Vol.2018-UBI-58 No.5
2018/5/19

バス運行状況ウェブサービス情報を用いた乗換案内アプリのための路線バス所要時間推定

綿貫 圭太[1,a)]　下坂 正倫[1,b)]

概要：近年，スマートフォンの普及に伴い，乗換案内アプリケーションの使用機会は増大している．乗換案内アプリケーションは路線バスを対象とするものも多いが，得られる所要時間はダイヤをそのまま使用したものであり，実際の運転状況を反映していない場合がある．本研究では，乗換案内アプリケーションへの応用を目的に，1 日以上先の路線バス所要時間に対して高精度な推定が可能な手法を構築する．長期の所要時間推定においては，直前の路線バス所要時間を推定に用いることができないため，曜日や時間帯といった要因から所要時間を推定することが必要となる．これを踏まえて，本研究では時間帯，曜日，天気などの説明変数から路線バス所要時間の確率分布を求める手法として低ランク双線形ガンマ回帰を提案する．既存の点推定を用いたモデルやポアソン回帰を用いたモデルが抱えていた運行時間の遅延に応じた分散の大きさを考慮できない問題を解決するものであり，アプリケーションへの応用を考えた場合に有益である．推定結果に応じた分散を考慮しているため，例えば「90％の確率で成功する乗換」をより頑健に導出することが可能である．この手法を 5 ヶ月程度の路線バス所要時間の実データを用いて，既存のバス所要時間推定にて使用されている手法である Random Forest と比較を行い，提案手法の有効性を検証する．

キーワード：路線バス所要時間予測，一般化線形モデル，低ランク双線形モデル，確率密度推定

Forecasting urban bus travel time for transit service using web-based bus location service data

Abstract: For recent years, opportunities for using transit applications are increasing with spread of smartphones. Many of transit applications provides bus transition based on bus diagram as well as train transition. However, due to the large difference between its actual travel time and the diagram, it causes low reliability of applications. In this paper, we deal with high-precision long-term bus travel time forecasting for more than 1 day ahead, that is suitable for transit applications. In this study, we propose a low-rank bilinear gamma regression to predict the probability of bus travel time from limited number of explanatory variables such as time of a day, week of days and weather conditions. Compared with the state-of-the-art methods on bus travel time prediction, our method provides both point estimation and probability density itself. Experimental results using web-based bus location service data spanning over 5 months show that our method performs well compared with the state-of-the-art methods including Random Forest.

Keywords: Bus travel time forecasting, Generalized linear model, Low rank bilinear model, Probability density estimation

1. 序論

近年，スマートフォンが普及するにつれて，公共交通機関の乗換案内アプリケーションが使用される機会は増加している．乗換案内により，利用者は外出先でも自分が乗るべき交通機関を適切に選択し，目的地に向かうことが可能である．乗換案内では，出発地と到着地の選び方によっては路線バスを利用するルートが提案される場合もある．現在日本で主流である乗換案内サービスの多くは路線バス会社が公表しているダイヤの所要時間をそのまま案内に使用

[1] 東京工業大学 情報理工学院 情報工学系
Department of Computer Science, School of Engineering, Tokyo Institute of Technology
[a)] watanuki@miubiq.cs.titech.ac.jp
[b)] simosaka@miubiq.cs.titech.ac.jp

している．しかし，一般に路線バスの所要時間は通行する道路の渋滞や乗り降りする乗客の数に影響を受けやすく，特に朝夕のラッシュ時において，ダイヤ通りの運行が行われることは少ない．従って，過去の運行実績などのデータを用いて路線バスの所要時間を推定し，それを乗換案内アプリケーションに表示することは，路線バスの利用者にとって有効である．

国土交通省が行ったアンケート[*1]によれば，路線バスの所要時間を得られた場合の効果には以下のものが挙げられる．

- 路線バスの利用客が目的地への行程をより細かく計画することが可能となる．路線バスから電車への乗り継ぎは，路線バスの所要時間に大きく影響を受けるため，これを正確に見積もることで，行程全体の所要時間を求めることが可能である．
- 見積もられた所要時間を乗車前あるいは乗車中に知ることによって，乗客の可処分時間を増加させることができ，ストレスを和らげることが可能である．

所要時間推定を乗換案内アプリケーションに応用するために推奨されることは以下である．

正確に推定できること　推定の精度は高いことが望ましい．精度の低い推定は，かえってアプリの利用者に混乱を与えてしまう．既存の路線案内のうち多くのもので用いられているダイヤを直接用いる手法は，実態より所要時間を短く見積もる傾向があり，この点において問題が指摘されうる．

長期での推定が可能であること　乗換案内アプリケーションには，1 ヶ月先といった長期の推定結果を表示する機能が存在し，そういった要求にも問題なく応答する必要がある．直前のバスの所要時間を説明変数に含むようなモデルは多く存在している [1], [2], [3], [4] が，こういったモデルは本研究の目的には合致しない．以降本論文では，「短期」を「すでに一部運行データが得られている日の推定」，「長期」を「運行データが得られている日の翌日以降の推定」と定義して使用する．

区間推定が可能であること　長期での推定においては，利用できる説明変数は限られているため，点推定による高精度での推定は難しい．また，交通機関の利用者は所要時間の分散や信頼区間に関して敏感であることが知られており [5]，これらを推定可能である手法が望ましい．既存の路線バス長期推定手法 [6], [7], [8] においては，説明変数に応じて分散を変更するような推定を行う手法は提案されなかった．本研究はこの問題を解決するモデルを提案する．

