

コンピュータ関数を用いた検索履歴データと乗降履歴データの関連性分析

鳥取大学 学生会員 ○小西 諒
 鳥取大学 正会員 桑野 将司
 鳥取大学 正会員 森山 卓
 鳥取学会 学生会員 細江 美欧

1. 背景と目的

近年、経路検索システムの検索履歴データを活用して、公共交通利用者の検索行動の特性を分析し、移動需要を把握する研究が進んでいる¹⁾²⁾。しかし、検索履歴は将来の利用希望を表すものであり、検索者が実際に公共交通を利用したかどうか不明である。検索件数と乗降件数の直接の関係はこれまでに十分に明らかにされていない。

検索履歴データには、通勤・通学などの日常的な移動需要だけでなく、出張・観光などの非日常的な移動需要も反映されていると考えられている。そのため、検索履歴データから将来の乗降件数を予測できれば、日常的な移動需要に加え、非日常的な移動需要も予測できる可能性があり、実用的な意義は大きい。

本研究では、鉄道路線「ことでん」を対象に経路検索システム「NAVITIME」を用いて検索された検索履歴データと、ことでんで使用されている交通系 IC カード「IruCa」の乗降履歴データの間を、コンピュータ関数によりモデル化する。そして、検索履歴データからの乗降件数の予測を試みる。

2. 使用データの概要

本研究では香川県で運行される鉄道路線「ことでん」を分析対象とし、2014年1月1日～同年12月31日までの12ヶ月間に収集された検索履歴データと乗降履歴データを用いる。各データの有効条件を以下に示す。

a) 「NAVITIME」検索履歴データ

出発地と目的地の双方で、ことでんの駅が指定され、かつ利用指定時刻がことでんの運行時間帯である5時台～24時台であることを条件に課す。得られた有効データ数は198,351件である。

b) 「IruCa」乗降履歴データ

キーワード 経路検索システム、交通系 IC カード、コンピュータ関数、潜在クラスモデル

連絡先 〒680-8550 鳥取県鳥取市湖山町南4丁目101 鳥取大学工学部 情報システム研究室

TEL 0857-31-5313

ことでんの運行時間に利用され、かつ移動所要時間が60秒以上であることを条件に課す。有効データ数は7,265,029件である。

3. 基礎集計分析

図1に日別の検索件数と乗降件数の同時分布を示す。図1より乗降件数約15,000件を境に2つの集団が存在していることがわかる。

図2に平日の検索件数と乗降件数の同時分布を示す。平日では、検索件数が350～399件、乗降件数が25,500～26,499件の日が最も多く、全247日(サンプル)のうち24日である。分布図からは、検索件数が少ない日ほど、乗降件数が多い傾向がみられる。平日では、通勤・通学などの交通系 IC カードの日常的な利用が多く、乗降件数も多いと考えられる。経路検索システムの利用者は普段利用しない経路や時間帯を対象に検索するという特性を踏まえると、普段の平日において検索件数は少ないことが考えられる。しかし、特異な平日においては、乗降件数が少ないにもかかわらず検索件数が多いということが考えられる。そのような特異日のデータに影響され負の関係が検出されていると思われる。

図3に土日及び祝日からなる休日の検索件数と乗降件数の同時分布を示す。休日では、検索件数が450～499件かつ乗降件数が7,500～8,499件の日、検索件数が550～599件かつ乗降件数が7,500～8,499件の日の割合が高い。休日では検索件数の増加に伴い乗降件数も増加する傾向がみられる。経路検索システムの利用特性から、観光・出張などの非日常的な外出行動が多い休日には利用が多くなると考えられる。一方、休日は交通系 IC カードの利用者の通勤、通学などの日常的な利用が少ないため乗降件数も少ない。その結果、休日の乗降件数の多くは非日常的な利用

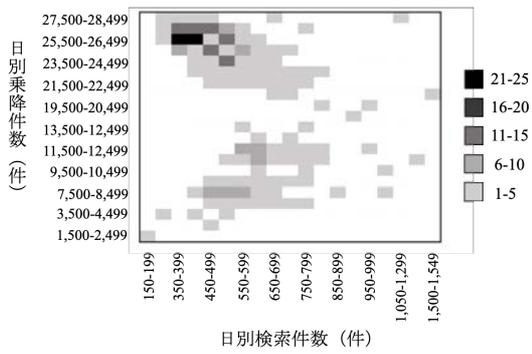
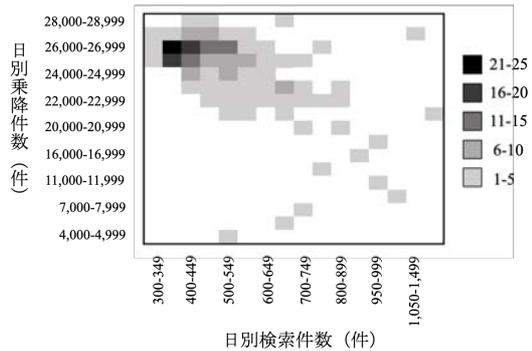
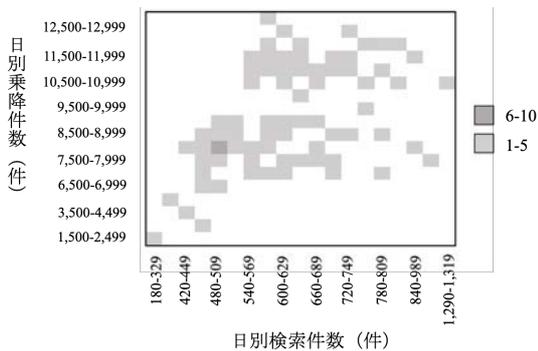


