修士論文
経済複雑性指標を用いた購買データ分析

谷口槙吾

主任指導研究教員　林幸雄

北陸先端科学技術大学院大学
金沢大学
（融合科学）

February 2021

目次

図一覧

表一覧

第1章　はじめに

1.1　研究背景

多様化する現代社会のマーケティングにおいては適切な市場分析に基づいて施作を決定する事が求められる。その中で購買データの分析は収益性を高める上で重要で、多面的な分析によって間違ったデータの解釈を防ぎ、感覚的には発見できない法則などを見つけ、効率の良い施作を決定する上で役立つ。現在、需要や売り上げの予測を統計的な手法や機械学習を用いて行う試みは多く為されているが、一方で購買者と商品の関係性についての分析は十分に行われていない。

1.2　研究の目的

本研究では、購買者と商品の関係性からどのような知見が得られるかを複雑度マップ、経済複雑性指標を用いて購買データ分析することで検討する。購買にそのまま複雑度マップ、経済複雑性指標を適用してもデータの粒度の違うため、有用な結果が得られないと考えられ、後述する2つのクラスタリング手法によって情報を集約し、それぞれの結果がどのように異なるか検討する。

第2章　関連研究

本研究では、頂点(ノード)集合$V = \{v_1, v_2, \ldots, v_N\}$とその2点を繋ぐ辺（エッジ）集合$E = \{e_1, e_2, \ldots, e_M\}$からなるデータの関係性を表すグラフを扱う。また、頂点集合を同じ集合内の頂点間には辺が存在しないように、二つの集合V, Uに分けられるような二部グラフを考える。但し、本研究で扱うグラフのエッジには向きと多重辺は無いものとする。

2.1　比較優位指数

国と輸出品の関係は重み付き二部グラフで表すことが出来る。国がある輸出品を輸出した量を表す重み付き二部グラフの隣接行列をX_{cp}とする。隣接行列X_{cp}のc行p列の成分は

$$X_{cp} = \begin{cases} \text{国}c\text{の製品}p\text{の輸出総額} & (\text{ある国}c\text{が製品}p\text{を輸出していた場合}) \\ 0 & (\text{ある国}c\text{が製品}p\text{を輸出していない場合}) \end{cases} \tag{2.1}$$

である。

X_{cp}から比較優位指数と呼ばれる、自国のある輸出品が他国に対して優位であるかを判定するための数値を算出する[3]。比較優位指数は

$$RCA_{cp} = \frac{X_{cp}}{\sum_c X_{cp}} / \frac{\sum_p X_{cp}}{\sum_{c,p} X_{cp}} \tag{2.2}$$

によって算出できる

RCA_{cp}は、ある国cにおける輸出品pが自国の輸出品の中で占める割合$\frac{X_{cp}}{\sum_c X_{cp}}$を世界全体の輸出品の割合$\frac{\sum_p X_{cp}}{\sum_{c,p} X_{cp}}$で割っている、この数値が1以上である場合、国々の平均よりも多く製品を輸出していることを表し、優位な輸出国と見なすことができる。RCAの処理によってX_{cp}から重み無し無向グラフの隣接行列M_{cp}を構成す

る。M_{cp}は

$$M_{cp} = \begin{cases} 1 & \text{if } RCA_{cp} \geq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

によって定義される[4]。

2.2 複雑度マップ

各国の産業の特徴を測るために、国の産業多様度$k_{c,0}$と輸出品の遍在度$k_{p,0}$を定義する[4]。

$$\text{国の産業の多様度 } k_{c,0} = \sum_{p} M_{cp} \tag{2.4}$$

$$\text{輸出品の遍在度 } k_{p,0} = \sum_{c} M_{cp} \tag{2.5}$$

上記により、多様度と遍在度を定義したが、多様度が同じ国でもありふれた製品を輸出しているか、希少なものを輸出しているかによって国の産業の特徴は異なってくる。一方、輸出品についても同様に、産出国が多様な製品を輸出しているのか、特定の製品のみを輸出しているだけかによって特徴は異なってくる。そこでk_cとk_pを互いの情報を用いて以下のように更新する。

$$k_{c,1} = \frac{1}{k_{c,0}} \sum_{p} M_{cp} \cdot k_{p,0} \tag{2.6}$$

$$k_{p,1} = \frac{1}{k_{p,0}} \sum_{p} M_{cp} \cdot k_{c,0} \tag{2.7}$$

式(2.4)式(2.6)により求めた$k_{c,0}$と$k_{c,1}$から国の産業の特徴を表す複雑度マップが得られる。

図2.1の$< k_{c,0} >$と$< k_{c,1} >$はそれぞれの平均値の線で区切られた4つの領域に別ける。多くの国は4つの領域中の左上か右下に属している。左上の領域は産業の多様性が低く、遍在度高い製品を多く輸出している日本、アメリカ、ドイツなどが属し、右下の領域には輸出の多様度が高く、遍在度の低い製品を多く輸出している途上国とされる南アメリカやアフリカの国々が属す。

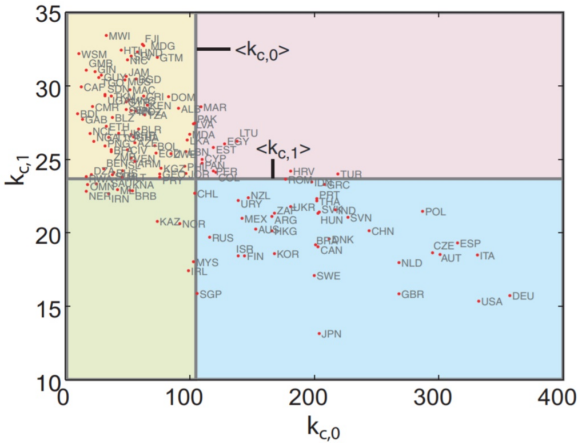


Fig. 2.1 国の複雑度マップ　([4]より転載)

2.3 経済複雑性指標

複雑度マップの両軸である$k_{c,0}$と$k_{c,1}$を求めたが、以降も互いの情報を用いて$k_{c,0}$と$k_{c,1}$を反復的に更新していくことで経済複雑性指標と商品複雑性指標が導出できる[4]。

$$k_{c,N} = \frac{1}{k_{c,0}} \sum_p M_{cp} \cdot k_{p,N-1} \tag{2.8}$$

$$k_{p,N} = \frac{1}{k_{p,0}} \sum_p M_{cp} \cdot k_{c,N-1} \tag{2.9}$$

ここで式(2.8)に式(2.9)を代入すると

$$k_{c,N} = \frac{1}{k_{c,0}} \sum_p M_{cp} \frac{1}{k_{p,0}} \sum_{c'} M_{c'p} \cdot k_{c',N-2} \tag{2.10}$$

$$k_{c,N} = \sum_{c'} k_{c',N-2} \sum \frac{M_{cp} M_{c'p}}{k_{c,0} k_{p,0}} \tag{2.11}$$