以上より，本論文では長期での推定が可能な過去の路線バス運行実績を利用した所要時間の信頼区間推定の手法に関して考察を行い，これが可能な手法を提案する．

本論文の貢献は以下にまとめられる．

- 既存の長期所要時間推定では行われていない，分散の大きさを考慮した所要時間の区間推定を可能とする低ランク双線形ガンマ回帰モデルを提案した．
- 平均と分散の関係による重み付き最小二乗法を用いた，離散な説明変数を持つ線形あるいは双線形のガンマ回帰モデルにおける Shape パラメータの推定方法を提案し，これを用いて最適な Shape パラメータの範囲を絞り込むことができることを示した．
- バス運行情報を表示するウェブサービスである「東急バスナビ」から，ソフトウェアを用いて取得した約 5 ヶ月間の路線バスデータを用いた実験によって，提案手法の点推定の性能が，絶対誤差の観点で既存手法である Random Forest に引けをとらず，さらに，単純に過去のデータを用いた手法より優れた区間推定を可能とすることを示した．

本論文の構成を述べる．2 節では関連研究に関して述べ，3 節では本研究にて取り組んだ問題の定式化とその問題に対する既存のアプローチを紹介する．4 節では本研究が提案する手法に関して議論を行い，その詳細に関して説明する．5 節では提案手法の有効性を示すために行う実験内容とその結果に関して述べる．最後に，6 節で本論文が行ったこととその貢献についてまとめを行う．

2. 関連研究

2.1 路線バス所要時間の定式化・分析

実際の運転行動に関して，遅延が発生した場合の回復運転や，早着した場合の時間調整を考量して定式化することでバスの所要時間を求める研究 [9] や，路線バスの運行の所要時間を，信号の数，乗客の数等の要因ごとに分析した研究 [10] が存在する．多くの研究において，一般の自動車と路線バスの違いとしてバス停留所での乗客乗り降りに伴う停留時間（Dwell time）が挙げられている [1], [9], [10], [11].

2.2 道路交通における所要時間の推定

所要時間推定に関しては，短期のものを中心に長期のものまで幅広く，その対象も路線バスに集中したものから一般の道路交通に至るまで存在する．いずれの研究でも，運行の曜日，時間帯，天気が説明変数としてよく用いられる．

路線バス所要時間の短期推定

Artificial Newral Network の誤差逆伝搬法を用いたもの [1]，k-近傍回帰アルゴリズムを用いたもの [2] や，加法モデルを用いたもの [3]，カルマンフィルタを用いたもの [4] が存在する．いずれの研究も，直前のバスの所要時間や，推定するバスのそれまでの運行実績との関連性に着目し，その値を説明変数として組み込むことにより，当日における推定精度を向上させることに成功している一方

[*1] http://www.mlit.go.jp/jidosha/busloca/01_houkokusyo.pdf

で，長期の推定は実現できないという問題を抱えている．

一般の道路交通の短期推定

個別の自動車ではなく道路に着目しており，直前に同じ道路を走った自動車の運行実績を説明変数に用いることで，精度を向上させるというアプローチをとるものは，路線バス案内にも利用することが可能である．このアプローチを用いている研究は数多く存在しており [12], [13], [14]，いずれの例でも，何分先を推定するのかを説明変数の形で与えている．従って，路線バス所要時間の短期推定に関する研究とは異なり，入力する過去の運転実績と推定する時間帯を独立に指定することが可能である．とはいえ，このようなモデルには 15 分以上の推定を難しいとするもの [13] や 2 時間以上の推定を難しいとするもの [14] が含まれており，本研究が目指す翌日以降の推定に適さないと考えられる．

本研究は，直前のバスの所要時間を用いることなく推定を行うことにより，長期の推定が難しいという問題を解決したモデルを提案する．

3. 問題設定と既存手法の課題

3.1 長期的な路線バス所要時間推定の定式化

本論文では，長期的な路線バス所要時間の点推定と区間推定という 2 つの問題を考える．長期的な推定に用いることのできる説明変数の定義として，以下を使用する．

出発時間 t_h 時 t_m 分 t_s 秒の場合は

$$\phi_t = t_h + \frac{t_m}{60} + \frac{t_s}{3600} \in [0, 24), \qquad (1)$$

と定義する．

曜日 出発した日の曜日に応じて，$\phi_d \in \{0, \ldots, 6\}$ とする．ただし，日曜日を $\phi_d = 0$，土曜日を $\phi_d = 6$ として，順序が変わらないよう対応づける．

国民の休日 真偽値として，$\phi_h \in \{0, 1\}$ とする．

天気 気象庁が公開している CSV データ[*2] にて用いられている表記[*3] をそのまま用い，$\phi_w \in \{1, 2, \ldots, 19, 22, 23, 24, 28, 101\}$ とする．

これらの説明変数をまとめた $\boldsymbol{x}$ を (2) で定義する．

$$\boldsymbol{x} = (\phi_t, \phi_d, \phi_h, \phi_w). \qquad (2)$$

目的変数である所要時間 y との組 $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$ を訓練データとして用い，モデルを構成する．

3.2 路線バス長期所要時間推定の既存研究におけるアプローチと問題点

長期の所要時間の点推定として，Mendes ら [6], [7], [8] は，本研究にて用いる出発時刻，曜日，天気，国民の休日などの情報に加えて，風速，気温，降水量，学校の休み時間，バスの運転手等の要因を説明変数として追加し，Random Forest, SVM, 射影追跡回帰を適用し比較を行い [7]，その後それらをアンサンブル学習により組み合わせた手法を提案した [8]．また，RReliefF を用いた特徴選択を行うことで，これらの変数が所要時間に与える影響を示した [6]．

