ProbeとTweetを用いたマルチリソースによる潜在交通状態推定

原　祐輔[1]・松田　耕史[2]・川崎　洋輔[3]・三谷　卓摩[4]・桑原　雅夫[5]

[1]正会員　東北大学大学院　情報科学研究科（〒 980-8579 仙台市青葉区荒巻青葉 6-6-06）
E-mail: hara@plan.civil.tohoku.ac.jp
[2]正会員　東北大学大学院　情報科学研究科（〒 980-8579 仙台市青葉区荒巻青葉 6-6-05）
E-mail: matsuda@ecei.tohoku.ac.jp
[3]正会員　東北大学大学院　情報科学研究科（〒 980-8579 仙台市青葉区荒巻青葉 6-6-06）
E-mail: kawasaki-y@plan.civil.tohoku.ac.jp
[4]正会員　東北大学大学院　情報科学研究科（〒 980-8579 仙台市青葉区荒巻青葉 6-6-06）
E-mail: mitani@plan.civil.tohoku.ac.jp
[5]正会員　東北大学大学院　情報科学研究科（〒 980-8579 仙台市青葉区荒巻青葉 6-6-06）
E-mail: kuwahara@plan.civil.tohoku.ac.jp

本研究は道路ネットワーク上の交通状態を推定するためにプローブカーデータと Twitter のツイートデータを用いる．ツイートデータから交通情報・地理情報を抽出するために，交通情報タグ付与器とジオパーズエンジンを開発し，テキストデータから交通情報と位置座標を取得する手法を構築した．また，プローブデータとソーシャルメディアデータという空間分解能やデータ空間が異なるデータを統一的に扱う交通状態推定モデルを構築し，シミュレーションデータでの検証によって有用性を確認した．最後に 2014 年 2 月に発生した甲信越地方での豪雪災害を対象としたケーススタディを行い，本手法の有用性を示した．

Key Words : *traffic state inference, probe car data, social media data, natural language processing, variational Bayes method*

1. はじめに

実世界で起こる交通現象や交通行動を観測する方法として，プローブカーやスマートフォンから得られる GPS 軌跡データは道路ネットワークの状況や渋滞延伸の様子を捉えられる点で有用である．また，道路インフラ側にある道路感知器と合わせることで，より詳細な交通状況を把握することが可能[1)]である．プローブデータや道路感知器データは交通流を直接的及び間接的に観測しているため，(1) 自由流状態と渋滞流状態を判別可能である，(2) 交通量等の総量に対する情報を有している (道路感知器)，(3) 各車両の加減速や渋滞の延伸/縮小情報などのミクロスコピックな情報を有している (プローブデータ)，(4) 各車両の経路や旅行時間などの利用者サイドの情報を有している (プローブデータ) といった特徴をもつ．

ある観測データを用いることは実世界の現象をあるデータ空間に埋め込むことに対応する．ある一つのデータリソースを用いたデータ空間に埋め込んだ場合，実用上は別現象として認識したい事象をデータからは識別不可能となることが多々ある．たとえば，自然渋滞や朝ピーク渋滞といった定常的な渋滞，イベント渋滞，事故渋滞，濃霧や積雪によるネットワークパフォーマンスの低下による渋滞などは人の目で実際の交通現象を見れば容易に判断可能であったとしても，プローブデータのみから識別することは難しい．そこで，本研究ではプローブデータから得られる有用な情報に加えて，「人の目で実際の交通現象を見れば容易に判断可能であった」データを利用することで，潜在的な交通状態を推定することを目的とする．

ソーシャルメディアから実世界の現象を把握しようとする研究は多く存在する．たとえば，Sakaki et al. (2010)[2)]は Twitter 上での人々のつぶやきと位置情報タグから，地震の発生源を特定した有名な研究である．また，Lee et al. (2010, 2011)[3),4)]は Twitter から各地域の通常時のつぶやき傾向を捉えることで，異常時判定し，地域イベントの検出を行うシステムを開発している．しかし，テキストベースのソーシャルメディアのみでは空間の分解能が粗いため，交通渋滞や交通事故のような局所的なイベントの場所を正確に検知することは難しい．

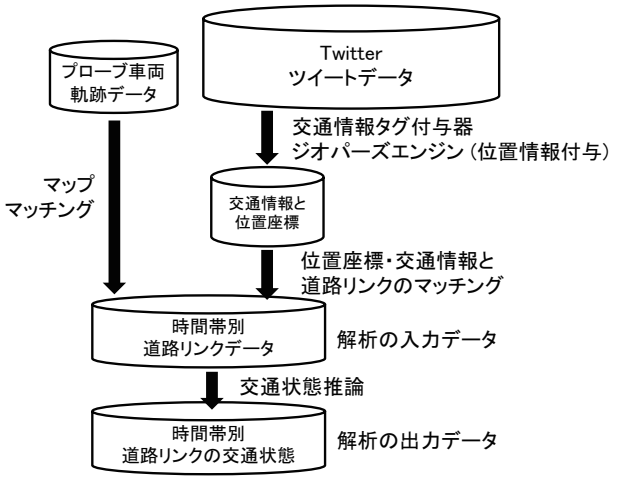
そこで，本研究はプローブデータとソーシャルメディアデータ (Twitter データ) の双方をデータリソースとして取り扱うことで，単一リソースのみでは把握が不可能であった交通状態推論を行う．プローブデータは交通現象の発生位置や正確な情報が得られる一方で，プ

ローブ観測が存在しない場合には道路リンク上で発生している現象を把握することができない．災害時などの異常時にプローブデータが存在しないのは単にプローブ車両が通過していないためなのか，道路自体が通行不可能であるのかを判別することは難しい．そのような場合にソーシャルメディアの情報は交通現象を人の目で判断し，つぶやかれているため，有用であろう．一方で，ソーシャルメディア上で「事故渋滞」とつぶやかれている場所はプローブデータから正確な位置が把握可能となるだろう．このように両データは実世界を観測する手法として補完的な関係性にある．

本研究におけるマルチリソースデータの取り扱いを具体的に述べる．まず，各道路リンクのリンク平均速度はプローブデータから算出し，Twitter からは交通情報と地理情報を含むツイートを抽出する．この情報抽出手法を新たに開発した点が本研究の一つ目の貢献である．次に，データ解像度が異なる 2 つのデータリソースを統一的に扱う統計モデルを構築し，各道路リンクの潜在的交通状態を推定する．マルチリソースデータを統一的に扱い潜在的交通状態を推論する手法を開発したことが本研究の二つ目の貢献である．本手法の検証として，最初にシミュレーションデータに対してモデルの検証を行い，次に，2014 年 2 月に発生した甲信越豪雪時の甲府市周辺の交通状態の推論をケーススタディとして実施する．

