第 57 回土木計画学研究発表会・講演集

混合分布モデルを適用した 到着分布モデルの推定に関する研究

宮内　弘太[1]・高田　和幸[2]

[1]学生会員　東京電機大学大学院　先端科学技術研究科（〒350-0394 埼玉県比企郡鳩山町石坂）
E-mail:18uda02@ms.dendai.ac.jp

[2]正会員　東京電機大学　理工学部（〒350-0394 埼玉県比企郡鳩山町石坂）
E-mail: takada@g.dendai.ac.jp

近年，ビックデータの出現により，データ解析をするための分析手法の提案が様々な分野で精力的に行われている．これらの手法の一つに混合分布モデルを用いた推定方法がある．これは推定の為に，複数の統計分布を適用しているというものである．この混合分布モデルの適用は，非常にモデルの表現力に富んでいるが，パラメータ推定が難しいという特徴を持つ．本研究では，この混合分布モデルの適用事例の一つとして鉄道利用者の到着分布に導入した．

近年の慢性化した遅延の影響を受けて到着状況が多様化したと考え，混合分布モデルの適用に至った．そこで適用を行いにあたり，首都圏の鉄道利用者を対象とし，通勤時間帯の到着状況について調査した．また，どのような混合分布モデルが鉄道利用者の到着分布と適用可能性が高いかについても検証した．本研究では，混合分布モデルへの適用可能性について言及する．

Key Words: *mixture distribution model, arrival distribution, big-data, pattern reconization*

1. はじめに

近年，ビックデータの出現により，精度良くデータ解析を行う為の手法の提案があらゆる分野で精力的に行われている．ビックデータを解析することで大部分で検知できるデータとは異なった特徴を持つデータが検出され，新たな情報として捉え分析を行うことが出来る [1]．ビックデータを解析する手法の一つに混合分布モデルを適用した分析手法がある．これはデータ中に存在する潜在変数が複数の確率分布に従っていると仮定をして推定をしたものである．その為，混合分布モデルを適用することにより，データのクラスタリングを容易に行うことが出来るという利点を持つ．クラスタリングを行う手法の一つとして k-means 法があるが，これは非確率的にデータをクラスタリングする手法であるのに対し，混合分布モデルでクラスタリングを行う場合は，潜在変数を用いてクラスタリングを行っている．すなわち分析者が任意にクラスタリングの基準を設定できる [2]．また近年では行動モデルの構築の際にも混合分布モデルが援用されている．例えば交通状況の変化が激しいような経路のドライバーの意思決定モデルを構築化の際には，混合分布モデルを使う事例が存在する [3]．また消費者の購買行動のモデル化に使われている [4]．多岐にわたる行動モデルへの適用可能性が示唆されている．

そこで本研究では，土木計画学における行動モデルへの適用可能性の一つとして，鉄道利用者の通勤時間帯における到着状況に着目をして，混合分布モデルを援用し，到着分布モデルの構築を行った．適用に至った理由として，近年，首都圏における通勤時間帯の鉄道ダイヤは，遅延が頻繁に発生している．その要因として相互直通運転や路線の延伸化が原因であるとされている．すなわち首都圏全体に鉄道網が形成されたことによって，遅延の被害を受けやすい構造になっているとされている．頻繁に起こる遅延の影響を受け，通勤時間帯の鉄道利用者の行動に変化が生じていることが判明した [5]．それは時刻表を信用せずに，あらかじめ遅延を見込んで余裕時間を設定して通勤をするようになったことである．すなわち余裕時間の設定の仕方によっては到着時刻が前後に変動するようになった．そこで本研究では，鉄道利用者の到着状況別のクラスタリングと遅延回避の為の潜在意識をそれぞれ混合分布モデルを適用し，従来のモデルよりも精緻化を図った．以下の二つの観点から分析を行った．

(a) アンケート回答者全員から見た到着分布

(b) 回答者個々人から見た到着分布

次に本稿の構成について示す．第 2 章では，混合分布モデルが使われている研究事例について整理する．第 3

章では，本研究で実施した首都圏の通勤時間帯における鉄道利用者へのアンケート調査の概要と基礎集計について説明する．第 4 章では，混合分布モデルのアルゴリズムについて説明する．また本研究では，混合分布モデルから拡張した異種混合分布モデルも分析に用いている．このアルゴリズムについても同様に説明したいと思う．第 5 章の前半部分では，アンケート回答者全員を考慮した際の到着分布モデルの混合分布モデルの適用可能性について検証する．また合わせて混合分布モデル適用の際の問題点の一つである，初期値依存の頑健性についても検証したいと思う．後半部分では，アンケート回答者個々人の観点から見た時の到着分布モデルについて検証する．最後に第 6 章では，本研究のまとめと今後の課題と展望について述べたいと思う．

2. 既往研究について

混合分布モデルが適用されている研究事例ついて整理すると，混合分布モデルの中で最も使われているモデルが，混合正規分布モデルである．これは複数の正規分布が混在していると仮定をしているものである．一般的にデータマイニング，パターン認識，機械学習，統計的解析，行動モデルの構築の際に使われている．そこで本研究では，行動モデルの構築の際に混合分布モデルが使われている研究事例について整理した上で，本研究の位置づけを明確にしたいと思う．