しかしながら，彼らによって提案されたいずれの手法でも，分散を考慮した区間推定を行うことはできない．

長期路線バス推定問題において分散を考慮する必要性を述べる．交通機関の利用者は所要時間が大きな分散を持つことに敏感であり [5]，また，路線バスの所要時間の分散は他の公共交通機関に比べて大きいことが知られている．この問題は，路線バス利用者が乗換案内アプリ等を用いて所要時間の分散や信頼区間を前もって知っておくことで軽減することが可能である．一般線形モデルやその他の点推定を行うモデルにおいても，残差が従う分布を固定された分散を持つ正規分布と仮定することによって信頼区間を求めることが可能であるが，一般に，平均値と分散が関係を持つ，右に裾の長い分布となることが知られている路線バスの所要時間分布 [15], [16] に正規分布を当てはめることは適当でない．同様に，損失関数についても考察する．一般に，推定された所要時間より遅く到着する現象は，早く到着する現象より重大であるといえる．このため，平方損失や絶対損失のような，正負の方向に対称な損失関数を仮定することの多い点推定は，乗換案内アプリケーションへの応用を目的とした，路線バス所要時間推定問題に対して不十分であると考えられる．損失関数に非対称な物を仮定するというアプローチは，分位点回帰 [17] により行われており，特定の分位点を回帰することに対応する．この手法で信頼区間を求めることができるため，一種の非対称な損失関数を用いることは区間推定と同一視できる．

以上より，分散の変化に関して考慮せず点推定のみを行う手法よりも，期待値に応じた分散を考慮した区間推定も可能である手法を用いることで，事前に利用者が信頼区間を高い精度で知ることができ，望ましいといえる．乗換案内アプリへの応用を考えると，例えば，“24 分” という点推定によって得られる結果に合わせて，信頼区間を用いた “20–30 分” という情報や，所要時間を過小に見積もることを避けた “90 %の確率で 28 分未満” という情報を得られる方が有益である．そこで，次節では点推定のみでなく，長期的な路線バス所要時間の頑健な区間推定を行うモデルを提案する．点推定と区間推定の評価に用いる指標は 5.3 節にて述べる．

4. 提案手法 : 低ランク双線形ガンマ回帰モデル

本研究は，バスの所要時間の特性や，その分散が乗客に与える影響を考慮した上で，所要時間の点推定だけでなく，分散を考慮した高精度な区間推定を目指すものである．高精度の区間推定を実現するためには，確率分布

[*2] http://www.data.jma.go.jp/gmd/risk/obsdl/index.php
[*3] http://www.data.jma.go.jp/gmd/risk/obsdl/top/help3.html

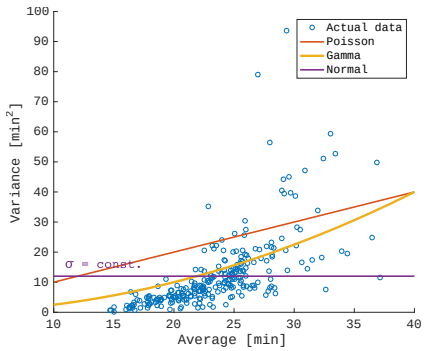

図 1 バス所要時間における平均と分散の関係の例

そのものを求めることが考えられる．確率分布推定の手法として，固定した説明変数に対して目的変数が従う分布を正規分布と仮定した一般線形モデルや，指数族分布に一般化した一般化線形回帰モデルが代表的である．一般化線形回帰の文脈では，正規分布は分散が平均値に対して不変である分布，ガンマ分布は定数 $k > 0$ と平均値 $\mathbb{E}[X]$ より分散 $\mathrm{Var}(X) = k\mathbb{E}[X]^2$ を持つ分布，ポアソン分布は $\mathrm{Var}(X) = \mathbb{E}[X]$ の関係を持つ分布として特徴付けられる [18]．**図 1** は，5.2.1 節にて示す路線バス所要時間データを，3.1 節にて示した説明変数に (10) の離散化を施してそれぞれ分割し，平均と分散を散布図としてプロットしたものである．この図より，分散が平均値に対して不変であるという一般線形モデルの仮定 [18] は，路線バスの所要時間データには適していないことがわかる．同様に，実際のデータはポアソン分布よりもガンマ分布によく従っているため，一般化線形ガンマ回帰を適用することが適当であると考えられる．そのほか，路線バスの所要時間が従う分布に関する研究 [15], [16] においても，指数分布族以外の分布として Burr XII 分布，指数分布族としてガンマ分布が挙げられている．

双線形回帰の手法を用いた場合，パラメータ数がベクトルから行列の次元に増加し，過学習のおそれがある．Shimosaka ら [19] はパラメータの行列を低ランク近似することで過剰適合を抑制する，低ランク双線形ポアソン回帰モデルを提案した．このモデルを適用することにより，今回の問題においても，Shimosaka らが取り組んだ都市動態分析の問題と同様に，過剰適合を抑制することが期待できる．

本節では低ランク双線形ポアソン回帰を元にした手法である，低ランク双線形ガンマ回帰を提案する．また，離散な説明変数に対して実データの平均と分散の関係を用いることにより，ガンマ回帰における Shape パラメータの推定を行う手法を提案する．

4.1 一般化線形ガンマ回帰モデル

提案手法である低ランク双線形ガンマ回帰モデルは，一般化線形モデルを二次形式に拡張し，パラメータ行列を低ランク化することにより定義される．

ガンマ分布は Shape パラメータ $\alpha > 0$，Scale パラメータ $\beta > 0$ を母数に持ち，確率密度関数が (3) で表される分布である．

$$\mathrm{Gam}(y|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} \exp\left(-\frac{y}{\alpha}\right). \tag{3}$$

