携帯電話の GPS ログデータを用いた,人々の行動パターンの分類

西村隆宏,秋山祐樹,金杉洋,Teerayut Horanont,柴崎亮介,関本義秀

Analysis and Evaluation of Human Lifestyle Pattern Using Mobile phone GPS in Japan

Takahiro Nishimura, Yuki Akiyama, Hiroshi Kanasugi, Teerayut Horanont, Ryosuke Shibasaki and Yoshihide Sekimoto

Abstract: In recent years, needs of Japanese customers are getting diversified. As a result, large retail store has advantage for other store because of considering customer's taste and lifestyle referred big data such as questionnaire survey and research on point of sale system (POS system). On the other hand, we can monitor locations of mobile phone's users regularly in broad area using GPS log data by mobile phones. Therefore, GPS log data by mobile phones is increasing use for research on person flow. In order to classify lifestyle pattern, information of each people's stay-points is very important. In this study, we focus on wayside of Tokyu Railway Corporation and extract person who live wayside of their railway based on stay-points. Moreover, we used the Cameo Code which can monitor regional characteristics published by Nikkei research Corporation to estimate user segmentation. Finally, we divide users to some group and can classify users in more detailed segmentation.

Keywords: GPS ログデータ(GPS log data), 分類(Clustering), 人の流れ(person flow)

1. はじめに

近年,顧客の嗜好の多様化により,Amazon や楽天をはじめとするオンラインショッピングの台頭や,少量生産少量消費社会へ移行する変化が起きている.その結果,局所的な需要に応える製品を発売する製造業が増えている.また,大型商業施設では,顧客の生活パターンや嗜好を考慮したマーケティングを行うことで競合と差をつけている.こうした情勢の変化により,既存商業地域における店舗の閉店や衰退が問題になっている.需要調査のためにアンケート調査やPOSデータ分析といった手法があるが,広域の結果を正確に得ることは難しい.

一方で携帯電話内蔵のGPSデータを大量に集計することによって,時々刻々と変化するユーザーの行動が明らかになりつつある.携帯電話の GPS データはユーザーの位置情報を定期的に蓄積し,大量の情報を広域に渡って収集できるそのため,人々の行動調査にはGPSデータを利用するものが近年見られるようになってきた.モバイル広告業界においても同様の傾向があり,ユーザーの位置情報を元に最適な広告表示を行うことでコンバージョン率を増加させる企業が増加している.また,山本ほか(2006)は新宿御苑内の行動パターンをGPS受信機の利用で明らかにした御苑内を複数のエリアに分け,通過状況を元に行動パターンの分類を行った.

本研究では,携帯電話のGPSデータを利用し,ユ

西村隆宏　〒277-8568　千葉県柏市柏の葉 5-1-5
東京大学大学院新領域創成科学研究科
Phone: 080-4337-5498
E-mail: nishimura@csis.u-tokyo.ac.jp

ーザーの行動パターンの分類を行う.行動パターンの分類を行い,POS データやアンケート調査と併用することでより細かな人々の分類を試みる.

2. 手法の概略

2.1 本研究で用いたデータセット

本研究では 3 つのデータセットを用いた. 1 は株式会社ゼンリンデータコム社提供の混雑統計®データである. このデータは,位置情報を最短5分間隔で取得した非集計データである. このデータの内,本研究では 2010 年 8 月 1 日から 2011 年 7 月 31 日の期間のデータを利用した. 二つ目は日経リサーチ社提供の Cameo コードである. Cameo コードとは,英国GMAP コンサルティング社が開発したエリアセグメントコードであり,国勢調査等を利用し,日本全国の町丁目を地域の消費水準,家族構成,住居形態などを元に分類している. Cameo コードは後述する人々の居住地属性の決定に利用した. 三つ目は,総務省統計局が提供している事業所・企業統計調査である. このデータは日本全国を 1/2 地域メッシュ(500m)に区切り,メッシュ内の事業所数と従業員数を集計した表である. 事業所・企業統計調査は後述する,人々の滞留点属性の決定に利用した.

