

変分ベイズ法に基づく部分的データによる人流分布推定方式

浅原 彰規^{1,a)} 寺本 やえみ^{1,b)}

受付日 2013年4月8日, 採録日 2013年10月9日

概要: 本報告では, 測位データ密度分布をリアルタイムで生成することを目的とし, 変分ベイズ法を用いた高速な点密度分布推定方式を提案する. 従来のカーネル密度分布推定法による測位データの密度分布可視化では分布計算の処理時間が測位データの数に比例するため, 多数の測位データを扱うと処理速度が非常に遅いという問題があった. そこで過去の分布情報を利用することにより, 一部の測位データだけの取得で全体の分布を精度良く推定できる方式を提案する. 提案方式について実験で評価したところ, 従来にくらべ取得データ数が1/10, 処理速度が約100倍になることが明らかとなった.

キーワード: GIS, 位置情報, 変分ベイズ法

People Flow Distribution Estimation Based on Variational Bayes Method

AKINORI ASAHARA^{1,a)} YAEMI TERAMOTO^{1,b)}

Received: April 8, 2013, Accepted: October 9, 2013

Abstract: A method to estimate a density distribution of positioning data is proposed in this paper. It takes very long time to estimate a density distribution of massive point-dataset with conventional kernel density estimators because processing time is in proportion to the size of the dataset with the conventional estimators. Thus a method in which the distribution is estimated accurately with using the past distribution was proposed. As the result of the experimental evaluation, it was confirmed that the amount of collected data and time for processing are reduced to 1/10 and 1/100 respectively.

Keywords: GIS, positioning data, variational Bayes method

1. 本研究の位置づけと目的

1.1 位置情報を収集する既存システムとその課題

近年, スマートフォンに代表される高度な機能を持った移動体端末の普及により, 人々の測位データを収集できる環境が整ってきた. ここでいう測位データとは, 測位装置によって得られる特定の人や物の位置と, 位置が計測された時刻を対応付けたデータをさす. 測位装置とはGPS (Global Positioning System) 測位をはじめとする, 対象の現在位置を計測し, 座標値として得る機能を持った装置を意味しており, 近年は無線LANを用いた測位 [1] や屋内設置型のGPS発信器 [2] など屋内でも利用できる測位方式も

知られている [3]. 測位データは, 人や物の動きの情報を得るのに有用であり, 道路の旅行時間推定による交通渋滞の検出や, 交通需要の把握による都市計画策定支援などに役立つ.

測位データを用いた応用の1つとして人の持つ移動体端末から測位データを収集し有益な情報を得るプローブパーソン [4] などの試みがある. 現在交通状況把握のために行われているアンケート調査では, 多大な費用をかけて, 多くの人々から移動の目的地, 出発地, 交通手段などの情報を収集している. その代わりに, 一般市民の協力者が所有するスマートフォンに情報を集めるソフトウェアを配信することによって, 低コストかつ高精度に, 交通需要や観光行動 [5], [6], [7], [8] などの人流把握ができる手段となるのである. スマートフォンの通信負荷や電力消費の大きさや, 人の位置というプライベートな情報を提供させるインセン

¹ 株式会社日立製作所
Hitachi Ltd., Kokubunji, Tokyo 185-8601, Japan
^{a)} akinori.asahara.bq@hitachi.com
^{b)} yaemi.teramoto.xy@hitachi.com

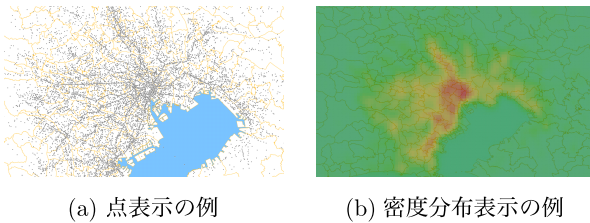


図 1 測位データの表示方法

Fig. 1 Positioning data visualizations.

タイプの提供手段、セキュリティなどの課題はあるが、それらを解決できれば、有用な情報が集まる。

プローブパーソンなどで収集された測位データを可視化し概要を把握できるようにする方法について、図 1 に例を示す。測位データは (a) のように点によって地図上に表示することができる。この表示では移動体の個別の動きが把握できる一方、高密度な地域では点が重なり合って把握困難となってしまう。プローブパーソンにおいては個人の動きは重要ではなく、その統計量、たとえば人口密度の把握が重要であるが、点表示は統計量の把握には適さないのである。そこで、(b) に示すような点の密度に基づく分布を表示することが考えられる。この表示は人の密度に応じて色の異なる矩形を描画したもので、緑→黄→赤の順に密度が多いことを示しており、一目で人の多い地域を把握できるようになっている。このような表示は一般的にヒートマップ表示などと呼ばれている。本表示の目的は人流の概況を可視化し、把握可能にすることである。これにより、たとえば過密地域を把握して保安・防災上の計画立案に役立てたり、人の流動を把握して交通需要の確認やマーケティングに使用したりすることが考えられる。また、点表示では個人の位置情報が露わになってしまうのに対し、分布表示では個人の位置は伏せたまま傾向だけを提示できるという点で、プライバシーの問題も少なくなるという利点もある。

移動体端末からリアルタイムに測位データを収集し密度分布表示すれば、リアルタイムな概況把握が可能になる。しかし、リアルタイムに収集した全データを取り扱うには処理時間がかかるという問題がある。たとえば、数百万人規模の人口を持つ都市と周辺地域の 1 千万人近くの人々について、その 5% の位置情報を 1 分に 1 度収集すると毎分五十万件の位置情報を処理しなければならない。これには 1 件あたり 120 μ s で処理を終えなければならないが、この性能は一般的なコンピュータでは達成困難である。

そこで、本研究ではすべての測位データを使用せず、密度分布を推定することを目指す。今回、移動体の群の分布をガウス分布でモデル化し、当該モデルを少数の測位データで補正することにより、少ない測位データから全体を推定する方式を提案する。今回、着目したのは移動体の位置の分布は急激には変化しないという点である。すなわち、時系列測位データの密度を逐次的に推定していく場合、

前時刻の推定結果は次の時刻の推定結果と類似すると想定できる。そこで、現在の時刻の分布は、直前の推定結果である密度分布をわずかに変更させたものであると考え、直前の密度分布推定結果を補正して現在の時刻の密度分布を推定するアプローチをとる。これにより、現在のすべての測位データを使用しなくても精度良く分布が推定できるのである。最後に、アンケートから推定された人の流れデータを用いて提案方式を評価し、その有効性を示す。

2. 従来方式の問題と提案方式のアプローチ

2.1 従来の密度分布推定法

空間的に分布する大規模データに対する統計解析手法は様々な目的のものがさかんに研究されている。たとえば、空間上に分布する点で気温などの値が得られたとき、その値の空間分布を推定する方法が研究されている [9]。一方、人の動きに着目した研究としては、人の移動パターンを見つける分析 [10], [11] や、移動パターンの確率モデルを用いた行き先予測 [12], [13], [14], [15], 移動パターンからの人物の属性推定 [16], [17] などが知られている。

本研究は点の密度分布の分析手法 [18] の研究である。密度分布の推定手法としては、大きく分けてパラメトリックな推定法とノンパラメトリックな推定法の 2 種類が知られている。パラメトリックな推定法とは、あらかじめ定められた確率密度分布関数などのパラメータを決定することによる確率密度分布推定方式である。一方、ノンパラメトリックな推定法では、測定されたデータ集合の数から確率密度分布を推定するなどデータに基づいて分布を推定する。

ノンパラメトリックな推定法のうち最も単純な方法の 1 つは、空間領域を矩形に分割し各矩形内の点の数を密度とするヒストグラムによる推定法である。ただし、矩形内に十分多数のサンプルデータがなければ推定結果の精度が低くなる、領域の境界で密度が不連続になるなどの問題があり、一般的にはカーネル密度推定法 [19], [20], [21] がよく用いられる。カーネル密度分布推定ではサンプルデータの周囲に分布を発生させ、それを重ね合わせた結果を密度分布とする。 N 個の各点 $\{\mathbf{x}_n\}$ を中心とするガウス分布を仮定したガウスカーネルを用いた確率分布 $P(\mathbf{x})$ は

$$P(\mathbf{x}) = \frac{1}{N} \sum_n \frac{1}{2\pi h^2} \exp\left(-\frac{|\mathbf{x} - \mathbf{x}_n|^2}{2h^2}\right), \quad (1)$$

となる。ここで、 h は分布の半径に相当する量である。

2 次元空間上の分布をより高速に生成できる方法として以下の Quadratic カーネルも用いられることがある。

$$P(\mathbf{x}) = \frac{1}{Nh^2} \sum_n Q\left(\frac{|\mathbf{x} - \mathbf{x}_n|}{h}\right) \quad (2)$$

$$Q(r) \equiv \begin{cases} \frac{6}{\pi} (1-r)^2 & r < 1 \\ 0 & r \geq 1 \end{cases} \quad (3)$$

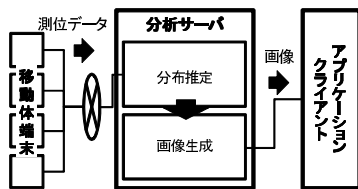


図 2 分布画像生成の手順

Fig. 2 Distribution production processes.

