メッシュ単位のデータ駆動型シミュレーションモデル作成手法に関する研究

森田仁美[1]・神谷啓太[2]・布施孝志[3]

[1]非会員 学士 (工学) 東京大学大学院 工学系研究科社会基盤学専攻（〒 113-8656 東京都文京区本郷七丁目 3-1）
E-mail: h-morita@trip.t.u-tokyo.ac.jp
[2]学生会員 修士 (工学) 東京大学大学院 工学系研究科社会基盤学専攻（〒 113-8656 東京都文京区本郷七丁目 3-1）
E-mail: kamiya@trip.t.u-tokyo.ac.jp
[3]正会員 博士 (工学) 東京大学教授 工学系研究科社会基盤学専攻（〒 113-8656 東京都文京区本郷七丁目 3-1）
E-mail: fuse@civil.t.u-tokyo.ac.jp

近年取得可能性が高まるメッシュデータを活用した予測・シミュレーションが期待されている．本研究では，メッシュ単位のシミュレーションモデルを作成する上で，メッシュ値の複雑なダイナミクスを定式化する点，有効な特徴量を利用して時系列予測を行う必要がある点に着目し，データ駆動型のシミュレーションモデル作成手法を構築した．具体的には，制約ボルツマンマシン（RBM）を用いてメッシュデータの特徴量を抽出し，多層パーセプトロン（MLP）を用いて一時刻先の特徴量を予測し，さらに RBM によって一時刻先のメッシュ値を予測する機構となっている．提案手法をミラノにおけるインターネット通信量のメッシュデータに適用し，作成したシミュレーションモデルの精度検証を行った．

Key Words : *simulation model, grid-based data, data-driven, multi-layer perceptron, restricted Boltzmann machine*

1. はじめに

時々刻々と変化する人の移動や活動を把握することは多くの分野で重要視されており，それらを再現・予測するためのシミュレーションモデルが利用されている．交通の分野を例にとると，交通需要予測や防災計画等に役立てることを目的とし，人の移動を推定する様々なモデルが考案されてきた．大鋳・小野木 (2008)[1] は，セル状に分割された空間を歩行者が移動するセルオートマトンモデルを用いて，混雑状況下における微視的な群集行動を再現した．よりマクロな人の移動を表現するものとして，藤井ら (1997)[2] は，アクティビティーベースのシミュレーションモデル PCATS によって一日の中の個人の生活行動の軌跡を詳細に再現した．また，情報通信の分野においては，ネットワーク上のトラフィック量を把握・制御するためのトラフィック予測手法が研究されており，気象情報やカレンダ情報などの外的要因を組み込んだインターネットトラフィック予測モデル [3] などがある．

他方，メッシュデータの蓄積が進んでいる．例えば，NTT ドコモは携帯電話基地局情報（Call Detail Records ; CDR）をもとに 1 時間ごとのメッシュ人口を推定しモバイル空間統計という名称で販売している [4] ほか，ミラノにおける SMS，電話，インターネット通信量の CDR が 10 分間隔のメッシュデータに集計されて公開・無償提供される例もある [5]．メッシュデータは匿名性が高く取得可能性も広がっていることから，今後さらなる活用が見込まれる．そこで，今後メッシュデータを活用した予測・シミュレーションが期待される．

メッシュ単位のシミュレーションモデル作成の上では，データの種類が異なればメッシュデータの生成過程が異なる点，同一のデータでも時間や場所によってメッシュ値のダイナミクスは大きく変動する点を考慮することが求められる．過去のメッシュ値が現在のメッシュ値に複雑に影響するためモデル構造の特定が容易ではなく，メッシュ値のダイナミクスの定式化が困難である．さらに，空間相関や冗長性，ノイズ等を考慮して有効な特徴量を抽出する必要がある．そこで，取得可能性が広がるメッシュデータから直接モデルを作成する，データ駆動型アプローチが有効であると考えられる．

以上の背景に基づき，本研究の目的は，メッシュ単位のデータ駆動型シミュレーションモデルを作成することである．ここで述べるシミュレーションモデルとは，ある時刻における対象地域内の全メッシュの観測値を入力とし，一時刻期先の全メッシュの観測値を予測するモデルである．本研究では，多層パーセプトロンおよび制約ボルツマンマシンを利用してシミュレーションモデルを作成する．ミラノにおけるインターネット通信量のメッシュデータを適用対象とし，モデルの