国外で行われている研究事例について整理すると，様々な行動モデルに混合分布モデルが適用されて分析が行われている．例えば Sumalee *et al.*[6)7)]らは，交通流動密度の解析に際に混合分布モデルを適用している．効率的に道路間の自動車を流すモデルを数理化した上で，実験を行うことで妥当性について検証している．また独自のアルゴリズムの提案を行っている．行動モデルへの適用に混合分布モデルの妥当性を検証した研究事例には，Nylund *et al.*[8)]らがある．混合分布モデルの個人差が生じているデータに対しての適用は有効であると説きながらも，混在している分布の要素数が最適に決まらない混合分布モデルが持つ特有の問題点に対して，潜在クラスモデル，因子混合分布モデル，成長混合分布モデルとそれぞれをサンプル数の条件を変えながら分析を行い，妥当性について検証した．

国内で行われている研究事例について整理すると，例えば交通事故解析を行なった研究[9)]がある．この研究では，低頻度と高頻度の事故発生率が混在するデータを取り扱い，混合ポアソンモデルを用いて事故件数モデルを推定している．その他、混合指数分布を用いた研究[10)]，混合正規分布を用いた研究[11)]がある．

以上国内外で行動モデルに混合分布モデルに適用されている研究事例について整理した．潜在クラスモデルとの違いについての検証や混合分布モデルを拡張した因子混合分布モデルや成長混合分布モデルについては研究事例が存在するが，混在している分布が異なった，異種混合分布モデルを適用して解析を行なった研究事例については見受けられなかった．

そこで本研究では鉄道利用者の到着状況に対して異種混合分布モデルについても適用しているが，構築したモデルは尤度の観点だけでなく，適用した際の初期値依存による頑健性についても同様に検証したいと思う．

3. 使用データ

(1) アンケート調査の概要と標本属性について

本研究では，株式会社マクロミルの調査モニターを活用して，鉄道の利用状況に関するアンケート調査を実施した．調査対象者は，首都圏の 1 都 3 県（東京・神奈川・千葉・埼玉）に居住する 15 歳以上の有職者（パート・アルバイトを含む）で，かつ週に 5 日以上通勤目的の為に鉄道を利用する者とした．調査は平成 27 年 2 月 23，24 日の 2 日間で行い，計 1000 名から有効回答を得た．**表-1** は，アンケートの調査概要である．この調査の回答データを用いて分析を行った．また乗車・降車・乗り換えに利用した駅，利用路線に関する回答結果に基づき，経路検索サイト（Yahoo!乗換案内，2015 年 4 月）を活用して，移動距離と所要時間に関するデータを作成した．

回答者の個人属性の内訳を**表-2** に示す．性別，年代，居住地について示す．大都市交通センサスと比較すると，男性の割合と 20 代の割合において差が生じているが，他の個人属性は特定の属性に偏ることなく標本が抽出されており，本調査は大都市交通センサスと概ね同じ条件でデータが抽出できたと考えられる．今回のアンケート調査は首都圏の一都三県の住民を対象にして行ったものである．標本の個人属性（性別・年齢・居住地）に大きな偏りがないことを，大都市交通センサスの構成割合との比較を通じて確認した．よって，分析を行うにあたり回答データに対して補正や重みづけをせずにそのまま使用した．

(2) 鉄道利用者の通勤時間帯の到着状況について

次に鉄道利用者の通勤時間帯の到着状況についての集計結果について説明する．**表-3** にアンケート調査内容について示す．Q.1 - Q.4 までは，回答者が普段の通勤で設定している時刻についてをお聞きしている．Q.5 については，到着がどれ位変動しているのかを感覚的に回答し

てもらうようにお聞きしている．

次に鉄道利用者の所要時間の認知状況を把握するために，理想の到着時刻と実際の到着時刻との乖離度合いに関するデータを作成した．アンケートでは鉄道利用者が自宅の最寄り駅を出発する出発時刻と，理想とする到着時刻を聞いている（**表-3**, Q2・Q4 参照），回答していただいた理想到着時刻を経路検索サイトに入力し，表示された出発時刻を調べ，理想とする到着時刻に着くための出発時刻を決定した．すると実際の出発時刻と路線検索サイトで調べた出発時刻との間には差が生じることが判明した．この差（理想の出発時刻－実際の出発時刻）が大きいほど，余裕時間が大きくなり，理想の到着時刻以前に到着できる確率が高まり，また反対に大きい場合には，理想の到着時刻以後に到着することを許容していることを示している．先ほど算出した余裕時間をアンケート回答者の通勤時の移動距離で割ることにより，単位 km あたりの余裕時間が算出される．単位距離 km 当たりを算出する理由として，鉄道利用者の理想の到着状況に対する早遅着状況の基準化を行う為に距離で割った．この時の階級と頻度の関係を**図-1** に示す．プラス側であれば，単位 km あたりの時間分だけ理想の到着時間と比較して遅着をしている．一方，マイナス側であれば，単位 km あたりの時間分だけ理想の到着時間と比較して早着をしている．図を見てみると半分以上の鉄道利用者が理想の到着時刻よりも早着をしていることが分かる．

しかし慢性的に遅延が発生している路線を利用している鉄道利用者は日々の経験から，遅延時間を織り込んで理想の到着時刻を設定しているとも考えられる．その場合，実際の到着時刻と理想の到着時刻が一致して，乖離が無いとも見てとれる．

本研究では，この場合の取扱について，あくまでも鉄道利用者が通勤時の到着時刻に対して余裕時間を考慮しているかどうかを調査している為，問題はないと考えている．鉄道利用者は通勤時に理想の到着時刻を設定していることが本調査より分かった．以上がアンケート回答者全体の到着状況についてまとめた．