カーネル密度分布推定法には、密度分布の値を得るためにサンプル数に比例する処理時間がかかるという問題がある。仮に 640×480 画素の画像を 10 万件のデータから生成する場合、 $640 \times 480 \times 10$ 万回つまり 30 億 7,200 万回もの計算が必要となり、処理速度が低下する要因となる。

パラメトリックな推定法としては、パラメータを持つ確率分布を用い、尤度が最大になるパラメータを決定する最尤法が知られている。最尤法でよく用いられるのは、観測不可能な内部確率変数を仮定する確率モデルを適用して精度を高める方法である。その場合、パラメータと同時に内部確率変数も推定が必要なため、内部確率変数の期待値推定とパラメータの最適化を交互に実行し収束させる EM アルゴリズム [22] が用いられる。内部確率変数を持つ確率モデルとして最も代表的な分布は混合ガウス分布である。混合ガウス分布は複数のガウス分布を組み合わせた分布であり、各確率変数が生成元のガウス分布を特定する内部変数を持つ。パラメトリックな推定法では、使用した確率分布が推定結果の制約となるため、適切かつ表現力のあるモデルを選択しなければ精度が低下してしまう。表現力の高いモデルは一般に複雑であるが、EM アルゴリズムを適用すると反復演算に処理時間がかかるという問題がある。

パラメトリック推定とノンパラメトリック推定の特徴を示すため、分布画像の生成手順の模式図を図 2 に示す。最初に移動体端末が自身の位置を測位データの形にして、分布推定を行う分析サーバへ送付する。分析サーバでは、密度分布を推定した後、密度分布の画像を生成するという 2 つの手順を実行する。アプリケーションクライアントは生成された画像を受け取り表示などをする。アプリケーションクライアントで画面解像度や地図の縮尺変更が起これば、画像生成だけを再度実行すればよい。

ノンパラメトリック推定においては、分布生成は全データを読み出すだけで実質的には何も処理を行わないが、画像生成はすべての測位データを用いて画素値を決定する処理になる。この処理は測位データの量が多いほど多くの時間を要する。パラメトリック推定では、分布推定処理に処理時間を要するかわりに、画像生成に要する時間は非常に小さくなる。図 2 のシステム構成を想定するとアプリケーションクライアント側で地図画面のズームなどを行うたびに画像生成だけを何度も呼び出すことになるが、その場合はパラメトリック推定のほうが高速に処理が実行できる。

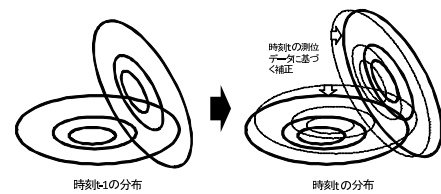


図 3 分布補正の概念図

Fig. 3 Distribution revise.

また、分析サーバの負荷を分散するために画像生成をアプリケーションクライアント側で実行することもできるが、ノンパラメトリック推定ではすべての測位データをアプリケーションクライアントへ送付する必要がある、膨大な通信量を要する。これは密度分布画像を生成する目的での比較であるが、別の情報、たとえば竜巻などの経路情報と重ねて被害者数の期待値を計算したい場合のように人口分布を分布としてそのまま用いる際にも、画像生成処理に相当する演算を繰り返し行うことになる。このように、パラメトリック推定はノンパラメトリック推定にはない有用な特徴を多数備えている。しかし、パラメトリック推定では分布推定に反復演算が含まれており、全体としての計算時間はノンパラメトリック推定よりもかかる場合が多い。ただし、入力となる測位データが多くなるとノンパラメトリック推定でも処理時間は増大する。

2.2 本研究のアプローチ

本研究では、パラメトリック推定を用い、人の密度分布に関する特徴を利用して分布推定に要する計算量を減らすことを目指す。これにより、パラメトリック推定の有用な特徴を維持しつつも、高速な分布推定ができるようになる。

本研究で着目したのは、移動速度の制約である。人はあまり高速には移動しないので、分布更新頻度が十分短ければ、更新前と極端に異なった分布にはならないはずである。そこで、前時刻の分布情報を活用して、前時刻の分布を新たなデータから補正するアプローチをとる。図 3 にその模式図を示す。まず、パラメトリックな推定法で時刻 $t-1$ の分布を推定する。次に、パラメータを少し補正することによって次時刻の測位データをよく説明する分布を求める。これを繰り返していけば、実際に取得する測位データは少量であっても、十分な精度が得られると考えられる。

本研究では、ベイズ推定を用いて上記のアプローチを実装する。前時刻の分布情報を事前分布と見なし、新たなデータを用いた事後確率分布推定によって前時刻の分布を補正することで、一部の情報だけを用いて全体の分布をリアルタイムに得ることを目指す。

3. 部分データによる分布生成方式

3.1 変分ベイズ法

変分ベイズ法 [23] は確率分布のパラメータの確率分布を

事前分布として導入し、パラメータの確率分布に独立性に関する近似を取り入れることにより、真の分布に十分近い分布を求める方法である。本研究では測位データの発生分布を一般的によく用いられる混合ガウス分布を用いて近似する。混合ガウス分布の分布関数を以下に示す。

$$P(\mathbf{x}|\{\pi_k\}, \{\mu_k\}, \{\Lambda_k\}) = \sum_k \pi_k \mathcal{N}(\mathbf{x}|\Lambda_k, \mu_k) \quad (4)$$

ここで、 $\mathbf{x}$ が観測データ（つまり、位置座標）であり、 $\mathcal{N}(\mathbf{x}|\Lambda, \mu)$ は分散共分散行列 Λ 、期待値 μ の 2 次元ガウス分布、 π_k は k 番目の要素の混合比である。この分布のパラメータ $\{\pi_k\}$ 、 $\{\mu_k\}$ 、 $\{\Lambda_k\}$ について以下の分布を事前分布として与える。

$$\begin{aligned} P(\{\pi_k\}, \{\mu_k\}, \{\Lambda_k\} | W_0, \nu_0, \beta_0) \\ = \text{Dir}(\{\pi_k\} | \alpha_0) \\ \times \prod_k \mathcal{N}(\mu_k | (\beta_0 \Lambda_k)^{-1}, \mathbf{m}_{0,k}) \mathcal{W}(\Lambda_k | W_0, \nu_0) \end{aligned} \quad (5)$$