2. 分析フレームワーク

本研究の分析フレームワークを**図–1** に示す．まず，プローブデータと Twitter データそれぞれに対して前処理を行う必要がある．プローブデータはデジタル道路地図 (DRM) に対してマップマッチングを行い，単位時間あたりの各リンク平均速度を算出する．Twitter データには，本研究の目的とは関連しない交通情報や地理情報を含まないツイートが多く存在するため，交通情報と地理情報を含むツイートのみを抽出する必要がある．また，地理情報は地名や施設名といったテキストデータではなく，位置座標として扱いたいため，地理表現から位置座標付与を行う．このような前処理が行われたデータは時間帯別道路リンク単位に集計されている．このデータを入力データとし，マルチリソースデータを統一的に扱う統計モデルを構築する．その解析結果として各時間帯のリンク単位の交通状態を推論することができる．統計モデルのの詳細は **3.** にて記述する．



図–1 本研究の分析フレームワーク

(1) プローブデータの前処理

プローブデータは各車両の数秒間隔の GPS 軌跡データである．これに対して，デジタル道路地図 (DRM) へのマップマッチングを行い，各車両のリンク単位経路データを作成する．次に，時間間隔を 15 分単位として，各リンクを通過した車両の平均リンク速度を算出する．

(2) Twitter データからの交通情報の抽出

Twitter データに対して，交通情報を抽出するための交通情報タグ付与器を構築する．既往研究では，Twitter データから有益な情報を抽出する方法として，トピックモデルを用いてトピックを抽出した後に，トピックとラベルの対応付けを行う研究[5]や鉄道運行トラブルに対して少量の教師データから SVM を構築し，同内容の抽出する研究[6]が存在する．本研究では，対象とする実世界現象が限定されていることから，後者の少量の教師データからタグ付与器を構築するアプローチを選択する．このタグ付与器では**表–1** に示すような交通手段，交通状態大分類，詳細交通状態，情報源 (一次情報 or 二次情報) を各ツイートに対して付与するための分類器である．

構築のために，まず教師データとして 5 万件のツイートに対して，交通手段や交通状態，情報源のタグ付与を手作業で行った．交通手段や交通状態については，1 ツイートに対して 1 件と限定せず，マルチタグ付与を許容した．次に教師データをもとに，交通情報タグ付与器の学習を行う．交通情報タグ付与器では内部でサポートベクターマシン (SVM) を利用している．ここで，内部で利用する SVM には大量のツイートに対して高速な処理を行うために，国立台湾大学で開発された LIBLINEAR[7] を利用した．LIBLINEAR は SVM のカーネルを線形カーネルにのみ限定する代わりに高速な学習・予測が行えるライブラリである．

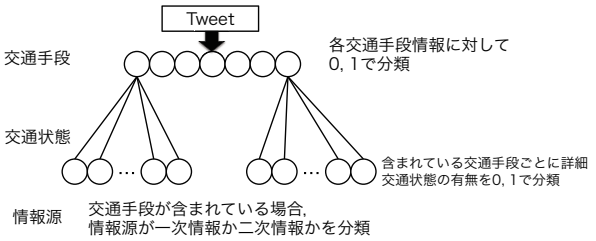
各ツイートの特徴量には MeCab でツイートを形態素

表–1 交通情報タグ一覧

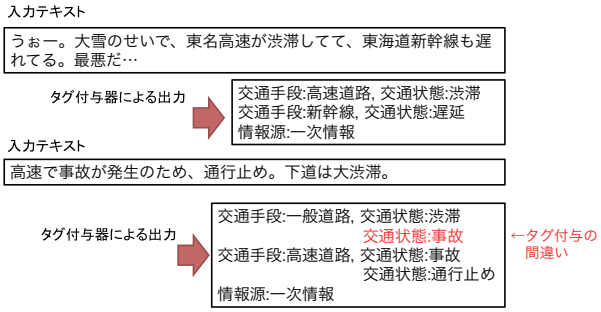
交通機関	交通状態	交通状態 (詳細)
一般道路	利用可能	交通量少，良路面状態
	障害発生	渋滞，混雑，工事，規制 悪路面状態，視界不良 事故，立ち往生，路面凍結
	利用不可	通行止め，道路冠水
	その他	歩道情報
高速道路	利用可能	通行可能
	障害発生	渋滞，事故，立ち往生 悪路面状態，チェーン規制 速度規制
	利用不可	通行止め
鉄道	利用可能	運行
	障害発生	遅延，事故
	利用不可	運休，運転見合わせ
	その他	車内混雑，駅構内混雑
新幹線	利用可能	運行
	障害発生	遅延，事故
	利用不可	運休，運転見合わせ
	その他	車内混雑，駅構内混雑
バス	利用可能	運行
	障害発生	遅延，事故，立ち往生
	利用不可	運休
	その他	車内混雑，待ち行列
高速バス	利用可能	運行
	障害発生	遅延，事故，立ち往生
	利用不可	運休
飛行機	利用可能	運行
	障害発生	遅延，引き返し 着陸先変更，滑走路閉鎖
	利用不可	欠航
	その他	空港足止め，空港混雑

解析後の表層形 (文中の表現そのまま) と原形 (動詞や形容詞を原形に直したもの) を用いた．あわせて特徴量ベクトルの長さは約 15 万 3000 である．コストパラメータとバイアス値は 10-fold cross-validation の値を用いて決定した．最適パラメータ値での 10-fold cross-validation 精度は 85%である．

また，マルチタグ付与に対応するために，**図–2** で示すツリー構造によって binary relevance として各交通手段・交通状態・情報源に対して 0, 1 でタグ付与を行っている．この場合，最小 8 個，最大 66 個の SVM を利用してタグ付与を行う．構築された交通情報タグ付与器を用いると，新たなツイートに対して，**図–3** に示すように，交通手段，交通状態，情報源などを付与することができる．このタグ付与器は当然のことながら，**図–3** に示すような解析誤りを示すこともある．



