G-38

# ID-POS データを用いた非会員顧客の年齢と性別の推定

## Estimation of Age and Gender of Non-Member Customers Using ID-POS Data

田井　紗瑛子 †　吉野 孝 †　貴志　祥江 ††　坂本　明一 ††　宮崎　裕之 ††　大西　剛 ††
Saeko Tai　Takashi Yoshino　Sachie Kishi　Akikazu Sakamoto　Hiroyuki Miyazaki　Takeshi Onishi

## 1. はじめに

新型コロナウイルスの感染拡大の影響を受け，ライフスタイルに欠かせない小売業界の需要が上がった．2020 年同様に，2021 年は外食自粛と内食の傾向が継続したので，2019 年と比べると，2021 年の小売業界の実績は 105.8 %にもなる [1]．コロナ禍に入る前の 2019 年以前の数年間は，小売店における消費支出は減少傾向にあった．これは，ドラッグストアをはじめとする他業態との競合や，消費者のライフスタイルの変化による買い物頻度の低下が考えられる [2]．また，人口減少に伴い，小売店の顧客数も減少している．特に，規模の小さな市区町村では人口減の影響で将来的に食品市場が現在の半分以下に縮小する自治体が出てくると予想されている [3]. そのため，小売業界には，生き残りをかけた改革が必要になる．

近年，スーパーマーケットにおいて，ポイントカードの導入率は 83.5 %にも上る [1]．ポイントカードを使用し，顧客の情報を ID-POS データと参照させることで，店内をどのように見て回ったのか示す動線に加え，買ったもの，買わなかったものを可視化できる [4]．それらの分析から見えてくる購買行動を基に，店内のレイアウトやプロモーション内容の企画・提案が可能になる．

しかし，全員が会員になっているわけではない．ポイントカードを煩わしいと思い作らない人や，提示しない顧客も多い．また，複数のポイントカードの利用も増加しており，非会員顧客の割合はますます増加している．そのため，ID-POS データの中に，年齢や性別などの情報が抜けてしまい，データを十分に活用できない．そこで，我々は，会員顧客の情報を分析して，非会員顧客の情報を推定することで，さらに ID-POS データの価値を向上できるのではないかと考えた．

本研究では，ID-POS データを用いて，会員顧客の購入情報を分析することで，非会員顧客の購入情報から，非会員顧客の年齢層と性別の推定を試みる．

## 2. 関連研究

石橋らは，商品 DNA と呼ばれる手法を用いて，販売する商品に関して質問紙調査等を行い,「健康志向」という消費者特性と商品特性を付与し，それを購買した消費派にも同様の特性があると考えた．スパース推定法という統計モデリングを行い，質問紙調査に回答していない消費者であっても，購買履歴データがあれば健康意識を推定できると報告した [5]．本研究では，ID-POS データのみを用いて，顧客情報を推定するため，情報収集する点が異なる．

下田らは，消費者の購買行動モデルを用いたシミュレーションを目的として，あるスーパーマーケットの購買データ，顧客データ，店舗データ，商品データを用いた購買行動モデルを行った．消費者エージェントが食料品を購入する際に，店舗までの距離や商品の価格，年齢や性別，そして家族構成などを考慮した上で，店舗と商品を選択していると考えられるので，購買データを用いて世帯構成や購買傾向の推定，さらには消費者エージェントの商品購入シミュレーションを実施した [6]．この研究では，非会員の推定は行っていないので，本研究とは推定したい目的が異なる．

## 3. 分析概要

### 3.1 使用データ

本研究では，株式会社オークワで収集された 2017 年から 2021 年の各年 8 月 21 日から 12 月 20 日の ID-POS データを使用して，分析を行った．

ID-POS データとは，商品をレジに通したときに取得する POS データに，だれが購入したのかわかる ID 情報を付与したものである[*1]．このデータの中には，日付，RECNO[*2]，加工コード[*3]，年代，性別，各商品の部門，ライン，クラスのコードや JAN コード[*4]，商品の販売価格[*5]などの詳細情報が記載されてある．

商品は，部門，AU，ライン，クラス[*6]で分類されてあり，該当する分類名が 1,276 種類記載されている．

本研究では，商品を部門，AU，ラインまで分類したデータを利用して分析を行った．

### 3.2 分析手法

最初に，CSV 形式の顧客の購買リストの情報が与えられている ID-POS データを読み込む．ID-POS データに記載されている購買情報を基に，顧客の性別，年齢の情報を学習し，非会員の情報を推定する．

### 3.3 分析方法

本研究では,「性別」と「年齢層」の推定を行った．
以下にデータの詳細を記載する．

(1) 性別

男性，女性の 2 種類の分類

(2) 年齢層

10〜20 代 (ヤング層), 30〜50 代 (ミドル層), 60〜80 代 (シニア層) の 3 つに分類

† 和歌山大学，Wakayama University
†† 株式会社オークワ，Okuwa Co., Ltd.