次に鉄道利用者個々人の到着状況について整理する．鉄道利用者が設定している理想の到着時刻から，100 回中何回早く着くのか，遅く着くのか，もしくは定刻通りに着いているのかを**表-3** の Q5 の回答結果より得た．その結果を用いてヒストグラムの作成を行った．**図-2** は，ある一例の到着状況についてを示している．本アンケート調査の回答者数は 1000 であるので，1000 個のヒストグラムが作成されたということである．以上個々人の到着状況について整理した．，その結果を用いて到着分布の作成を行った．中山 [12]の研究成果よると，旅行時間に対する統計分布の当てはめには，正規分布や対数正規分布，指数分布が使われていると指摘がされている．そこ

表-1　アンケートの調査概要

実施日時	2015年2月23・24日
調査人数	1000人
調査方法	インターネット調査
対象者	・一都三県（東京，埼玉，千葉，神奈川）に居住する方 ・15歳以上の有職者の方 ・週に5日以上鉄道を利用している方
個人属性	性別，年齢，職業，年収
鉄道の利用状況に関する質問	・通勤時の起終点駅（路線・駅名） ・通勤時の利用経路（乗換駅） ・鉄道利用頻度 ・認知している所要時間
鉄道遅延経験	運行遅延に関する認識 （遅延遭遇頻度・遅延による損失時間）

表-2　本調査の標本属性

	第12回大都市交通センサス	本調査
男女比		
男性，女性	65.1%，34.9%	78.3%，21.3%
年代		
10代，20代， 30代，40代， 50代，60代以上	0.4%，16.7%， 25.1%，28.4%， 18.2%，11.3%	0.0%，5.6%， 28.5%，26.9%， 24.3%，14.7%
居住地		
東京都，埼玉県， 千葉県，神奈川県	42.0%，19.0%， 17.0%，22.0%	49.2%，14.5%， 13.1%，23.2%
所要時間		
30分未満 30分以上~1時間未満 1時間以上~1時間30分 1時間30分～2時間	5.0% 37.0% 53.9% 4.1%	4.2% 35.5% 56.3% 4.0%

表-3　アンケートの調査内容（一部抜粋）

通勤時の出発・到着時間に関する設問	回答例	
Q1.始業時刻に間に合うように職場へ到着するためには遅くとも何時までに自宅を出発しなければいけませんか?	7:10	
Q2.最初に乗車する駅に何時の電車に乗っていますか?	7:00	
Q3.最後に降りる駅に何時までに着かなければなりませんか?	8:20	
Q4.最後に降りる駅に何時に着くのが理想的ですか?	8:00	
理想の到着時間との乖離状況について	回答例 （早着／遅着）	
Q5.あなたが理想としている到着時間から前後1分以内に到着しているのは100回中何回ですか?	30回	
前後1分以上で駅に着くのは100回中何回ですか?	25回	15回
前後2分以上で駅に着くのは100回中何回ですか?	15回	5回
前後3分以上で駅に着くのは100回中何回ですか?	3回	1回
前後4分以上で駅に着くのは100回中何回ですか?	3回	0回
前後5分以上で駅に着くのは100回中何回ですか?	3回	0回
前後10分以上で駅に着くのは100回中何回ですか?	0回	0回
前後20分以上で駅に着くのは100回中何回ですか?	0回	0回
前後30分以上で駅に着くのは100回中何回ですか?	0回	0回

で本研究では，比較的数式モデルの援用がしやすい，正規分布と指数分布が混在していると仮定をした．対数正規分布については，今後の課題とする．次に正規分布と指数分布がいくつ混合されているのが最適なのかを GAP 統計量 [13]を用いて判断した．GAP 統計量とは解析対象データのコンパクト性と分離性をそれぞれ定量的に評価したもので ある．この二つの評価値を足し合わせたものを GAP 統計量とし，最大値となるクラスタ数が最適なクラスタ数となる．使用データには，鉄道利用者の全体の到着分布のクラスタ数を調べる為には，**図-1** の結果を，個々の到着分布を調べる際には**図-2** の結果を用いた．**図-3** に全体で見た時とここで見た時のクラスタ数別の GAP 統計量の値を示す．なお，ここで見た時の Gap 統計量は，1000 人のうちの平均値を示している．本推定では，クラスタ数を 1 から 5 まで設定をして，それぞれの GAP 統計量を算出した．その結果，どちらのケースでも，クラスタ数が 3 の時が最も良いという結果になった．すなわち，鉄道利用者の到着状況は 3 タイプに分けられる事が推定結果より分かった．次節では，混合分布モデルの数式展開と異種混合分布モデルへの拡張について言及していくこととする．

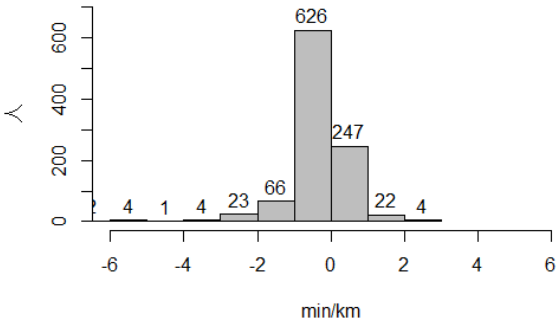

図-1　アンケート回答者の距離当たりの余裕時間
(n＝1000)