図–2 タグ付与器の分類フロー



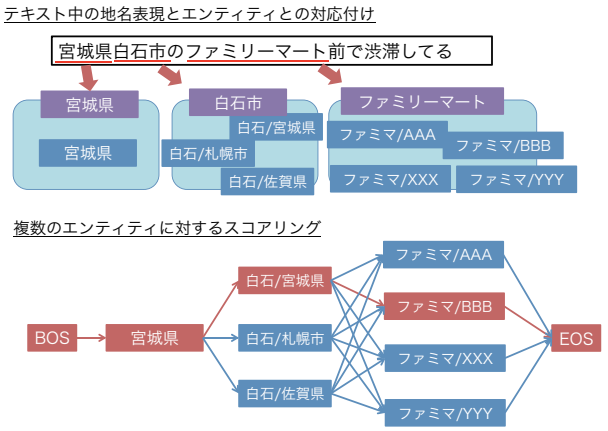
図–3 交通情報タグ付与器の分析結果例

(3) Twitter データからの地理情報の抽出

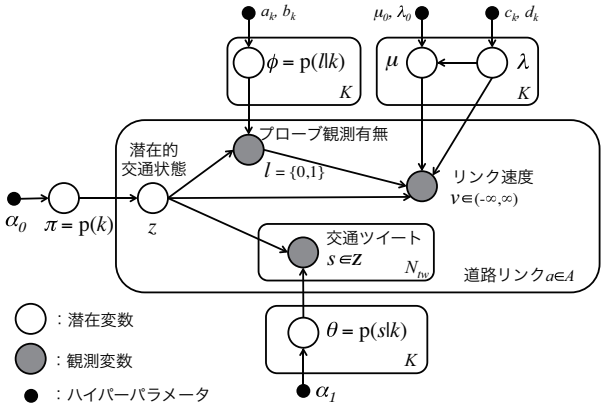
Twitter データからの地理情報抽出には現在，東北大の乾・岡崎研が開発しているジオパーズ (Geo Parse) エンジンを用いた．ジオパーズとはテキスト中の場所を参照する表現を認識し，実際の場所に対応づける技術である．本研究で用いるジオパーズエンジンは具体的にはテキスト内の地理情報に対して，位置座標を付与することができる．ジオパーズは一般に (1) 地理表現の認識，(2) 地理的曖昧性の解消の 2 つの要素技術から構成される．正確な地理表現の認識には様々なバリエーションがある地理表現を認識するためにカバレッジの高い辞書を構築することが重要であり，地理的曖昧性の解消のためには文脈を認識し，多数の候補の中から適切なものを選択することが重要である．

本研究で用いるジオパーズエンジンは地理表現の認識には Web から収集した大規模辞書を用いることで対処している．具体的にはランドマークには Yahoo!ロコ 850 万エントリ，鉄道・駅には国土数値情報から 5 万エントリ，道路にはデジタル道路地図 165 万エントリ，住所には街区レベル位置参照情報 15 万エントリを利用している．

ツイートに対する位置座標の付与は以下の手続きで行う．まず，街区レベル位置参照情報を用いて，日本全国の地名表現とその指し示すエンティティ（実存する概念）の対応を接尾辞木形式の辞書に変換する．次に，処理対象の各ツイートから，前述の辞書に含まれる地名表現を共通接尾辞探索を用いて最左最長一致で取り出



図–4 地理表現の曖昧性解消と位置座標付与



図–5 モデル構造のグラフィカルモデル

し，対応するエンティティの位置座標を付与する．ただし，曖昧性が存在し，複数のエンティティが対応しうるような地名表現に対しては，同一ツイート内でのエンティティ間の距離が最短となるような経路解釈[8]を動的計画法により探索し，その解釈におけるエンティティの座標を付与する．一連のプロセスを**図–4** に示す．

以上より，プローブデータ，ツイートデータともに，道路リンク単位の同一時間・空間分解能に集約することができる．このような 15 分集計道路リンク単位の平均リンク速度とツイートによる交通情報を入力データとして，次章にて交通状態の推論モデルを構築する．

3. リンク速度と交通・地理情報ツイートの生成モデル

本研究で考える確率的生成モデルは道路リンク間は独立，時間帯も独立という各道路リンク単位での単純な生成モデルを考える．そのため，一つのデータはある時間帯のある道路リンクを表し，プローブ観測有無 l，リンク速度 v，交通状態ツイート数 $\boldsymbol{s}$ が観測されているとする．交通状態ツイート数 $\boldsymbol{s}$ とは該当リンク，該当時間帯における (通行可能ツイート数，通行障害ツイート数，通行不可ツイート数，その他ツイート数) のベクトルを表す．以降，$I = 4$ として i 番目の交通状態ツイート数を s_i で表す．

潜在的交通状態 z は K 種存在すると仮定する．これは多項分布 (パラメータ $\boldsymbol{\pi}$) から決定し，その事前分布はディリクレ分布 (ハイパーパラメータ $\boldsymbol{\alpha}_0$) である．各リンクに対し，プローブ観測有無は事前分布がベータ分布である二項分布 (パラメータ $\boldsymbol{\phi}_k$) と潜在的交通状態から生成される．次に，潜在的交通状態と観測有無，平均 μ_k, 精度 λ_k の正規分布からリンク速度が生成される．ツイート側は，潜在的交通状態と多項分布 (パラメータ $\boldsymbol{\theta}_k$) から交通状態ツイート数 $\boldsymbol{s}$ が生成される．該当リンク・該当時間帯に対応する総ツイート数が N_{tw} のとき，多項分布から N_{tw} 回生成した結果として，交通状態ツイート数 $\boldsymbol{s}$ が得られる．このような生成モデルは**図–5** として変数間の関係性を表現することができる．

この生成モデルの直感的なイメージを説明しよう．仮にリンクの交通状態が自由流と渋滞流の 2 つの状態しかないと仮定する．平均リンク速度が 15 km/h の場合と 60 km/h の場合があったとき，我々は前者は渋滞流である確率が高いと推論し，後者は自由流である確率が高いと推論する．これは実際には以下の条件付き確率

$$p(k|l,v) \propto p(v|k,l)p(l|k)p(k) \tag{1}$$

を考えた上で，観測されているので $l = 1$ として，

$$p(k = \text{"渋滞"}|v = 15) > p(k = \text{"自由"}|v = 15)$$
$$p(k = \text{"自由"}|v = 60) > p(k = \text{"渋滞"}|v = 60)$$

と推論しているのに他ならない．この場合，**図–5** のグラフィカルモデルで示す確率モデルを考えていることは，ツイートの生起情報を利用することで，

$$p(k|l,v) \propto p(v|k,l)p(l|k)p(\boldsymbol{s}|k)p(k) \tag{2}$$

と更に，潜在的交通状態 k を推論するための条件 $p(\boldsymbol{s}|k)$ を追加していることになる．たとえば同じ渋滞であっても「事故」が含まれるツイートが多い場合と「積雪」が含まれるツイートが多い場合では潜在的交通状態を新たに判別可能だろう．このように，プローブのみでは判別できなかった (条件付き確率間に差を生み出せなかった) 交通状態を追加的なデータリソースによって判別可能となる．これが本研究の確率モデルの特徴である．