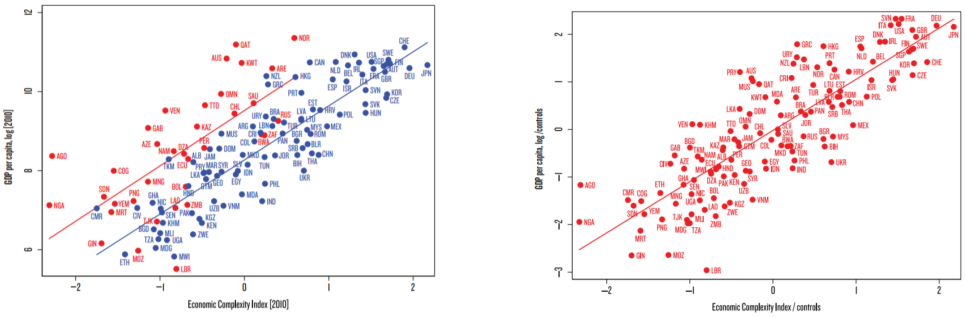
となる。上式の$\widetilde{M}_{cc'} = \sum_p \frac{M_{cp}M_{c'p}}{k_{c,0}k_{p,0}}$とおけば

$$k_{c,N} = \sum_{c'} \widetilde{M}_{cc'} k_{c',N-2} \tag{2.12}$$

と書きなおせる。式(2.12)は$k_{c,N} = k_{c',N-2} = 1$の時に成立する。これは$\widetilde{M}_{cc'}$の最大固有値に対応する固有ベクトルである。この固有ベクトルの各成分は全て1のため、2番目に大きい固有値に対応するベクトルを見る。これは最大分散量を捉えるベクトルであり、経済の複雑性を捉えていると考えられる。従って経済複雑性指標(ECI)は

$$ECI = \frac{\vec{K} - < \vec{K} >}{\mathrm{stdev}(\vec{K})} \tag{2.13}$$

と定義する。式(2.13)で$\vec{K}$は二番目に大きい固有値に対応する$\widetilde{M}_{cc'}$の固有ベクトル、stddevは標準偏差、<>は平均を表す。商品複雑性指標は式(2.9)に式(2.8)を代入し、同じように式変形し固有ベクトルを算出することで求められる。 ECIは石油を除けば、一人あたりのGDPと相関が見られることから、経済指標としてある程度の妥当性はあると言える。しかしECIはそれ以外にも、1人あたりのGDPがECIから期待される値より低い国では国が急成長する傾向にあると示されている。つまり、ECIは現在のGDPだけではなくこれから取得を生み出す国の産業の潜在能力を表す指標であると考えられている。



(a)ECIと一人当たりGDP
(赤の国は産油国)

(b)石油を除いたECIと
一人当たりGDP

Fig. 2.2 ECIと一人当たりGDPの散布図 ([4]より転載)

2.4 経済複雑性指標と正規化カットの対応

経済複雑性指標の計算は以下の正規化カットと同等であるという事が知られている。[5] 非負の重みを持つ重み付き無効グラフ$G = (V, E)$を考える。このグラフGの隣接行列をSとして、頂点iの重みをd_i対角成分$D = \{d_1, d_2, \ldots, d_N\}$を以下のよ

うに定義する。

$$d_i = \sum_{j \in V} S_{ij} \tag{2.14}$$

頂点集合Aの重み総和は

$$\mathrm{vol(A)} = \sum_{i \in A} d_i \tag{2.15}$$

と書ける。 一方、グラフを二つに分割する方法として最小カット問題を解く。最小カット問題とは グラフを二つの集合に分割する時にカットした辺の重み和が最小となるような分割を最小カットと呼ぶ。 しかしながら通常の最小カット問題では一つの頂点とそれ以外の頂点というような分割が得られる事が多い。そこでなるべく二つの頂点集合の頂点数が同じとなるような制約を設けた以下の最小正規化カットを定義する。

$$\mathrm{Ncut}\,(A, \bar{A}) = \left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(\bar{A})} \right) \sum_{i \in A, j \in \bar{A}} S_{ij} \tag{2.16}$$

上記の目的関数(2.16)を最小化することで最小正規化カットを達成できる。

$$\min_A \mathrm{Ncut}(A, \bar{A}) = \min_y \frac{y^T (D - S) y}{y^T D y} \tag{2.17}$$

しかしながらこれは整数制約を持つNP困難問題である。そこで、一般化固有値方程式の2番目に小さい固有値に対応する固有ベクトルを求める事で実数値緩和した近似解を得る事が知られている。[6]

$$(D - S)y = \lambda D y \tag{2.18}$$

上式(2.18)のyは

$$y = D^{-1/2} z \tag{2.19}$$

として代入する事で

$$D^{-\frac{1}{2}} (D - S) D^{-\frac{1}{2}} z = \bar{L}_S z = \lambda z \tag{2.20}$$

が得られる。上式(2.20)の$\bar{L}_S = D^{-\frac{1}{2}} (D - S) D^{-\frac{1}{2}}$は隣接行列の$S$の一般化正規化ラプラシアンとして呼ばれている。一般化正規化ラプラシアンの最小固有値に対応する最小固有ベクトルは全て0であるため、2番目に小さい固有値に対応する固有ベクトルが上式の一般化固有値問題の解となる。

$$y^{[2]} = D^{-1/2} z^{[2]} \tag{2.21}$$

$y^{[2]}$、$z^{[2]}$はそれぞれ2番目に小さい固有値に対応する固有ベクトルを表す。

ところで前節で示した$\widetilde{M}_{cc'}$は

$$\widetilde{M}_{cc'} \equiv \sum_p \frac{M_{cp}M_{c'p}}{k_{c,0}k_{p,0}} = \frac{1}{k_{c,0}} \sum_p \frac{M_{cp}M_{c'p}}{k_{p,0}} \tag{2.22}$$

という形に変形する事が出来る。 更に$\widetilde{M}$を行列形式で表記する。

$$\tilde{M} = D^{-1}MU^{-1}M' \tag{2.23}$$

D^{-1}は多様度$k_{c,0}$からなる対角行列の逆行列を表し、U^{-1}は多様度$k_{p,0}$からなる対角行列の逆行列を表す。またはM'はMの転置行列を表す。

上式(2.23)はそれぞれの国の産業がどの程度似ているかを示す類似度ネットワークに対して多様度で重みづけを行っているとも解釈する事が出来る。そこで、$S = MU^{-1}M'$として

$$\tilde{M} = D^{-1}S' \tag{2.24}$$

とする。 式(2.18)の両辺に$D^{-\frac{1}{2}}$を掛けると

$$D^{-\frac{1}{2}}(D - S)D^{-\frac{1}{2}}z = \bar{L}_S z = \lambda z \tag{2.25}$$