構造を可変とした上でシミュレーションモデルの予測精度に関する比較・分析を行う．さらに，時間帯別に精度を比較することで作成したシミュレーションモデルの特徴についても考察を行う．

本稿の構成は以下の通りである．まず第２章で，多層パーセプトロンおよび制約ボルツマンマシンについて概説した後，本研究の提案手法を述べる．第３章で実データへの適用結果とその考察を示し，最後に第４章で本研究の成果と今後の課題を述べる．

2. 提案手法

前述の通り，メッシュ単位のシミュレーションモデルを作成する上では，メッシュ値の複雑なダイナミクスを定式化する点，有効な特徴量を抽出して時系列予測を行う点を考慮する必要がある．これらの要件を満たすために，本研究では多層パーセプトロン（multi-layer perceptron ; MLP）と制約ボルツマンマシン（restricted Boltzmann machine ; RBM）を組み合わせた手法を構築する．この章では，MLP と RBM について概説した後，両者を組み合わせた提案手法を説明する．

(1) 多層パーセプトロン

MLP はニューラルネットワークを多層にした構造で，前の層のユニットの出力が次の層のユニットの出力となって情報が左から右に一方向に伝播する．図-1(a) のように中間層を 1 つ持つ構造の MLP を考え，各変数を右肩に層の番号（$l = 1, 2, 3$）を付けて表す．入力 $\mathbf{x}$ が与えられたとき，中間層（$l = 2$）と出力層（$l = 3$）のユニットの出力はそれぞれ

$$\mathbf{u^{(2)}} = \mathbf{W^{(2)}x} + \mathbf{b^{(2)}}, \quad \mathbf{z^{(2)}} = \mathbf{f(u^{(2)})} \tag{1}$$

$$\mathbf{u^{(3)}} = \mathbf{W^{(3)}z^{(2)}} + \mathbf{b^{(3)}}, \quad \mathbf{z^{(3)}} = \mathbf{f(u^{(3)})} \tag{2}$$

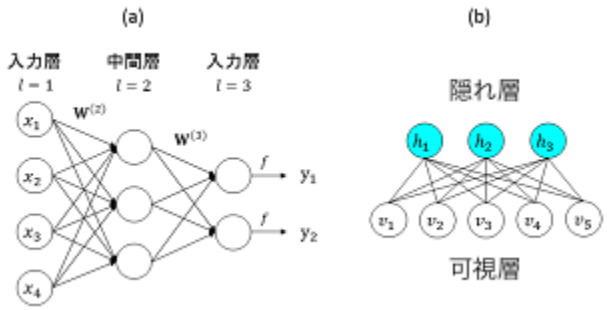
と計算され，最終的な出力は $\mathbf{y} = \mathbf{z^{(3)}}$ となる．上式において $\mathbf{W^{(l)}}$ は各層間の重み，$\mathbf{b^{(l)}}$ は各層のユニットのバイアスであり，これらのパラメータを全てまとめて $\mathbf{w}$ と表記する．また，f は活性化関数である．訓練データ $\mathbf{x}_n$ の目標出力を $\mathbf{d}_n$ とするとき，MLP の学習は，誤差関数

$$E(\mathbf{w}) = \frac{1}{2}\sum_n \|\mathbf{y}(\mathbf{x}_n; \mathbf{w}) - \mathbf{d}_n\|^2 \tag{3}$$

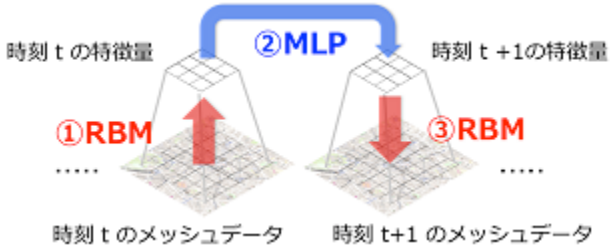
を最小化する $\mathbf{w}$ を求めることで行う．

(2) 制約ボルツマンマシン

RBM は，図-1(b) のように可視変数と隠れ変数を用いてネットワークを定義する．可視変数はデータの全成分に対応する確率変数である．また，隠れ変数はデータとは直接関係はないがネットワークの状態を支配する確率変数であり，この隠れ変数の集合をデータを低次元化した特徴量と捉えることができる．本研究では RBM をメッシュデータに適用するため，可視変数に連続値をとりうる Gaussian-Bernoulli RBM を利用する．