図1 日別の検索件数と乗降件数の同時分布

図2 日別の検索件数と乗降件数の同時分布
(平日)図3 日別の検索件数と乗降件数の同時分布
(休日)

が占めるため、検索件数が多い日に乗降件数も多くなるという正の関係が検出されたと考えられる。

以上の基礎集計分析から、検索件数の増減と乗降件数の増減の関係性は、平日と休日で異なることがわかる。本研究では、正の相関が見られる休日データのみを対象に分析を行う。

4. 潜在クラスモデルを用いたコピュラ関数の概要

a) コピュラ関数

本研究では、検索履歴データと乗降履歴データの関係性を表現するために、コピュラ関数³⁾⁴⁾を用いる。本研究では、異なる依存構造を表現する4つのコピュラ、1) 正規コピュラ、2) クレイトン・コピュラ、

3) ガンベル・コピュラ、4) フランク・コピュラを候補とし、各コピュラについてモデルを推定する。そして情報量基準(AIC)を用いて、最も当てはまりが良いコピュラを選択する。

b) 潜在クラスを用いた2変量コピュラモデル

本研究では、多峰分布を表現できる、潜在クラスを用いたモデルを構築する。

潜在クラス数を k とし、第 h クラスに属する帰属確率を $P_h (h = 1, \dots, k)$ 、第 h クラスを表現する2変量コピュラを $C(u_1, u_2)_h$ とすると、潜在クラスを考慮する2変量コピュラ関数は式(1)で表される。

$$C(u_1, u_2) = \sum_{h=1}^k P_h C(u_1, u_2)_h \quad (1)$$

第 h クラスに属するコピュラの対数尤度関数を $l(\psi_1, \psi_2, \alpha)_h$ とすると、帰属確率 P_h は式(2)で表される。

$$P_h = \frac{\exp(l(\psi_1, \psi_2, \alpha)_h)}{\sum_{h=1}^k \exp(l(\psi_1, \psi_2, \alpha)_h)} \quad (2)$$

式(2)の帰属確率 P_h を用いて、全ての潜在クラスに関わる対数尤度関数を式(3)で表す。本研究では、式(3)で表す対数尤度関数を最大化するパラメータを、EMアルゴリズムにより求める。

$$\ln L = \sum_{h=1}^k P_h l(\psi_1, \psi_2, \alpha)_h \quad (3)$$

4. 推定結果

a) コピュラ関数の決定

本研究では、1) 正規コピュラ、2) クレイトン・コピュラ、3) ガンベル・コピュラ、4) フランク・コピュラの4つのコピュラ関数を用いたモデルの推定結果を比較する。なお、事前分析の結果から、周辺分布には潜在クラスを考慮するワイブル分布を仮定する。

潜在クラス数別に4種類のコピュラを仮定したモデルの推定結果を表1に示す。表1では、潜在クラス別での適合度が最も高いコピュラの組み合わせを示している。推定結果から、クラス数4以上は検証せず、クラス数が2のフランク・コピュラとクレイトン・コピュラの組み合わせを最適なモデルとして選択する。

表1 モデル推定結果

潜在クラス数	コピュラ	AIC lnL 平均帰属確率
なし	クレイトン	-26.831 18.42
2	フランク (クラス1) クレイトン (クラス2)	-76.557 48.28 50.51% : 49.49%
3	フランク (クラス1) クレイトン (クラス2) クレイトン (クラス3)	-55.409 42.70 32% : 36% : 32%

表2 推定結果から得られたパラメータ値

コピュラ	パラメータ推定値	t値
フランク・コピュラ (クラス1)	1.8807	4.16 ***
クレイトン・コピュラ (クラス2)	1.7781	1.67 *