上式2.25に式(2.24)を代入すると

$$\tilde{M}D^{-\frac{1}{2}}z = (1 - \lambda)D^{-\frac{1}{2}}z \tag{2.26}$$

となる。よって、$\widetilde{M}$の固有値問題は以下に帰着する。

$$\tilde{M}\tilde{y} = \tilde{\lambda}\tilde{y} \tag{2.27}$$

上記の二つの式(2.26)と式(2.27)から

$$\tilde{\lambda} = 1 - \lambda \tag{2.28}$$

$$\tilde{y} = D^{-\frac{1}{2}}z \tag{2.29}$$

が導かれる。よって、正規化カットの$\bar{L}_s$の二番目に小さい固有値は$\widetilde{M}$の二番目に大きい固有値と対応する事、さらに正規化カットの解は近似解は二番目に小さい固有値に対応する固有ベクトルECIは二番目に大きい固有値に対応する固有ベクトルであるから

$$\tilde{y}^{[2]} = y^{[2]} = D^{-\frac{1}{2}}z^{[2]} \tag{2.30}$$

を求めれば良い。

よって、国の類似度ネットワークの隣接行列Sに対する正規化カットの近似解とECIは同等である。ここでは省略するが、商品の類似度ネットワークの隣接行列Sに対する正規化カットの近似解とPCIは同等であることも示されている[5]。

2.5 購買データに対する経済複雑性指標の適用

複雑度マップと複雑性指標を購買データに対して適用して、ユーザーや商品の知見を得る事を試みる。関連研究[4]では数万ある輸出品を200程度のカテゴリに集計した取引データを使用しているが購買データの場合細かなカテゴリ分けが為されているものは少なく、またユーザーも数千 数万という数存在する。そのため、関連研究の手法をそのまま適用するとユーザーや商品の知見を得る事が難しいと考えられる。図(2.3)は購買データに関連研究の手法をそのまま適用して得た複雑度マップの一例である。多くのユーザーが左下の領域に集まっていることが見て取れるが、これは多くのユーザーは一部の多様な商品を購入しているユーザーと比べると多様度が低く、またその購買商品の平均の遍在度が低い。つまり全体から見れば、ニッチな商品を主に購入していく傾向があるユーザーが大半を占めるということである。

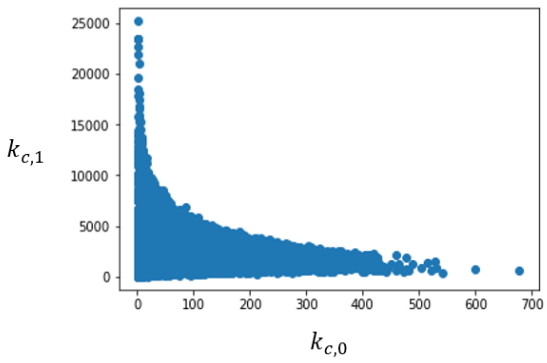


Fig. 2.3 そのままの購買データから得たユーザーの複雑度マップ

第3章　提案手法

本章では、経済複雑性指標を購買データに対して適用し、ユーザーと商品に関する知見を得るためにユーザーと商品についての二部グラフX_{cp}の頂点を集約する方法を提案する。その際、ユーザーを購買行動に基づいてクラスタリングするRFMと、ユーザーと商品の購買パターンによってクラスタリングするNMFの二つを以下説明する。

3.1　RFMによるユーザークラスタリング

RFMとは、ユーザーをRecency（直近の購入日　Frequency(購入頻度) Monetary（購買金額の3つの指標に基づいて分ける手法である[7]。表(3.1)のような基準を設けて、それぞれの指標ごとにユーザーのランクをつける。この場合はランクが5までの3つの指標により5×5×5の125通りのセグメントにユーザーがランク分けされる事となる。

Table 3.1 RFMのクラス分け例

ランク	Recency	Frequency	Monetary
1	上位~20%	上位~20%	上位~20%
2	上位20~40%	上位20~40%	上位20~40%
3	上位40~60%	上位40~60%	上位40~60%
4	上位60~80%	上位60~80%	上位60~80%
5	上位80%~	上位80%~	上位80%~

3.2　非負値行列因子分解によるクラスタリング

非負値行列分解(NMF)では、図(3.1)に示すようにI×Jサイズの非負値行列をYをI×Kの非負値行列HとK×Jの非負値行列Uで近似する事が目的である[8]。

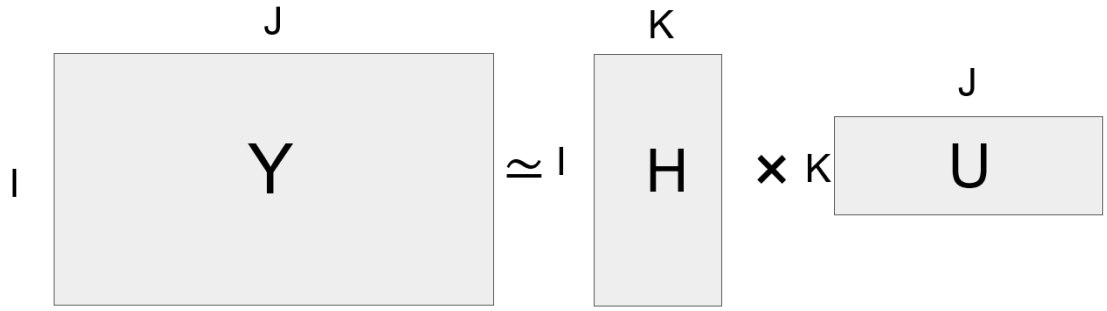


Fig. 3.1 NMFによる行列分解

行列HとUには、その中の要素が全て非負となるような制約を設けている。それらの要素を非負に制限する事で数値の解釈が行いやすく、元の行列Yの中にある頻出パターンを抽出することが可能となる。 NMFの目的を達成するためのアルゴリズムは様々だが、本稿では広く用いられているmultiplicative update rulesについて説明する。またNMFにはいくつか最適化基準が存在するが本稿では、実験で用いたKLダイバージェンスを例として説明をする。KLダイバージェンスを最適化基準とした場合、目的関数は次のようになる。

$$\begin{aligned} &\text{minimize} \sum_{\omega,t}\left(Y_{\omega,t}\log\frac{Y_{\omega,t}}{\sum_i H_{\omega,i}U_{i,t}} - Y_{\omega,t} + \sum_i H_{\omega,i}U_{i,t}\right) \\ &\text{subject to} \quad H_{\omega,i} \geq 0, \quad U_{i,t} \geq 0 \end{aligned} \tag{3.1}$$