図–1 (a) 多層パーセプトロンと (b) 制約ボルツマンマシン



図–2 シミュレーションモデルの作成手法

可視変数，隠れ変数の値をそれぞれ v_i，h_j のように書き．エネルギー関数を次のように定義する．

$$\Phi(\mathbf{v}, \mathbf{h}, \theta) = -\sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_i \sum_j w_{ij} \frac{v_i}{\sigma_i} h_j \tag{4}$$

ここで，$\{a_i\}$，$\{b_j\}$，$\{w_{ij}\}$ はそれぞれ，可視ユニットのバイアス，隠れユニットのバイアス，およびユニット間の結合の重みである．これらのパラメータをまとめて θ と表す．また $\{\sigma_i\}$ は可視変数が従うガウス分布の標準偏差である．式 (4) のエネルギー関数の式を用いて，全ユニットの状態 $\{\mathbf{v}, \mathbf{h}\}$ は確率分布

$$p(\mathbf{v}, \mathbf{h}|\theta) = \frac{1}{Z(\theta)} \exp\{-\Phi(\mathbf{v}, \mathbf{h}, \theta)\} \tag{5}$$

によって与えられる．$Z(\theta)$ はモデル分布が確率分布の条件 $\sum_{\mathbf{z}} p(\mathbf{z}|\theta) = 1$ を満たすための規格化定数であり，分配関数と呼ばれる．

以上から，条件付き分布は次のように計算される．

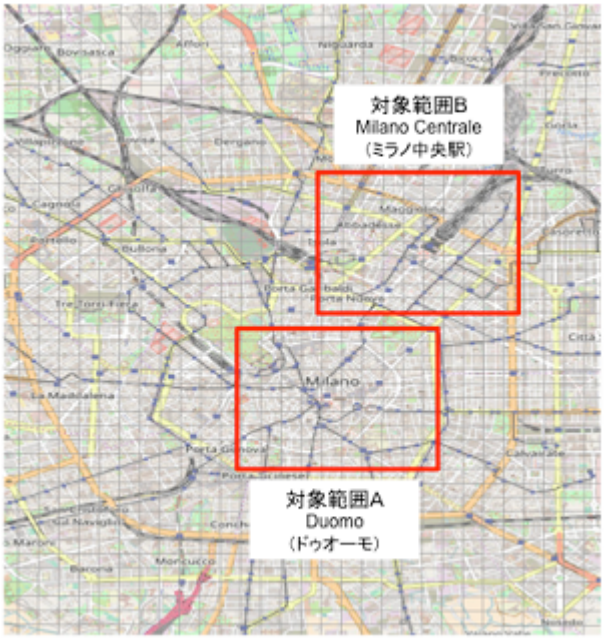
$$p(v_i|\mathbf{h}) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(v_i - a_i - \sum_j w_{ij} h_j)^2}{2\sigma_i^2}\right\} \tag{6}$$

$$p(h_j = 1|\mathbf{v}, \theta) = \sigma\left(b_j + \sum_i w_{ij} \frac{v_i}{\sigma_i}\right) \tag{7}$$

RBM の学習は，対数尤度関数

$$\log L(\theta) = \sum_n p(\mathbf{v_n}|\theta) \tag{8}$$

を最大化するパラメータ θ を求めることで行う．



図–3 対象範囲

(3) シミュレーションモデルの作成手法

MLP と RBM を組み合わせたシミュレーションモデル作成手法の概要を図-2 に示した．まず RBM を用いてメッシュデータの特徴量を抽出し，次に MLP を用いて一時刻先の特徴量を予測し，最後に RBM を用いて特徴量からメッシュ値に復元することで，ある時刻の全メッシュ値から一時刻の全メッシュ値の予測を行う．ここで，RBM の隠れ層のユニット数および MLP の中間層のユニット数は手動で設定する必要があるため，ユニット数を可変とした実験を行い，可能な限り簡潔なモデル構造かつ高い精度を担保するという観点から各ユニット数を決定することとする．