一般化線形モデルをガンマ分布に適用するには，Shape パラメータをハイパーパラメータ $\alpha = \alpha_0$ として固定し，Scale パラメータ β を，非線形変換 $\boldsymbol{\varphi}$ を用いて説明変数 $\boldsymbol{z} = \boldsymbol{\varphi}(\boldsymbol{x})$ に対して変動させることによって実現できる．そして，リンク関数に対数関数 $\ln(\cdot)$ を用いれば，(4) が得られる．ただし，パラメータベクトルを $\boldsymbol{m}$ としている．

$$\ln(\alpha_0\beta) = \boldsymbol{m}^{\top}\boldsymbol{z}. \tag{4}$$

4.2 低ランク双線形ガンマ回帰モデルの定義

本研究では，(2) にて定義した説明変数 $\boldsymbol{x}$ を用いて推定を行うが，今回のモデルにて扱いやすくするため，5.4 節に示す非線形変換 $\boldsymbol{\varphi}_d, \boldsymbol{\varphi}_s$ を行った $\boldsymbol{d} = \boldsymbol{\varphi}_d(\boldsymbol{x}) \in \mathbb{R}^p, \boldsymbol{s} = \boldsymbol{\varphi}_s(\boldsymbol{x}) \in \mathbb{R}^q$ を用いる．

双線形回帰の枠組みでは，パラメータ行列 $\boldsymbol{W} = \mathbb{R}^{p\times q}$ を用いて，(5) のように表現される [19]．

$$\ln(\alpha_0\beta) = \boldsymbol{d}^{\top}\boldsymbol{W}\boldsymbol{s}. \tag{5}$$

(5) において，行列 $\boldsymbol{U} \in \mathbb{R}^{p\times k}, \boldsymbol{V} \in \mathbb{R}^{q\times k}$ を用い，$\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^{\top}$ とおくことで低ランク近似を行うことができる．これにより，パラメータ数の増加を抑制し，過学習を抑制することができる．$k \in \mathbb{N}$ はハイパーパラメータとなる．

4.3 低ランク双線形ガンマ回帰モデルの学習

低ランク双線形ガンマモデルにおける対数尤度は，(6) によって与えられる．

$$\begin{aligned}\ln L(\boldsymbol{U},\boldsymbol{V}) = \sum_{i=1}^{N}\Big[&\alpha_0\ln\alpha_0 + (\alpha_0-1)\ln y_n \\ &- y_n\alpha_0\exp(-\boldsymbol{d}_n{}^{\top}\boldsymbol{U}\boldsymbol{V}^{\top}\boldsymbol{s}_n) \\ &- \ln\Gamma(\alpha_0) - \alpha_0\boldsymbol{d}_n{}^{\top}\boldsymbol{U}\boldsymbol{V}^{\top}\boldsymbol{s}_n\Big].\end{aligned} \tag{6}$$

最適化の対象となる 2 つの行列 $\boldsymbol{U}, \boldsymbol{V}$ を同時に最適化する場合，尤度関数は凸ではなくなってしまう．一方で，Shimosaka ら [19] は片方を固定した状態でもう片方を最適化することを繰り返すことによって，凸最適化問題に落とし込み，この問題を解決している．本研究においても，同様の手法を用いて最適化を行った．

正則化には (7) のフロベニウスノルムを用いた．

$$\|\boldsymbol{A}\|_2 = \sqrt{\sum_i\sum_j a_{i,j}^2}. \tag{7}$$

フロベニウスノルムを用いた正則化項を (8) に示す．ここで，正則化の強さを表す $\lambda > 0$ はハイパーパラメータとなる．

$$\Omega(\boldsymbol{U}, \boldsymbol{V}) = \lambda(\|\boldsymbol{U}\|_2^2 + \|\boldsymbol{V}\|_2^2). \quad (8)$$

(8), (6) によって表される対数尤度と正則化項を足した目的関数 (9) の最小化を行う．

$$\hat{\boldsymbol{U}}, \hat{\boldsymbol{V}} = \arg\min_{\boldsymbol{U},\boldsymbol{V}} \{-\ln L(\boldsymbol{U}, \boldsymbol{V}) + \Omega(\boldsymbol{U}, \boldsymbol{V})\}. \quad (9)$$

ハイパーパラメータは α_0, k, λ の 3 つであり，交差検定により決定する．

4.4 Shape パラメータの決定手法

本研究では，所要時間の平均値だけでなく，分散も考慮した区間推定を行うことを目標としている．そこで，平均が固定されたときに分散を正しく推定できるような枠組みを用いて，ハイパーパラメータである Shape パラメータを決定する手法について述べる．基本的なアイデアは，説明変数が同一となる条件下において目的変数の標本分散がモデルの推定結果から求められる分散と合致するように Shape パラメータを調整することである．このとき，説明変数が同一の条件下であるとするためには，何らかの離散化が必要となる．本研究の説明変数は連続変数を含んでいるため，四捨五入の関数 round(·) を用いて (10) により離散化を行っている．

$$\boldsymbol{x}' = (\mathrm{round}(\phi_t), \phi_d, \phi_h, \phi_w) \in \mathbb{N}^4. \quad (10)$$

ガンマ回帰のモデルでは，この散布図において Shape パラメータ $\alpha_0 > 0$ を用いた以下の関係が成り立つことを仮定している [18].

