データフュージョンによる
行動データマイニングのための基礎分析

日下部　貴彦[1]・朝倉　康夫[2]

[1]正会員　東京工業大学助教　大学院理工学研究科土木工学専攻
（〒152-8552 東京都目黒区大岡山2-12-1-M1-20）
E-mail:t.kusakabe@plan.cv.titech.co.jp
[2]正会員　東京工業大学教授　大学院理工学研究科土木工学専攻
（〒152-8552 東京都目黒区大岡山2-12-1-M1-20）
E-mail:asakura@plan.cv.titech.co.jp

交通系ICカードによる処理データは，料金収受にともなって収集されることから，長期間かつ連続的な収集が可能であるという特徴がある．このような特徴は，これまで交通計画で用いられてきたパーソントリップ調査などの調査データにはない特徴であり，交通需要の変動をモニタリングし，施策を検討する際に有用であると考えられる．一方，ICデータは，行動調査を目的としたデータではないために，行動分析で必要な項目が必ずしも観測されていないという面もある．本研究の目的は，交通系ICカードデータをパーソントリップ調査データを用いたベイズ推定により拡張する方法を構築することである．この方法をデータマイニングの一手法である可視化技術とともに実装することにより，データに含まれる交通需要変動の解釈を容易にすることを意図している．

Key Words : *Smart card data, Person trip survey, Datamining, Data fusion, Behavioural analysis*

1. はじめに

交通系ICカードによる処理データ（以下，ICデータと呼ぶ）は，料金収受を目的としたデータ収集を行っていることから，長期間かつ継続的なデータ収集が可能であり，その特徴を活かした分析方法の構築が期待されている．特に，改札を通過したICカード利用者全員の改札通過時刻及びカードを識別できるID情報が記録されることから，これまでに比べて需要の変動要因をより詳細に把握できるのではないかと期待できる．例えば，ID情報が付加されていることにより，従来の磁気券による自動改札機によるデータなどでは分析することができない利用者の利用頻度などの項目を分析できるようなってきているからである．しかし，一方で，ICデータは，交通行動を収集することを目的としたデータではないことから，交通目的やトリップの出発地，目的地など，従来の交通行動を収集することを目的としたデータで収集されている項目を観測できない．このようなことから，ICカード利用者の行動の変動の要因を知ることは必ずしも容易ではなく，分析者が推測するしかないのが現状である．例えば，Morencyら[1]は，クラスタ分析を用いて利用者の利用パターンを分類しているが，その解釈は分析者が行うものとなっている．日下部・朝倉[2][3]は，潜在クラスモデルや隠れマルコフモデルを用いて利用者の行動パターンを抽出しているが抽出されたパターンの解釈には課題が残っている．

既往の交通行動を交通計画で用いられてきたパーソントリップ調査などの交通行動の調査データは，被験者の負担やコストの面から長期間かつ連続的なデータ収集には不向きであった．一方で，ICデータは長期的かつ連続的なデータ収集が可能である．このような特徴は，例えば，データマイニング手法の一手法である可視化技術[4]などと組み合わせて交通需要の変動をモニタリングすることで活用できると考えられる．また，施策や調査の実施の動機付けや検討を行う際に既往のデータと比べてより詳細な変動をとらえられるという点で有用であると考えられる．

本研究の目的は，パーソントリップ調査データ（以下PTデータと呼ぶ）を用いてベイズ推定により交通系ICカードデータを拡張する方法を構築し，可視化技術に実装することである．この方法により，データに含まれる交通需要変動をトリップ目的ごとの変動として解釈することにより容易にし，より直観的にモニタリングを行うことを意図している．