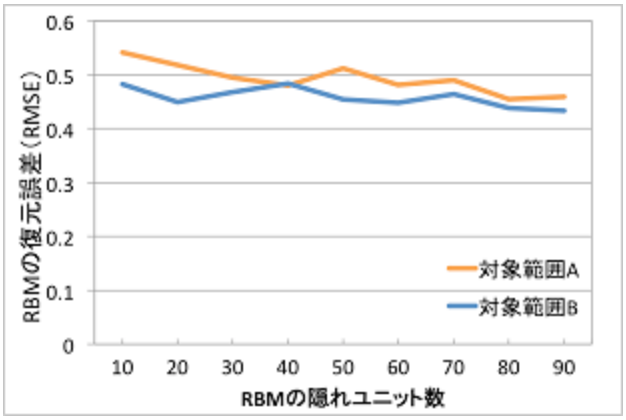
3. 提案手法の適用と精度検証

(1) 適用対象データ

本研究で適用対象としたメッシュデータは，携帯電話会社 Telecom Italia が計測したミラノにおける 2013 年 11 月から 12 月の 2ヶ月間のインターネット通信量データである．各メッシュの値を 10 分間隔の通信量の 1 時間集計値とし，休日と年末（12 月 23 日以降）を除いた平日 35 日分，全 5040 データを使用した．

各メッシュのサイズは 235m 四方であり，対象範囲として 10 × 10=100 メッシュの A：Duomo（ドゥオーモ）周辺，B：Milano Centrale（ミラノ中央駅）周辺の 2 か所を選定した（図-3）．

なお、学習のしやすさの観点から，あらかじめ各時刻のデータの平均を 0、分散を 1 にする正規化を行った．



図–4 隠れ変数を可変とした場合の RBM による復元誤差

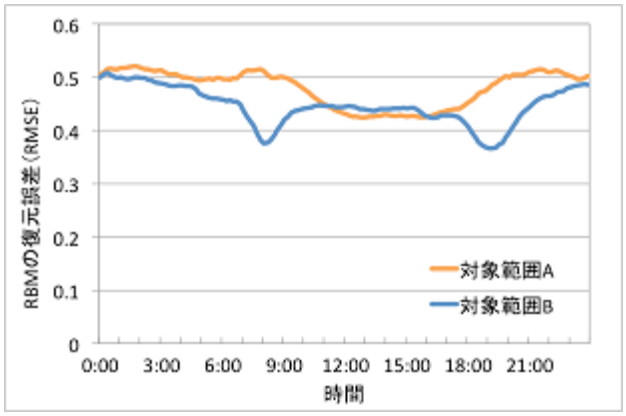
(2) 制約ボルツマンマシンのユニット数決定のための実験

ここでは，データから特徴量を抽出する機構である RBM の隠れユニット数を決定するための実験の内容とその結果を示す．

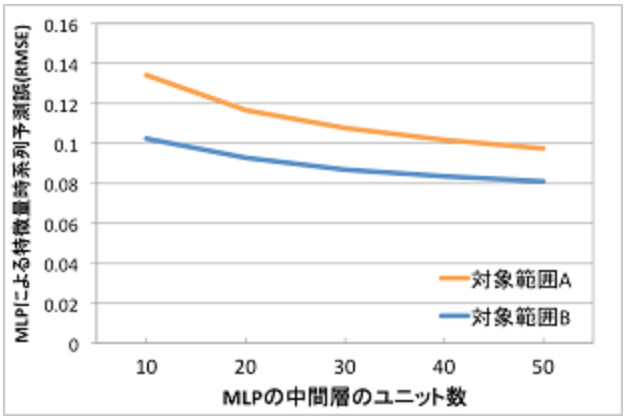
RBM の可視層のユニット数は 100 であるため，隠れ層のユニット数は 100 より小さい値で設定可能である．２つの対象エリアについて隠れ層のユニット数を，10 から 90 まで 10 ごとに変化させて実験を行った．メッシュデータを隠れ層，可視層の順にサンプリングし，その期待値を用いてメッシュデータを復元した結果と元のデータとの RMSE（Root Mean Squared Error，二乗平均平方根誤差）で RBM の性能を評価した．これは１メッシュあたりの復元誤差であり，値が小さいほど RBM の性能が高いことを意味する．RMSE の算出は交差確認法によって行った．以降の実験においても同様に交差確認法を用いる．

RBM の復元誤差を各隠れユニットについて求めて図-4 に示した．モデルが扱いやすい隠れユニット数が小さなものから順に検討していく．対象範囲 A の場合，隠れユニット数 40 まで低下した誤差が，ユニット数を 50 にすると増加する．その後 60 で低下し，70 で再び増加する．一方，対象範囲 B では，ユニット数が 20 から 40 にかけて誤差が増加し，その後 60 まで減少，70 にすると再び増加する．対象範囲 A，B ともに前後の隠れユニット数より誤差が小さい 60 に決定する．