$$\mathrm{Var}(y) = \frac{1}{\alpha_0}\mathbb{E}[y]^2. \quad (11)$$

各点に割り当たっているデータ数を N_i，総データ数を $N = \sum_i N_i$ とすれば，この平面上における平均二乗誤差は (12) によって与えられる．

$$\mathrm{MSE}(\alpha_0) = \frac{1}{N}\sum_i N_i \left(\mathrm{Var}[y_i] - \frac{\mathbb{E}[y_i]^2}{\alpha_0}\right)^2. \quad (12)$$

低ランク双線形ガンマモデルにおいて，Shape パラメータに (12) を最小化するような値 (13) を用いれば，これは Shape パラメータの良い近似となると考えられる．

$$\hat{\alpha_0} = \arg\min_{\alpha_0} \mathrm{MSE}(\alpha_0). \quad (13)$$

5. 路線バス長期所要時間予測に関する各手法の性能比較実験

5.1 実験の目的

本手法の点推定の性能と区間推定の性能を既存手法と比較することにより，点推定の性能に関してはそのままに，区間推定に関しては他の手法を上回ることを，実データにモデルを適用することによって示す．

図 2　玉 07 の通行ルート

表 1　説明変数の一覧

説明	記号	データ型	例
出発時間	ϕ_t	実数値	10 時 8 分 0 秒
曜日	ϕ_d	シンボリック	月曜日，火曜日，···
国民の祝日	ϕ_h	真偽値	true, false
天気	ϕ_w	シンボリック	快晴，曇り，雨，···

5.2 使用するデータ

5.2.1 路線バスの所要時間データ

バス運行情報を表示するウェブサービスである東急バスナビ[*4]から運行データを自動取得し，これを訓練データとした．

2017 年 5 月 1 日から 2017 年 11 月 13 日の 197 日間のデータを用いる．対象路線は東急バスの玉 07（東京都世田谷区内，成城学園前駅から二子玉川駅）であり，全体のデータのうち正しく取得できた 13592 回の運行データを使用した．データの取得当時，対象路線の走行距離全長は 5.25 km または 5.43 km であり，平日には 1 日 166 本のバスが運行されていた．所要時間は凡そ 20–35 分である．玉 07 の路線を図 2 に赤線で示した．ただし，緑色で表現した部分は平日朝の一部の便のみが使用するルートである．

5.2.2 天気データ

天気に関しては，気象庁の Web サイトからダウンロードすることが可能である気象データを用いた．従って，天気の分類は気象庁のものに準じている．

5.3 路線バス所要時間推定問題の評価指標

本研究が取り組む問題の性質をもとに，本実験の評価指標を以下のように定める．

(1) 所要時間の点推定．

[*4] http://tokyu.bus-location.jp/blsys/navi

真の所要時間を表す関数を $f(\boldsymbol{x})$ としたときに，(14) で表す平均絶対誤差を小さくする $\hat{f}$ が望ましい．

$$\mathrm{MAE} = \mathbb{E}\left[\left|f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})\right|\right]. \tag{14}$$

(2) 所要時間の区間推定．

真に所要時間が従う分布の累積分布関数を $P(y|\boldsymbol{x})$ とした時にこの分布に「近い」累積分布関数 $\hat{P}(y|\boldsymbol{x})$ が望ましい．本研究では，路線バス所要時間推定問題の先行研究 [20] でも用いられているコルモゴロフ–スミルノフ検定 [21] の統計量（KS 統計量）を指標として用いる．KS 統計量は (15) で与えられる．

$$\mathbb{E}_{\boldsymbol{x}}\left[\max_y\left\{\left|\hat{P}(y|\boldsymbol{x}) - P(y|\boldsymbol{x})\right|\right\}\right]. \tag{15}$$

5.3.1 点推定性能の MAE による比較

点推定の性能を低ランク双線形ガンマ回帰モデル，低ランク双線形ポアソン回帰モデル，線形ガンマ回帰モデル，Random Forest に関して (14) で表現される平均絶対誤差（MAE, [min]）を用いて比較する．MAE の算出には $K = 5$ の K-交差検定を用いる．ガンマ分布を用いたモデルの点推定の値は，ガンマ分布のモードである $(\alpha_0 - 1)\beta$ を使用する．

5.3.2 区間推定性能の KS 統計量による比較

確率分布の推定に関して，実データとの当てはまりを，(15) によって表される KS 統計量を用いて，低ランク双線形ガンマモデル，低ランク双線形ポアソンモデル，そして 3.2 でも述べた過去のデータモデルの 4 種に対して比較を行う．

しかし，実際には，(15) を直接計算することは難しいため，(10) のように離散化を行っている．これを用いて，(15) は，(17) のように近似できる．

$$\mathbb{E}_{\boldsymbol{x}}\left[\max_y\left\{\left|\hat{P}(y|\boldsymbol{x}) - P(y|\boldsymbol{x})\right|\right\}\right] \tag{16}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{n} N_i \max_y\left\{\left|\hat{P}(y|\boldsymbol{x}'_i) - P(y|\boldsymbol{x}'_i)\right|\right\}. \tag{17}$$

真の累積確率分布を表す $P(y|\boldsymbol{x}'_i)$ を，離散化した条件 $\boldsymbol{x}'_i$ を満たすテストデータの集合 $\{y_{\boldsymbol{x}'_i}\}$ を用いて，(19) のように表現する．ただし，$\mathrm{card}(\cdot), \mathrm{card}'(\cdot)$ はともに集合の要素数を表す関数であり，$\mathrm{card}'(\cdot)$ に関しては，ゼロ除算を防ぐために (18) のようにおいている．

$$\mathrm{card}'(\{y\}) = \max\{\mathrm{card}(\{y\}), 1\}, \tag{18}$$

$$P(y|\boldsymbol{x}'_i) = \frac{\mathrm{card}(\{y_{\boldsymbol{x}'_i};\ y_{\boldsymbol{x}'_i} \le y\})}{\mathrm{card}'(\{y_{\boldsymbol{x}'_i}\})}. \tag{19}$$