ここで、 α_0 、 W_0 、 $\mathbf{m}_0$ 、 ν_0 、 β_0 は事前分布のパラメータであり、ハイパーパラメータと呼ばれるアприオリに与えられるパラメータである。また、 $\mathcal{W}(W, \nu)$ は分散 W 、自由度 ν のウィシャート分布、 $\text{Dir}(\{\pi\} | \alpha_0)$ は α_0 デリクレ分布である。ガウス–ウィシャート分布はガウス分布の、デリクレ分布は多項分布の共役事前分布である。共役事前分布とは観測値に基づきベイズ推定を行った結果が、再び事前分布と同じ形になる分布であり、結果的に推定結果はハイパーパラメータの更新として表現される。なお、ガウス分布の中心（期待値）の事前分布であるガウス分布は、 $\mathbf{m}_0$ が期待値、 $\beta_0 \Lambda_k$ （元のガウス分布の分散の β_0 倍）が分散である。なお、これらの表記は文献 [24] に記載のものと同一のものをを用いた。ハイパーパラメータ α_0 、 β_0 、 $\mathbf{m}_0$ 、 W_0 に対し、 $\{\mathbf{x}_n\}$ を推定に用いることで、

$$\alpha_k = \alpha_0 + N_k \quad (6)$$

$$\beta_k = \beta_0 + N_k \quad (7)$$

$$\mathbf{m}_k = \frac{1}{\beta_k} (\beta_0 \mathbf{m}_0 + N_k \bar{\mathbf{x}}_k) \quad (8)$$

$$W_k^{-1} = W_0^{-1} + S_k N_k + \frac{N_k \beta_0}{N_k + \beta_0} (\mathbf{m}_0 - \bar{\mathbf{x}}_k)(\mathbf{m}_0 - \bar{\mathbf{x}}_k)^T \quad (9)$$

$$\nu_k = \nu_0 + N_k \quad (10)$$

のように更新される。ここで N_k 、 $\mathbf{x}_k$ 、 S_k はそれぞれ

$$N_k = \sum_n r_{n,k}, \quad (11)$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_n r_{n,k} \mathbf{x}_n, \quad (12)$$

$$S_k = \sum_n r_{n,k} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T \quad (13)$$

である。この式中の $r_{n,k}$ は負担率と呼ばれる量であり、

$$r_{n,k} = \frac{\rho_{n,k}}{\sum_n \rho_{n,k}}, \quad (14)$$

$$\begin{aligned} \rho_{n,k} \\ \equiv \tilde{\pi}_k |\tilde{\Lambda}_k|^{\frac{1}{2}} \exp \left[-\frac{1}{\beta_k} - \frac{1}{2} (\mathbf{x}_n - \mathbf{m}_k)^T W_k (\mathbf{x}_n - \mathbf{m}_k) \right] \end{aligned} \quad (15)$$

$$\ln \tilde{\pi}_k \equiv \ln \psi(\alpha_k) - \ln \psi \left(\sum_k \alpha_k \right) \quad (16)$$

$$\ln |\tilde{\Lambda}_k| \equiv \ln \psi \left(\frac{\nu_k}{2} \right) + \ln \psi \left(\frac{\nu_k - 1}{2} \right) + 2 \ln 2 + \ln |W_k|. \quad (17)$$

と計算される。ここで $\psi(\cdot)$ はダイガンマ関数である。負担率は各サンプルからの各混合ガウス分布に対する所属の度合いに相当し、あるサンプルが特定のガウス分布に完全に所属するとしたとき負担率は 1 となる。よって、負担率の和 N_k は、 k 番目のガウス分布に所属するサンプルの数を一般化したものに相当する。この負担率の計算にハイパーパラメータが含まれており、ハイパーパラメータの更新式に負担率が必要であるが、この計算は繰返し処理によって収束させることができる。したがって、ハイパーパラメータに適切な初期値を与えて負担率を計算し、次に負担率に基づきパラメータを更新、再び新たなハイパーパラメータから負担率を計算し、と繰り返すことにより、最適なハイパーパラメータが得られる。このハイパーパラメータを用いた予測分布は以下ようになる。

$$P(x|W_k) = \frac{1}{\sum_k \alpha_k} \sum_k \alpha_k \text{st}(\mathbf{x}|\mathbf{m}_k, L_k, \nu_k - 1), \quad (18)$$

$$L_k = \frac{\beta_k (\nu_k - 1)}{1 + \beta_k} W_k \quad (19)$$

ここで、 $\text{st}(\cdot)$ は t 分布である。以上の計算が一般的な変分ベイズ法による混合ガウス分布推定の手続きである。

3.2 カルバックライブラ情報量による計算打ち切り

本研究では、上記手続きにおいてアприオリに与えられたハイパーパラメータ α_0 、 β_0 、 $\mathbf{m}_0$ 、 W_0 を前時刻の分布推定結果に基づき合成するアプローチをとる。結果的に、現時刻の測位データを一部だけ用いたとしても、全測位データを用いて推定したのに近い分布が得られるのである。

ここで問題となるのは、現時刻のデータのうち、使用するデータ数をどのようにして決定するかである。今回、人流分布の変化が激しくないという前提のもと、人流分布の変化が激しいときほど多くの測位データを用い、変化が緩やかな場合には少数の測位データを用いるようにする。図 4 にこの処理の概念図を示す。まず、直前の時刻の分布をもとに事前分布を構成する（図中、ステップ 1）。同時に、入力サンプルをいくつかのブロックに分割する。次に、

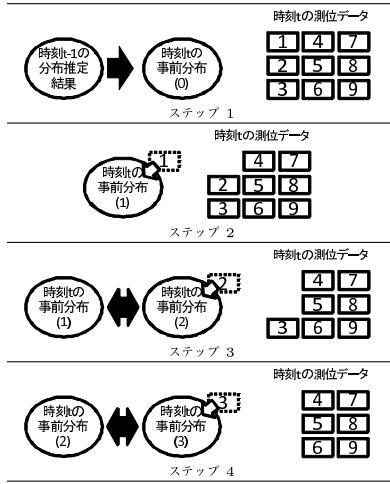


図 4 逐次投入の概念図

Fig. 4 Serial inputation.

ステップ2のようにそれらのブロックのうち1つを事前分布に投入し事後推定する．さらに，ステップ3，4に示すように直前のステップの結果に対して新たな測位データを投入し事後推定する．この際，直前のステップの結果と比較して，分布の変化があまり大きくなかった場合には，測位データが十分な量投入されたと判断して，計算を打ち切る．これにより，測位データ数を削減できる．この際，確率分布の変化を評価する量として，以下の $KL(p(x)||q(x))$ によって定義されるカルバックライブラ情報量（以降，KL情報量）を用いる．

$$KL(p(x)||q(x)) = \int p(x) \ln \frac{q(x)}{p(x)} dx \quad (20)$$

KL 情報量は複数の確率分布間の距離に相当する量である．今回のケースに KL 情報量を適用すると

$$\begin{aligned} K &\equiv KL(p(\{\theta_i\}|x_{m+1})||p(\{\theta_i\}|x_m)) \\ &= \int p(\{\theta_i\}|\{\mathbf{x}_n\}_m) \ln \frac{p(\{\theta_i\}|\{\mathbf{x}_n\}_{m+1})}{p(\{\theta_i\}|\{\mathbf{x}_n\}_m)} \prod_i d\theta_i \end{aligned} \quad (21)$$

ここで $\{\theta_i\} = \{\{\pi_k\}, \{\Lambda_k\}, \{\mu_k\}\}$ とし， $\{\mathbf{x}_n\}_m$ は m 番目に投入される測位データのブロックを意味する．この K が十分小さいとき，確率分布が収束したと見なして計算を打ち切ればよい．式中の $p(\{\theta_i\}|\{\mathbf{x}_n\}_m)$ は変分ベイズ法の推定結果 $q(\{\theta_i\}|\{\mathbf{x}_n\}_m) = q(\{\pi_k, \Lambda_k, \mu_k\}|\{\alpha_k, \beta_k, \mathbf{m}_k, \nu_k, W_k\})$ に置き換えられるので，

$$\begin{aligned} K &= \int q(\{\theta_i\}|\{\alpha_k, \beta_k, \mathbf{m}_k, \nu_k, W_k\}) \\ &\quad \times \ln \frac{q(\{\theta_i\}|\{\alpha'_k, \beta'_k, \mathbf{m}'_k, \nu_k, W'_k\})}{q(\{\theta_i\}|\{\alpha_k, \beta_k, \mathbf{m}_k, \nu_k, W_k\})} \prod_i d\theta_i \end{aligned} \quad (22)$$

と書ける．ここで $\{\alpha_k, \beta_k, \mathbf{m}_k, \nu_k, W_k\}$ は $\{\mathbf{x}_n\}_m$ を用いた場合の推定結果， $\{\alpha'_k, \beta'_k, \mathbf{m}'_k, \nu_k, W'_k\}$ は $\{\mathbf{x}'_n\}_{m+1}$ を用いた場合の推定結果とした．この K は式 (17) の定義を用い