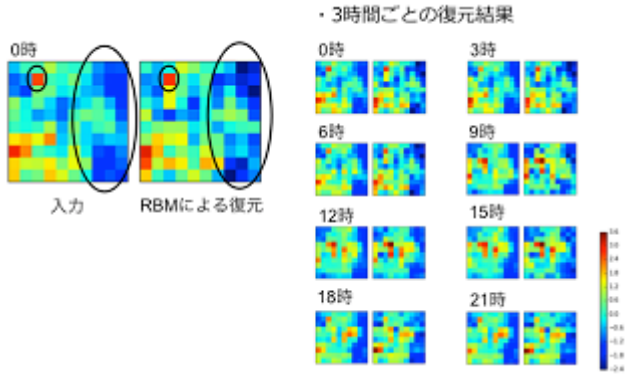
隠れユニット数が 60 のモデルにおいて，時間帯別の詳細な復元誤差を図-5 に示す．RMSE の平均で 0.48 の精度で復元できたことが確認できる．また，時間帯によって RBM の復元誤差が大きく変動することが読み取れる．深夜から早朝にかけての時間帯は対象範囲 A，B ともに誤差が大きくなっており，日中に比べて通信量が少ないこのような時間帯は RBM による特徴量の抽出が難しく，日による通信量の変動が予測誤差を増加



図–5 隠れ変数が 60 の場合の RBM による時間帯別復元誤差



図–7 中間層のユニット数を可変とした場合の MLP による特徴量の予測誤差



図–6 対象範囲 A における RBM による時間帯別の復元結果

させやすいという可能性が考えられる．鉄道駅を中心とする対象範囲 B を見ると，8 時頃と 19 時頃がとりわけ復元誤差が小さくなっている，これは通勤のため駅の利用数が多いと思われる時間帯であり，RBM によってその時間固有の有意な特徴量を抽出できたからと予想される．

対象範囲 A における，ある一日の観測データと，それを隠れユニット数が 60 の RBM によって復元した結果を 3 時間ごとに図-6 に示す．各時刻において左側が入力の観測データ，右側が RBM による復元結果である．メッシュ値が小さい青い範囲は青く復元され，メッシュ値が局所的に大きい赤いメッシュは赤く復元されており，大域的な特徴と局所的な特徴を捉えられていることが読み取れる．全体的に RBM を用いてメッシュデータから有効な特徴量を抽出できたといえる．

(3) 多層パーセプトロンのユニット数決定のための実験

特徴量の時系列変化を学習する機構である，MLP の中間層の構造を決定するための実験の概要とその結果を示す．

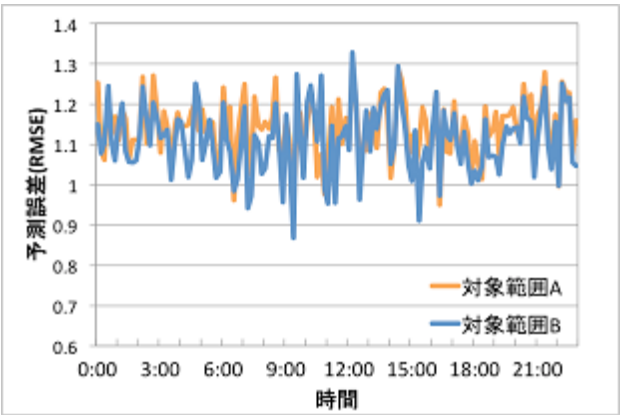
ここでは，隠れユニット数が 60 の RBM を用いて各時刻のメッシュデータから特徴量を抽出したデータを使用した．各時刻の特徴量データを入力，入力の一時間先の特徴量データを出力として MLP の学習を行った．MLP は中間層が 1 つの構造とし，中間層のユニット数を 10 から 50 まで 10 ごとに変化させて実験を行った．MLP の性能は，MLP による出力とその時刻の実際の特徴量データとの RMSE によって評価する．

中間層のユニット数を可変とした MLP の予測誤差を図-7 に示した．中間層のユニット数を増やすたびに誤差は小さくなり特徴量の時系列予測精度が向上したことがわかる．これは，ユニット数を増やすことでより複雑な特徴量変化を表現可能になることを意味している．モデルの扱いやすさという観点も踏まえて，MLP の中間層のユニット数を 40 とする．