* : 10%有意, ** : 5%有意, *** : 1%有意

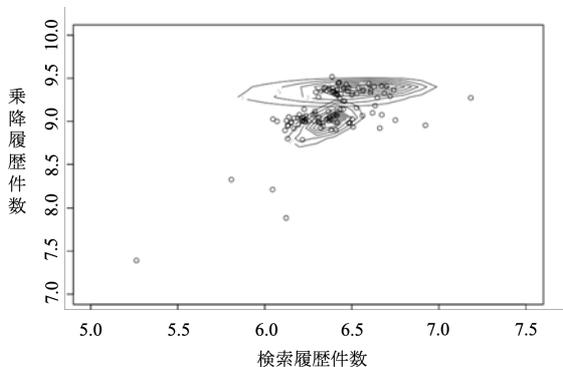


図4 選択したコピュラの同時密度関数

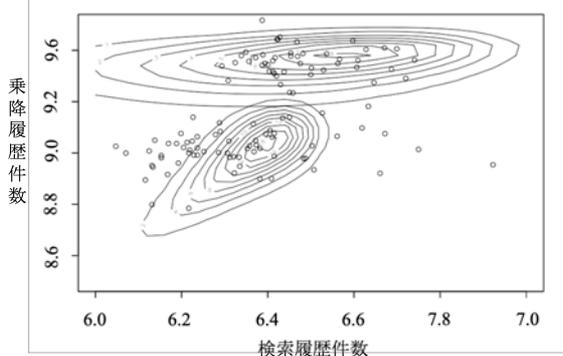


図5 選択したコピュラの同時密度関数 (拡大図)

モデル適合度が最も高いと判断されたフランク・コピュラ×クレイトン・コピュラのモデルに対して、推定されたパラメータの値を表2に示す。また、パラメータの値から算出される検索件数と乗降件数の同時確率密度を図4に、図4を拡大した図を図5に示す。図4、図5にある2つの分布の内、上側の分布がクラス1に属するフランク・コピュラを、下側の分布がクラス2に属するクレイトン・コピュラを表している。

クラス1に属するフランク・コピュラの同時確率分布の形状は、分布の中心から左右に広がっている。また、中心から右側の方が等高線の間隔が狭く、左側に比べて依存度合いが高いことを示している。確率密度は検索件数が6.5~6.7かつ乗降件数が9.3~9.4のときにピークを持つ。

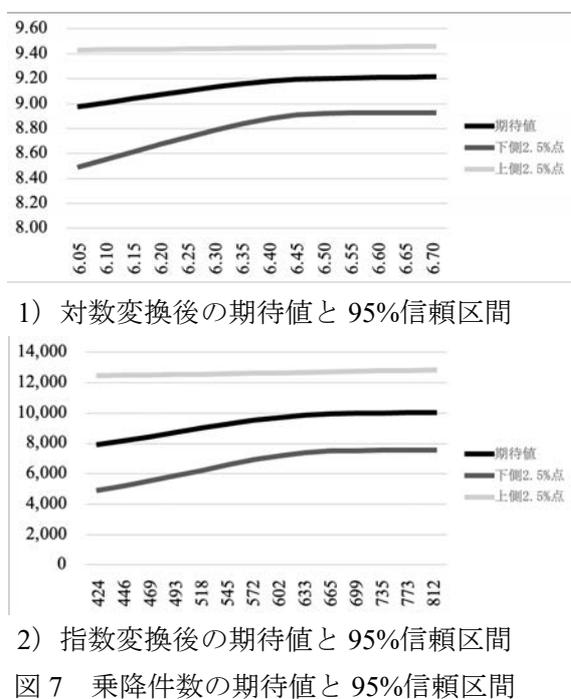
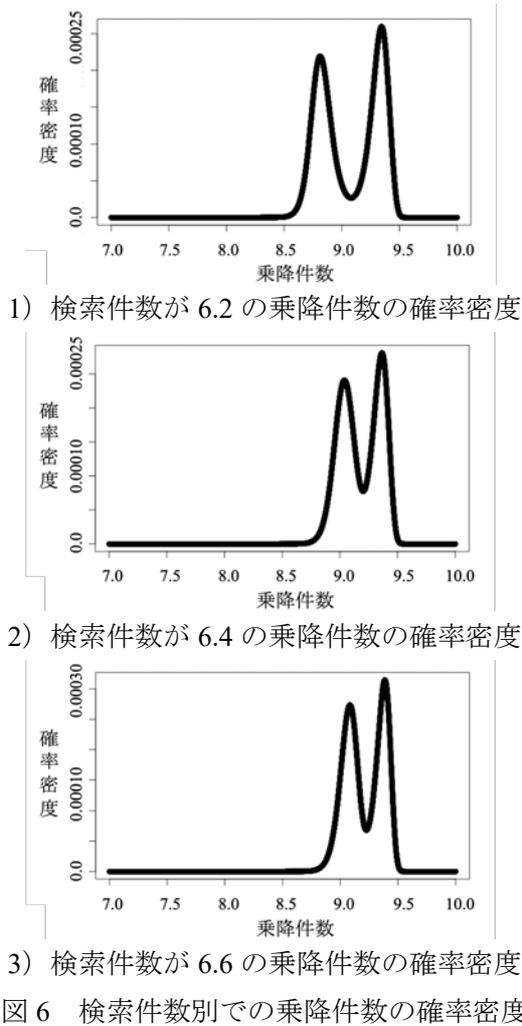
クラス2に属するクレイトン・コピュラの同時確率密度は、分布の左下側と右上側で等高線が広く、分布の中心で狭い。検索件数が6.4~6.5かつ乗降件数が9.0あたりのときに確率密度はピークを持つ。