4. 変分近似による更新式の導出

本研究で提案する**図–5** の確率モデルのパラメータ推定には変分ベイズ法[9),10),11)]を用いる．これは求めたいパラメータの事後分布をいくつかの分布に分解することで近似する方法であり，EM アルゴリズムのベイズ的な拡張として解釈することができる．以降の導出はややテクニカルな式変形を含む．

(1)　モデルの仮定，条件付き分布，事前分布の設定

まず，本手法の仮定や事前分布の設定，条件付き分布の導出を行う．すべての確率変数の同時分布はグラフィカルモデルから以下のように分解される．

$$p(\boldsymbol{v},\boldsymbol{l},\boldsymbol{s},\ \boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\pi}) = p(\boldsymbol{v}|\boldsymbol{l},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\lambda})p(\boldsymbol{\mu}|\boldsymbol{\lambda})p(\boldsymbol{\lambda}) \\ p(\boldsymbol{l}|\boldsymbol{\phi},\boldsymbol{Z})p(\boldsymbol{\phi})p(\boldsymbol{s}|\boldsymbol{Z},\boldsymbol{\theta})p(\boldsymbol{\theta})p(\boldsymbol{Z}|\boldsymbol{\pi})p(\boldsymbol{\pi}) \quad (3)$$

ここで，事後分布の変分近似を考える．これが変分ベイズ法の仮定である．これは潜在変数に対して，パラメータとそれ以外の潜在変数が独立であるという仮定することに対応する．

$$q(\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\pi}) = q(\boldsymbol{Z})q(\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\pi}) \quad (4)$$

以降，事後分布は $q(\cdot)$ で表す．

条件付き分布をそれぞれ記述する．混合比 $\boldsymbol{\pi}$ が与えられたときの $\boldsymbol{Z}$ の条件付き分布は

$$p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}} \quad (5)$$

二項パラメータ $\boldsymbol{\phi}$ と潜在変数 $\boldsymbol{Z}$ が与えられたときの観測有無データ $\boldsymbol{l}$ の条件付き分布は

$$p(\boldsymbol{l}|\boldsymbol{\phi},\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\phi_{k1}^{l_{n1}z_{nk}}\cdot\phi_{k0}^{l_{n0}z_{nk}} \quad (6)$$

多項パラメータ $\boldsymbol{\theta}$ と潜在変数 $\boldsymbol{Z}$ が与えられたときのツイート状態データ $\boldsymbol{s}$ の条件付きデータは

$$p(\boldsymbol{s}|\boldsymbol{\theta},\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\frac{m_N!}{s_1!\cdots s_I!}\prod_{i=1}^{I}\theta_{ki}^{s_{ni}z_{nk}} \quad (7)$$

正規分布パラメータ $\boldsymbol{\mu},\boldsymbol{\lambda}$，観測有無データ $\boldsymbol{l}$，潜在変数 $\boldsymbol{Z}$ が与えられたときの速度データ $\boldsymbol{v}$ の条件付き分布は

$$p(\boldsymbol{v}|\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{l},\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}N(v_n|\mu_k,\lambda_k)^{l_{n1}z_{nk}} \quad (8)$$

である．

事前分布の設定を行う．混合比 $\boldsymbol{\pi}$ の事前分布はディリクレ分布を用いると，

$$p(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0)\prod_{k=1}^{K}\pi_k^{\alpha_0-1} \quad (9)$$

である．ここで，$\boldsymbol{\alpha}_0 = (\alpha_0,\ldots,\alpha_0)$ とする．観測有無の二項分布パラメータ $\boldsymbol{\phi}_k$ の事前分布はベータ分布を用いると，

$$p(\boldsymbol{\phi}_k) = Beta(\boldsymbol{\phi}_k|a_k,b_k) = \frac{\Gamma(a_k+b_k)}{\Gamma(a_k)\Gamma(b_k)}\phi_{k0}^{a_k-1}\phi_{k1}^{b_k-1} \quad (10)$$

$$p(\boldsymbol{\phi}) = \prod_{k=1}^{K}p(\boldsymbol{\phi}_k) \quad (11)$$

である．ツイート生成分布パラメータ $\boldsymbol{\theta}$ の事前分布はディリクレ分布を用いると，

$$p(\boldsymbol{\theta}_k) = Dir(\boldsymbol{\theta}_k|\boldsymbol{\alpha}_1) = C(\boldsymbol{\alpha}_1)\prod_{i=1}^{I}\theta_{ki}^{\alpha_1-1} \quad (12)$$

$$p(\boldsymbol{\theta}) = \prod_{k=1}^{K}p(\boldsymbol{\theta}_k) \quad (13)$$

である．正規分布パタメータ μ_k, λ_k の事前分布はガウス・ガンマ分布を用いると，

$$p(\mu_k|\lambda_k) = N(\mu_k|\mu_0,(\lambda_0\lambda_k)^{-1}) \quad (14)$$

$$p(\lambda_k) = Gam(\lambda_k|c_k,d_k) \quad (15)$$

である．以上が事前分布の設定である．

(2)　最適な事後分布の導出

潜在変数 $\boldsymbol{Z}$ の最適な事後分布の対数は観測データと潜在変数の同時分布の対数を考え，それぞれの期待値をとったものと等しい．そのため，

$$\begin{aligned}\ln q^*(\boldsymbol{Z}) &= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\pi}}\left[\ln p(\boldsymbol{v},\boldsymbol{l},\boldsymbol{s},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\lambda},\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\pi})\right] \\ &= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\lambda}}\left[\ln p(\boldsymbol{v}|\boldsymbol{l},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\lambda})\right] + \mathbb{E}_{\boldsymbol{\phi}}\left[\ln p(\boldsymbol{l}|\boldsymbol{Z},\boldsymbol{\phi})\right] \\ &\quad + \mathbb{E}_{\boldsymbol{\theta}}\left[\ln p(\boldsymbol{s}|\boldsymbol{Z},\boldsymbol{\theta})\right] + \mathbb{E}_{\boldsymbol{\pi}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})\right] + const\end{aligned} \quad (16)$$