過去のデータモデルは，(19) から累積確率分布を求めるモデルである．しかし，この手法では，説明変数の多い $\boldsymbol{x}'$ をそのまま用いると精度が出ないことが考えられる．そのため，国民の祝日，天気を説明変数に含めない説明変数

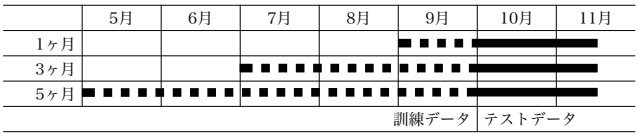

図 3 訓練データとテストデータの配分

$\boldsymbol{x}'' = (\mathrm{round}(\phi_t), \phi_d) \in \mathbb{N}^2$ を用いたモデルとの比較も行った．KS 統計量を用いた区間推定性能の比較では，真の累積確率分布を (19) に近似する性能が問題点となる．このような観点から，テストデータを十分にとるために**図 3** のように訓練データとテストデータを配分し，実験を行った．

5.3.3 Shape パラメータ推定手法の KS 統計量による性能評価

低ランク双線形ガンマ回帰に関して，ハイパーパラメータ λ, k を固定した状態で α_0 のハイパーパラメータ探索を行い，KS 統計量において精度の高い結果を残す Shape パラメータ $\hat{\alpha}_0$ が，平均と分散の重み付き最小二乗法を用いた Shape パラメータ $\bar{\alpha}_0$ と合致しているかを確認する．本研究では，後に示す $\bar{\alpha}_0$ 周辺である $\alpha_0 \in \{25, 30, \ldots, 60\}$ を探索し，その中で最も良い KS 統計量を出したものを $\hat{\alpha}_0$ とする．

5.4 説明変数の表現

本研究が扱うモデルの説明変数 $\boldsymbol{x} = (\phi_t, \phi_d, \phi_h, \phi_w)$ の要素の一覧を**表 1** に示した．これを，一般化線形モデルの一種である双線形ガンマモデルにおいて扱いやすい 2 つのベクトルの形式 $(\boldsymbol{d}, \boldsymbol{s})$ へと変換する．

$\boldsymbol{s}$ を時間帯を表現するベクトルに割り当てる．このとき，j 番目の要素 $(\boldsymbol{s})_j$ を，(20) により定義する．

$$(\boldsymbol{s})_j = \mathcal{N}(\phi_t | t_w i, t_{\sigma^2}); \tag{20}$$

$$j = i - \left\lfloor \frac{t_b}{t_w} \right\rfloor + 1, i = \left\lfloor \frac{t_b}{t_w} \right\rfloor, \ldots, \left\lceil \frac{t_e}{t_w} \right\rceil. \tag{21}$$

ただし，$t_b, t_e \in [0, 24)$ はそれぞれバスの始発出発時刻と最終出発時刻であり，t_w はベクトルの 1 つの要素に割当てられる時間の幅である．t_w, t_{σ^2} はハイパーパラメータとなる．

曜日に関しては，(22) のように，月曜日–日曜日と，国民の祝日かつ平日の 8 種類に分類し，1-of-K 表現とする．

$$\phi_{\mathrm{HW}} = \begin{cases} 7 & \text{if } \ \phi_d \in \{1, \ldots, 5\} \wedge \phi_h = 1 \\ \phi_d & \text{otherwise} \end{cases}. \tag{22}$$

これを 1-of-K 表現にしたものを $\boldsymbol{\varphi}_{\mathrm{HW}} \in \{0, 1\}^8$ とする．天気に関しては，ϕ_w を次のように，快晴・晴・薄曇・曇とそれ以外で区別し，1-of-K 表現にした $\boldsymbol{\varphi}_c = \{0, 1\}^2$ を定義する．

$$\boldsymbol{\varphi}_c = \begin{cases} (1, 0)^\top & \text{if } \ \phi_w \in \{1, 2, 3, 4\} \\ (0, 1)^\top & \text{otherwise} \end{cases}. \tag{23}$$

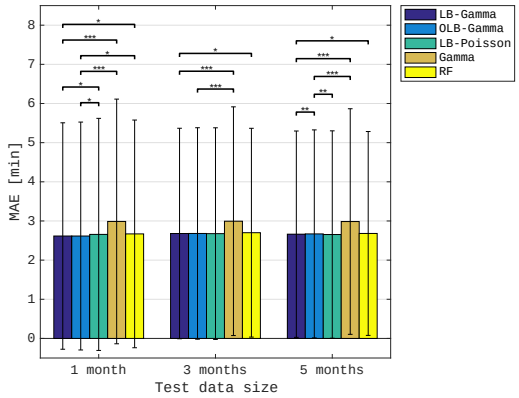

図 4　訓練データ量ごとの各手法の点推定の MAE の比較

(24) により，$(\boldsymbol{d}, \boldsymbol{s})$ の 2 つのベクトルを双線形モデルの説明変数とすることができる．

$$\boldsymbol{d} = \boldsymbol{\varphi}_{\mathrm{HW}} \otimes \boldsymbol{\varphi}_c \in \{0,1\}^{16}. \tag{24}$$