分解された行列の積$\sum_i H_{\omega,i}U_{i,t}$と元の行列$Y_{\omega,t}$の差を最小にする事が目的となるが、これは非線形最適化問題であり直接最適化することは難しい。そこで補助関数法と呼ばれる最適化手法によって最適化を行う。

3.3 補助関数法

目的関数を$D(\theta)$とし

$$D(\theta) = \min_{\bar{\theta}} G(\theta, \bar{\theta}) \tag{3.2}$$

を満たすような上限$G(\theta, \bar{\theta})$を補助関数として定義する。補助関数を定義した後に

$$\begin{aligned} \bar{\theta}^{(n+1)} &= \underset{\bar{\theta}}{\text{argmin}} G\left(\theta^{(n)}, \bar{\theta}\right) \\ \theta^{(n+1)} &= \underset{\theta}{\text{argmin}} G\left(\theta, \bar{\theta}^{(n+1)}\right) \end{aligned} \tag{3.3}$$

を交互に更新する事で目的関数$D(\theta)$の停留点を得ることができる。[8]

3.4 補助関数の定義

式(3.1)で示した目的関数の補助関数を定義する。目的関数は以下のように式変形できる。

$$\begin{aligned}\mathcal{D} :&= \sum_{\omega,t}\left(Y_{\omega,t}\log\frac{Y_{\omega,t}}{\sum H_{\omega,k}U_{k,t}} - Y_{\omega,t} + \sum_k H_{\omega,k}U_{k,t}\right)\\ &= \sum_{\omega,t}\left(Y_{\omega,t}\log Y_{\omega,t} - Y_{\omega,t}\log\sum H_{\omega,k}U_{k,t} - Y_{\omega,t} + \sum_k H_{\omega,k}U_{k,t}\right)\end{aligned} \tag{3.4}$$

変形した式(3.4)の第二項$-Y_{\omega,t}\log\sum H_{\omega,k}U_{k,t}$は負の対数関数であり凸関数であるから、Jensenの不等式から

$$-\log\sum_k H_{\omega,k}U_{k,t} \leq -\sum_k \lambda_{k,\omega,t}\log\frac{H_{\omega,k}U_{k,t}}{\lambda_{k,\omega,t}} \tag{3.5}$$

より補助関数

$$G := \sum_{\omega,t}\left(Y_{\omega,t}\log Y_{\omega,t} - Y_{\omega,t}\sum_i \lambda_{i,\omega,t}\log\frac{H_{\omega,i}U_{i,t}}{\lambda_{i,\omega,t}} - Y_{\omega,t} + \sum_i H_{\omega,i}U_{i,t}\right) \tag{3.6}$$

が得られる。補助関数を定義した後、以下の更新のステップを導出する。

$$\lambda_{i,\omega,t} \leftarrow \underset{\lambda_{i,\omega,t}}{\mathrm{argmin}} G = \frac{H_{\omega,i}U_{i,t}}{\sum H_{\omega,i'}U_{i',t}} \tag{3.7}$$

によって得られた$\lambda_{i,\omega,t}$を

$$H_{\omega,k} \leftarrow \underset{H_{\omega,k}}{\mathrm{argmin}} G = \frac{\sum_t Y_{\omega,t}\lambda_{k,\omega,t}}{\sum_t U_{k,t}} \tag{3.8}$$

$$U_{k,t} \leftarrow \underset{U_{k,t}}{\mathrm{argmin}} G = \frac{\sum_\omega Y_{\omega,t}\lambda_{k,\omega,t}}{\sum_\omega H_{\omega,k}} \tag{3.9}$$

に代入して更新を繰り返していくことで、目的関数を最小化することが出来る。[8]

第4章　実験・評価

RFMとNMFを、公開されている購買データ[Sou][UCI]に対して適用し複雑度マップ、経済複雑性指標を算出する。その結果と売り上げの関係性について検討する。

4.1　実験データ

実験データは卸売り業者のB2B取引データ[UCI]とある地域の世帯ごとの購買データ[Sou]を使用する。

B2B取引データ

- 取引数 106万件
- ユーザー数　5563人
- 商品数　4856品

ある地域の世帯ごとの購買データ

- 取引数 250万件
- 世帯数 2500世帯
- 商品数 92339品

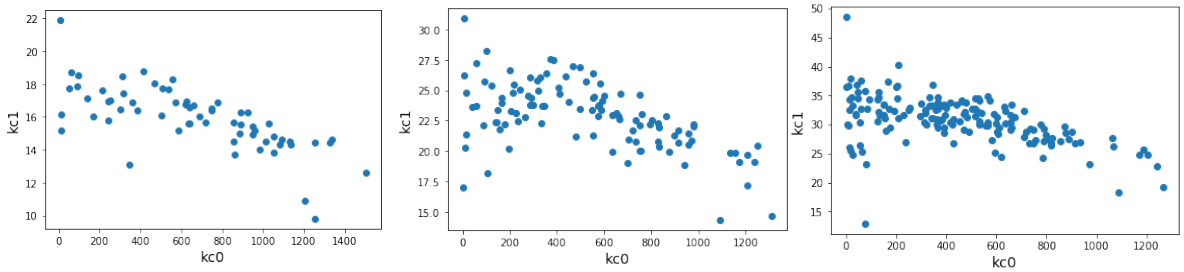
4.2　複雑度マップの実験

上記の購買データから作成したユーザーと商品の二部グラフに対して、RFMとNMFを適用した結果を示す。RFMは64通り、125通り、216通りのランクが出来るようにクラスタリングを行い、NMFも比較のために、同様の数を基底数に指定しクラスタリングを行った。

NMFでは、ユーザー、商品どちらか片方だけを情報集約をした場合と両方の情報を集約した場合で複雑度マップの計算を行った。ここで情報集約とはクラスタリングで同クラスタに所属するユーザーまたは商品を1つの頂点と見なし、同じ頂点に接続されている重み付き辺を足し合わせて1つにすることを指す。

4.3 B2Bデータの複雑度マップ

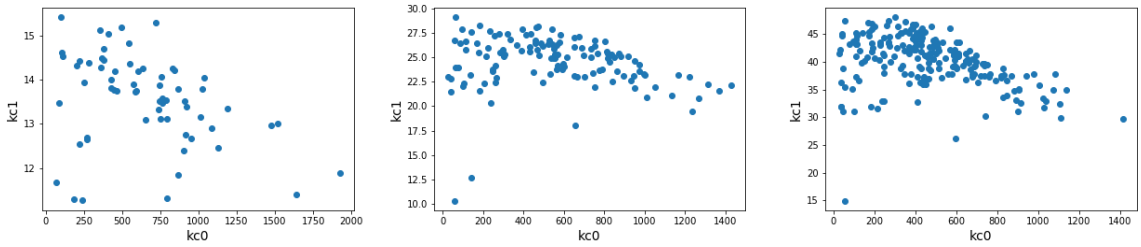
図(4.1)にRFMによってユーザーの情報を集約したユーザーの複雑度マップを示す。ランク数64では国の複雑度マップ同様に左上から右下にクラスタが分布する形となったがランク数を増やしていくにつれて左下のニッチな商品を多く買うクラスタが見られるようになった。