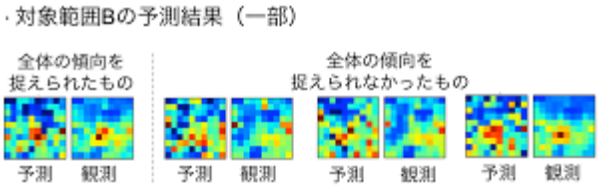
(4) シミュレーションモデルの予測精度の検証

RBM の隠れ層のユニット数を 60，MLP の中間層のユニット数を 40 に設定し，両者を組み合わせて一時間後の全メッシュ値の予測を行って精度を RMSE によって評価する．時間帯別の予測精度を図-8 に示した．ここから，どの時間においても RMSE の平均で 1.1 の予測精度であることが確認された．また，対象範囲 B において，ある一日においての予測結果と実際の観測データとの比較を図-9 に示す．大域的な傾向を捉えられた時間が存在した一方で、特徴を捉えられなかった時間が多い結果となった．RBM 自体の復元精度が RMSE で 0.48 であったのが，MLP によって特徴量の時系列変化を予測する機構を挟んだことで予測誤差が RMSE で 1.1 に増幅していることから，特徴量からメッシュ値への復元において MLP による予測誤差が増幅された可能性が考えられる．

作成したシミュレーションモデルの予測誤差の原因はいくつか考えられる．まず，本研究のモデルは時間帯を考慮せずに 1 時間後の全メッシュ値を予測するもの



図–8 作成したシミュレーションモデルの時間帯別の予測誤差



図–9 対象範囲 B における予測結果の一部

となっている．しかし現実においては，時間帯によってメッシュ値のダイナミクスは大きく変動し同一のモデルでは十分に表現できない可能性がある．さらに，本研究で学習に用いたのはメッシュ値の観測データのみであり，気象情報やユーザー属性情報といった通信量に変動を与える外部要因をモデルに組み込んでいないことも予測誤差の原因と考えられる．

4. おわりに

本研究では，メッシュデータを用いてデータ駆動型シミュレーションモデルを作成した．具体的には，RBM を用いてメッシュデータの特徴量を自動で抽出して低次元化した上で，MLP を用いて特徴量の時系列変化の予測を行った．そして，提案手法をミラノにおけるインターネット通信量のデータに適用した．予測精度に関する比較・分析を行い，可能な限り簡潔なモデル構造かつ高い精度を担保するという観点から RBM と MLP それぞれのユニット数を決定した．最後に，RBM と MLP を組み合わせて作成したシミュレーションモデルの予測精度を分析した．RBM については復元誤差が大きい時間帯が存在したものの全体の傾向を復元できる学習が行え，有意な特徴量を抽出できたことを確認した．一方で，RBM と MLP を組み合わせた全メッシュ値の予測においては，大きな誤差が生じることを確認した．

今後の課題としては，予測精度の向上が挙げられる．まず，時間帯を考慮したモデルを構築することが考えれる．また，本研究の提案手法においては RBM の可視層に複数の観測データを設定することが可能である．メッシュ値に変動を与える気象情報等の複数のデータを統合することで，より本質的な特徴量を抽出することができ，予測精度の向上が期待できる．さらに，人口分布データ等の他のメッシュデータに対する適用可能性も検討する必要がある．

参考文献

1) 大鋳史男, 小野木基裕: セルオートマトン法による避難流動のシミュレーション, 日本オペレーションズ・リサーチ学会和文論文誌, Vol.51, pp.94-111, 2008
2) 藤井聡, 大塚祐一郎, 北村隆一, 門間俊幸: 時間的空間的制約を考慮した生活行動軌跡を再現するための行動シミュレーションの構築土木計画学研究・論文集, No.20(2), pp.189-192, 1997.
3) 秋月俊寛, 市野将嗣, 甲藤二郎, 小松尚久: ロジスティック回帰分析法を用いたトラヒック予測手法に関する一検討, 電子情報通信学会, CQ 研究会, CQ2012-58, pp.7-12, 2012.
4) NTT ドコモ: モバイル空間統計に関する情報, https://www.nttdocomo.co.jp/corporate/disclosure/mobile_spatial_statistics/, (参照 2017425).
5) Dandelion: Open Big Data, https://dandelion.eu/datamine/open-big-data/, (参照 2017-4-25).
6) 岡谷貴之: 深層学習, 講談社, 2015.