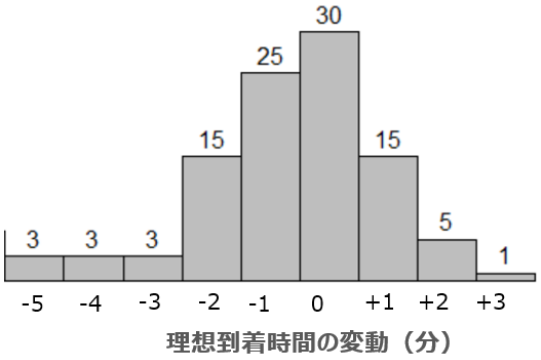

図-2　Q5 の回答結果を用いたヒストグラム
（一部抜粋）

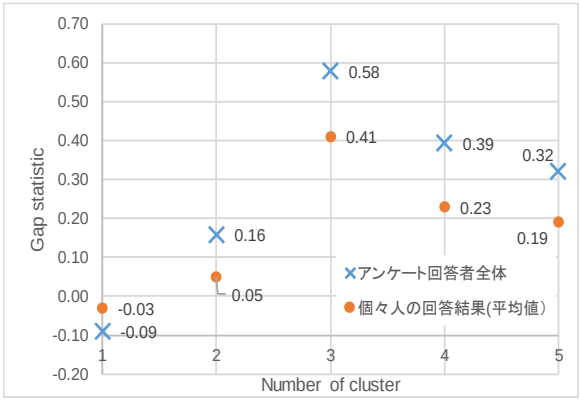

図-3　クラスタ数別の GAP 統計量

4.　到着分布への混合分布モデルの適用可能性

(1)　混合分布モデルについて

前章で到着状況を定性的に考察した結果，鉄道利用者の到着分布については，複数の分布が混在している可能性が明らかとなった．そこで本研究では，混合分布モデルを用いて到着分布の推定を試みる．

なお混合分布モデルで到着分布を推定することにより，多様な集団がどのように混ざり合っているかを定量的に求めることが出来る．今日では，主に画像解析やビッグデータからの将来予測に活用されている [14]．

混合分布モデルは以下の式で表される．

$$p(x|\theta)=\sum_{k=1}^{K}\pi_k\ f_k(x|\theta_1,\theta_2,\theta_3...\theta_k) \qquad (1)$$

ここで，p は混合分布の確率密度関数，x は観測データ，K は混合されている分布の総数，π は分担率，fk はクラスタ k の確率密度関数，θ k クラスタ k を構成するパラメータを指す．なお，分担率の合計は 1 となる．

本論では，正規分布のみで表す正規混合分布と，異なる分布で表す異種混合分布による推定を行った．

(2)　混合分布モデルの算出式

a)　正規混合モデル

正規分布のみで構成された混合分布モデルは一般的に正規混合モデル（Gaussian Mixture Model）と呼ばれ，混合分布モデルにおいて頻出するモデルである．正規混合分布の確率密度分布は，複数の正規分布を重ね合わすことでモデル化できるものである．

したがって，K 個の正規分布が存在すると仮定した時，正規混合モデルは以下の式で表される [15]．

$$p(x|\pi,\mu,\sigma^2)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\sigma_k^2) \qquad (2)$$

$$\sum_{k=1}^{K}\pi_k=1$$

式(2)の対数尤度関数を算出し，正規分布要素の平均

μ_kに関して微分して 0 と置くと(3)式が得られる．

$$\sum_{n=1}^{N}\frac{\pi_k N\left(x_n\middle|\mu_k,\sigma^2 k\right)}{\Sigma_j \pi_j N\left(x_n\middle|\mu_j,\sigma^2 j\right)}\Sigma_k^{-1}(x_n-\mu_k)=0 \qquad (3)$$
$$\frac{\pi_k N\left(x_n\middle|\mu_k,\sigma^2 k\right)}{\Sigma_j \pi_j N\left(x_n\middle|\mu_j,\sigma^2 j\right)}=\gamma(z_{nk})$$

しかし，このままでは求めたいパラメータを陽に解くことができないため，まず以下のように潜在変数として負担率 γ を定義する．ここで，z_{nk} は n 番目の観測データが，クラスタ k に属しているかを表す変数である．ベルヌーイ試行に従い，クラスタ k に属している場合は 1，属していない場合は 0 となる． (4)式中の π_k を z_k=1 となる事象の事前確率，$\gamma(z_k)$を x が観測したときの事後確率とすると，負担率というのは，観測データが与えられた下での z の条件付き確率である．ベイズの定理を用いて以下のように得られる．

$$\gamma(z_k)=p\left(z_k=1\middle|x\right)$$
$$=\frac{p(z_k=1)p\left(x\middle|z_k=1\right)}{\sum_{j=1}^{K}p\left(z_j=1\right)p\left(x\middle|z_j=1\right)}$$
$$=\frac{\pi_k N\left(x\middle|\mu_k,\sigma^2{}_k\right)}{\sum_{j=1}^{K}\pi_j N\left(x\middle|\mu_j,\sigma^2{}_j\right)} \qquad (4)$$

この式(4)を式(3)に代入し，式(5)が導出される．

$$\sum_{i=1}^{N}\gamma_k x_i=\mu_k\sum_{i=1}^{N}\gamma_k \qquad (5)$$

さらに，k 番目の分布に所属するデータ x の数を N_k と置き，パラメータの最尤推定量を以下に示す．

$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma_k x_i$$
$$\sigma^2{}_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma_k\left(x_i-\mu_k\right)^2 \qquad (6)$$