(a)ランク数64　(b)ランク数125　(c)ランク数216

Fig. 4.1 B2Bデータのユーザー複雑度マップ(RFM)

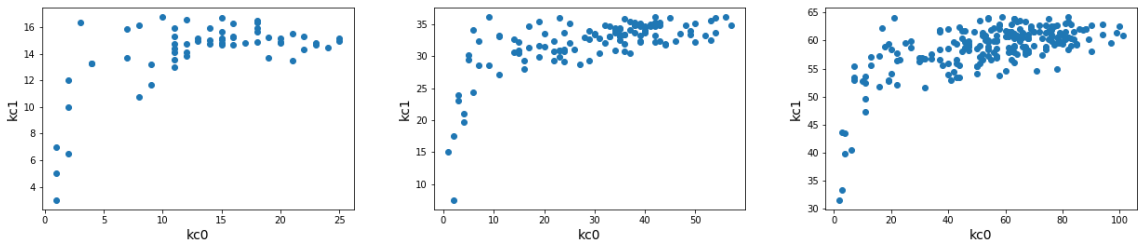
図(4.2)にNMFによって得たクラスタに基づき、ユーザーのみ情報集約を行い、算出した複雑度マップを示す。RFMの場合とは異なり、ランク数を64でも左下に分布するニッチな購買層が確認できる。これはNMFでは購買商品パターンを元にクラスタリングを行うため、レアな購買パターンのユーザーが1つのクラスタに分類されるためであると考えられる。さらにランク数を増やしていくと左下に突出したクラスタが確認された。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

Fig. 4.2 B2Bデータのユーザー複雑度マップ(NMFによりユーザーのみ情報集約)

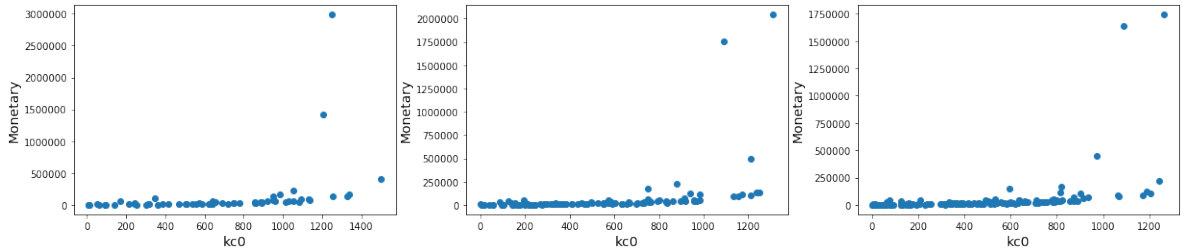
図(4.3)にNMFによって得たクラスタに基づき、ユーザーと商品の情報集約を行い、算出したユーザーの複雑度マップを示す。RFM、NMFでユーザーのみ情報を集約させて場合のどちらの結果とも異なり、右下にクラスタが存在しない形となった。つまり多様であるがニッチな購買傾向を持つユーザーがいないという事を示す。一部購買量の少ないユーザークラスタを除いて多くのユーザークラスタが同じ商品クラスタにおいて優位であるためこのような結果になったと考えられる。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

Fig. 4.3 B2Bデータのユーザー複雑度マップ(NMFによりユーザーと商品の情報集約)

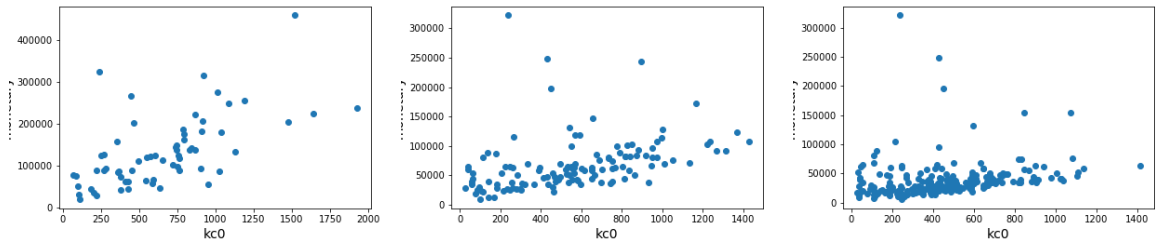
RFMによってユーザーの情報集約を行った時の、多様度と購買額の散布図を図(4.4)に示す。どのランク数でも、多様度が高いクラスタが購買額の大半を占める結果となった。RFMはユーザーの購買額や回数でランクを付けているため1つの売り上げに全体の売り上げが大部分が集中しやすく優位に購入しているものも多様になると考えられる。



(a)ランク数64 (b)ランク数125 (c)ランク数216

Fig. 4.4 B2Bデータの多様度と購買額の散布図(RFM)

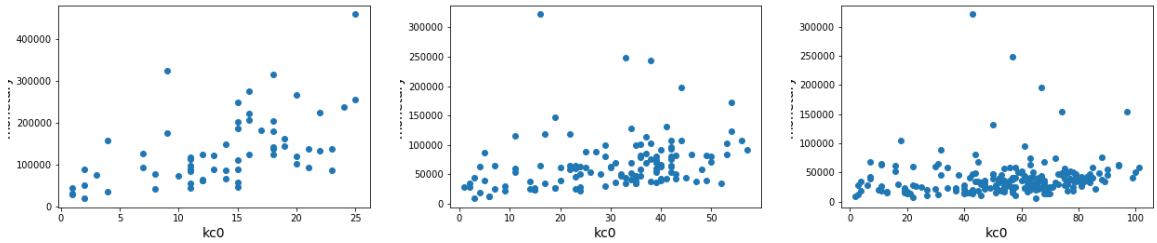
NMFによって得たクラスタに基づき、ユーザーの情報集約を行った時の多様度と購買額の散布図を図(4.5)に示す。どのクラスタ数でも多様度が中程度のクラスタが一番購買額が高かった。これは特定の商品を大量に購入するユーザーが存在するためだと考えられる。また、一部のクラスタを除いて、多様度に比例して売り上げが伸びる傾向にあるのが確認できる。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.5 B2Bデータの多様度と購買額の散布図(NMFによりユーザーのみ情報集約)