2. 分析方法

本研究では，ICカード利用者の行動文脈を把握するために，PTデータを活用したICカード利用者のトリップ目的の推定方法を構築する．この方法で想定しているICデータで観測されている項目は，乗降駅，乗降時間，カードIDであり，これらの項目が継続的に観測されていることを想定している．一方，PTデータは，ある1日について，トリップの目的，出発地，目的地とともに，乗降駅，乗降時刻を記録しているものである．これらのデータの共通の項目である乗降駅，乗降時刻を用いることで，それぞれ片方のデータでしか観測ができていないトリップ目的と利用頻度などの関係をベイス推定により推定する方法を構築する．具体的には，乗降駅・乗降時刻を元にトリップ目的を推定するモデルをPTデータから作成し，そのモデルをICデータの乗降駅・乗降時刻に適用することにより，ICデータのトリップ目的を推定する．

第一節では，PTデータから作成するモデルについて述べ，第二節では，ICデータでのトリップ目的推定方法について述べる．第三節では，トリップ目的推定の枠組みを用いてICデータで観測されるトリップの頻度とトリップ目的の関係を分析する方法について述べる．

(1) PTデータによるトリップ目的推定モデル

ICデータとPTデータで共通するデータ項目は，乗降駅と乗降時刻である．本研究では，対象駅での降車時刻と，その駅で降車した時刻から次の乗車の時刻までの間隔を入力値とした推定方法を構築する．なお，本研究ではこの間隔を乗降間隔と呼ぶ．鉄道利用者は，降車してから，目的地まで何らかの交通手段を使って行き，何らかの活動を行った後に，また降車駅まで戻ってくるような場合が想定されることから，乗降間隔は，鉄道利用者のトリップ目的や目的地にある程度依存していると考えられる．

乗降間隔は，ICデータを各IDについて日付毎に降車時刻順に並び替えた後，前後のトリップを比較するという手順で求める．図-1は，ICデータの一例を示したものである．並び替えが済んだICデータから乗降間隔を求める手順は，まず，対象となる降車駅のレコードを探索したのち，それ以降のレコードで，その降車駅と一致する乗車駅をもつレコードを探索する．それらのレコードの降車時刻と乗車時刻の差分より乗降間隔を求めるという手順である．この際，同一IDの同一日に降車駅と一致する乗車駅を含むレコードが見つからない場合には，「復路なし」という分類とする． なお，PTデータで乗降間隔を求める際も，ICデータと同様に扱うことで乗降間隔を求めることができる．

ICデータ

ID	日付	乗車駅	乗車時刻	降車駅	降車時刻
A25687DK	2007/10/12	A	7:10	C	7:23
A25687DK	2007/10/12	D	19:22	A	19:36
B68677DS	2007/10/13	A	7:11	C	7:23
⋮					
B67732RR	2007/10/11	A	7:16	C	7:23
B67732RR	**2007/10/13**	**A**	**7:11**	**B**	**7:18**
B67732RR	**2007/10/13**	**B**	**20:20**	**A**	**20:28**
B67732RR	2007/10/14	A	7:05	B	7:13

乗降間隔
同一IDによる，同一日に降車駅と乗降駅が一致するトリップ

図-1 ICデータと乗降間隔

降車時刻Tと乗降間隔Dは離散的に扱う．降車時刻Tは，5～24時台までの2時間間隔で設定し，乗降間隔Dは，0～13時間台の1時間間隔と14時間以上，復路なしの分類とする．

降車時刻Tと乗降間隔Dからそのトリップの目的Aが起こる確率$p_p(A|T,D)$を推定するモデルをPTデータを用いて構築する．この際，PTデータで観測されているデータ数では，ICデータで観測されていると考えられる降車時刻Tと乗降間隔Dの組み合わせが十分に観測されていないことから，$p_p(A|T)$と$p_p(A|D)$に独立性を仮定する．$p_p(A|T)$と$p_p(A|D)$は，それぞれ，PTデータを集計することで求めることができる．このとき，$p_p(A|T,D)$は，

$$p_p(A|T,D)=\frac{p_p(A|T)p_p(A|D)}{\sum_A p_p(A|T)p_p(A|D)} \quad (1)$$

として求めることができる．

(2) ICデータのトリップ目的推定方法

ICデータでは，降車時刻Tと乗降間隔Dを観測している一方で，トリップの目的Aは観測されていない．利用者のトリップ目的がAである確率は，降車時刻Tと乗降間隔Dを式(1)に適用して求める．