[*1] 匿名データである
[*2] レシートに 1 枚ずつ付けられた番号
[*3] RECNO に紐づき，会員である顧客を識別するための番号
[*4] 商品を識別するために付けられた番号
[*5] 特売や見切りの情報を含む
[*6] POS データは階層で分類されており，部門・AU・ライン・クラスの順に細かくなっている

表 1: 同年 9 月を学習データとした場合の精度

| 学習データの年月 | 推定するデータの年月 | 性別の精度 | 年齢層の精度 |
|---|---|---|---|
| 2019 年 9 月 | 2019 年 10 月 | 0.78(k=17) | 0.58(k=17) |
| | 2019 年 11 月 | 0.79(k=15) | 0.59(k=17) |
| | 2019 年 12 月 | 0.79(k=15) | 0.59(k=19) |
| 2020 年 9 月 | 2020 年 10 月 | 0.78(k=15) | 0.57(k=17) |
| | 2020 年 11 月 | 0.76(k=13) | 0.57(k=15) |
| | 2020 年 12 月 | 0.78(k=15) | 0.57(k=17) |
| 2021 年 9 月 | 2021 年 10 月 | 0.78(k=19) | 0.59(k=22) |
| | 2021 年 11 月 | 0.79(k=19) | 0.61(k=22) |
| | 2021 年 12 月 | 0.79(k=19) | 0.63(k=25) |

表 2: 前年 9 月を学習データとした場合の精度

| 学習データの年月 | 推定するデータの年月 | 性別の精度 | 年齢層の精度 |
|---|---|---|---|
| 2018 年 9 月 | 2019 年 10 月 | 0.80(k=21) | 0.58(k=13) |
| | 2019 年 11 月 | 0.80(k=25) | 0.58(k=17) |
| | 2019 年 12 月 | 0.79(k=18) | 0.58(k=17) |
| 2019 年 9 月 | 2020 年 10 月 | 0.79(k=17) | 0.58(k=15) |
| | 2020 年 11 月 | 0.78(k=15) | 0.58(k=15) |
| | 2020 年 12 月 | 0.80(k=21) | 0.57(k=15) |
| 2020 年 9 月 | 2021 年 10 月 | 0.77(k=15) | 0.57(k=17) |
| | 2021 年 11 月 | 0.76(k=13) | 0.58(k=15) |
| | 2021 年 12 月 | 0.78(k=15) | 0.57(k=17) |

本稿では，8 月 21 日から 9 月 20 日[*7]のデータを 9 月のデータ，9 月 21 日から 10 月 20 日のデータを 10 月のデータ，10 月 21 日から 11 月 20 日のデータを 11 月のデータ，11 月 21 日から 12 月 20 日のデータを 12 月のデータとしている．

### 3.4 調査項目

今回は，2 つの調査を行ったので，それぞれを「調査項目 1」,「調査項目 2」とする．

(1) 調査項目 1

精度が最も高くなるときの学習データは何か

(2) 調査項目 2

非会員の割合はどうなっているのか

## 4. 調査項目 1：精度の調査

もし，各店舗や，各月のデータを学習データとして調査し，精度に変化が見られないのであれば，学習データの固定が可能と考えた．そこで，会員割合の高い和歌山県にある A 店のデータを用いて，精度を調べた．

### 4.1 予備検討

#### 4.1.1 商品の分類方法

今回は，商品を部門，AU，ラインまでに分類して行った．

#### 4.1.2 予備検討

商品を部門，AU までに分類して行った場合の精度は，性別の精度が 0.78(k=17)，年齢層の精度が 0.61(k=19) であった．

商品を部門，AU，ライン，クラスまでに分類して行った場合は，性別の精度が 0.78(k=25), 年齢層の精度が 0.61(k=39) であった．

どちらも 2021 年 9 月の A 店のデータを学習モデルとした場合，同じ店舗の 2021 年 10 月のデータを推定した結果である．どちらも精度に大きな違いは見られなかった．

