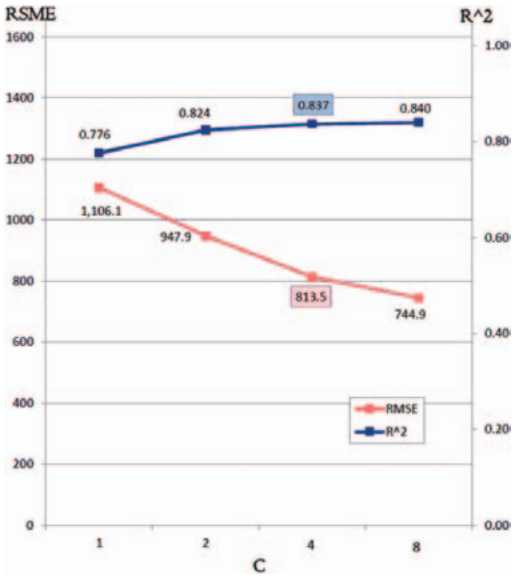


図12　線形カーネルにおけるCの影響

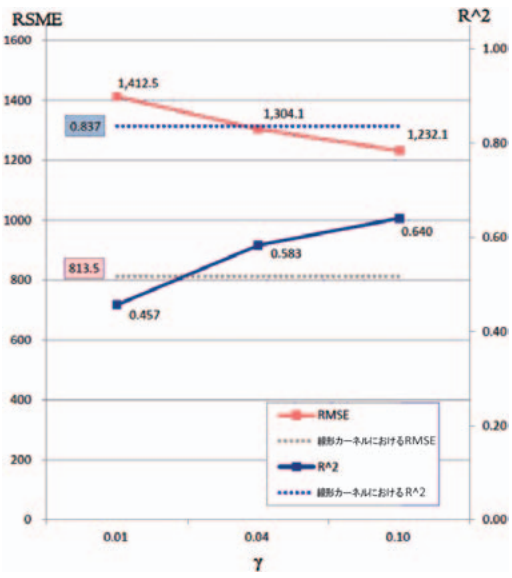


図13　RBFカーネルにおけるγの影響

図12からCを大きくしたとき，評価指標値がともに向上することがわかった。しかし，2章2節で述べたようにCは誤りに対するペナルティを表しており，大きくすると過学習が懸念されること，及び改善の効果は，$C=4$以降は鈍化していることを考慮し，$C=4$を最適なパラメータとした。なお，Cはεと同様にカーネル関数に依存するものではないため，以降の実験においては固定して行うこととした。

●RBFカーネルについて

$\varepsilon=0.1$，$C=4$とし，γに関して値を変化させた結果を図13に示す。

図13からγを変化させても線形カーネルにおける評価指標値を上回らなく，人口予測においては，RBFカーネルは適していないことがわかった。

●多項式カーネルについて

$\varepsilon=0.1$，$C=4$とし，また，$\alpha=0$，$p=3$（ともにデフォルト値）として，γに関して値を変化させた結果を図14に示す。図14からγを大きくすることで評価指標値が向上し，$\gamma=0.10$において線形カーネルにおける評価指標値を上回ることがわかった。

そこで$\varepsilon=0.1$，$C=4$，$\gamma=0.10$として，αとpに関して値を変化させた結果を図15に示す。図15からαとpともに大きくても線形カーネルにおける評価指標値を下回ることはなく，大きいほど，評価指標値が向上することがわかった。しかしながら，式（2.6）からpを大きくしすぎるとき，カーネル関数が複雑になり，最