ICデータに記録されている，降車時刻Tと乗降間隔Dの利用者のシェアを$p_s(T,D)$とするとき，トリップ目的Aの利用者のシェアは，

$$p(A)=\sum_T\sum_D p_p(A|T,D)p_s(T,D) \quad (2)$$

とすることで推定することができる．

(3) ICデータのトリップ目的と利用頻度の分析方法

ある利用者の利用頻度kは，ICデータからのみ観測することが可能である．そこで，改札通過する利用者の属性が，利用頻度k，降車時刻T，乗降間隔Dである確率を$p_s(k,T,D)$とする．このとき，式(2)のときと同様に考えると，改札通過する利用者の属性が，利用頻度kで目的がAである確率は，

$$p(k,A)=\sum_T\sum_D p_p(A|T,D)p_s(k,T,D) \qquad (3)$$

と表すことができる．また，ある利用目的を持った利用者の利用頻度$p(k|A)$は，ベイズの式の事後確率として表現でき，式(2)と式(3)を用いて，

$$p(k|A)=\frac{p(k,A)}{p(A)} \qquad (4)$$

と表すことができる．

3. 推定結果と可視化による分析例

第三章では，分析に用いたデータについて示したのち，PTデータから作成する乗降間隔からトリップ目的を推定するモデルの推定結果について述べる．第三節では，トリップ目的を推定モデルをICデータに適用し，トリップ目的毎の長期的な変動を可視化技術を用いて分析する．

(1) データ

本章では，都市部に路線をもつ鉄道会社A社a駅で観測されたICデータを用いる．対象の期間は，2007年10月1日から2009年5月31日までの20ヶ月411日間の平日である．この間に観測されたa駅を乗降するトリップ数は，6,913,979トリップであった．PTデータは，2002年に実施された第4回京阪神パーソントリップ調査によるデータを用いる．

(2) PTデータによるモデルの推定結果

図-2は，PTデータより乗降間隔Dからトリップ目的Aを推定するモデル$p_p(A|D)$の推定結果である．推定結果より，乗降間隔が5時間以下の場合には，自由目的のトリップが半数以上を占めていることがわかる．通学目的のトリップは，8時間の場合に半数を占めている．8時間以上の場合には，通勤目的のトリップが7割以上を占めており，13時間以上では，9割以上を通勤目的のトリップが占めている．帰宅目的のトリップは，PTデータでは滞在時間が5時間のときと，同一日に復路がない場合にのみ観測されている．同一の日付で復路がないトリップは，帰宅目的のトリップ以外のトリップでも観測されており，それらの割合は，37.1%を占めている．これは，本研究で対象としている路線以外に平行する他路線を利用者が利用して帰宅しているためと考えられる．

推定結果をICデータに適用する際には，いずれかの乗車間隔Dでシェアが大きいトリップ目的の変動は，より正確にICデータのトリップ目的のシェアの推定結果に反映されると考えられる．一方で，業務目的などシェアが小さいトリップ目的では，そのトリップ目的の変動よりも他のトリップ目的の変動による影響が大きくなってしまうことからICデータから変動を捉えることは難しいと考えられる．例えば，乗車間隔が14時間以上の場合には，通勤目的である確率が92.3%である一方で業務目的は5.2%であることから，実際のICデータトリップで，14時間以上の業務目的のトリップが増えた場合でも，その92.3%は通勤目的のトリップが増えたことの影響として推定されるからである．一方で，実際のICデータトリップで，14時間以上の通勤目的のトリップが増えた場合では，推定結果でも，その92.3%は通勤目的のトリップとして推定されるので，業務目的のトリップが増えた場合と比較してより正確にその変動を捉えることができると考えられる．