[*7] データ提供先の企業のデータ管理方式のため

表 3: 前年同月を学習データとした場合の精度

| 学習データの年月 | 推定するデータの年月 | 性別の精度 | 年齢層の精度 |
|---|---|---|---|
| 2017 年 9 月 | 2018 年 9 月 | 0.78(k=23) | 0.59(k=22) |
| 2017 年 10 月 | 2018 年 10 月 | 0.77(k=17) | 0.59(k=21) |
| 2017 年 11 月 | 2018 年 11 月 | 0.78(k=19) | 0.60(k=15) |
| 2017 年 12 月 | 2018 年 12 月 | 0.79(k=21) | 0.60(k=17) |
| 2020 年 9 月 | 2021 年 9 月 | 0.78(k=23) | 0.59(k=19) |
| 2020 年 10 月 | 2021 年 10 月 | 0.77(k=17) | 0.59(k=21) |
| 2020 年 11 月 | 2021 年 11 月 | 0.78(k=19) | 0.60(k=17) |
| 2020 年 12 月 | 2021 年 12 月 | 0.79(k=21) | 0.60(k=17) |

### 4.2 分析内容と手順

以下の手順より，分析データを構築する．

(1) 推定したい月のデータより前の ID-POS データの読み込み

(2) 商品を部門，AU，ラインまで分類し，POS データを RECNO ごとにグルーピングし，RECNO ごとのベクトルを作成

(3) 推定したい月の ID-POS データを読み込み，(2) 同様にデータを加工

(4) (2) のデータを k 近傍法を用いて学習

(5) k 近傍法を用いて，(3) で読み込んだデータの中から，会員のデータを用いて，性別，年齢層のそれぞれの精度を測定

### 4.3 結果と考察

今回，ID-POS データを用いて k 近傍法で分析して抽出した精度の結果を表 1〜表 4 に示す．なお，本研究において，精度を算出する際のデータは, 各月のデータをランダムに 10 %抽出して利用しているため，毎回結果は異なるが，大きな違いはなかった．

表 1 では，同じ年の 9 月を学習データとした場合の精度の結果を示している．表 2 では，前年の 9 月を学習データとした場合の精度の結果を示している．表 3 では，前年の同じ月を学習データとした場合の精度を示している．表 4

表 4: 他の店舗の学習データを使用した場合の精度

| 学習データ | 推定するデータ | | 2021 年 10 月 | 2021 年 11 月 | 2021 年 12 月 |
|---|---|---|---|---|---|
| 和歌山県 B 店 | 和歌山県 B 店 | 性別 | 0.71(k=15) | 0.77(k=17) | 0.70(k=25) |
| 2021 年 9 月 | | 年齢層 | 0.56(k=17) | 0.56(k=15) | 0.55(k=17) |
| 和歌山県 A 店 | | 性別 | 0.72(k=17) | 0.70(k=17) | 0.70(k=15) |
| 2021 年 9 月 | | 年齢層 | 0.55(k=21) | 0.55(k=20) | 0.56(k=20) |
| 愛知県 C 店 | 愛知県 C 店 | 性別 | 0.73(k=21) | 0.73(k=29) | 0.70(k=25) |
| 2021 年 9 月 | | 年齢層 | 0.57(k=17) | 0.51(k=25) | 0.55(k=17) |
| 和歌山県 A 店 | | 性別 | 0.72(k=13) | 0.72(k=23) | 0.72(k=15) |
| 2021 年 9 月 | | 年齢層 | 0.59(k=17) | 0.56(k=17) | 0.57(k=17) |
| 岐阜県 D 店 | 岐阜県 D 店 | 性別 | 0.76(k=19) | 0.76(k=19) | 0.76(k=21) |
| 2021 年 9 月 | | 年齢層 | 0.58(k=19) | 0.57(k=19) | 0.56(k=17) |
| 和歌山県 A 店 | | 性別 | 0.76(k=19) | 0.76(k=19) | 0.76(k=17) |
| 2021 年 9 月 | | 年齢層 | 0.55(k=21) | 0.57(k=17) | 0.57(k=17) |

では，他店の同じ年の 9 月のデータを学習データとした場合の精度を示している．

表 1 と表 2 の結果より，同じ店舗の同じ年の 9 月を利用した場合と，前年の 9 月を利用した場合は，結果にあまり差はないことがわかった．表 3 の結果より，新型コロナウイルスの影響を受けて推定の精度に差がないことがわかった．また，表 4 の結果より，他店の同年 9 月のデータを学習データとした場合にも，精度に大きな違いは見られない．

この結果より，各精度に差は見られないので，調査項目 1 に関しては，学習データに大きく依存しないことがわかった．

今回の結果から，性別の精度が高いので，購入商品に性別が密接に関係している可能性がある．年齢層の精度が十分ではないので，現在の手法では，購入商品と年齢の関係性を十分には捉えられていない可能性がある．今後，来店する顧客の時間を考慮するために，購買時間帯のデータを特徴量に加えるなどの手法を用いることで，年齢層の精度の向上を目指す．