このパラメータと潜在変数をについて，反復計算により求める手法が EM アルゴリズムである．

b) 異種混合分布モデルの算出式

鉄道利用者の到着状況は正規分布だけで構成されているわけではないことを第 3 章で仮定をした．そこで，式(3)を拡張し，正規分布と正規分布以外の分布が混合されたモデルを構築する．具体的には，式(2)に指数分布を追加することで，新たに異種混合モデルを定式化する．

$$p\left(x\middle|\pi,\mu,\sigma^2,\lambda\right)=\sum_{k=1}^{K}\pi_k N\left(x\middle|\mu_k,\sigma^2{}_k\right)+\pi_E f\left(x\middle|\lambda\right) \qquad (7)$$

正規混合モデルと同様に潜在変数を置き，最尤推定量を導く．以下に，潜在変数を用いて求めるパラメータをまとめる．

$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma_k x_i$$
$$\sigma^2{}_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma_k\left(x_i-\mu_k\right)^2$$
$$\lambda=\frac{N_E}{\sum_{i=1}^{N}\gamma_E x_i} \qquad (8)$$
$$\pi_k=\frac{1}{N}\sum_{i=1}^{N}\gamma_k=\frac{N_G}{N}$$
$$\pi_E=\frac{1}{N}\sum_{i=1}^{N}\gamma_E=\frac{N_E}{N}$$

ただし，観測データ x が負である場合，指数分布に所属するかどうかの変数 z_E は 0 となる．そのため，以下のように場合分けを行う．

$$if\ \ x\geq 0\begin{cases}\gamma_k=\dfrac{\pi_k N\left(x\middle|\mu_k,\sigma^2{}_k\right)}{\sum_{j=1}^{K}\pi_j N\left(x\middle|\mu_j,\sigma^2{}_j\right)+\pi_E f\left(x\middle|\lambda\right)}\\ \gamma_E=\dfrac{\pi_E f\left(x\middle|\lambda\right)}{\sum_{j=1}^{K}\pi_j N\left(x\middle|\mu_j,\sigma^2{}_j\right)+\pi_E f\left(x\middle|\lambda\right)}\end{cases} \qquad (9)$$

$$if\ \ x<0\begin{cases}\gamma_k=\dfrac{\pi_k N\left(x\middle|\mu_k,\sigma^2{}_k\right)}{\sum_{j=1}^{K}\pi_j N\left(x\middle|\mu_j,\sigma^2{}_j\right)}\\ \gamma_E=0\end{cases} \qquad (10)$$

以上の式(7)から式(10)に対し，EM アルゴリズムによる反復計算によってパラメータの推定を行なった．そこで次節では，実際に混合分布モデルのアルゴリズムを適用して，分析を行う．鉄道利用者の到着分布に対する混合分布モデルの適用の可能性について検証する．

5． 推定結果

(1) 鉄道利用者全体から見た時

a) 分布形状と EM アルゴリズムの初期値の設定

初めに初期値の中心が 0 分となる正規分布 1 が存在すると仮定した．本研究では，4 種類の混合分布モデルと単一な正規分布モデルを推定し，BIC 基準により説明力を検証した．分担率については無情報を表現するため，設定した分布に均等割りした．

b) パラメータの推定結果

到着状況を乗車距離(㎞)で基準化した到着分布モデルを，混合分布モデルに適用して推定を行った．
推定結果を**表-4** に示す．正規混合分布と 3 つの異種混合分布，比較のために正規分布モデルを推定した．各分布

表-4 混合分布モデルのパラメータの推定結果（アンケート回答者全員で一つの到着分布を仮定した時）

		正規分布	正規混合分布 (N1, N2)	異種混合分布 1 (N, E)	異種混合分布 2 (N1, N2, E)	異種混合分布 3 (N1, N2, N3, E)
正規分布 N_1	μ_1	-0.285	-0.164	-0.308	-0.249	-0.019
	σ_1^2	0.699	0.152	0.749	0.113	0.002
正規分布 N_2	μ_2	-	-0.866	-	-0.959	0.089
	σ_2^2	-	2.914	-	2.760	0.141
正規分布 N_3	μ_3	-	-	-	-	-0.055
	σ_3^2	-	-	-	-	7.66
指数分布	λ	-	-	9.284	3.208	5.412
分担率	π_1	1.00	0.827	0.815	0.676	0.401
	π_2	-	0.172	-	0.171	0.257
	π_3	-	-	-	-	0.083
	π_E	-	-	0.185	0.152	0.154
評価指標	BIC	10,062	1,363	2,951	1,074	8,903

のパラメータ，分配率，さらに構築した分布の適合度を表すベイズ情報量規準 BIC を示した．尚，BIC の値が小さいほど適切なモデルである．

単一の正規分布を仮定した場合の BIC に着目すると，他と比較して値が大きい．すなわち，単一の分布よりも，複数の分布が混在した，混合分布モデルを適用した推定の方が優位であることが読み取れる．

次に正規混合分布モデルと異種混合分布モデル 1-3 を比較すると，異種混合分布モデル 2 が最も有意であることが表-4 から見て取れる．また混在している分布数が最も多い，異種混合分布モデル 3 に着目すると，BIC の値が大きいことより，多すぎても適切ではないことが読み取れる．

異種混合分布モデル 2 について考察する．この異種混合分布モデルに含まれる分布の形状は，図-4 の通りである．分布の分担率（π_1,π_2,π_E）は，それぞれ 0.676，0.171，0.152 となった．特段小さな分担率がないことから，到着分布には複数の分布が混在しており，混合分布の適用が適切であることが示された．