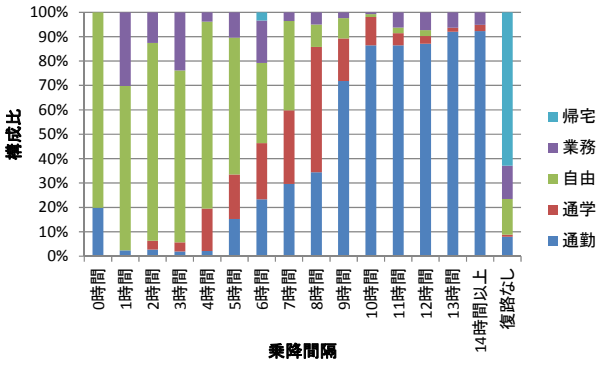

図-2 $p_p(A|D)$の推定値

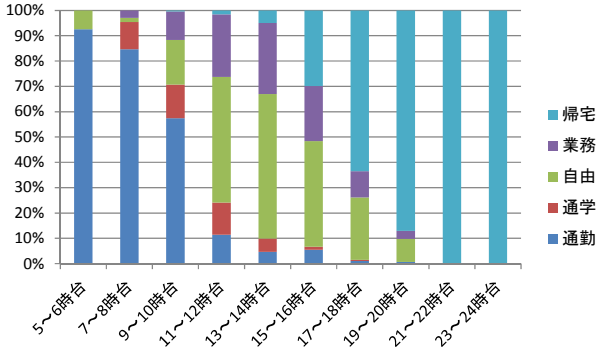

図-3 $p_p(A|T)$の推定値

図-3は，PTデータより降車時間帯Tからトリップ目的Aを推定するモデル$p_p(A|T)$の推定結果である．推定結果より，8時台までの降車したトリップでは通勤目的のトリップが80%以上を占めていることがわかる．11時台から15時台は，6割以上のトリップが自由又は業務トリップであり，17時以降は帰宅目的のトリップが大半を占め，21時以降はすべてのトリップが帰宅目的となっている．

(3) ICカードデータへの適用による長期的変動の分析

図-4は，式(2)のモデルによって推定したトリップ目的毎のトリップ数を示している．この駅では，通勤目的のトリップが多く，帰宅，私用，業務，通学の順に続くことがわかる．私用以外の目的のトリップでは， お盆や年末年始にトリップ数が減少していることがわかる．帰宅目的のトリップは他の目的のトリップに比べて，4月から9月にかけてのばらつきが大きくなっている．これは，この路線の他の駅近辺にあるイベント施設で行われているイベントの期間に一致しており，その帰宅の乗客が他の社局の路線への乗り換えているものを捉えたものだと考えられる．

図-5～7は，トリップ目的別に，各日の時間帯別のトリップ数を可視化したものである．通勤トリップは，8時台をピークに9時台まで多く見られることがわかる．私用目的のトリップは，11時～14時台に多く見られると共に17時台にも増えることが読み取れる．帰宅目的のトリップは，15時以降にトリップの増加が見られ，17時以降に顕著に増加している．また，帰宅目的のトリップは，4月から9月にかけて21時以降にトリップ数の大きな増加が見られる日がある．この増加は，図-4の帰宅目的のトリップで見られた大きなばらつきが発生する日と一致しており，イベントからの帰宅交通が中心であることと推察される．