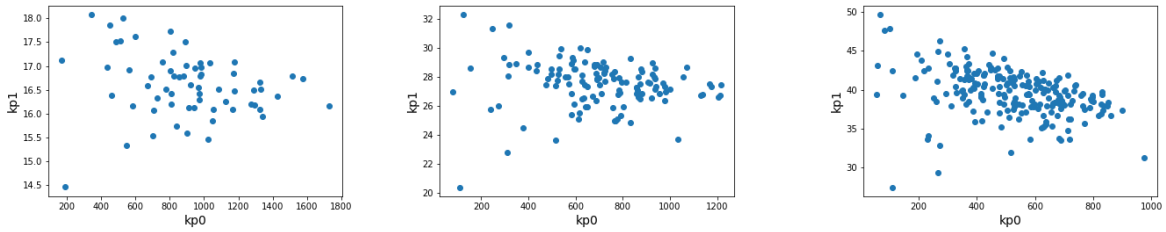
NMFによって得たクラスタに基づき、ユーザーと商品の情報集約を行った時の多様度と購買額の散布図を図(4.6)に示す。多様度と購買額の関係はユーザーの複雑度マップとは異なり、両方の情報集約をしてもNMFでユーザーのみ情報集約した場合と結果は大きく異ならないことがわかる。



(a)クラスタ数数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.6 B2Bデータの多様度と購買額の散布図(NMFにより両方の情報集約)

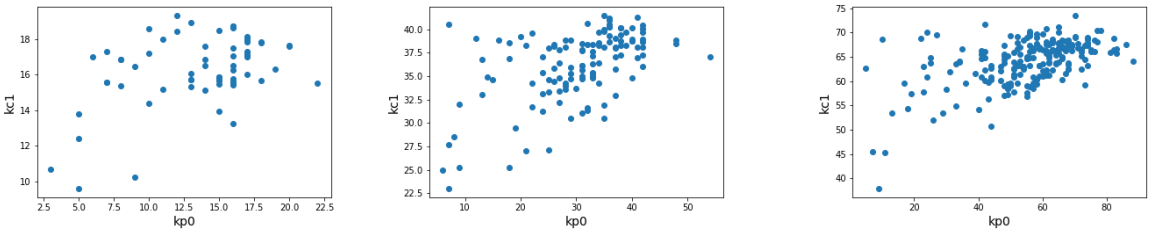
NMFによって得たクラスタに基づき、商品のみ情報集約を行い、算出した商品の複雑度マップを図(4.7)に示す。どのクラスタ数でも左下に突出したクラスタが存在する事が確認できる。商品を購買するユーザーのパターンに基づいてクラスタリングが行われるため、一部ユーザーしか購入しないニッチな商品が、ひとつのクラスタに分類されやすいためであると考えられる。しかし左下を除けば、ユーザー同様に左上から右下に分布するような結果となっている。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.7 B2Bデータの商品複雑度マップ(NMFで商品のみ情報集約)

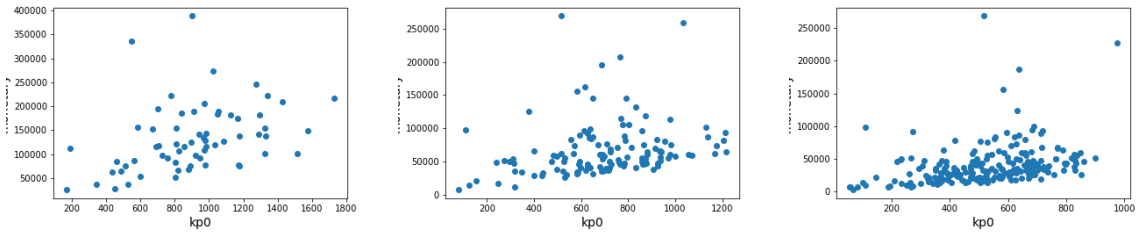
NMFによって得たクラスタに基づき、商品の複雑度マップを算出した結果を図(4.8)に示す。ユーザーの複雑度マップ同様、ユーザーと商品どちらも情報集約を行うと右下以外に広くクラスタが分布する結果となった。



(a)クラスタ数64　(b)クラスタ数125　(b)クラスタ数216

Fig. 4.8 B2Bデータの商品複雑度マップ(NMFで両方の情報集約)

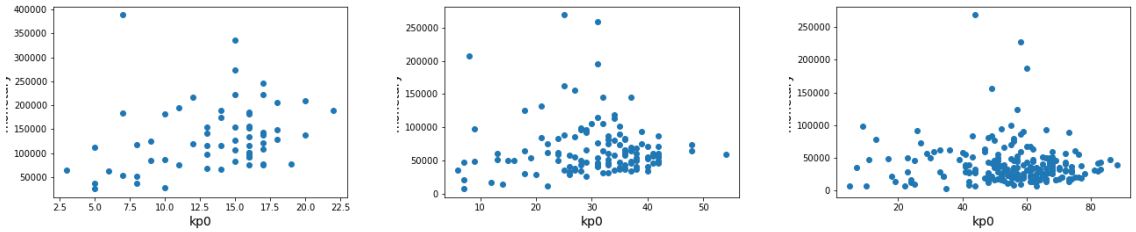
NMFによって得たクラスタに基づき、商品のみ情報集約を行い算出した売り上げと遍在度の散布図を図(4.9)に示す。売り上げと遍在度でも購買額と多様度の関係同様に、相関が確認できるが、特に売り上げの多いクラスタは遍在度が中程度であることが見て取れる。これは一部ユーザーが売り上げの多くを占める商品クラスタであると考えられる。



(a)クラスタ数64　(b)クラスタ数125　(b)クラスタ数216

Fig. 4.9 B2Bデータの商品の遍在度と売り上げ(NMFで商品のみ情報集約)

NMFによって得たクラスタに基づき、商品とユーザー両方の情報集約を行い、算出した売り上げと遍在度の散布図を図(4.10)に示す。商品のみの場合と比べ、遍在度の売り上げの関係は大きく変わらなかった。

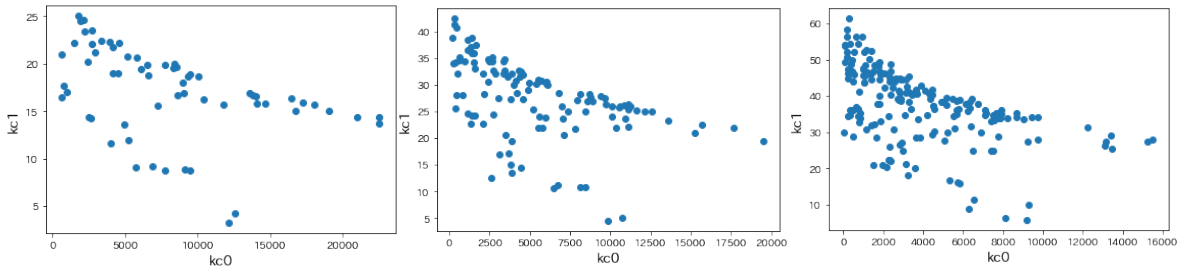


(a)クラスタ数64　(b)クラスタ数125　(b)クラスタ数216

Fig. 4.10 B2Bデータの商品の遍在度と売り上げ(NMFで両方情報集約)