が成り立つ．ここで，それぞれの項は依存関係がないので，個別に考えることができる．式 (8) より，

$$\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\lambda}}\left[\ln p(\boldsymbol{v}|\boldsymbol{l},\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\lambda})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}l_{n1}\left(\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\lambda}}\left[\ln N(v_n|\mu_k,\lambda_k)\right]\right) \quad (17)$$

式 (6) より，

$$\mathbb{E}_{\boldsymbol{\phi}}\left[\ln p(\boldsymbol{l}|\boldsymbol{Z},\boldsymbol{\phi})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}l_{n1}\mathbb{E}_{\boldsymbol{\phi}}\left[\ln\phi_{k1}\right] + z_{nk}l_{n0}\mathbb{E}_{\boldsymbol{\phi}}\left[\ln\phi_{k0}\right] \quad (18)$$

式 (7) より，

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\ln p(\boldsymbol{s}|\boldsymbol{Z},\boldsymbol{\theta})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\sum_{i=1}^{I}s_{ni}z_{nk}\mathbb{E}_{\boldsymbol{\theta}}\left[\ln\theta_{ki}\right] + \ln(m_N!) - \ln\prod_{i=1}^{I}s_i\right) \quad (19)$$

式 (5) より，

$$\mathbb{E}_{\boldsymbol{\pi}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})\right] = \sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\mathbb{E}_{\boldsymbol{\pi}}\left[\ln\pi_k\right] \quad (20)$$

である．以上より，事後分布

$$\begin{aligned}\ln q^*\left(\boldsymbol{Z}\right) &= \sum_{n=1}^{N}\sum_{k=1}^{K}z_{nk}\Big\{l_{n1}\left(\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\lambda}}\left[\ln N(v_n|\mu_k,\lambda_k)\right]\right) \\ &\quad + l_{n1}\mathbb{E}_{\boldsymbol{\phi}}\left[\ln\phi_{k1}\right] + l_{n0}\mathbb{E}_{\boldsymbol{\phi}}\left[\ln\phi_{k0}\right] \\ &\quad + \sum_{i=1}^{I}s_{ni}\mathbb{E}_{\boldsymbol{\theta}}\left[\ln\theta_{ki}\right] + \mathbb{E}_{\boldsymbol{\pi}}\left[\ln\pi_k\right]\Big\} + const\end{aligned}$$

が得られる．ここで，

$$\ln q^*(\boldsymbol{Z}) \equiv \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \rho_{nk} + const$$

とおくと，

$$q^*(\boldsymbol{Z}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}}$$

が得られる．ここで，

$$r_{nk} = \frac{\rho_{nk}}{\sum_{k=1}^{K}\rho_{nk}}$$

とおくと，潜在変数 $\boldsymbol{Z}$ の事後分布は

$$q^*(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} r_{nk}^{z_{nk}} \tag{21}$$

という多項分布として得られる．また，$\mathbb{E}[z_{nk}] = r_{nk}$ より，$N_k = \sum_{n=1}^{N} r_{nk}$ とおく．これは潜在変数 $Z = k$ であるデータの個数として解釈できる．

次に，パラメータの変分事後分布を考える．これは以下のように大きく 4 つのパートに分解できる．

$$\begin{aligned}
&\ln q^*(\boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\pi}) \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{v}, \boldsymbol{l}, \boldsymbol{s}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\lambda}, \boldsymbol{\phi}, \boldsymbol{\theta}, \boldsymbol{\pi})\right] + const \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{v}|\boldsymbol{l}, \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\lambda})\right] + \ln p(\boldsymbol{\mu}|\boldsymbol{\lambda}) + \ln p(\boldsymbol{\lambda}) \\
&+ \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{l}|\boldsymbol{\phi}, \boldsymbol{Z})\right] + \ln p(\boldsymbol{\phi}) \\
&+ \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})\right] + \ln p(\boldsymbol{\pi}) + const
\end{aligned} \tag{22}$$

まず，$\boldsymbol{\pi}$ に依存する項を考える．最適な事後分布の対数は式 (5), (9) より

$$\begin{aligned}
\ln q^*(\boldsymbol{\pi}) &= \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{\pi})\right] + \ln p(\boldsymbol{\pi}) + const \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \ln \pi_k + (\alpha_0 - 1)\sum_{k=1}^{K} \ln \pi_k + const \\
&= \sum_{k=1}^{K} (N_k + \alpha_0 - 1) \ln \pi_k + const
\end{aligned}$$

として得られる．ここで，両辺の指数を取ると，$\boldsymbol{\pi}$ の事後分布は

$$q^*(\boldsymbol{\pi}) \propto Dir(\boldsymbol{\pi}|\boldsymbol{\alpha}) \tag{23}$$

というディリクレ分布として得られる．ここで，$\boldsymbol{\alpha} \in \mathbb{R}^K$ の各要素は $\alpha_k = \alpha_0 + N_k$ である．

同様のアプローチにより，$\boldsymbol{\theta}$ に依存する項から，$\boldsymbol{\theta}_k$ の事後分布は

$$q^*(\boldsymbol{\theta}_k) \propto Dir(\boldsymbol{\theta}_k|\boldsymbol{\beta}) \tag{24}$$

というディリクレ分布として得られる．ここで，$\boldsymbol{\beta} \in \mathbb{R}^I$ の各要素は $\beta_i = \alpha_1 + N_{ki}$ である．$N_{ki} \equiv \sum_{n=1}^{N} r_{nk} s_{ni}$ は潜在変数が k であるときの状態 i のツイート総数として解釈できる．

$\boldsymbol{\phi}$ に依存する項から，$\boldsymbol{\phi}_k$ の事後分布は

$$q^*(\boldsymbol{\phi}_k) \propto Beta(\boldsymbol{\phi}_k|a_k + N_{k0}, b_k + N_{k1}) \tag{25}$$

というベータ分布が得られる．ここで，$\sum_{n=1}^{N} r_{nk} l_{n0} \equiv N_{k0}$, $\sum_{n=1}^{N} r_{nk} l_{n1} \equiv N_{k1}$ であり N_{k0}, N_{k1} は潜在変数が k で未観測 ($l = 0$) のデータ個数，観測 ($l = 1$) のデータ個数として解釈できる．