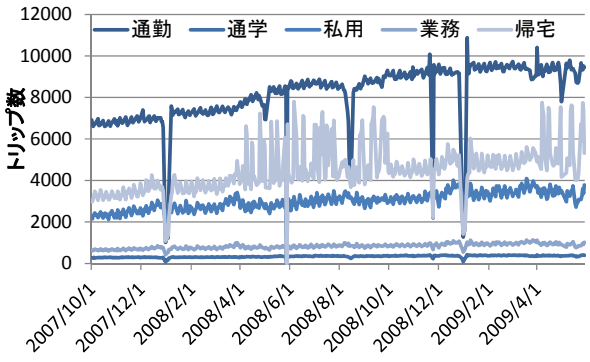

図-4　トリップ目的の推定結果

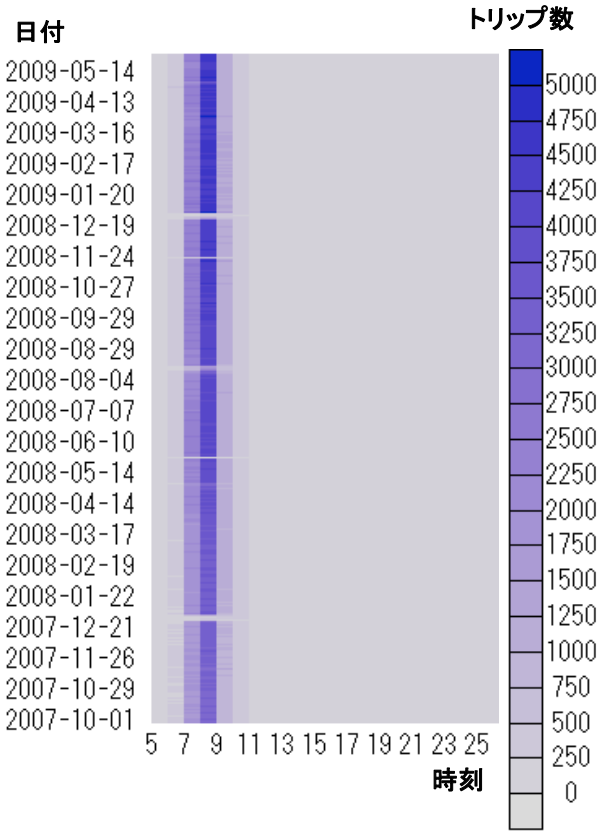

図-5　時間帯別のトリップ目的の推定結果（通勤目的）

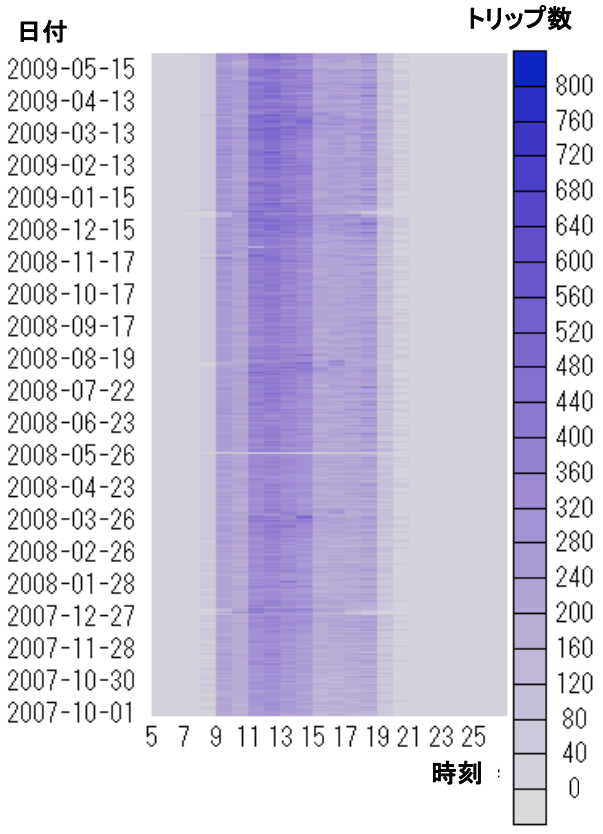

図-6　時間帯別のトリップ目的の推定結果（私用目的）

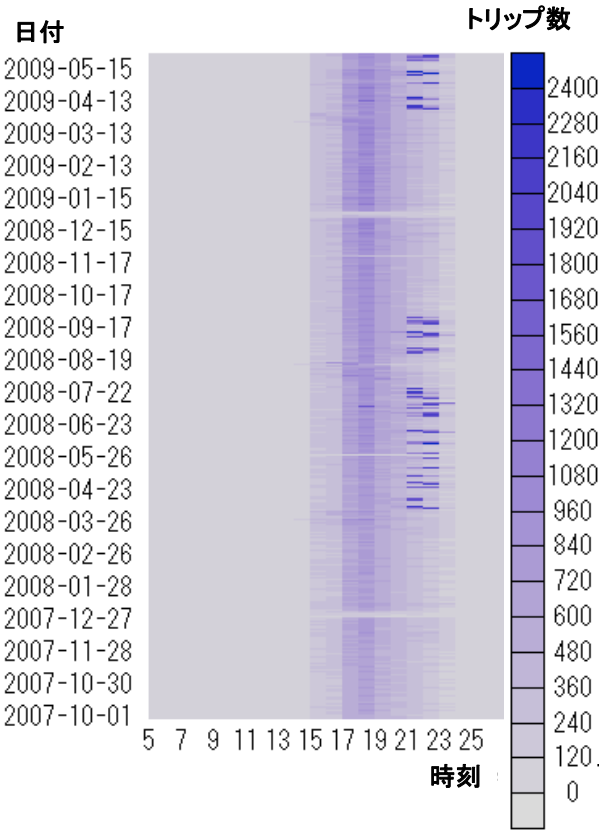

図-7　時間帯別のトリップ目的の推定結果（帰宅目的）

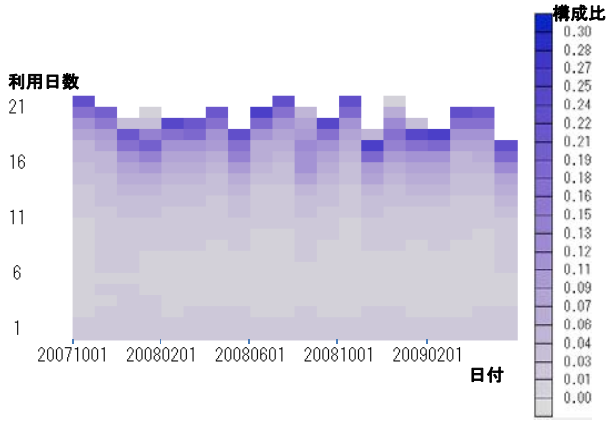

図-8　トリップ目的別利用日数の推定結果（通勤目的）

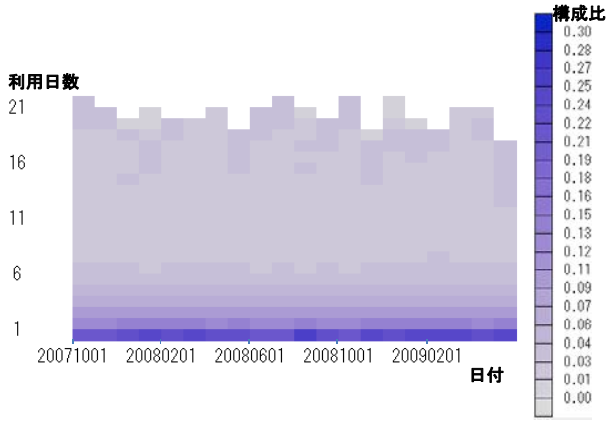

図-9　トリップ目的別利用日数の推定結果（私用目的）

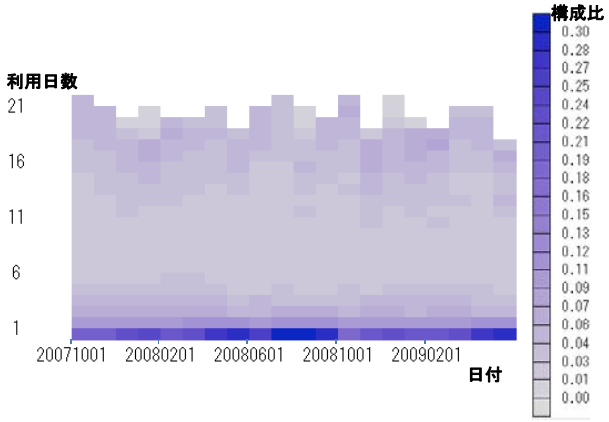

図-10　トリップ目的別利用日数の推定結果（帰宅目的）

図-8～10は，式(4)により求めたトリップ目的ごとの利用日数の月別の分布を示している．通勤目的のトリップでは，利用日数のピークは平日の日数と一致している月が多いことがわかる．年末年始がある12月，1月とお盆のある8月では，最頻値が出現する日数が平日の日数よりも小さくなっている．また，最頻値の値が小さくなっていることからばらつきも大きくなっていることがうかがえる．私用目的のトリップでは，最頻値は，月1日の利用となっており，どの月も約2割以上を占めている．また，通勤目的のトリップのように，大きく傾向の変わる月は見られない．帰宅目的のトリップでも，最頻値は，月1日の利用となっているとともに，月の平日の日数と一致する部分に弱いピークが見られる．この弱いピークの傾向は，通勤目的のトリップでの傾向に似ていることがわかる．このことは，他の駅への通勤からの帰宅トリップが影響しているものと考えられる．