4.4 世帯ごとのデータの複雑度マップ

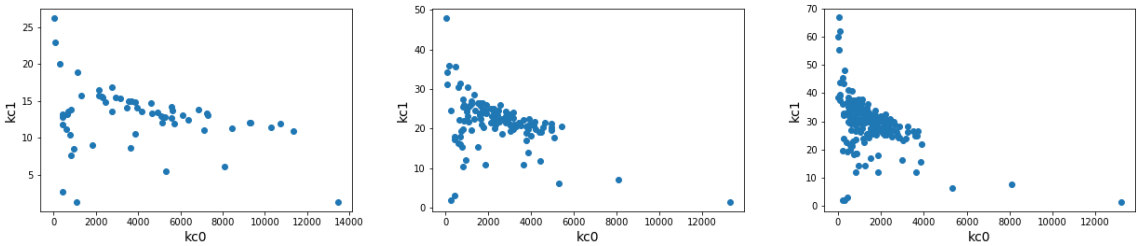
図(4.11)に世帯ごとの購買データからRFMによってユーザーの情報集約を行い、算出したユーザーの複雑度マップを示す。B2Bのデータ RFMから算出したユーザーの複雑度マップと異なり、左下に分布するニッチなクラスタが存在し、2方向に枝分かれするような形となっている。この左下の部分は、ガソリンを購入しているユーザーが属すクラスタであった。



(a)ランク数64 (b)ランク数125 (c)ランク数216

Fig. 4.11 世帯データのユーザー複雑度マップ(RFM)

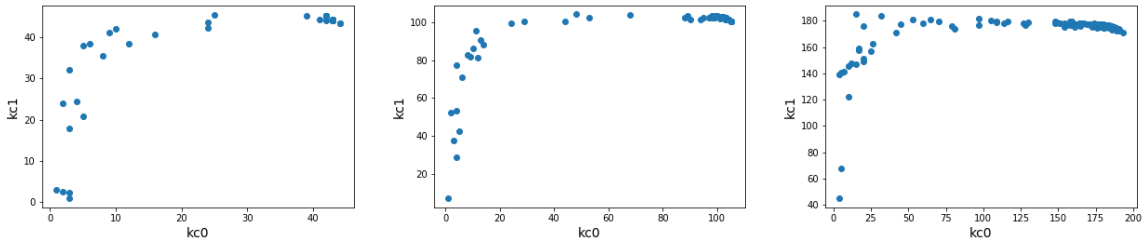
図(4.12)にNMFによって得たクラスタに基づき、ユーザーのみ情報集約を行い、算出したユーザーの複雑度マップ示す。クラスタ数を増やすほど左側に多くのクラスタが集まり、突出して多様度の高いクラスタが出来る事が確認できた。このクラスタはガソリンを多く購入するユーザーが属すクラスターであった。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.12 世帯データのユーザー複雑度マップ(NMFでユーザーのみ情報集約)

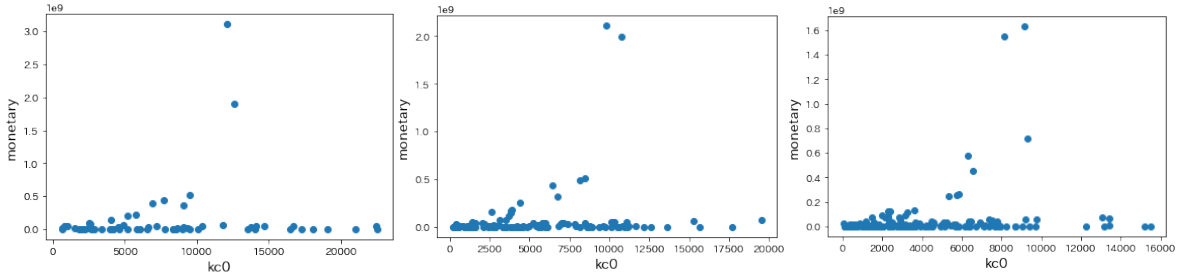
図(4.13)にNMFによって得たクラスタに基づき、ユーザーと商品について情報集約を行い、算出したユーザーの複雑度マップ示す。両方の情報集約を行うと複雑度マップはB2Bのデータと時と同様に、右下にはクラスタが分布しない形となった。



(a)クラスタ数64 (b)クラスタ125 (c)クラスタ数216

Fig. 4.13 世帯データのユーザー複雑度マップ(NMFで両方の情報集約)

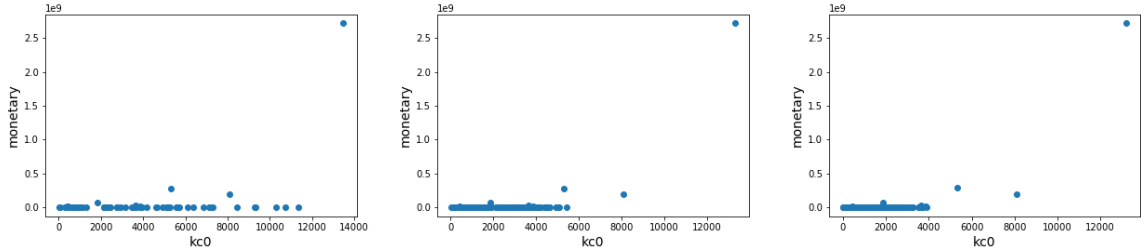
図(4.14)にRMFによってユーザーの情報集約を行った時の多様度と購買額の散布図を示す。多様度が中程度に購買額の半分以上を占めるクラスタが確認されたが、このクラスタは購買額の内の多くをガソリンが占めるクラスタであった。



(a)ランク数64 (b)ランク数125 (c)ランク数

Fig. 4.14 世帯データの多様度と購買額の散布図(RFM)