2.2 解析方法

2.2.1 滞留点の推定と居住地エリア判定

まず,混雑統計のデータから滞留点を推定した. 滞留点の算出方法は秋山ほか、羽田野ほかの方法を元に行った.推定された滞留点から,各人の居住地エリアの推定を行った.居住地エリアの推定に関しては Horanont ほかの方法を元に行った.居住地エリア推定の結果から,解析に利用したサンプルを抽出した. 対象地域は東京急行電鉄の沿線から 2km 圏内とし、そこを居住地エリアとする人々を分析対象として抽出した.

2.2.2 居住地エリアの属性と滞留点属性の決定

滞留点の推定と居住地エリア判定により,各人の滞留点を居住地エリアとその他の滞留点に分類できる.この内,居住地エリア属性の決定には Cameo コードを利用し,表 1 は Cameo コードの分類の説明とコード別所属割合である.表 1 より,グループ No.6 のデータの所属人数が非常に少ない. そのため本研究では解析対象とせずにグループ No が 5 までの人々を対象に解析を行った.

表 1 Cameo コードの分類と所属人数

グループNo	グループ特徴	所属人数
1	裕福な単身・二人世帯の多い都会地域	1141
2	裕福な中高年の多い地域	554
3	裕福なファミリーの多い地域	68
4	比較的裕福な単身者の多い地域	212
5	ホワイトカラー・2世帯住宅の多い地域	105
6	平均的な中高年の多い地域	4

また,その他の滞留点属性の決定には事業所・企業統計調査からメッシュを5つに分類したものを用いた.分類には非階層的クラスタリング手法の一つである k-means 法を用いた.K-means 法はあらかじめクラスタ数を決定しなければいけない欠点があるが,大規模データの分類を高速で効率的に行える利点がある.クラスタ数を変えて数回分類を行い,最適な分類数を決定した.

2.2.3 各居住地属性のユーザークラスタリング

以上の操作により,各人の滞留点は居住地属性のよる Cameo コードと商業コードに変換される. 次に各人の商業コード訪問頻度を求めた.訪問頻度に変換することでデータが正規化され,ユーザーの分類が行える.Cameo コード別に階層的クラスタリングを行った.距離の算出にはユークリッド距離を用い,クラスタリングにはウォード法を用いた.この操作により,デンドログラムが 5 つ作成される.

2.2.4 各居住地属性の標準行動の決定

Cameo コード毎に作成された5つのデンドログラムから,適切な位置で切断し,クラスタ数を決定した.切断した各デンドログラムから,最もクラスタが大きいものを標準行動クラスタと命名し,それ以外を非標準行動クラスタと命名した.このような命名の理由は,Cameo コードのグループ No. がその地域に居住している人の大部分の傾向を

示すからである.この特性を利用して最もサイズが大きいクラスタはその居住者コードの行動パターンを示していると仮定した.

2.2.5 クラスタ間の類似判定とユーザー属性の再決定

2.2.4 の操作により各 Cameo グループの標準行動パターンが決定された.次に全てのクラスタに対し,標準特徴量を算出した.標準特徴量とは,クラスタに属する人々の各商業コードの頻度平均である.各クラスタの標準特徴量を元に,標準行動クラスタと非標準行動クラスタの類似度を求めた.類似度にはコサイン類似度を用いた.これは,特徴ベクトルの角度の大きさによって類似度を判定するものであり,クラスタや文書の類似度を求めるのに頻繁に利用される手法である.標準行動クラスタと非標準行動クラスタ間の類似度をとることにより,非標準行動クラスタの行動パターンがどの Cameo コードの行動パターンと似ているのか推定できる.

3. 解析結果

解析対象の人々の居住地エリアを黄緑色で示し,Cameo コード別に色分けした結果は図 1 の通りである.