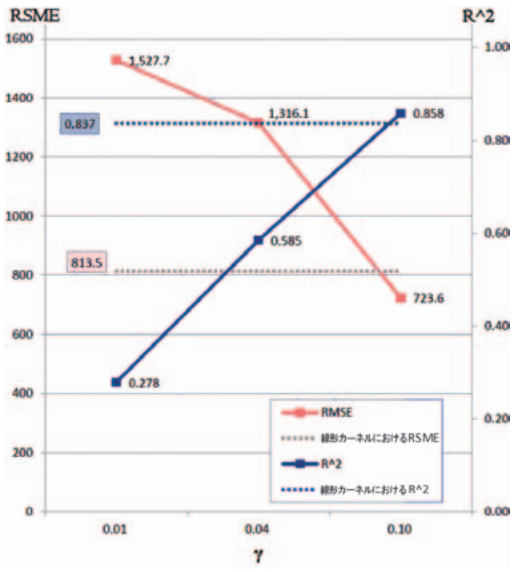


図14　多項式カーネルにおけるγの影響

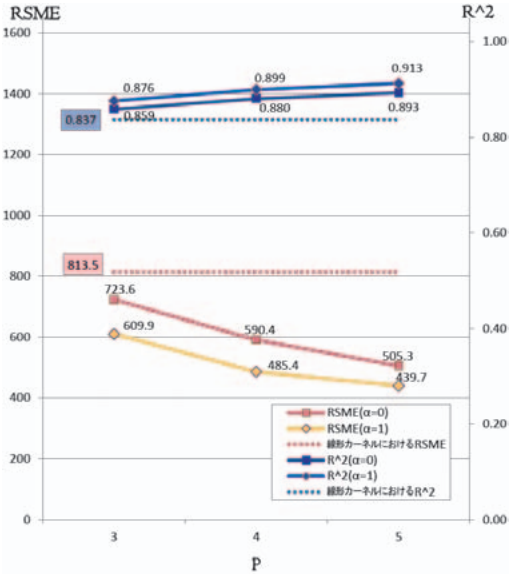


図15　多項式カーネルにおけるαとpの影響

適化に非常に時間がかかることから$\alpha = 1$及び$p = 5$を最適なパラメータとした。

4.2　3種類の地域データセットによる人口推定

4章1節の結果と過学習を抑えた汎用性を踏まえ，人口予測モデルを多項式カーネル（$C = 4$，$\varepsilon = 0.1$，$\gamma = 0.1$，$\alpha = 1$，$p = 5$）とするSVRにより構築した。

3種類の地域データセットを用いて行った人口推定の実験結果を表4に示す。表4から，考案した近傍地物値と近傍距離値の対数を地域特徴量に取り入れたNo.3の地域データセットが最良の結果となった。No.3は，土地利用の面積値を近傍地物値とし，施設の数・長さを近傍距離値の対数とし，国土計画の近傍地物値としたものである。地域から土地利用種の面積などをそのまま地域特徴量としたNo.1の地域データセットに対し，R^2において0.09ポイント，RSMEにおいて19.702ポイントの改善となった。

つまり，人口予測において，No.1のように地域に存在することによる事実をそのまま特徴とすることより，地域の周辺地域の状況を考慮することが重要であることがわかった。これは，例えば，ある

表4　地域データセット別の評価指標比較

評価指標	No.1 (Baseline)	No.2	No.3
決定係数 R^2	0.898	0.897	0.907
平均二乗誤差 RSME	451.496	453.752	431.794

地域を考えるとき，当該地域だけでなく周辺地域も含めて考える我々がとる行動と同様のことであり，妥当なものであると考えられる。

しかしながら，地域ごとに実人口と推定人口における差の絶対値（以下，誤差という）を観察してみると，かなり大きなものも確認される。実人口を階級分けして，0人及び1,000人単位で区分けし集計した誤差の平均を図16に示す。図16から人口が多い地域ほど誤差が大きくなることがわかる。また，図17に地域ごとの誤差がどのようになっているかを示す。名古屋市周辺などの人口の多い地域で大きな誤差が生じていることがわかる。これは，図7で示した通り，地域データセットの人口の状況に大きな偏りがあることに起因している。つまり，SVRによる予測モデルを構築する際に，人口が少ない地域が過多のため，訓練データセットに収録される地域は，人口の少ない地域に偏ってしまい，人口の少ない地域の予測性能を向上することを目指した学習となったためと考えられる。

5. むすび

市町村計画や住宅マスタープランなどを策定するとき，地方自治体では，公的統計の結果を地域の面積などを利用した按分値によって，地域人口を簡易的に求めることは行われる。しかしながら，人口のみを利用しており，本来，考察しなければならない社会環境の変化により生じる人の動きを予測することは困難と