4.おわりに

本研究では，データに含まれる交通需要変動をトリップ目的ごとの変動として解釈するために，ICデータとPTデータをベイズ推定により融合し，分析する手法を構築した．この方法を可視化技術に適用することにより，ICデータでは直接観測できないトリップ目的を把握でき，変動の要因をより直観的に捉えることが可能となった．このようなシステムを活用することで，従来の可視化手法と比較して，より詳細に鉄道沿線のイベントや集客施設による影響などによる変動の要因を推測することができると考えられる．また，特異な変動が見られた場合には，変動があった利用者の利用目的を考慮した上で，より詳細な調査を検討するとともに， 改善施策やマーケティングの検討への展開も見込めるだろう．

参考文献

1) Morency, C., Trépanier, M. and Agard, B.: Measuring Transit Use Variability with Smart-card Data, Transport Policy, Vol.14, No.3, pp.193–203, 2007.
2) 日下部貴彦, 朝倉康夫：生存時間モデルによる交通系ＩＣカードデータの分析, 第 29 回交通工学研究発表会論文報告集, pp.273-276, 2009.
3) 日下部貴彦, 朝倉康夫：IC カードデータを用いた時系列交通行動解析手法の構築, 第 30 回交通工学研究発表会論文報告集, pp.245-248, 2010.
4) 日下部貴彦, 中島良樹, 朝倉康夫：可視化技術をもちいた交通系 IC カードデータの分析, 土木計画学研究・講演集, Vol.39, CD-ROM, 2009.

（2012. 5. 7 受付）

BEHAVIOURAL DATAMINING OF SMART CARD DATA: DATA FUSION APPROACH

Takahiko Kusakabe and Yasuo Asakura

The aim of this study is to develop a datamining methodology to estimate the behavioural contexts of trips and to find their changes observed in smart card data. In order to estimate the behavioural contexts from smart card data, data fusion methodology using person trip survey data is developed. Person trip survey data is used for interpretation of the behavioural contexts. As the results of the analysis, the methodology can illustrate the travellers who cause the change of the demand by seeing the share of the trip purpose and the relationship between the trip frequency and the estimated trip purpose. These results show that time series changes of trip purpose that is difficult to obtain from person trip survey data.