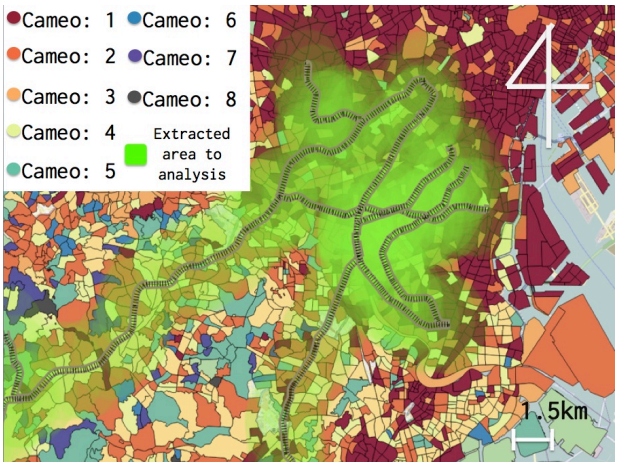

図 1 解析対象の人々の居住エリアと Cameo コードによる町丁目の分類

さらに,表 2 は滞留点を変換し,商業コードで集計したデータの一部であり,表 3 は各居住地属性のクラスタリング結果である.表 3 において,太字が標準行動クラスタを示す.

表 2 集計表の一部

UserID	Cameo	1	2	3	4	5
358	2	31	88	214	54	129
528	2	0	7	264	357	4
811	1	71	32	133	23	58
1290	1	15	40	484	67	82
1305	1	29	30	136	7	687
2028	1	288	205	428	3	171
2069	2	39	69	342	488	404
2115	4	14	77	664	121	69
2142	4	14	92	983	635	159
3381	1	3	14	371	43	274

表 3 より,クラスタ数に大きな偏りが見られる.ここで標準クラスタ間の類似傾向をみるために相関行列を作成した.結果は表 4 の通りである.

表 3 居住地属性毎のクラスタリング結果

居住地属性	人数	居住地属性	人数
1-1	**495**	3-3	17
1-2	138	3-4	15
1-3	179	4-1	54
1-4	330	**4-2**	**95**
2-1	153	4-3	31
2-2	163	4-4	78
2-3	87	4-5	54
2-4	**169**	5-1	21
2-5	82	**5-2**	**48**
3-1	**20**	5-3	44
3-2	16		

表 4 標準クラスタ間の相関行列

	1	2	3	4
2	0.172540596			
3	0.783662468	0.585968344		
4	0.991754303	0.136298728	0.717427681	
5	0.066505029	−0.41734219	−0.25536902	0.165158776

表 4 より,Cameo コード 1 番と 4 番（以下 1 番と 4 番）の行動傾向が似ていることがわかる.表 1 より,1 番は比較的な裕福な単身者,2 人世帯の多い都会地域であり,4 番は比較的な裕福な単身者の多い地域であるため,1 番と 4 番の標準行動クラスタは裕福な単身者の行動パターンであるとわかる.一方 1 番の非標準行動クラスタの一つは 2 人世帯の行動パターンを表していると推定される.ま

た,3 番と 4 番の標準行動クラスタの平均特徴量の相関がやや高いことから,裕福な人々も似た行動パターンをとることが推定される.さらに年齢によって行動パターンが変化することも表 4 の Cameo コード 1 番と 2 番の相関係数を見ることでわかる.次に各クラスタの平均特徴量から,非標準行動クラスタと各標準行動クラスタの類似度の表を示す.