b) 検索件数による乗降件数の予測結果

検索件数と乗降件数の依存関係に、フランク・コピュラおよびクレイトン・コピュラを仮定したモデルの推定結果を用いて、1) 検索件数を6.2とした場合、2) 検索件数を6.4とした場合、3) 検索件数を6.6とした場合の乗降件数の確率密度を図6に示す。なお、ここでの検索件数は対数値であるため、スケールを戻すと、それぞれ検索件数が493件、602件、735件のときに対応する。

確率密度の比較結果から、検索件数の値によって2つの分布のピークの離れ度合いが異なることが明らかとなった。また、検索件数の増加に伴い2つの確率密度が接近するが、検索件数が6.4以降ではほとんど変化しないことがわかる。

図6に示す確率密度から得られた乗降件数の期待値と95%信頼区間の関係を、図7のグラフに示す。図7では、それぞれ、1) 元の検索履歴データを対数変換した場合での期待値と95%信頼区間、2) 「1)」による推定値を指数変換し、実際の乗降件数のスケールに戻した期待値と95%信頼区間を示している。1)に着目すると、検索履歴データの平均値である6.38を境に、期待値の増加の傾きが変化することがわかる。また、95%信頼区間の下側2.5%点と上側2.5%点に着目すると、6.3に近づくにつれて信頼区間の幅が狭まり、それ以降ほとんど横ばいになっている。これは、図5で示すように、検索履歴件数の増加に伴い2つの確率密度が接近し、検索履歴件数が6.3以上では、確率密度の形状がほとんど変化しないためである。2)も同様に、検索履歴データの平均値である602.63件を境に乗降件数の期待値の傾きが変化しており、全体的に1)とほとんど同じ形状となっている。



以上の結果から、検索件数の増加に伴い乗降件数の期待値も増加することがわかる。また、検索件数の平均値である6.38以下では、推定される乗降件数の95%信頼区間の幅が広く、モデルの予測精度は低い

ことが考えられる。一方で、6.38以上では乗降件数と95%信頼区間ともにほとんど横ばいで推移しており、モデルの予測精度は比較的高いことが考えられる。

5. 結論

本研究は経路検索システム「NAVITIME」の検索履歴データと鉄道路線ことでんで使用されているICカード「IruCa」の乗降履歴データを用いて、両者の関係性を分析した。正の相関が見られる休日データのみに着目し、得られたモデルを用いて乗降件数の期待値を算出したところ、検索件数の増加に伴い乗降件数の期待値も増加する結果が得られた。95%信頼区間に着目すると、検索履歴データの平均値付近を境に期待値と95%信頼区間の両方でグラフの傾きが変わっており、検索件数が平均値以上のときに、より良い精度で乗降件数を予測できることがわかった。本研究によって、検索履歴データと乗降履歴データの相互依存関係がコンピュータ関数によって表現可能である点や、検索件数が平均値に近づくと95%信頼区間の幅が狭まることが示された。

今後の重要な課題として、対象期間を延長することによってサンプル数を増やす、あるいはことでん全体ではなく路線や起終点を限定したうえでモデルを構築するなど、予測精度を向上させるための方法について検討する必要がある。

参考文献

- 1) 伊藤昌毅, 見生元気, 川村尚生, 菅原一孔: 乗換案内サービス利用履歴から探る公共交通の利用状況, 第9回電子情報通信学会ヒューマンプロープ研究会, CD-ROM, 2013.
- 2) 香川喬之, 桑野将司, 福山敬, 谷本圭志, 川村尚生: バス経路検索履歴データを用いた移動特性の分析, 交通工学論文集, 第2巻, 第2号(特集号A), pp.A_115-A_124, 2016.
- 3) 戸坂凡展, 吉羽要直: コピュラの金融実務での具体的な活用方法の解説, 日本銀行金融研究所, 金融研究, pp115-162, 2005.
- 4) 桑野将司, 藤原章正, 塚井誠人, 張峻屹, 岩本真由子: コピュラを用いた自動車保有期間と走行距離の同時決定モデルの開発, 土木学会論文集D, Vol.66, No. 1, pp. 54-63, 2010.