## 5. 調査項目 2：非会員の推定

学習データによって，精度の値は変化しないので，非会員の推定を下記のデータを利用して非会員の推定を行った．

**学習データ**

和歌山県の A 店の 2020 年 9 月のデータ

**推定するデータ**

和歌山県の A 店の 2021 年 9 月のデータ

### 5.1 分析内容と手順

以下の手順より，分析データを構築する．

(1) 調査項目 1 と同様に，推定したい月のデータより前の ID-POS データの読み込み

(2) POS データを，RECNO ごとにグルーピングし，RECNO ごとのベクトルを作成

(3) 推定したい月の ID-POS データを読み込み，(2) 同様にデータを加工

(4) (2) のデータを k 近傍法を用いて学習

(5) k 近傍法を用いて，(3) で読み込んだデータの中から，非会員のデータを用いて，性別，年齢層を推定し，会員，非会員の割合を算出

### 5.2 結果と考察

図 1 に非会員の性別の推定結果を, 図 2 に非会員の年齢層の推定結果を示す．

図 1 より，男性の非会員の割合は全体の 1.5 %であることがわかる．また，男性の 8.1 %は非会員であり，91.9 %は会員であった．女性の場合，女性の 23.1 %が非会員で，76.9 %が会員であった．女性のほうが非会員の割合が高いことが図 1 より読み取れる．これは，女性は会員カードとは別のポイントカードを提示している可能性が考えられる．

図 2 より，ヤング層の割合は全体の 3 %未満であった．ヤング層の非会員の割合は全体の 1 %未満であった．ヤング層の非会員の割合は 1.9 %，会員の割合は 98.1 %であった．ミドル層の非会員の割合は 25.3 %，会員の割合は 74.7 %であった．シニア層の場合，非会員の割合は 17.8 %，会員の割合は 82.2 %であった．ヤング層の非会員の割合が一番小さく，ミドル層の非会員の割合が一番大きく，シニア層の非会員も一定数いることがわかった．ただし，年齢層の推定精度は不十分であることを考慮する必要がある．

今回の結果より，男性の非会員の割合が低いので，そもそも買い物に行く習慣があまりないのか，あるいはスーパーマーケットではあまり買い物をせず，コンビニエンスストアなどで済ましてしまっているのではないかと考えられる [7]．また，ヤング層全体の割合が 3 %とかなり低いので，こちらも，男性同様にあまりスーパーマーケットで買い物をしていないと考えられる．

## 6. おわりに

本研究では，ID-POS データを用いて，非会員の年齢層と性別の推定をした．

今回は k 近傍法を用いたが，今後は，他の手法の利用を検討し，精度の向上を目指す．

## 参考文献


[1] 全国スーパーマーケット協会:2022 年版 スーパーマーケット白書，入手先:<http://www.super.or.jp/wp-content/uploads/2021/02/NSAJ-Supermarket-hakusho2022.pdf>(参照日：2022 年 7 月 11 日).

[2] 全国スーパーマーケット協会:2019 年版 スーパーマーケット白書 第 1 章，入手先:<http://www.super.or.jp/wp-content/uploads/2019/02/hakusho2020-1.pdf>(参照日：2022 年 7 月 12 日).

[3] 全国スーパーマーケット協会:2014 年版 スーパーマーケット白書 第 2 章，入手先:<http://www.super.or.jp/wp-content/uploads/2014/02/supermarket-hakusho2015-2.pdf>(参照日：2022 年 7 月 12 日).

[4] 中村 綾乃，吉野 孝，松山 浩士，貴志 祥江，大西 剛: 客動線分析のための ID-POS データを用いたエージェントシミュレーションシステムの提案，情報処理学会論文誌，Vol.63, No.1, pp.56–65 (2022).

[5] 石橋 敬介, 尾崎 幸謙：購買履歴データを用いた商品特性の付与と消費者特性の推定，日本行動計量学会抄録集，Vol.47, pp.176–177 (2019).

[6] 下田稜，小山竜兵，杉浦孝典，島田匠都，村井詩音，矢田勝俊，原田拓弥，李皓：スーパーマーケットの POS データに基づく消費者の購買行動モデル構築，第 18 回社会システム部会研究会，pp.175–178 (2019).

[7] ニャック：「利用頻度が高いのはどっち? 品揃えのスーパー vs 便利さのコンビニ」Sirabee リサーチ：入手先:<https://sirabee.com/2019/03/16/20162015585/>(参照日：2022 年 7 月 19 日).