以下，3 つの分布の特性について考察を加える．
分担率が 0.676 と推定された正規分布 N1 は，**図-4** において赤色の破線で示されているが，ほぼ理想としている到着時刻通りに到着できているタイプである．次に，分担率が 0.171 と推定された正規分布 N2 は，**図-4** で青色の破線で示されており，理想としている到着時刻より前に到着するタイプである．さらに指数分布は二つの分布で説明しきれていない部分を説明しているものである．この場合の指数部分は到着時刻が遅着型タイプの人であることが分かる．この時の遅着型タイプというのは指数分布であるので時間通りに着くことはほとんどないというタイプである．つまり理想の到着時刻は，あくまでも目安の一つとしか考えていないというタイプがこれに当てはまる．また到着分布が指数分布の形状を呈するのは，

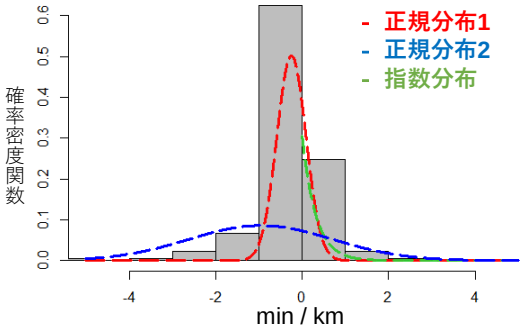

図 - 4 到着状況を表した異種混合分布モデル

表 - 5 初期値の設定による頑健性の確認

～初期分配率の設定に関する頑健性の確認～			
	Case1	Case2	Case3
初期値 π_1	0.8	0.1	0.1
収束値 π_1	0.676	0.676	0.676
初期値 π_2	0.1	0.8	0.1
収束値 π_2	0.171	0.171	0.171
初期値 π_E	0.1	0.1	0.9
収束値 π_E	0.152	0.152	0.152
～初期パラメータの設定に関する頑健性の確認～			
	Case4	Case5	Case6
初期値 $\mu_1\mu_2$	N(0, 1)	N(0, 100)	N(0, 1000)
収束値 μ_1	-0.249	-0.249	-0.249
収束値 μ_2	-0.959	-0.959	-0.959
初期値 $\sigma_1\sigma_2$	N(0, 1)	N(0, 100)	N(0, 1000)
収束値 σ_1	0.113	0.113	0.113
収束値 σ_2	2.760	2.760	2.760
初期値 λ	U(0, 1)	U(0, 10)	U(0, 100)
収束値 λ	3.208	3.208	3.208

鉄道が時刻表通り運行されていると信じて行動している場合である．従って，このような行動をとる利用者が 1.5 割程度存在していることを示唆するものである．

c) 初期値に対する頑健性の確認

ここでは，パラメータの推定結果の初期値依存について考察する．分担率，期待値，分散値の初期値を変動させて，パラメータの推定値の頑健性を確認した．結果を表-5 に示す．本調査では Case1 – Case 6 までを行った．Case 1 – Case 3 は分配率に対しての頑健性の調査を行い，Case 4 – Case 6 は期待値，分散値の頑健性についての調査である．

まず初めに Case 1 – Case 3 の結果に着目すると，分配率の初期値にどの様な値を設定しても，同じ収束値となることが推定結果より判明した．ただし例外として分配率を 0，0.5，0.5 や 1.0，0，0 などの 0 の値を入れてしまうと収束値が算出されないことが分かった．その他の場合は初期値の値を変化させても収束値の値が変わるといった事は無かった．

Case 4 – Case 6 の結果に着目すると，Case 4 – Case 6 は初期値の値の取り方を徐々に広くしている．その結果，どのような値を取っても収束値には，変化の影響がないことが判明した．

以上，構築した異種混合分布モデルのアルゴリズムの初期値に対する頑健性は，分析者が故意に不適当な値を設定しない限りは，分析ができることが堪忍できた．次節では，個々人から見た時の異種混合分布モデルの適用結果について説明する．

(2) 鉄道利用者個々人から見た時

a) 分布形状と EM アルゴリズムの初期値の設定

次に鉄道利用者個々人の回答結果に着目をして分析を行う．(1)と設定は概ね同じ条件で分析を行った．初めに初期値の中心が 0 分となる正規分布 1 が存在すると仮定した．ただし，(1)で 2 つの正規分布と 1 つの指数分布が混在しているいう事が明らかになったので，今回の分析ではその設定条件のみで行う事とする．分担率についても(1)と同様に無情報を表現するため，均等割りした．

b) パラメータの推定結果

鉄道利用者自身が設定している理想としている到着時刻に対して「早く到着する（早着認知）」「ほぼ時刻通りに到着する（定刻認知）」「理想到着時刻は目安程度にしか考えておらず遅刻をしなければ良い（遅着認知）」の 3 タイプの考え方が潜在的にあると仮定し，これらの認知状況を表す分布を，パラメータと共に以下のように仮定した．