最後に $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$ に依存する項から最適な変分事後分布を導出する．ここで，以下は必ず成り立つ．

$$q(\boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_{k=1}^{K} q(\mu_k, \lambda_k) \tag{26}$$

$$q^*(\mu_k, \lambda_k) = q^*(\mu_k|\lambda_k) q^*(\lambda_k) \tag{27}$$

μ_k, λ_k の事後分布は式 (14), (15) より，

$$\begin{aligned}
&\ln q^*(\mu_k, \lambda_k) \\
&= \ln N(\mu_k|\frac{N_{k1}\bar{v}_{nk} + \lambda_0\mu_0}{\lambda_0 + N_{k1}}, \{(\lambda_0 + N_{k1})\lambda_k\}^{-1}) \\
&+ \ln Gam(\lambda_k|c_k + \frac{N_{k1}+1}{2}, d_k + \frac{N_{k1}\lambda_0(\bar{v}_{nk} - \mu_0)^2}{2(N_{k1} + \lambda_0)}) \\
&+ const
\end{aligned} \tag{28}$$

となる．ここで，$\bar{v}_{nk} \equiv \frac{\sum_{n=1}^{N} z_{nk} l_{n1} v_n}{N_{k1}}$ である．以上より，λ_k が与えられたもとでの μ_k の条件付き事後分布は平均 $\hat{\mu}_k \equiv \frac{N_{k1}\bar{v}_{nk} + \lambda_0\mu_0}{\lambda_0 + N_{k1}}$，精度 $\hat{\lambda}_k \equiv (\lambda_0 + N_{k1})\lambda_k$ の正規分布である．また λ_k の事後分布は

$$\begin{aligned}
&q^*(\lambda_k) \\
&= Gam(\lambda_k|c_k + \frac{N_{k1}+1}{2}, d_k + \frac{N_{k1}\lambda_0(\bar{v}_{nk} - \mu_0)^2}{2(N_{k1} + \lambda_0)})
\end{aligned} \tag{29}$$

というガンマ分布である．

最後に，μ_k の事後分布を求めるために $q^*(\mu_k|\lambda_k)$ を λ_k に対して周辺化する．一次元正規分布を精度に対して積分消去したものはスチューデントの t 分布として知られている．ここで，自由度 $\nu_k \equiv \frac{\hat{c}_k}{(\lambda_0 + N_{k1})\hat{d}_k}$, 精度 $\lambda'_k \equiv 2\hat{c}_k$ と新たにパラメータを定義すると，μ_k の事後分布は

$$q^*(\mu_k) = St(\mu_k|\hat{\mu}_k, \lambda'_k, \nu_k) \tag{30}$$

のスチューデントの t 分布となる．

(3) パラメータの変分事後分布の更新

これまで導出した各パラメータの変分事後分布を用いてパラメータの更新を行う．その際に $\mathbb{E}[z_{nk}] = r_{nk}$ が必要となる．これは ρ_{nk} を正規化すれば得られる．ρ_{nk} の定義を再掲すると，

$$\begin{aligned}
\ln \rho_{nk} = &l_{n1}\left(\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\lambda}}\left[\ln N(v_n|\mu_k, \lambda_k)\right]\right) \\
&+ l_{n1}\mathbb{E}_{\boldsymbol{\phi}}\left[\ln \phi_{k1}\right] + l_{n0}\mathbb{E}_{\boldsymbol{\phi}}\left[\ln \phi_{k0}\right] \\
&+ \sum_{i=1}^{I} s_{ni}\mathbb{E}_{\boldsymbol{\theta}}\left[\ln \theta_{ki}\right] + \mathbb{E}_{\boldsymbol{\pi}}\left[\ln \pi_k\right]
\end{aligned} \tag{31}$$

であるので，それぞれの項の計算を行う．

式 (23) より，ディリクレ分布の対数期待値は

$$\mathbb{E}[\ln \pi_k] = \psi(\alpha_0 + N_k) - \psi(K \cdot \alpha + N) \tag{32}$$

である．ここで $\psi(\cdot)$ はディガンマ関数である．同様に

式 (24) より，

$$\mathbb{E}[\ln \theta_{ki}] = \psi(\alpha_1 + N_{ki}) - \psi(I \cdot \alpha_1 + \sum_{i=1}^{I} N_{ki}) \quad (33)$$

が得られる．ベータ分布に対しても同様に式 (25) より，

$$\mathbb{E}[\ln \phi_{k1}] = \psi(a_k + N_{k0}) - \psi(a_k + N_{k0} + b_k + N_{k1})$$
$$\mathbb{E}[\ln \phi_{k0}] = \psi(b_k + N_{k1}) - \psi(a_k + N_{k0} + b_k + N_{k1})$$

が得られる．最後に，μ_k, λ_k に関して

$$\begin{aligned}\mathbb{E}_{\mu_k,\lambda_k}\ & [\ln N(v_n|\mu_k,\lambda_k)] \\ &= \frac{1}{2}\left\{\psi(c_k + \frac{N_{k1}+1}{2}) \right. \\ &\quad - \ln\left(d_k + \frac{N_{k1}\lambda_0(\bar{v}_{nk}-\mu_0)^2}{2(N_{k1}+\lambda_0)}\right) \\ &\quad \left. - \frac{1}{(\lambda_0+N_{k1})} - \frac{(\hat{\mu}_k - v_n)^2 \hat{c}_k}{\hat{d}_k}\right\} + const \quad (34)\end{aligned}$$

が得られる．以上より，$\ln \rho_{nk}$ が求まるので，これを用いて r_{nk} が求まる．この r_{nk} を用いて，潜在変数の更新と各パラメータの更新を順次行う．

5. モデルの挙動分析

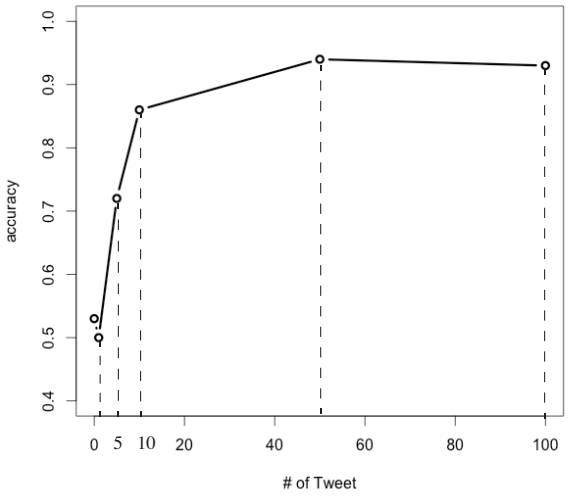
次に，本研究の生成モデルに基づきデータを生成し，本研究の推定手法による推論精度の検証を行う．特に，プローブのみが観測データである場合に比べて，ツイートの存在が精度向上に与える影響についての分析を行う．