5.5　実験結果

$K = 5$ の K-交差検証により，各手法のハイパーパラメータを設定した．

5.5.1　点推定性能の MAE による比較結果

1 ヶ月，3 ヶ月，5 ヶ月の学習データを用いたモデルによる MAE の比較を図 **4** に示す．ただし，エラーバーは標準偏差を表し，凡例では低ランク双線形ガンマ回帰，双線形ポアソン回帰，線形ガンマ回帰，Random Forest をそれぞれ “LB-Gamma”，“LB-Poisson”，“Gamma”，“RF” と表現している．また，それぞれの指標の平均値に関して 2 標本の片側 t 検定を行った．手法 m_a, m_b が指標 s において $s(m_a), s(m_b)$ の性能を示す場合，$\mathbb{E}[s(m_a)] < \mathbb{E}[s(m_b)]$ を帰無仮説とした片側 t 検定である．

1 ヶ月から 5 ヶ月までのいずれの場合，いずれの統計量においても，低ランク双線形ガンマ回帰は Random Forest を $p < 0.05$ の有意水準において上回る性能を見せた．線形ガンマ回帰は，本研究で比較した他のいずれの手法からも $p < 0.001$ の有意差をつけ下回った．このため，双線形形式へと拡張することにより，表現力の高いモデルを構築することができているといえる．

5.5.2　区間推定性能の KS 統計量による比較結果

KS 統計量の比較を図 **5** に示す．凡例においては，低ランク双線形ガンマモデル，低ランク双線形ポアソンモデル，過去のデータモデル，天気と国民の祝日を用いない過去のデータモデルを，それぞれ “LB-Gamma”，“LB-Poisson”，“Historical 1”，“Historical 2” と表現している．エラーバーは (25) による重み付き標準偏差を表している．

$$\sqrt{\mathrm{Var}_{\boldsymbol{x}}\left[\max_y \left\{\left|\hat{P}(y|\boldsymbol{x}) - P(y|\boldsymbol{x})\right|\right\}\right]}. \tag{25}$$

いずれの手法でも，訓練データ量が増加するにつれて性能が向上していることがわかる．そして，1 ヶ月から 5 ヶ月までのいずれの場合でも，低ランク双線形ガンマ回帰を用いた場合がもっとも性能に優れているといえる．また，過去のデータモデルとの比較から，路線バスの所要時間分布への近似として，ガンマ分布への近似は一定の効果を示したといえる．

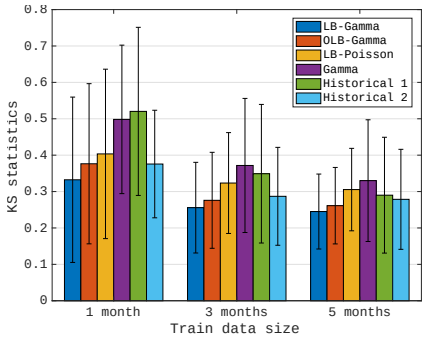

図 5　訓練データ量ごとの各手法の KS 統計量の比較

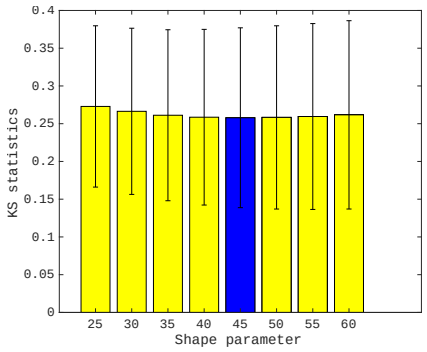

図 6　5 ヶ月分の訓練データを用いた低ランク双線形ガンマ回帰における Shape パラメータごとの KS 統計量

5.5.3　Shape パラメータ推定手法の KS 統計量による評価結果

重み付き最小二乗法によって導出された Shape パラメータは，$\bar{\alpha}_0 = 40.36$ である．先述した範囲での Shape パラメータ探索の結果を図 **6** に示す．特に，$\bar{\alpha}_0$ を青色で示した．これにより，$\hat{\alpha}_0 = 45$ が明らかとなり，$\bar{\alpha}_0 = 40$ における KS 統計量の値はその次に優れたものであった．このように，$\bar{\alpha}_0$ の値は，そのまま用いることも十分考慮に入れられる上，この周辺に $\hat{\alpha}_0$ も存在していたことから，もっとも優れた α_0 の探索範囲を狭めることができたといえる．

6.　結論

本研究では，乗換案内アプリケーションへの応用を目的に，長期の路線バス所要時間の推定を行った．長期の推定においては，利用することのできる説明変数が限られており，その推定精度には限界がある．そこで，本研究では確率分布そのものを求めることに着目し，これを高精度で求めることを目標とした．所要時間が従う分布に関して，様々な分布を平均と分散の関係という点において比較を行い，その中でガンマ分布が最適であるという結論を得た．そして，所要時間がガンマ分布に従うという仮定のもと，

低ランク双線形ガンマ回帰を提案した．上記の提案手法と既存手法を，東京都の路線バスの 1 つの路線における実際の運行データを用いて，点推定，区間推定の両面で比較することにより，提案手法が，低ランク双線形ポアソン回帰，Random Forest，単純に過去のデータからヒストグラムを構成したモデル等の比較手法を上回ることを示した．そのほか，平均と分散の関係から，重み付き最小二乗法を用いて Shape パラメータを求める方法を提案し，実験では最適な値を得ることはできなかったものの，それに近い値をハイパーパラメータの探索なしに得ることができた．

今回は始点から終点までの所要時間推定のみを行なったが，それ以外の部分的な所要時間を求める問題も存在する．停留所の数を n 個としたときに，乗車地点と降車地点の選び方は $n(n-1)/2$ 通りである．それぞれについて所要時間を求めることは効率的ではないため，より効率的な方法を模索していく必要がある．

参考文献

[1] Jeong, R. and Rilett, R.: Bus arrival time prediction using artificial neural network model, *Proceedings of Trans. on Intelligent Transportation Systems*, IEEE, pp. 988–993 (online), DOI: 10.1109/ITSC.2004.1399041 (2004).