・早着認知…到着時刻の平均が理想到着時刻よりも早い時間となる正規分布$N(\mu_1, \sigma_1)$ただし$\mu_1 < 0$

・定刻認知…到着時刻の平均が理想到着時刻となる正規

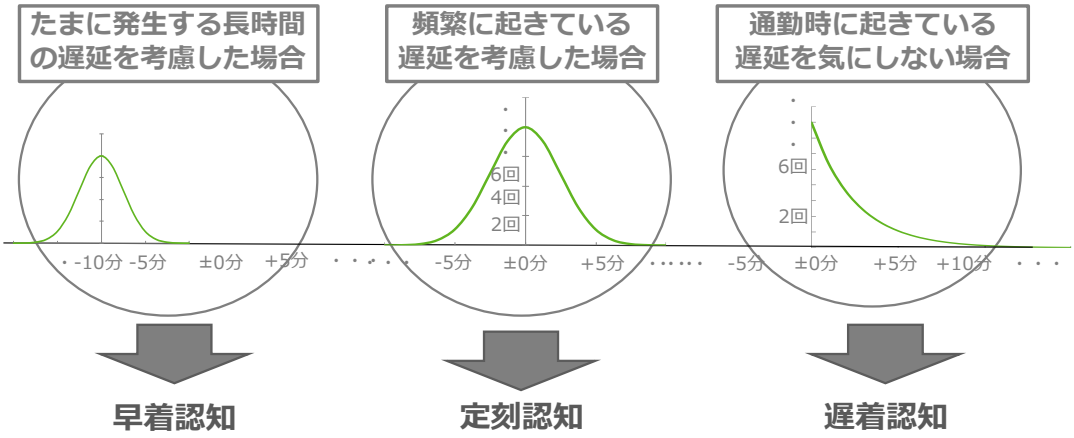

図 - 5　遅延に対する認知状況と到着分布の関係性

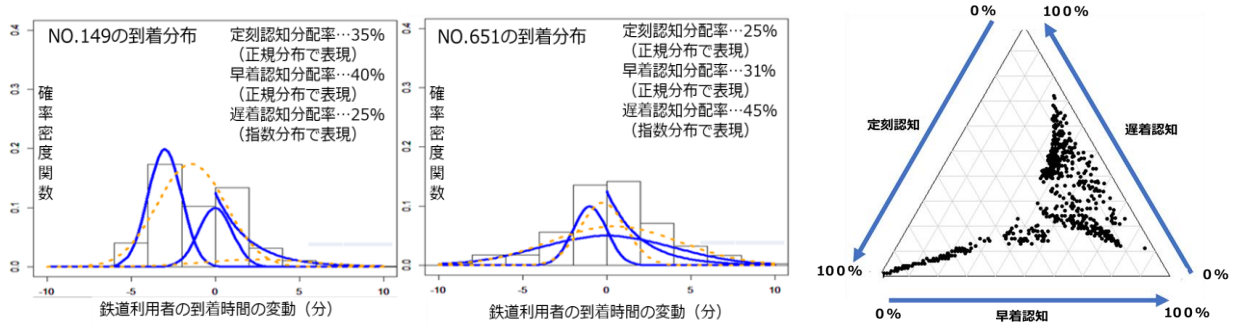

図 - 6　到着状況を表した異種混合分布モデル　(一例)と全回答者の分配率の内訳結果

分布$N(0,\sigma_2)$

・遅着認知…理想到着時刻はあくまでも到着の目安としか考えていない場合を暫定的に指数分布$E(\lambda)$で仮定

（一般的に理想到着時刻よりも早く着くことはないと考えられる為，指数関数を描くような形となる．）

図-5 に遅延に対する認知状況と到着分布の関係性を表す概念図を示すこれらの認知状況が組み合わさって構成されていると考えることが出来る．**図-6** の右図に異種混合分布モデルを適用した一例を示す．鉄道利用者が気にしている遅延の認知状況によって分布の形状や分配率に差が生じていることが読み取れる．**図-6** の左図には，アンケート回答者全員分の異種混合分布モデルを適用した際の分配率を示している．特徴として，電車での移動距離が少ない，鉄道通勤者は定刻認知の割合がかない高いということが明らかになった．

６.まとめと今後の展望

近年，ビックデータの出現により新たな分析手法の考案があらゆる分野で精力的に行われている．精緻なモデルを構築する上で，単一の統計分布だけでは不十分であるという考えの基，本研究では異種混合分布モデルの提案をし，行動モデルへの応用を図った．混合分布モデルの適用自体は先行研究を整理すると，あらゆる分野で行われていることが判明した．特にビックデータ解析や多様化した行動データへの分析アプローチとして使われているが，異なる分布が混在していると仮定した，異種混合分布モデルの適用事例は，ほとんどないことが国内外の先行研究を整理することで明らかになった．そこで本研究では，混合分布モデルのアルゴリズムを拡張した異種混合分布モデルのアルゴリズムを構築し，鉄道利用者の理想の到着時刻に対する到着分布への適用を行った．

到着分布モデルの適用を行うにあたり，二つの観点から分析を行った．一つが鉄道利用者全体で一つの到着分布を仮定した時，混在している分布ごとに鉄道利用者の分類を行うこと，そして二つ目が，個々人の到着分布に着目をして，個々人が遅延回避の潜在意識に対する認知状況の把握を行うことである．

以上の二つの分析を行う上で，鉄道利用者の到着状況が複数の確率分布で表現できるとした時，いくつの分布が混在しているのかを調べる必要がある．そこで Gap 統計量で判断をした．Gap 統計量が最大となるクラスタ数が適切な混在している分布の数となるが，どちらも 3 つが最適な分布数という結果が得られた．適切な分布形については，モデルの推定を通じて検証した．分析の結果，正規分布の形をしたのが 2 つ，指数分布が 1 つとした時の異種混合分布モデルが最も良いという結果が得られた．分布の特徴に着目してみると，正規分布 1 は，理想の到着時刻にほとんどの確率で到着をしているというタイプであった．これが最も多いタイプで回答者の 67%がこのタイプであった．正規分布 2 は，比較的理想の到着時刻よりも早く着くタイプであった．これが回答者の 17%がこのタイプであった．最後に指数分布の確率で到着しているタイプは，時刻表を完全に信用して行動しているという事が明らかになった．ここまでをまとめると，鉄道利用者の到着状況を表す到着分布を混合分布モデルで表すことにより，鉄道利用者の到着に対する考え方は全員同じではなく異なっているということが示された．