て簡単に書き表せ，

$$\begin{aligned} &\int q(\{\theta_i\}|\{\alpha_k, \beta_k, \mathbf{m}_k, \nu_k, W_k\}) \\ &\quad \times \ln q(\{\theta_i\}|\{\alpha'_k, \beta'_k, \mathbf{m}'_k, \nu_k, W'_k\}) \prod_i d\theta_i \\ &= \ln \alpha_0 + \sum (\alpha_0 - 1) \ln \tilde{\pi}_k - \ln \beta'_k \\ &\quad - \frac{\beta'_k}{2} \left(\frac{2}{\beta_k} + \nu_k (\mathbf{m}_k - \mathbf{m}'_k)^T W_k (\mathbf{m}_k - \mathbf{m}'_k) \right) \\ &\quad + \ln B(W'_k, \nu'_k) + \frac{1}{2} (\nu'_k - 1) \ln |\tilde{\Lambda}_k| - \sum_{i,j} \frac{W'^{-1}_{i,j}}{W^{-1}_{i,j}} \end{aligned} \quad (23)$$

となる．ただし， $B(\cdot)$ は以下のとおりである．

$$B(W, \nu) \equiv |W|^{-\frac{1}{2}} \left(2^\nu \pi^{\frac{1}{2}} \Gamma\left(\frac{\nu}{2}\right) \Gamma\left(\frac{\nu-1}{2}\right) \right)^{-1} \quad (24)$$

3.3 事前分布の生成

逐次的に測位データを投入する方法として最も単純に考えられるのは，前ステップで算定された推定結果の事後分布をそのまま次の事前分布とする，逐次ベイズ推定的な方法である．たとえば β に関する更新式 (7) について，時刻 t の寄与率の和 $N_{t,k}$ を用いた推定結果 $\beta_{t,k} = \beta_0 + N_{t,k}$ を，そのまま寄与率と $N_{t+1,k}$ の時刻 $t+1$ における推定処理の事前分布とした場合， $\beta_{t+1,k} = \beta_0 + N_{t,k} + N_{t+1,k}$ となり，これを繰り返していくと $\beta_{t+n,k} = \beta_0 + N_{t,k} + N_{t+1,k} + \dots + N_{t+n,k}$ と際限なく大きな値になる．すると，新しいデータの寄与 $N_{t+n,k}$ が相対的に小さくなり推定結果にほとんど影響しなくなってしまう．本来，時刻 t の分布という事象と時刻 $t+1$ の分布という事象は別の事象なので，1つの事象の観測値を追加するような事後推定は適さないのである．そこで変分ベイズ法のパラメータ更新式に登場する負担率の和 $N_{t,k}$ に着目する．負担率とは各測位データの混合分布の各成分 (k と表記されている) に対する寄与の大きさを表しており，分布に寄与するサンプル数（測位データ数）を連続化したものに相当する．つまり，前述の β の更新式において，新しい測位データの寄与が相対的に小さくなるのは，過去のあらゆる時刻のすべての測位データ数の累積値がハイパーパラメータに含まれているためと解釈される．

そこで本研究では，新しい測位データを追加するとき同時に過去の測位データの寄与を減衰させ，ハイパーパラメータに包含されている測位データ数を一定に保つようにする．まず，時刻 t の測位データセットを十分大きな M 個のブロックに分解する．ここで各ブロックに含まれる測位データ数を $n_t^{(1)}, n_t^{(2)}, \dots, n_t^{(i)}, \dots, n_t^{(M)}$ とする．これらのブロックを逐次投入し，事後確率推定を行う．このとき， j 番目までのブロックを投入した時点での投入済み測位データの全測位データに対する比率は $\frac{\sum_{i=1}^j n_t^{(i)}}{\sum_{i=1}^M n_t^{(i)}}$ となる．そこで，以下のように減衰率 $\gamma^{(j)}(t)$ を定義する．

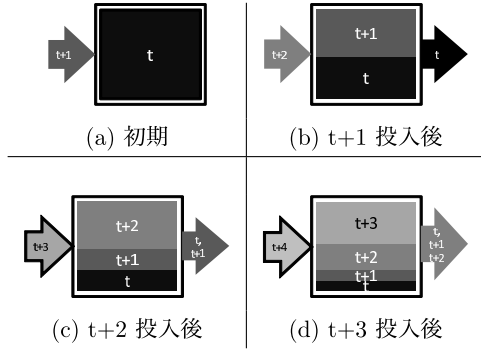


図 5 逐次投入における負担率調整の模式図

Fig. 5 Responsibility control with serial imputation.

$$\gamma_t^{(j)} = 1 - \frac{\sum_i^j n_t^{(i)}}{\sum_i^M n_t^{(i)}} \quad (25)$$

この減衰率を用いて、過去の測位データの数を決減させてから新たに現時刻の測位データを投入する。これによって、推定されるハイパーパラメータに含まれる過去のデータの寄与はおおむね一定に保たれると考えられる。

この手順の効果を図 5 を用いて説明する。図中 (a) の四角形の箱には時刻 t の寄与が詰め込まれている。これは、 t における測位データの負担率のみから事前分布が作られている状態を意味している。測位データの寄与は (a), (b), (c) で箱に左から入る矢印で示されているように随時投入され、それにともない、箱内の寄与は投入されたのと同程度程度の寄与が押し出される。図中、箱から右へ出る矢印は前時刻の寄与投入により押し出された分の寄与を示しており、その量は投入データの全データに対する比率 $\gamma_t^{(j)}$ に従い、箱内に既存の寄与を同比率で減衰させたのと同程度である。すると、(d) にあるように古い時刻の寄与は指数関数的に減衰していくことになる。つまり、現在時刻の箱の中にある寄与のみを保存し、新たなデータ投入時にその投入比率に応じて減衰させるだけで、古いデータの寄与を急激に減らしていくことができるのである。

この減衰の手続きを適用すると、時刻 t の推定結果を利用して生成される時刻 $t+1$ の事前分布は以下ようになる。

$$\alpha_{0,t+1}^{(0)} = \alpha_0 + \gamma_t^{(0)} N_{t,k} \quad (26)$$

$$\beta_{0,t+1}^{(0)} = \beta_0 + \gamma_t^{(0)} N_{t,k} \quad (27)$$

$$\mathbf{m}_{0,t+1}^{(0)} = \frac{1}{\beta_{0,t+1}^{(0)}} \left(\beta_0 \mathbf{m}_0 + \gamma_t^{(0)} N_{t,k} \bar{\mathbf{x}}_{t,k} \right) \quad (28)$$

$$W_{0,t+1}^{-1(0)} = W_0^{-1} + \gamma_t^{(0)} S_{t,k} N_{t,k} + \frac{\gamma_t^{(0)} N_{t,k} \beta_0}{\gamma_t^{(0)} N_{t,k} + \beta_0} (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})(\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})^T \quad (29)$$

$$\nu_{0,t+1}^{(0)} = \nu_0 + \gamma_t^{(0)} N_{t,k} \quad (30)$$

ここで時刻 t において全データを用いた変分ベイズ法の推定結果があるとしており、 $N_{t,k}$, $\bar{\mathbf{x}}_{t,k}$, $S_{t,k}$ はそれぞれ時