[2] Chang, H., Park, D., Lee, S., Lee, H. and Baek, S.: Dynamic multi-interval bus travel time prediction using bus transit data, *Transportmetrica*, Vol. 6, No. 1, pp. 19–38 (online), DOI: 10.1080/18128600902929591 (2010).

[3] Kormaksson, M., Barbosa, L., Vieira, M. R. and Zadrozny, B.: Bus travel time predictions using additive models, *Proceedings of ICDM*, IEEE, pp. 875–880 (online), DOI: 10.1109/ICDM.2014.107 (2014).

[4] Bai, C., Peng, Z. R., Lu, Q. C. and Sun, J.: Dynamic bus travel time prediction models on road with multiple bus routes, *Computational Intelligence and Neuroscience*, Vol. 2015, p. 63 (online), DOI: 10.1155/2015/432389 (2015).

[5] Bates, J., Polak, J., Jones, P. and Cook, A.: The valuation of reliability for personal travel, *Transportation Research Part E*, Vol. 37, No. 2-3, pp. 191–229 (online), available from ⟨https://doi.org/10.1016/S1366-5545(00)00011-9⟩ (2001).

[6] Moreira, J. M., Soares, C., Jorge, A. M. and de Sousa, J. F.: The effect of varying parameters and focusing on bus travel time prediction, *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, pp. 689–696 (2009).

[7] Moreira, J. M., Jorge, A. M., de Sousa, J. F. and Soares, C.: Comparing state-of-the-art regression methods for long term travel time prediction, *Intelligent Data Analysis*, Vol. 16, No. 3, pp. 427–449 (online), DOI: 10.3233/IDA-2012-0532 (2012).

[8] Moreira, J. M., Jorge, A. M., de Sousa, J. F. and Soares, C.: Improving the accuracy of long-term travel time prediction using heterogeneous ensembles, *Neurocomputing*, Vol. 150, No. PB, pp. 428–439 (online), DOI: 10.1016/j.neucom.2014.08.072 (2015).

[9] Lin, W.-H. and Zeng, J.: Experimental study of real-time bus arrival time prediction with GPS data, *Transportation Research Record*, Vol. 1666, pp. 101–109 (online), DOI: 10.3141/1666-12 (1999).

[10] Tirachini, A.: Estimation of travel time and the benefits of upgrading the fare payment technology in urban bus services, *Transportation Research Part C*, Vol. 30, pp. 239–256 (online), DOI: 10.1016/j.trc.2011.11.007 (2013).

[11] Chen, M., Liu, X., Xia, J. and Chien, S. I.: A dynamic bus-arrival time prediction model based on APC data, *Computer-Aided Civil and Infrastructure Engineering*, Vol. 19, No. 5, pp. 364–376 (online), DOI: 10.1111/j.1467-8667.2004.00363.x (2004).

[12] Wu, C. H., Ho, J. M. and Lee, D. T.: Travel-time prediction with support vector regression, *Trans. on Intelligent Transportation Systems*, Vol. 5, No. 4, pp. 276–281 (online), DOI: 10.1109/TITS.2004.837813 (2004).

[13] Zhang, Y. and Haghani, A.: A gradient boosting method to improve travel time prediction, *Transportation Research Part C*, Vol. 58, pp. 308–324 (online), DOI: 10.1016/j.trc.2015.02.019 (2015).

[14] Huang, L. and Barth, M.: A novel loglinear model for freeway travel time prediction, *Proceedings of ITSC*, pp. 210–215 (online), DOI: 10.1109/ITSC.2008.4732620 (2008).

[15] Ma, Z., Ferreira, L., Mesbah, M. and Zhu, S.: Modeling distributions of travel time variability for bus operations, *Journal of Advanced Transportation*, Vol. 50, No. 1, pp. 6–24 (online), DOI: 10.1002/atr.1314 (2016).

[16] Susilawati, S., Taylor, M. A. P. and Somenahalli, S. V. C.: Distributions of travel time variability on urban roads, *Journal of Advanced Transportation*, Vol. 47, No. 8, pp. 720–736 (online), DOI: 10.1002/atr.192 (2013).

[17] Koenker, R. and Bassett Jr, G.: Regression quantiles, *Econometrica: journal of the Econometric Society*, pp. 33–50 (online), available from ⟨https://www.jstor.org/stable/1913643⟩ (1978).

[18] Smyth, G. K. and Verbyla, A. P.: Adjusted likelihood methods for modelling dispersion in generalized linear models, *Environmetrics*, Vol. 10, No. 6, pp. 695–709 (1999).

[19] Shimosaka, M., Maeda, K., Tsukiji, T. and Tsubouchi, K.: Forecasting urban dynamics with mobility logs by bilinear poisson regression, *Proceedings of UbiComp*, ACM, pp. 535–546 (online), DOI: 10.1145/2750858.2807527 (2015).

[20] Rahman, M. M., Wirasinghe, S. C. and Kattan, L.: Analysis of bus travel time distributions for varying horizons and real-time applications, *Transportation Research Part C*, Vol. 86, No. December 2017, pp. 453–466 (online), DOI: 10.1016/j.trc.2017.11.023 (2018).

[21] Massey Jr, F. J.: The Kolmogorov-Smirnov test for goodness of fit, *Journal of the American Statistical Association*, Vol. 46, No. 253, pp. 68–78 (online), DOI: 10.1080/01621459.1951.10500769 (1951).