シミュレーションの設定を**表–2** に示す．潜在的交通状態として 10 種類を想定し，表に示すパラメータに基づき，プローブ観測有無，平均リンク速度，10 種のツイート観測数のデータを生成する．各データの組を 1000 個生成し，データを生成したクラス (潜在的交通状態) とモデルの推論したクラス (交通状態) の一致度を計算した．このような手続きを 50 回繰り返し，生成クラスと推論クラスの一致度の平均を示す．

図–6 に各リンクに対するツイート数と平均的な推論精度の結果を示す．ツイートがない場合 (ツイート数 0 の場合)，生成クラスと推論クラスの一致度は 0.5 程度であり，プローブ速度のみでデータを生成した潜在的交通状態を正確に判別することは難しいことがわかる．しかし，ツイート数が 5 個の場合は精度は 0.7，10 個の場合は精度が 0.85 となり，ツイート数の増加に応じて推論精度が向上していることがわかる．このように，少量のツイートが存在するだけで，潜在的な交通状態の推論精度が大幅に向上することがシミュレーションデータによって示された．

表–2 シミュレーションパラメータ設定

パラメータ	1	2	3	4	5	6	7	8	9	10
生成確率 π	0.05	0.05	0.1	0.1	0.1	0.1	0.1	0.15	0.15	0.1
速度平均 μ	0	10	22	30	35	40	48	55	62	70
分散逆数 λ	100	1	0.5	0.5	0.5	0.3	0.2	0.1	0.05	0.05
観測確率 ϕ	0	0.1	0.2	0.3	0.5	0.5	0.6	0.7	0.8	0.9
ツイート θ_1	0	0	0.01	0	0	0	0	0	0	0
ツイート θ_2	0	0.01	0.28	0	0.08	0.43	0	0.17	0.02	0.14
ツイート θ_3	0.18	0.04	0.13	0.03	0.05	0.14	0.01	0	0.54	0.34
ツイート θ_4	0.21	0.02	0	0.08	0.09	0.05	0	0.25	0.08	0
ツイート θ_5	0.11	0.14	0	0.02	0.02	0.02	0.59	0.01	0.07	0.02
ツイート θ_6	0.05	0.22	0.32	0.52	0.51	0.03	0.06	0.03	0.14	0
ツイート θ_7	0	0.01	0	0	0	0.26	0.02	0.19	0	0
ツイート θ_8	0.43	0.14	0.13	0.21	0.03	0.02	0.20	0.11	0.03	0.48
ツイート θ_9	0	0.40	0.1	0.11	0.21	0.03	0.01	0.14	0.10	0
ツイート θ_{10}	0	0	0	0	0	0	0	0.08	0	0



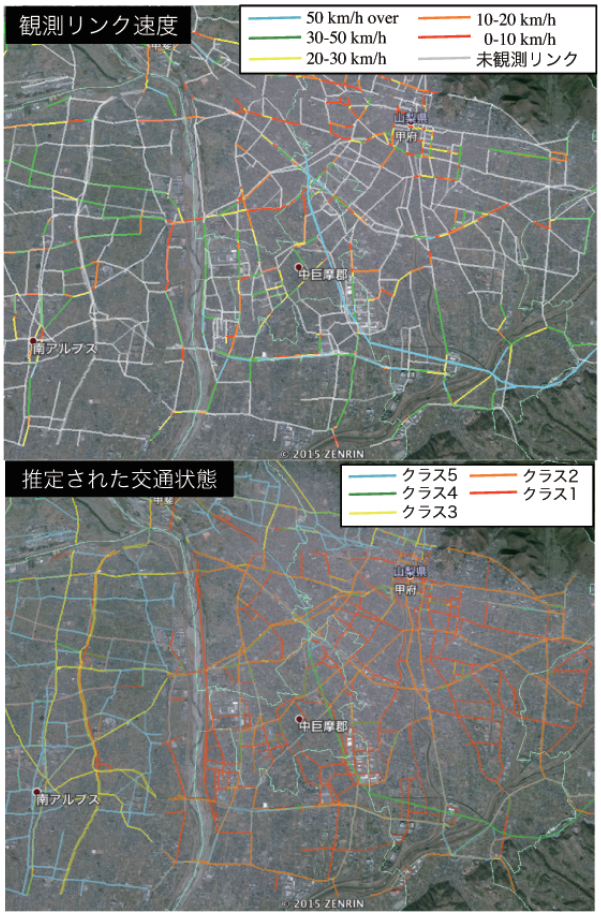
図–6 推論精度とツイート数の関係

6. 実データによるケーススタディ

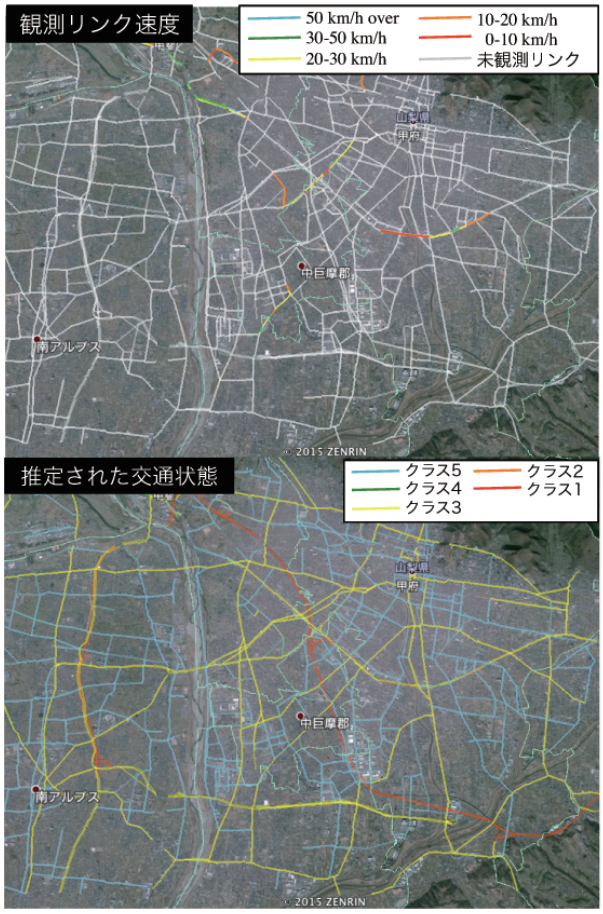
(1) 対象期間・地域とデータ概要

シミュレーションデータに対する検証結果から，本手法を用いることで，潜在的交通状態の推論が可能であることを確認した．最後に，実際のプローブデータとツイートデータを用いた提案手法のケーススタディを行う．ケースとして，2014 年甲信越地方における豪雪災害時の交通状態推定を取り上げる．対象期間は 2014 年 2 月 14 日から 16 日の 3 日間，対象地域は甲府市周辺の 20km 四方のエリアである．