表 5 類似度表

		標準行動クラスタ				
		1	2	3	4	5
非標準行動クラスタ	1−2	0.440057096	0.394977235	0.574737847	0.453138992	**0.684569182**
	1−3	0.60794733	0.587165794	0.624735585	**0.74786782**	0.742123266
	1−4	0.640440617	0.576376588	**0.845200894**	0.632930027	0.844997233
	2−1	0.634997425	0.57307574	**0.834950105**	0.640575383	0.809195296
	2−2	0.753115334	0.749637714	0.668308916	**0.848935887**	0.696439975
	2−3	0.370151509	0.346963435	0.43090645	0.559080583	**0.635917779**
	2−5	0.542224353	0.49552683	0.674992157	0.582674012	**0.790831433**
	3−2	0.57108514	0.555324512	0.570306216	**0.722317848**	0.693600007
	3−3	0.46687456	0.408551054	0.661963574	0.503661592	**0.802674947**
	3−4	0.893994748	**0.899456585**	0.741455329	0.858869817	0.591716972
	4−1	0.878003913	**0.892117356**	0.687964703	0.847221066	0.52833559
	4−3	0.518645689	0.447255419	**0.783647374**	0.505825537	0.694596312
	4−4	0.587999818	0.540486722	0.724137077	0.682511701	**0.87307659**
	4−5	0.493198106	0.482821685	0.477848181	**0.661652448**	0.62530982
	5−1	0.444493579	0.431433052	0.446955997	0.622969781	**0.629056185**
	5−2	0.896688727	**0.897977069**	0.764573474	0.877915268	0.63694907

表 5 より,非標準行動クラスタの大部分はクラスタ自身の Cameo 居住地コードと異なる地域の標準行動クラスタと類似度が高いことがわかる.つまり,表 3,表 5 から地域のマイノリティグループがわかる.例えば,Cameo コード 1 番の地域には平均的な中高年者が 12%,裕福なファミリーが 15%含まれているとわかる.しかし,東京急行電鉄沿線居住者のみ対象にしたため,Cameo コードの地域数の差が非常に大きい.沿線環境が比較的裕福な地域であることは図 1 で明らかなので,比較的裕福でない人々が含まれる割合を示せない.ただ,裕福な人々に限定すると大まかな傾向を把握できる.

4. おわりに

本研究では東京急行電鉄沿線に居住地エリアを持つ人々を対象に携帯電話の GPS ログデータから推定された滞留点を利用し,地域の居住者傾向と,滞留点の商業コードを元に行動パターンの分類を行った.その結果,裕福な人々に限定して言えば分類が行えた.しかし,沿線環境が裕福な人々が多く居住している地域が多いため,比較的裕福でない人々の行動パターンは推定が困難だった.また,地域の分類に事業所数と従業員数のデータだけでは十分な分類ができたといえず,詳細な分類が必要になる.今後の展望として,日本全国を対象に同様の解析を行うことや事業所・企業統計調査の他にもテレポイントデータという電話帳の電子化データを用いて地域の分類を行い,滞留点の滞留時間を考慮に入れることで,より詳細な人々の分類が行えると考えられる.

謝辞

本研究を進めるにあたり, 株式会社ゼンリンデータコムは混雑統計®データを, 株式会社日経リサーチはエリアセグメントコード Cameo を提供して頂きました. ここに感謝の意を申し上げます.

参考文献

[1] 特許庁、特許審決公報 管理番号 1207226

[2]山本泰裕,伊藤弘,小野良平,下村彰男(2006) : GPS を用いた新宿御苑における利用者の行動パターンに関する研究,日本造園学会誌 69, 601-604

[3] Akiyama, Y., Takada, T. and Shibasaki, R., 2013, "Development of Micropopulation Census through Disaggregation of National Population Census", CUPUM2013 conference papers, 110.

[4] 羽田野真由美・上山智士・秋山祐樹・Horanont Teerayut・柴崎亮介, 2012 年, 「GPS データを用いた商業集積地来訪者の行動パターン抽出方法の検討」, 第２１回地理情報システム学会講演論文集（CD-ROM, F-3-4）

[5] Teerayut Horanont (2010) : “A Study on Urban Mobility and Dynamic Population Estimation by Using Aggregate Mobile Phone Sources”, 東京大学大学院工学研究科社会基盤学専攻,博士論文

[6] 秋山祐樹,Teerayut Horanont,柴崎亮介 (2013) : 大規模人流データを用いた商業地域における来訪者数の時系列分析,第 22 回地理情報システム学会講演論文集掲載予定