刻 t における式 (13) の値を示している。この事前分布のもとでの推定結果は以下ようになる。

$$\alpha_{t+1,k}^{(0)} = \alpha_0 + \gamma_t^{(0)} N_{t,k} + n_{t,k}^{(0)} \quad (31)$$

$$\beta_{t+1,k}^{(0)} = \beta_0 + \gamma_t^{(0)} N_{t,k} + n_{t,k}^{(0)} \quad (32)$$

$$\mathbf{m}_{t+1,k}^{(0)} = \frac{1}{\beta_{t+1,k}^{(0)}} \left(\beta_0 \mathbf{m}_0 + \gamma_t^{(0)} N_{t,k} \bar{\mathbf{x}}_{t,k} + n_{t,k}^{(0)} \bar{\mathbf{x}}_{t,k}^{(0)} \right) \quad (33)$$

$$W_{t+1,k}^{-1(0)} = W_0^{-1} + \gamma_t^{(0)} S_{t,k} N_{t,k} + s_{t,k}^{(0)} n_{t,k}^{(0)} + \frac{\gamma_t^{(0)} N_{t,k} \beta_0}{\gamma_t^{(0)} N_{t,k} + \beta_0} (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})(\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})^T + \frac{n_{t,k}^{(0)} \beta_0^{(0)}}{n_{t,k}^{(0)} + \beta_0^{(0)}} (\mathbf{m}_{0,t}^{(0)} - \bar{\mathbf{x}}_{t,k}^{(0)})(\mathbf{m}_{0,t}^{(0)} - \bar{\mathbf{x}}_{t,k}^{(0)})^T \quad (34)$$

$$\nu_{t+1,k}^{(0)} = \nu_0 + \gamma_t^{(0)} N_{t,k} + n_{t,k}^{(0)} \quad (35)$$

ここで、

$$n_{t,k}^{(0)} = \sum_n r_{t,n,k}^{(0)}, \quad (36)$$

$$\bar{\mathbf{x}}_{t,k}^{(0)} = \frac{1}{n_{t,k}^{(0)}} \sum_n r_{t,n,k} \mathbf{x}_{t,n}^{(0)}, \quad (37)$$

$$s_{t,k}^{(0)} = \sum_n r_{t,n,k}^{(0)} (\mathbf{x}_{t,n}^{(0)} - \bar{\mathbf{x}}_{t,k}^{(0)})(\mathbf{x}_{t,n}^{(0)} - \bar{\mathbf{x}}_{t,k}^{(0)})^T \quad (38)$$

であり、 $\mathbf{x}_{t,n}^{(0)}$, $r_{t,n,k}^{(0)}$ はそれぞれ n 番目のブロックに含まれる測位データとその負担率を意味する。

次にこの推定結果と初期の分布との間の KL 情報量が大きかった場合、新たにデータを追加する。このとき、事前分布のハイパーパラメータを

$$\alpha_{0,t+1}^{(1)} = \alpha_0 + \gamma_t^{(1)} N_{t,k} + n_{t,k}^{(0)} \quad (39)$$

$$\beta_{0,t+1}^{(1)} = \beta_0 + \gamma_t^{(1)} N_{t,k} + n_{t,k}^{(0)} \quad (40)$$

$$\mathbf{m}_{0,t+1}^{(1)} = \frac{1}{\beta_{0,t+1}^{(1)}} \left(\beta_0 \mathbf{m}_0 + \gamma_t^{(1)} N_{t,k} \bar{\mathbf{x}}_{t,k} + n_{t,k}^{(0)} \bar{\mathbf{x}}_{t,k}^{(0)} \right) \quad (41)$$

$$W_{0,t+1}^{-1(1)} = W_0^{-1} + \gamma_t^{(1)} S_{t,k} N_{t,k} + s_{t,k}^{(0)} n_{t,k}^{(0)} + \frac{\gamma_t^{(1)} N_{t,k} \beta_0}{\gamma_t^{(1)} N_{t,k} + \beta_0} (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})(\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})^T + \frac{n_{t,k}^{(0)} \beta_0^{(0)}}{n_{t,k}^{(0)} + \beta_0^{(0)}} (\mathbf{m}_{0,t}^{(0)} - \bar{\mathbf{x}}_{t,k}^{(0)})(\mathbf{m}_{0,t}^{(0)} - \bar{\mathbf{x}}_{t,k}^{(0)})^T \quad (42)$$

$$\nu_{0,t+1}^{(1)} = \nu_0 + \gamma_t^{(1)} N_{t,k} + n_{t,k}^{(0)} \quad (43)$$

とすればよい。これを繰り返すと i 回目の更新における事前分布は

$$\alpha_{0,t+1}^{(i)} = \alpha_0 + \gamma_t^{(i)} N_{t,k} + \sum_j^{i-1} n_{t,k}^{(j)} \quad (44)$$

$$\beta_{0,t+1}^{(i)} = \beta_0 + \gamma_t^{(i)} N_{t,k} + \sum_j^{i-1} n_{t,k}^{(j)} \quad (45)$$

$$\begin{aligned} \mathbf{m}_{0,t+1}^{(i)} &= \frac{1}{\beta_{0,t+1}^{(i)}} \left(\beta_0 \mathbf{m}_0 + \gamma_t^{(i)} N_{t,k} \bar{\mathbf{x}}_{t,k} + \sum_j^{i-1} n_{t,k}^{(j)} \bar{\mathbf{x}}_{t+1,k}^{(j)} \right) \\ & \quad (46) \end{aligned}$$

$$\begin{aligned} W_{0,t+1}^{-1(i)} &= W_0^{-1} + \gamma_t^{(i)} S_k N_k + \sum_j^{i-1} s_{t+1,k}^{(j)} n_{t+1,k}^{(j)} \\ & \quad + \frac{\gamma_t^{(i)} N_{t,k} \beta_0}{\gamma_t^{(i)} N_{t,k} + \beta_0} (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k}) (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})^T \\ & \quad + \sum_j^{i-1} \frac{n_{t+1,k} \beta_{0,t+1}^{(j)}}{n_{t+1,k} + \beta_{0,t+1}^{(j)}} \\ & \quad \times (\mathbf{m}_{0,t+1}^{(j)} - \bar{\mathbf{x}}_{t+1,k}) (\mathbf{m}_{0,t+1}^{(j)} - \bar{\mathbf{x}}_{t+1,k})^T \quad (47) \end{aligned}$$

$$\nu_{0,t+1}^{(i)} = \nu_0 + \gamma_t^{(i)} N_{t,k} + \sum_j^{i-1} n_{t,k}^{(j)} \quad (48)$$

となる。ここで、更新式を得るために i 番目と $i+1$ 番目の式の差をとる。 β_0 に関する式は

$$\beta_{0,t+1}^{(i+1)} = (\gamma_t^{(i+1)} - \gamma_t^{(i)}) N_{t,k} + \beta_{0,t+1}^{(i)} + n_{t,k}^{(i)} \quad (49)$$

となる。ここで $\beta_k^{(i)} = \beta_0^{(i)} + n_k^{(i)}$ なので

$$\beta_0^{(i+1)} = \beta_{t,k}^{(i)} + (\gamma_t^{(i+1)} - \gamma_t^{(i)}) N_{t,k} \quad (50)$$

と記述できる。同様にして

$$\alpha_{0,t+1}^{(i)} = \alpha_{t,k}^{(i)} + (\gamma_t^{(i+1)} - \gamma_t^{(i)}) N_{t,k} \quad (51)$$

$$\begin{aligned} \mathbf{m}_{0,t+1}^{(i)} &= \frac{1}{\beta_{0,t+1}^{(i)}} \left(\beta_{t+1,k}^{(i)} \mathbf{m}_{t+1,k}^{(i)} + (\gamma_t^{(i+1)} - \gamma_t^{(i)}) N_{t,k} \bar{\mathbf{x}}_{t,k} \right) \\ & \quad (52) \end{aligned}$$

$$\begin{aligned} W_{0,t+1}^{-1(i)} &= W_{t,k}^{-1(i)} + (\gamma_t^{(i+1)} - \gamma_t^{(i)}) N_{t,k} S_{t,k} \\ & \quad + \left(\frac{\gamma_t^{(i+1)} N_{t,k} \beta_0}{\gamma_t^{(i+1)} N_{t,k} + \beta_0} - \frac{\gamma_t^{(i)} N_{t,k} \beta_0}{\gamma_t^{(i)} N_{t,k} + \beta_0} \right) \\ & \quad \times (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k}) (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})^T \quad (53) \end{aligned}$$

$$\nu_{0,t+1}^{(i)} = \nu_{t,k}^{(i)} + (\gamma_t^{(i+1)} - \gamma_t^{(i)}) N_{t,k} \quad (54)$$