利用するデータは本田技研工業から提供された対象エリア内のプローブデータと東北大学乾・岡崎研究室から提供された 2014 年 2 月の日本語ツイート約 9 億 8000 万件である．マップマッチング後のプローブデータは 3564 トリップ，交通情報・地理情報抽出後のツイートデータは 4536 ツイートである．



図–7 2014 年 2 月 14 日 (金) 8:00-8:15 の甲府周辺



図–8 2014 年 2 月 16 日 (日) 8:00-8:15 の甲府周辺

(2) 推論結果

提案手法により，各道路リンクの 15 分単位の交通状態推定を行う．潜在的交通状態の数は予め 5 つとして設定した．本研究で用いたデータは 2 月 14 日から 16 日の 3 日間という短い期間のため，以下の結果は現時点ではモデルとデータの動作確認の意味合いが強い．

モデルによる推定結果では 5 つの潜在交通状態のクラスはそれぞれ平均リンク速度やツイート傾向が異なるクラスとして推定された．たとえば高速道路の場合，平均リンク速度が 81km/h, 48.1 km/h, 31.3 km/h, 23.3 km/h, 9.1 km/h の 5 つのクラスが推定されている．速度が高いものは自由流状態の交通状態を表していると考えられるが，遅い速度のクラスはそれぞれ自然渋滞や積雪による速度低下，事故渋滞，通行止めなどを表していると考えられる．これらのクラスの考察は，速度のみならずツイートの傾向などを見ることで解釈することが必要である．

結果の例として，まだ積雪がそれほどひどくなかった 2014 年 2 月 14 日 (金) の午前 8 時の様子と積雪が 1m 以上となっている 2 月 16 日 (日) の午前 8 時の様子を示す．**図–7** は 2 月 14 日の様子であり，上図はプローブのみのリンク速度の表示 (白色は未観測リンク)，下図は推定された交通状態ごとに色で表示した結果である．**図–8** は 2 月 16 日の結果を同様に示している．本手法の特徴として，プローブが観測されていない道路リンクであっても，ツイートの傾向から地域全体の交通状態の様子を推論することができる．たとえば，同じプローブ未観測リンクであっても，14 日の未観測リンクと 16 日の未観測リンクでは異なる交通状態として推定されていることがわかる．このように，マルチリソースデータを用いることによって，1 つのデータリソースの欠損を補間したり，より詳細な交通状態を把握することができることが示された．

7. おわりに

本研究の成果をまとめる．まず，Twitter データから交通状態や地理状態を抽出するための交通情報タグ付与器とジオパーズエンジンを開発した．次に，これらのデータを統一的に扱う確率的生成モデルを構築し，変分ベイズ法によってモデルの学習を行った．次に，シミュレーションデータから 1 つの道路リンクの交通状態を正しく推論するには 5 から 10 個程度のツイートでも大きな情報量を持つことを示した．最後に甲信越豪雪を対象に，実際のプローブデータとツイートデータを用いて，潜在的交通状態の推論を行った．

本研究の手法に関する今後の課題を示す．まず，一点

目として，交通情報タグ付与器とジオパーズエンジンの更なる精度の向上が必要である．二点目として，提案する確率的生成モデルは道路リンクごとに i.i.d. として仮定したモデリングをしているが，現実的には時空間的な相関関係が存在する．三点目として，潜在的交通状態の各クラスの解釈方法について，簡易的な方法を考案する必要がある．四点目として，より長期間の実観測データを用いて提案手法の有用性を確認する必要がある．

謝辞:
本研究は文部科学省 委託事業「実世界ビッグデータの利活用のための高性能データ融合解析技術の研究開発」，情報通信研究機構 高度通信・放送研究開発委託研究「ソーシャル・ビッグデータ利活用・基盤技術に関する研究開発」の助成を受けたものです．また，本研究を進めるにあたって東北大学 乾健太郎教授，岡崎直観准教授からは Twitter データの提供と有益なコメントを頂きました．記して感謝致します．

参考文献

1) Mehran, B., Kuwahara, M. and Naznin, F.: Implementing kinematic wave theory to reconstruct vehicle trajectories from fixed and probe sensor data, *Transportation Research Part B*, Vol.20, Issue 1, pp.144–163, 2012.
2) Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, *Proc. 19th International Conference on World Wide Web (WWW'10)*, pp.851–860, 2010.
3) Lee, R. and Sumiya, K.: Measuring geographical regularities of crowd behaviors for Twitter-based geosocial event detection, *Proc. 2nd ACM SIGSPATIAL Inter- national Workshop on Location Based Social Networks (LBSN ' 10)*, pp.1–10, 2010.
4) Lee, R., Wakamiya, S. and Sumiya, K.: Discovery of Unusual Regional Social Activities using Geo-tagged Microblogs, *World Wide Web (WWW) Special Issue on Mobile Services on the Web*, Vol.14, No.4, pp.321–349, 2011.
5) 山本修平・佐藤哲司: 実生活 Tweet に対する局面の階層的推定法, 第 6 回データ工学と情報マネジメントに関するフォーラム, 2014.
6) 土屋圭・豊田正史．喜連川優: マイクロブログを用いた鉄道の運行トラブル発生期間および付帯情報の抽出, 第 6 回データ工学と情報マネジメントに関するフォーラム, 2014.
7) Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. and Lin, C.-J.: LIBLINEAR: A library for large linear classification, *Journal of Machine Learning Research*, Vol. 9, pp.1871–1874, 2008.
8) Leidner, J.L.: Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names, Boca Raton, FL, USA, Universal Press, 2008.
9) Jaakkola, T. and Jordan, M.I.: Bayesian parameter estimation via variational methods, *Statistics and Computing*, Vol.10, pp.25–37, 2000.
10) Jakkola, T.: Tutorial on variational approximation methods, In Opper, M. and Saad, D. (eds), *Advances in Mean Field Methods*, pp.129–159, MIT Press, 2001.
11) MacKay, D.J.C.: *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

(2015. 4. 24 受付)

LATENT TRAFFIC STATE ESTIMATION USING PROBE DATA AND TWEET DATA

Yusuke HARA, Koji MATSUDA, Yosuke KAWASAKI, Takuma MITANI, Masao KUWAHARA