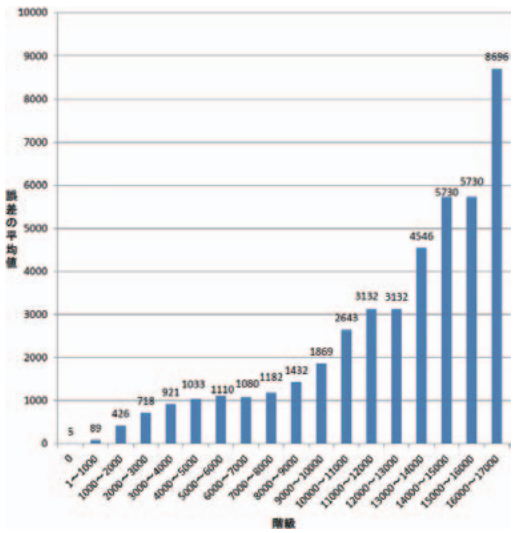


図16　3次メッシュ予測人口の誤差の分布

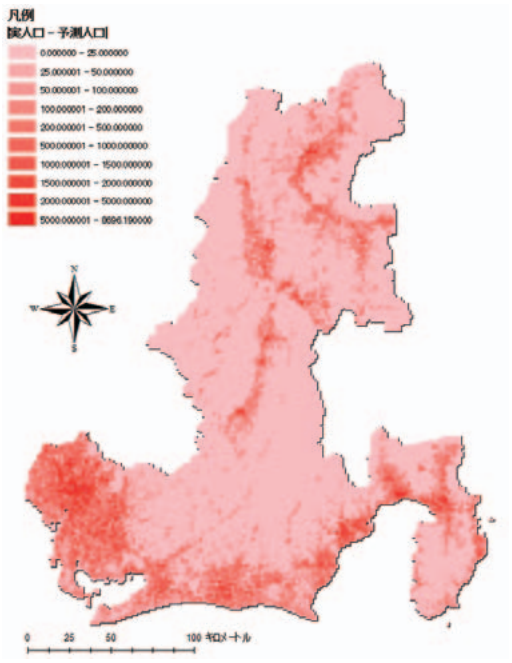


図17　予測人口における誤差の状況

なっている。そこで，本稿では，地方政策や社会活動の結果から人口を把握することを目指し，地域における社会環境の状況と人口の関係を機械学習手法により明らかにすることを試みた。提案手法により，人口を使わずにある程度人口の予測ができることから，例えば，「ここに駅ができたら。。。」「ここに工場ができたら。。。」などの社会環境を変化が生じたときの人口がどのように変化するかを実例から把握することも可能となるであろう。

今後の課題として，地域の人口予測の性能を高めなければならない。そのためには，地域の人口の著しい偏りへの対応，地域特徴量の見直し，機械学習手法の選定や調整などが挙げられる。一つ目の人口の少ない地域に傾斜して学習がされたことに対しては，地域人口の偏りの補正をすることで対応ができる。類似値の少ない人口の多い地域を外れ値として除外することや，分割すること，どのような人口範囲においても同数となるようにバランスよく地域を選択すること（例えば，データが過多である人口範囲内の地域はアンダーサンプリングをする，過小である場合は，地域の複製などにより増加させるオーバーサンプリング）を行うことも考えていきたい。二つ目の地域特徴量については，今回は公的データの地域人口と関連があると想定されるものを主観的に選択したが，国土データ基盤には様々なものがあるので取り入れていきたい。また，作成時期が異なっているものもあったため，時間の統一についても正確に考慮をしていかなければならない。三つ目として，予測モデルの機械学習手法にSVRを利用したため，予測性能を左右するパラメータを選定しなければならなかったが，それらの値による最適化には，さらに詳細な値を組み合わせて確認する必要がある。また，回帰分析手法も様々な手法があるため最適な手法を探索しなければならない。

最後に，今回は現在の3次メッシュにおける現在人口を目的変数として人口推計を行ったが，目的変数を将来人口とする，5歳階級別人口とすることなども可能であり，予測できることの意義を活かし，地域に効果のある人口予測へ発展していきたいと考えている。

参考文献

i　小西純，現状把握のための小地域統計データの利用と共有，法政大学日本統計研究所研究所報，Vol.40，pp.33-48，2010.

ii　澤田貴行，蒋湧，国勢調査を利用した任意地域の人口算出，愛知大学情報メディアセンター紀要COM，VoL40，pp.1-15，2015.

iii　堤盛人，村上大輔，嶋田章，“我が国の三大都市圏を対象とした住宅地価分布図の作成”，「GIS－理論と応用」，22（2），1-11. 2014

iv　地域メッシュ統計

http://www.stat.go.jp/data/mesh/gaiyou.htm

v 統計に用いる標準地域メッシュおよび標準地域メッシュ・コード（昭和48年行政管理庁告示第143号）

vi A.J.Smoda and Schoelkopf, A tutorial on support vector regression,NeuroCOLT2 Technical Report,NC2-TR-1998-030,1998.

vii V. Vapnik, The Nature of Statistical Learning Theory; Statistics for Engineering and Information Science, Springer, 1995.

viii “政府統計の総合窓口” http://e-stat.go.jp/SG2/eStatGIS/page/download.html（2015年4月6日参照）

ix “国土数値情報 ダウンロードサービス”, http://nlftp.mlit.go.jp/ksj/（2015年5月27日参照）

x 国土計画 http://www.mlit.go.jp/kokudoseisaku/kokudokeikaku_tk3_000008.html

xi “土地利用3次メッシュデータ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-L03-a.html（2015年5月27日参照）

xii “土地利用細分メッシュデータ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-L03-b.html（2015年5月27日参照）

xiii “標高・傾斜度3次メッシュデータ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-G04-a.html（2015年5月27日参照）

xiv “道路密度・道路延長メッシュデータ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N04.html（2015年5月27日参照）

xv “公共施設データ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P02-v4_0.html（2015年5月27日参照）

xvi “避難施設データ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P20.html（2015年5月27日参照）

xvii “バス停留所データ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-P11.html（2015年5月27日参照）

xviii “駅データ”, http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N02-v2_2.html（2015年5月27日参照）

xix “農業地域データ” http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-A12.html

xx “森林地域データ” http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-A13.html（2015年5月27日参照）

xxi “自然保全地域データ” http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-A11.html（2015年5月27日参照）

xxii “自然公園地域データ” http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-A10-v3_1.html（2015年5月27日参照）

xxiii “都市地域データ”
http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-A09.html（2015年5月27日参照）
xxiv Chang, C.C.；Lin, C.J.：“LIBSVM -- A Library for Support Vector Machines”, http://www.csie.ntu.edu.tw/~cjlin/libsvm/（2015年5月21日参照）.
xxv “libsvm -NET packaging of libsvm using IKVM”, https://code.google.com/p/libsvm-net/（2015年5月21日参照）.