となる。このように、 $i+1$ 番目の事前分布は直前の推定結果 $\alpha_{t,k}^{(i)}$, $\beta_{0,t+1}^{(i)}$, $\mathbf{m}_{t+1,k}^{(i)}$, $W_{t,k}^{-1(i)}$, $\nu_{t,k}^{(i)}$ と減衰率 $\gamma_t^{(i+1)}$, $\gamma_t^{(i)}$ から算出することができる。この更新式を KL 情報量が十分小さくなるまで繰り返した後、最後に得られる結果が時刻 $t+1$ の推定結果となる。推定結果を得たのが $i=l$ のときであったとすると、

$$N_{t+1,k} = \alpha_{t,k}^{(l)} - \alpha_0 = \beta_{t,k}^{(l)} - \beta_0 \quad (55)$$

$$\bar{\mathbf{x}}_{t+1,k} = \frac{1}{N_{t+1,k}} \left(\beta_{t+1,k}^{(l)} \mathbf{m}_{t+1,k}^{(l)} - \beta_0 \mathbf{m}_0 \right) \quad (56)$$

$$\begin{aligned} S_{t+1,k} &= \frac{1}{N_{t+1,k}} \left(W_{t,k}^{-1(l)} - W_0^{-1} - \frac{N_{t+1,k} \beta_0}{N_{t+1,k} + \beta_0} \right. \\ & \quad \times (\mathbf{m}_0 - \bar{\mathbf{x}}_{t+1,k}) (\mathbf{m}_0 - \bar{\mathbf{x}}_{t+1,k})^T \left. \right) \quad (57) \end{aligned}$$

のように定義した $N_{t+1,k}$, $\bar{\mathbf{x}}_{t+1,k}$, $S_{t+1,k}$ を用いて

$$\alpha_{0,t+1,k} = \alpha_0 + N_{t+1,k} \quad (58)$$

$$\beta_{0,t+1,k} = \beta_0 + N_{t+1,k} \quad (59)$$

$$\mathbf{m}_{0,t+1,k} = \frac{1}{\beta_{0,t+1,k}} (\beta_0 \mathbf{m}_0 + N_{t+1,k} \bar{\mathbf{x}}_{t+1,k}) \quad (60)$$

$$\begin{aligned} W_{0,t+1,k}^{-1} &= W_0^{-1} + S_{t+1,k} N_{t+1,k} + \frac{N_{t+1,k} \beta_0}{N_{t+1,k} + \beta_0} \\ & \quad \times (\mathbf{m}_0 - \bar{\mathbf{x}}_{t+1,k}) (\mathbf{m}_0 - \bar{\mathbf{x}}_{t+1,k})^T \quad (61) \end{aligned}$$

$$\nu_{0,t+1,k} = \nu_0 + N_{t+1,k} \quad (62)$$

と書き換えることができる。この形は変分ベイズ法における更新式と同等の形になっている。したがって、次の時刻 $t+2$ に対しても同様の更新手続きが実行できる。なお、ここで定義した値は以下のようにも書ける。

$$N_{t+1,k} = \gamma_t^{(l)} N_{t,k} + \sum_j^l n_{t+1,k}^{(j)}, \quad (63)$$

$$\bar{\mathbf{x}}_{t+1,k} = \gamma_t^{(l)} N_{t,k} \bar{\mathbf{x}}_{t,k} + \sum_j^l n_{t+1,k}^{(j)} \bar{\mathbf{x}}_{t+1,k}^{(j)}, \quad (64)$$

$$S_{t+1,k} = \gamma_t^{(l)} S_{t,k} N_{t,k} + \sum_j^l s_{t+1,k}^{(j)} n_{t+1,k}^{(j)} + \delta W \quad (65)$$

ここで用いた δW は

$$\begin{aligned} \delta W &= \frac{\gamma_t^{(i)} N_{t,k} \beta_0}{\gamma_t^{(i)} N_{t,k} + \beta_0} (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k}) (\mathbf{m}_0 - \bar{\mathbf{x}}_{t,k})^T \\ & \quad - \frac{N_{t+1,k} \beta_0}{N_{t+1,k} + \beta_0} (\mathbf{m}_0 - \bar{\mathbf{x}}_{t+1,k}) (\mathbf{m}_0 - \bar{\mathbf{x}}_{t+1,k})^T \\ & \quad + \sum_j^l \frac{n_{t+1,k} \beta_{0,t+1}^{(j)}}{n_{t+1,k} + \beta_{0,t+1}^{(j)}} (\mathbf{m}_{0,t+1}^{(j)} - \bar{\mathbf{x}}_{t+1,k}) \\ & \quad \times (\mathbf{m}_{0,t+1}^{(j)} - \bar{\mathbf{x}}_{t+1,k})^T \quad (66) \end{aligned}$$

である。式 (13) の定義を参照すると、 N および $\bar{\mathbf{x}}$ については過去のデータの寄与を減衰させて加算をとった自然な形になっている。 S についても同様であるが、 δW の項だけが残っている。ただし、 δW は逐次計算にともなって発生する誤差であり、 N や n が十分大きい場合には、ほとんど計算には影響しない。したがって、上記計算式による更新で分布が推定できる。

4. 実験

4.1 実験条件

提案方式の評価のための実験を行った。表 1 に実験条

表 1 使用データ

Table 1 Summary of the dataset for the experiment.

項目名	詳細
使用データ	人の流れ PJ [25] 東京都市圏 (H20) の位置情報 (緯度経度)
件数	毎分約 60 万人分
時間	午前 7 時～午前 9 時

件を示す．今回用いたのは，アンケートにより収集された人がいつどこへ移動したかという情報に対して交通網の経路探索などの内挿処理を施した位置情報のデータ [25] であり，1 分おきに全人物がどの地点にいたかが格納されている．このデータのうち午前 7 時から午前 9 時までのデータについて分布を生成する．なお，各データは緯度経度で表記されているため，初期時刻のデータの平均座標値を原点にし，外接矩形の大きさが 10.0 になるように正規化したものを入力データとした．

今回は以下のようにして評価を行った．まず，各時刻の人の流れデータをランダムに並べ替え，最初の 50% のデータから分布を推定し，後半 10% のデータを対象に精度評価を行った．なお，後半 10% を評価対象としたのは，データ件数が多すぎて処理時間が過剰にかかるためで，10% つまり 6 万点のデータを対象とすれば評価には十分な数が得られていると考えられる．精度の評価基準としては，各点に関する予測分布の値の対数の平均値，つまり N 点の測位データ $\mathbf{x}_n$ について

$$Ev = \frac{1}{N} \sum_{n=1}^N \ln P(\mathbf{x}_n) \quad (67)$$

によって定義される Ev を精度とした．なお，この値は予測分布から $\mathbf{x}_n$ が得られる同時観測確率を規格化したものを意味している．比較のため今回の対象領域 10.0×10.0 の範囲の一樣分布をこの評価基準で評価すると， $Ev = \frac{1}{N} \sum_{n=1}^N \ln P(\frac{1}{10^2}) \simeq -4.61$ となる．

提案方式では，ガウス分布の混合数を 50 とし，測位データの逐次投入ブロックの大きさは 1,500 とした．ハイパーパラメータの初期値は $\alpha_0 = 2.1$ ， $\beta_0 = 1.0$ とし， W_0 は対角成分が 0.4，対角成分以外は 1.0×10^{-10} ， $\nu_0 = 2.2$ とした．また， $\mathbf{m}_0$ は初期時刻の点からランダムに選択した座標値を用いた．なお，計算を始める初期時刻を対象時刻の 3 時間前の午前 4 時とし，逐次投入にともなう密度分布更新が十分に行われた後の状態を評価対象としている．比較対象として，ガウスカネル法，Quadratic カーネル法に関しても同様の評価を行った．ガウスカネル法の平滑化パラメータは $h = 0.15$ とした．また，Quadratic カーネル法の平滑化パラメータは $P(\mathbf{x}_n) = 0$ とならないように $h = 3$ とした．これは $P(\mathbf{x}_n) = 0$ ならば $Ev = -\infty$ であり比較に適さないためである．なお，実行環境としては Java (JDK7 Update 13) で実装されており，Intel Core

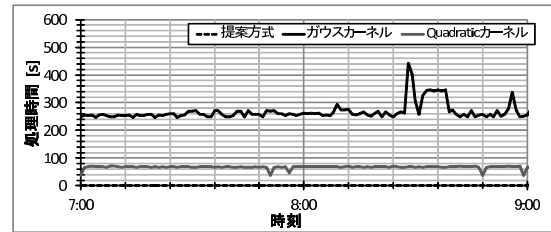


図 6 処理速度の比較

Fig. 6 Contrast of processing speed.