次に個々人の回答データに対して異種混合分布モデルを適用することにより，その回答者が潜在的にどの分布が最も気にしているという事を推定できることが明らかになった．これは，分析者の方で事前に分布の形状を設定しておくことで明らかにすることが出来る．本研究では，遅延に対する回避状況を正規分布と指数分布を仮定をして異種混合分布モデルの適用を行った．その結果，モデルの分配率は回答者によって大きく異なっている事が判明した．

以上の 2 つのアプローチ法からの推定結果を見ると，異種混合分布モデルを適用した行動モデルは，マクロ視点から見た場合のような分類手法としてミクロ視点から見た場合のような個人間の潜在意識を調べるのに適用できるのではないかという結論に至った．

本研究を踏まえての今後の展望として，正規分布と指数分布以外にも混合できるようにしていきたいと思う．具体的に言うと，データの中身や量によって混在してくる分布はかなり違ってくると考えられる．そこでどのようなデータでも対応できるようにすることが行うべきこととしての一つ目である．二つ目にどのような異種混合分布モデルの限界適用性に関する検証である．今回の鉄道利用者の到着分布には異種混合分布モデルの適用が最も良いという結論が得られたが，必ずしも異種混合分布モデルで与えれば良いという訳ではないと考えられる．具体的には，仮想のデータから始めていき，最終的には実際の回答データやビックデータ等を用いて検証を行いたいと考えている．

参考文献

1) 藤巻遼平，森永聡：ビックデータ時代の最先端データマイニング，NEC 技法，Vol.65，No.2，pp.81-85，2012.
2) 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-,元田浩,栗田多喜夫,樋口知之,松本裕治,村田昇 訳,丸善出版(株)，2012.
3) 羽藤英二，朝倉康夫，平井千智：不確実性下の意思決定を考慮した逐次的情報参照モデル，土木学会論

文集，No.660，IV-49，pp.27-37，2010.
4) 李成，山本俊之，森川高行：情報探索アプローチと費用便益アプローチを統合した買い物場所選択集合の拡大過程に関する実証分析，土木学会論文集 D 部門，Vol.63，No.1，pp45-54，2007.
5) 高田和幸，鈴木孝典，藤生慎：鉄道の遅延時間を考慮した出発時刻決定行動に関するモデル分析，土木学会論文集 D 部門，Vol.68，No.，I_1071-1077，2012.
6) A.Sumaleea，R.X.Zhonga，T.L.Pana，W.Y.Szetob：Stochastic cell transmission model (SCTM): A stochastic dynamic traffic model for traffic state surveillance and assignment，Transportation Research Part B: Methodological，Volume 45, Issue 3，pp.507-533，2011.
7) Agachai Sumalee, Tianlu Pan, Renxin Zhong, Nobuhiro Uno, Nakorn Indra-Payoong：Dynamic stochastic journey time estimation and reliability analysis using stochastic cell transmission model: Algorithm and case studies, Transportation Research Part C Emerging Technologies, Volume 35, pp.263-285, 2013.
8) Karen L. Nylund, Tihomir Asparouhov Muthén Muthén, Bengt O. Muthén：Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study，Structural Equation Modeling: A Multidisciplinary Journal，Volume 14, Issue 4，pp.535-569，2007.
9) 熊谷徹，赤松幹之：混合ポアソンモデルを用いた高速道路の事故件数解析 - モバイル機器による状況に応じた安全情報の呈示を目指して - ，モバイル学会・学会誌，vol.1(2)，2011.
10) 奥村誠，塚井誠人，カルロス　ナパ　フォンカセ，吉村充功：空港退出自動車交通量に関する混合ハザードモデルの EM 推定，土木計画学研究・論文集，Vol.25，2008
11) 中西航，布施孝志：角度データに着目した歩行者動線分析手法に関する基礎的検討，交通工学論文集（特集号 A），2015
12) 中山晶一朗：道路の時間信頼性に関する研究レビュー，土木学会論文集 D3(土木計画学)，Vol.67，2011.
13) Tibshirani, R., Walther, G., Hastie, T：Estimating the number of clusters in a dataset via the gap statistic. Journal of Royal Statistical Society, Vol.63，Part.2, pp.411-423, 2001.
14) 上田修功：ビッグデータを活かす機械学習技，NTT 技術ジャーナル，31-35, 2013.
15) C. F. J. Wu.：On the convergence properties of the EM algorithm, Annals of Statistics,Vol. 11, pp. 95-103, 1983.

（？ 受付）

ESTIMATION OF ARRIVAL DISTRIBUTION MODEL APPLYING MIXTURE DISTRIBUTION MODEL

Kota MIYAUCHI, Kazuyuki TAKADA

Recent years, "big-data" appeared by Internet of Things (IoT). At present, this analysis method is energetically performed by various fields. Even out of these, there are the estimation methods that applying mixture distribution model. It means that applying the multiple distributions for estimation. The feature of mixture distribution has more expression than single distribution, the estimation of the parameter is a difficulty. In this study, the estimation method of mixture distribution model proposes to apply the arrival distribution for railway commuters. The reason for applying is the arrival situation for railway commuters became the variety. Thus, the object focus on the railway commuters in Tokyo metropolitan area, it was surveyed for their arrival situation on commuter time. It was verified that how mixture distribution model matches the arrival distribution for railway commuters. In this study, it refers to the possibility applying the mixture distribution model.