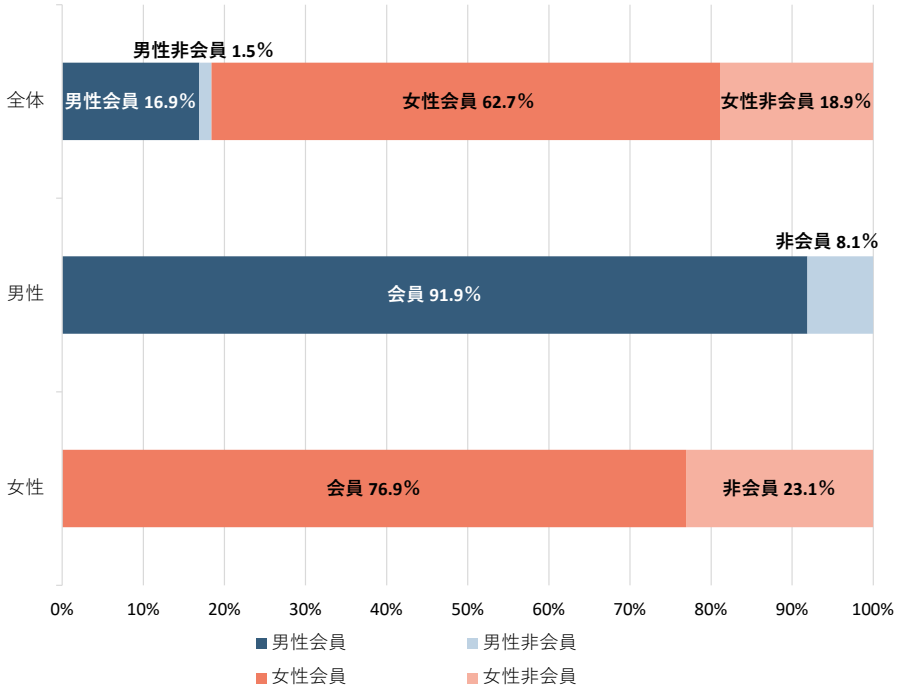


図 1: 性別の推定結果

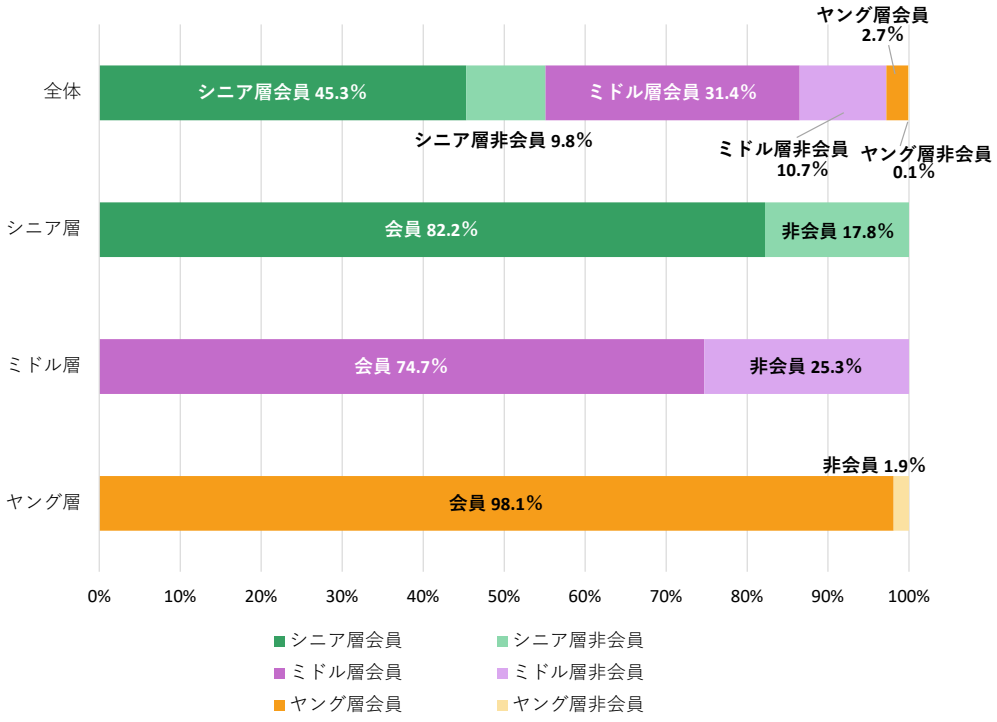


図 2: 年齢層の推定結果