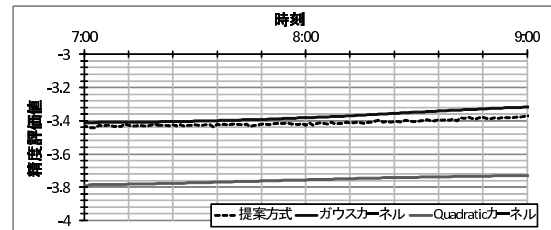


図 7 分布推定精度の比較

Fig. 7 Contrast of distribution accuracy.

i7-3980K 3.2 GHz, RAM 32.0 GB のコンピュータを使用した．

4.2 実験結果

図 6 に処理速度の比較を示す．この処理速度としては，測位データから 150×100 画素の画像を生成するのにかかる時間を計測して用いた．この計算では横軸の時刻における測位データから画像生成するのにかった時間を縦軸とした．提案方式は 0.5～1.0 秒程度，平均 0.67 秒で処理を終えているのに対し，従来方式はガウスカネルで 250～270 秒で平均 260 秒，Quadratic カーネルでも 30～70 秒，平均 67 秒程度の処理時間を要している．提案方式の処理速度は従来のいずれの方式に比べても十分高速である．1 分のデータを処理するのに 1 分以上の時間をかけるとリアルタイムではなくなるが，提案方式だけがその要求を満たしており，提案方式は従来比 100 倍以上高速であったといえる．

図 7 に精度に関する実験結果を示す．精度に関しては，ガウスカネルが最も良い結果を示した．Quadratic カーネルは処理時間が高速化された反面，精度はガウスカネルよりも低い値となった．一方，提案方式はガウスカネルより精度が低いものの，Quadratic カーネルよりは大きく改善されている．図 8 に実際に生成された分布の比較を示す．本評価で生成した画像は 150×100 画素の画像であるが，これを $1,500 \times 1,000$ 画素の背景地図 [26] に重畳した結果を示している．(a) が提案方式，(b) はガウスカネル法，(c) は Quadratic カーネル法で生成された画像である．ガウスカネル法で想定するガウス分布はつねに円形であるのに対し，提案方式におけるガウス分布では楕円状の分布を許している．そのため，提案方式の生成画像は鉄

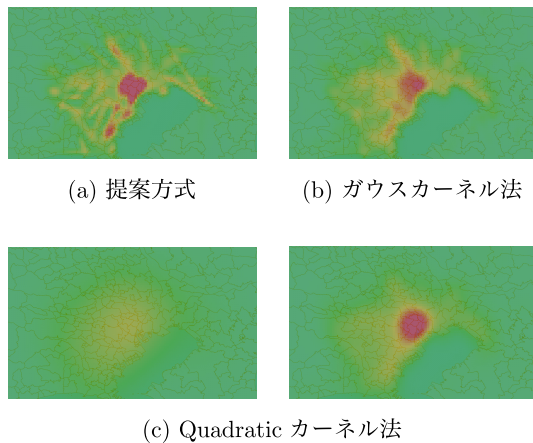


図 8 生成された分布画像の比較

Fig. 8 Contrast of rendered distributions.

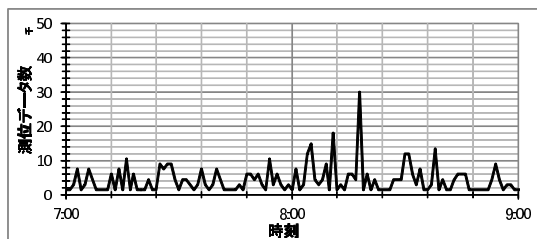


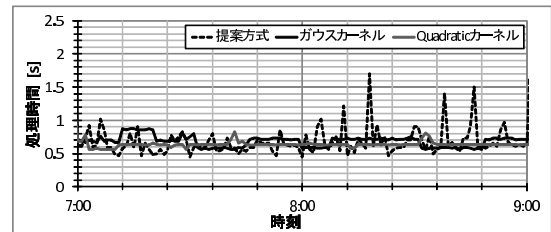
図 9 提案方式の使用データ数

Fig. 9 Number of used data with the proposed method.

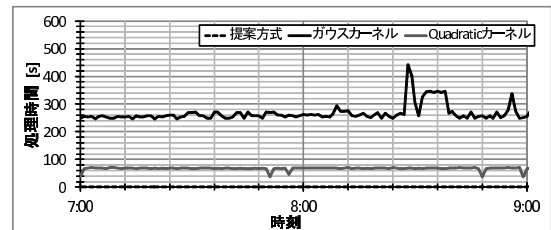
道路線などに沿って線状の分布が得られている。また、(c)の左側は精度評価に用いたパラメータ $h = 3$ の Quadratic カーネル法の結果であるが、薄く全体に広がってしまっている。これは、予測分布が 0 にならないように h を決めたためであり、たとえば $h = 1$ にすると右側の画像が得られ、他の方式と近い分布となる。ただし、このときの精度指標は $-\infty$ である。このように Quadratic カーネル法で生成された分布は予測分布として扱うには適さない。ただし、可視化のためにパラメータ調整すれば有用である。

4.3 考察

実験結果から、提案方式は従来よりも非常に高速に処理ができると判明した。これは、提案方式のデータ数削減の影響が大きいと考えられる。実際には、通信などにかかる時間が発生することを考慮すれば、提案方式の効果はさらに大きくなると推定される。図 9 に提案方式において使用されたデータ数を示す。平均 4,500 件程度データを使用しており、最大でも 30,000 件程度、すなわち全体の 1/10 まで使用データを削減したことになる。今回、逐次投入されるブロックは 1,500 件ずつであるので、平均的には 2~3 回逐次投入が行われたといえる。なお、精度について、全方式とも 9:00 に近づくにつれて値が大きくなっているが、これは当該時間帯が通勤時間帯であり、人口が都心部に密集してきているために予測がしやすくなったということが原因と考えられる。また提案方式は従来のガウスクーネル



(a) 使用データ削減時の従来方式速度



(b) 使用データ削減時の従来方式精度

図 10 データ削減による精度低下

Fig. 10 Decrease in accuracy due to the data reduction.

法にくらべてこの精度向上が低く差が大きくなっているという挙動を示している。これは、提案方式の処理中で過去のデータを使用しているため、人口分布が急激に変化した場合でも変化前のデータの寄与が残り、分布の変化に追従するのが遅くなっているためであると考えられる。

従来方式で使用するデータ数を削減する方式としてランダムにデータを間引く方法が考えられる。そこで、提案方式と同程度の速度になるまで測位データを間引きしてから従来方式を適用した場合の結果を図 10 に示す。このグラフに表示した提案方式は、比較のため図 6 で示した結果を再掲したものである。ガウスクーネル法で用いたのは 1,100 件、Quadratic カーネル法で用いたのは 3,000 件であり、(a) に示すとおり、それぞれ平均 0.68 秒、0.64 秒で処理を終えており、提案方式 (0.67 秒) と同程度の処理速度になった。このとき、(b) のとおり提案方式が最も高精度な結果が得られた。なお、ガウスクーネル法、Quadratic カーネル法では、データを削減した結果、前節の評価時に用いたパラメータでは精度が $-\infty$ になってしまった。これは、点間引きによって極端に尤度の低い点が現れてしまい、計算機精度を下回ったためである。そこで、ガウスクーネル法では $h = 3.0$ 、Quadratic カーネル法では $h = 5.0$ としたときの結果を示している。以上のように、提案方式と同程度の速度で処理ができるように間引きを行うと、精度が著しく低下していることが確認でき、提案方式は間引きを行っても精度を保つ方式であるといえる。

図 11 に提案方式の処理時間の内訳を示す。提案方式はパラメトリックな推定方式であり、密度分布生成と画像生成の 2 つの手順それぞれで処理時間がかかる。画像生成の処理時間はおよそ 0.5 秒で一定しているのに対し、分布推定はほとんど時間がかからないこともあるが 1 秒以上の時

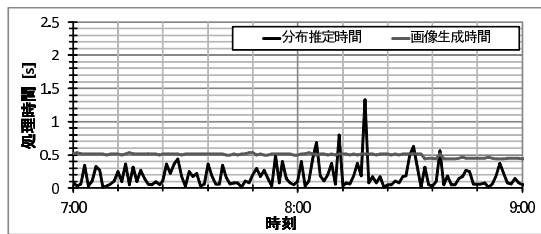


図 11 提案方式における処理時間の内訳

Fig. 11 Breakdown of process-time by the proposed method.

間がかかることもあった。これは、データの逐次投入による影響であり、データを多数投入する必要がある場合には処理時間がやや大きい、計算が打ち切れた場合は極端に小さくなっているといえる。

5. おわりに

本研究では、移動体端末から収集された測位データの密度分布を高速に精度良く推定する方法として、変分ベイズ法に基づき一部の測位データだけを利用して全体分布を推定する分布推定方式を提案した。また、提案方式について東京都市圏の人の流れデータを用いて評価したところ、従来に比べて約 100 倍の速度で分布が推定できた。さらに全データの半数を用いて生成された分布と残りのデータの比較による精度評価を行い、提案方式は従来方式と遜色ない精度が得られ、実際に生成される分布画像も従来と大きくは異ならないことが確認できた。提案方式において分布生成に用いる測位データは全体のおよそ 1/10 程度でよいことが判明した。

本研究で対象としたのは最も基本的な密度分布推定であるが、今後、空間上に分布する量的データに関する推定処理 [9] についても同様のアプローチをとれる可能性があり、それは今後の課題である。

参考文献

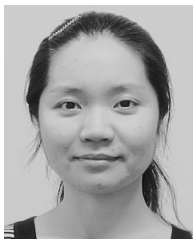
- [1] 伊藤誠悟, 吉田廣志, 河口信夫: locky.jp: 無線 LAN を用いた位置情報・測位ポータル, 情報処理学会研究報告, 2005-MBL-34(4), pp.25-31 (2005).
- [2] Manandhar, D., Kawaguchi, S., Uchida, M., Ishii, M. and Tomohiro, H.: IMES for mobile users: Social implementation and experiments based on existing cellular phones for seamless positioning, *Proc. Int. Symposium on GPS/GNSS 2008* (2008).
- [3] 興梠正克, 大隈隆史, 蔵田武志: 歩行者ナビのための自蔵センサモジュールを用いた屋内測位システムとその評価, モバイル学会シンポジウムモバイル 08 予稿集, pp.151-156 (2008).
- [4] 羽藤英二: モデルとデータ: プローブパーソン調査の課題と今後, 土木計画学研究・講演集, Vol.27, p.259 (2003).
- [5] 杉野勝敏, 矢野晋哉, 羽藤英二, 朝倉康夫: プローブパーソン調査を用いた観光行動分析, 土木計画学研究・講演集, Vol.32, p.294 (2005).
- [6] Asakura, Y. and Hato, E.: Tracking survey for individual travel behaviour using mobile communication instruments, *Transportation Research Part C: Emerging*

- Technologies*, Vol.12, No.3-4, pp.273-291 (2004).
- [7] 長尾光悦, 川村秀憲, 山本雅人, 大内 東: 観光動態情報の獲得を意図した GPS ログデータマイニング (社会システムにおける知能: 社会システムのモデル化), 情報処理学会研究報告, 2004-ICS-135(2), pp.7-12 (2004).
- [8] 鈴木恵二, 松原 仁, 鈴木克也: はこだてまちナビにおける観光客動態の推定, 情報処理学会研究報告, UBI, Vol.2008, No.40, pp.41-46 (2008).
- [9] 矢島美寛, 平野敏弘: 時空間大規模データに対する統計的解析法, 統計数理, Vol.60, No.1, pp.57-71 (2012).
- [10] Chudova, D., Gaffney, S., Mjolsness, E. and Smyth, P.: Translation-invariant mixture models for curve clustering, *Proc. 9th ACM SIGKDD*, pp.79-88, ACM (2003).
- [11] 浅原彰規, 寺本やえみ, 丸山貴志子, 柴崎亮介: 時空間ネットワーク解析を用いた主動線抽出, 情報処理学会論文誌, Vol.54, No.1, pp.135-145 (2013).
- [12] 山田直治, 磯田佳徳, 南 正輝, 森川博之: B-15-23 GPS 搭載携帯電話を用いた移動経路履歴に基づく訪問地予測方式 (B-15. モバイルマルチメディア通信, 一般セッション), 電子情報通信学会総合大会講演論文集, Vol.2010, No.1, p.639 (2010).
- [13] 田中宏平, 寺田 努, 岸野泰憲, 西尾章治郎: カーナビゲーションシステムのための走行履歴と運転状況を用いた車両目的地予測手法, 情報処理学会論文誌, Vol.51, No.1, pp.141-151 (2010).
- [14] 浅原彰規, 佐藤暁子, 丸山貴志子: 混合自己回帰隠れマルコフモデルによる歩行者行き先予測, 情報処理学会論文誌, Vol.53, No.1, pp.342-351 (2012).
- [15] Asahara, A., Maruyama, K., Sato, A. and Seto, K.: Pedestrian-movement prediction based on mixed Markov-chain model, *Proc. 19th ACM SIGSPATIAL GIS 2011*, pp.25-33, ACM (2011).
- [16] 松尾 豊, 岡崎直観, 中村嘉志, 西村拓一, 橋田浩一, 中島秀之: 位置履歴からのユーザ属性の推定 (モバイルコンピューティング), 情報処理学会論文誌, Vol.48, No.6, pp.2106-2117 (2007).
- [17] 小磯貴史, 服部可奈子, 吉田琢史, 今崎直樹: 歩行者動線分析システムを用いた大型家電量販店での行動分析, 情報処理学会研究報告, 2003-UBI-2(11) (2003).
- [18] 貞広幸雄: 空間集計データにおける面補間法の推定精度評価, 都市計画, Vol.49, No.2, pp.75-81 (2000).
- [19] Silverman, B.W.: *Density Estimation for Statistics and Data Analysis*, Chapman and Hall/CRC (1986).
- [20] Simonoff, J.S.: *Smoothing methods in statistics*, Springer (1996).
- [21] 田中耕市: 1990 年代の東京 23 区における都市密度変化と土地利用転換: 事務所建築物と集合住宅を対象として, 地学雑誌, Vol.117, No.2, pp.479-490 (2008).
- [22] Bishop, C.M.: *Pattern Recognition and Machine Learning*, Springer (2006).
- [23] 伊庭幸人, 汪 金芳, 田栗正章, 手塚 集, 樺島祥介, 上田修功: 統計科学のフロンティア 11 計算統計 I 確率計算の新しい手法, 岩波書店 (2006).
- [24] C.M. ビショップ: パターン認識と機械学習 (下) ベイズ理論による統計的予測, pp.187-195, シュプリンガー・ジャパン (2008).
- [25] 東京大学空間情報科学センター (共同研究「ダイナミックデータ統合可能な都市空間情報基盤の研究」): 人の流れプロジェクト, 入手先 (<http://pflow.csis.u-tokyo.ac.jp/index-j.html>).
- [26] 国土地理院: 国土基盤地図情報 1/25000, 入手先 (<http://www.gsi.go.jp/kiban/index.html>).



浅原 彰規 (正会員)

2002 年北海道大学理学部物理学科卒業。2004 年同大学大学院理学研究科物理学専攻修士課程修了。同年 (株) 日立製作所入社。以来、中央研究所にて空間情報システムの研究に従事。電子情報通信学会会員。



寺本 やえみ (正会員)

2003 年早稲田大学理工学部電子情報通信学科卒業。2005 年同大学大学院理工学研究科情報・ネットワーク専攻修士課程修了。同年 (株) 日立製作所入社。中央研究所にてテキスト解析システムおよび空間情報システムの研究に従事。2013 年 4 月より、(株) 日立製作所情報・通信システム社にてビッグデータ分析事業に従事。