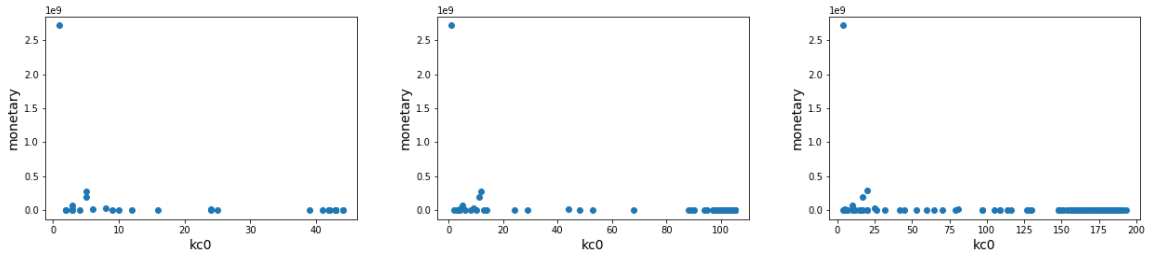
図(4.15)にNMFによって得たクラスタに基づき、、ユーザーのみ情報集約を行った時の、多様度と購買総額の散布図を示す。多様度が高いクラスタが全体の購買額の7割を占めていたが、ガソリンを購入していたユーザーがほとんどこのクラスタに属していた。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

Fig. 4.15 世帯データの多様度と売り上げ散布図(NMFでユーザーのみ情報集約)

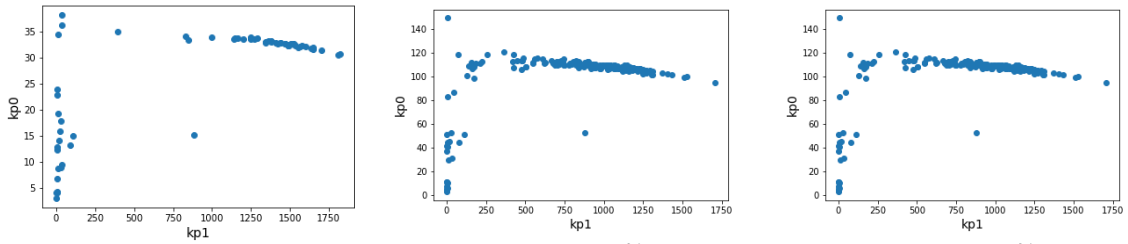
図(4.16)にNMFによって得たクラスタに基づき、、ユーザーと商品の情報集約を行った時の多様度と購買額の散布図を示す。多様度の低いクラスタが全体の購買額の多くを占めるが、このクラスタの購買の内訳のほとんどがガソリンである。そのため、ガソリン購入者の属すクラスタはガソリンの属す商品クラスタと他数個の商品クラスタでしか優位ではなかったため、多様度が低くなったと考えられる。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

Fig. 4.16 世帯データの多様度と売り上げ散布図(NMFで両方の情報集約)

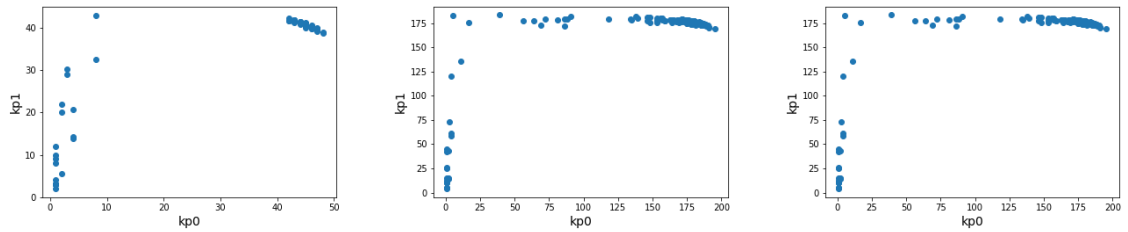
図(4.17)にNMFによって得たクラスタに基づき、商品のみ情報集約を行い算出した商品の複雑度マップ示す。多様度の低いクラスタとそのクラスタを優位に購入しているユーザーの多様度平均が、同程度のクラスタがほとんとであることが確認できる。



(a)クラスタ数64 (b)クラスタ数125 (b)クラスタ数216

Fig. 4.17 世帯データの商品複雑度マップ(NMFで商品のみ情報集約)

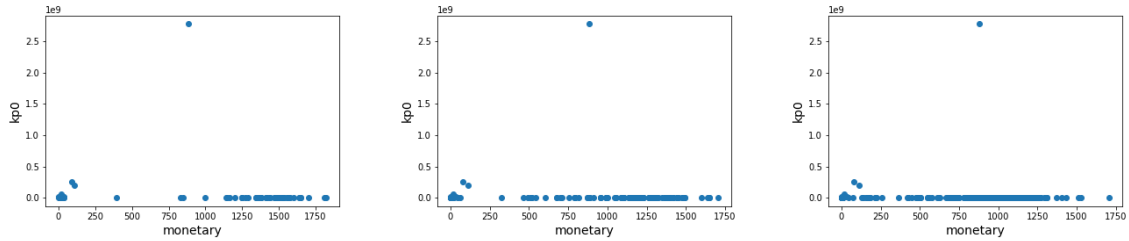
図(4.18)にNMFによって得たクラスタに基づき、ユーザーと商品の両方について情報集約を行い、算出した商品の複雑度マップ示す。NMFによって商品のみ情報集約した場合と大きく異ならない事が確認できる。



(a)クラスタ数64 (b)クラスタ数125 (b)クラスタ数216

Fig. 4.18 世帯データの商品複雑度マップ(NMFで両方の情報集約)

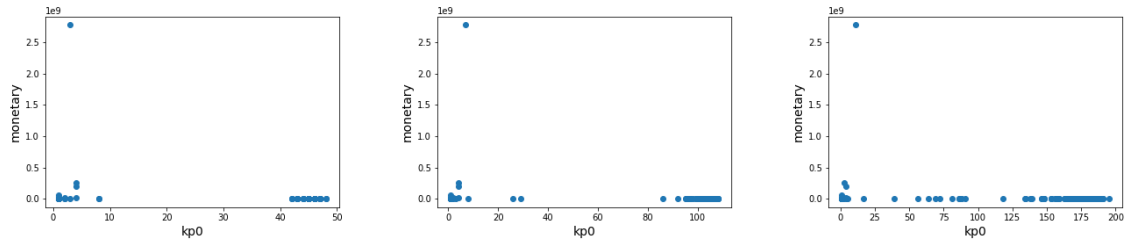
NMFによって得たクラスタに基づき、商品のみ情報集約を行い算出した遍在度と売り上げの散布図を図(4.19)に示す。遍在度が中程度のクラスタが売り上げの大半を占めているが、これはガソリンの属すクラスタであった。



(a)クラスタ数64　(b)クラスタ数125　(b)クラスタ数216

Fig. 4.19 世帯データの遍在度と売り上げ(NMFで商品のみ情報集約)

NMFによって得たクラスタに基づき、商品とユーザーの情報集約を行い、算出した遍在度と売り上げの散布図を図(4.20)に示す。遍在度の低いクラスタが売り上げの大半を占めているが、これはガソリンの属すクラスタであった。ガソリンを購入しているユーザーが1つのクラスタに集約されたため、ガソリンの属すクラスタの遍在度が低くなったと考えられる。



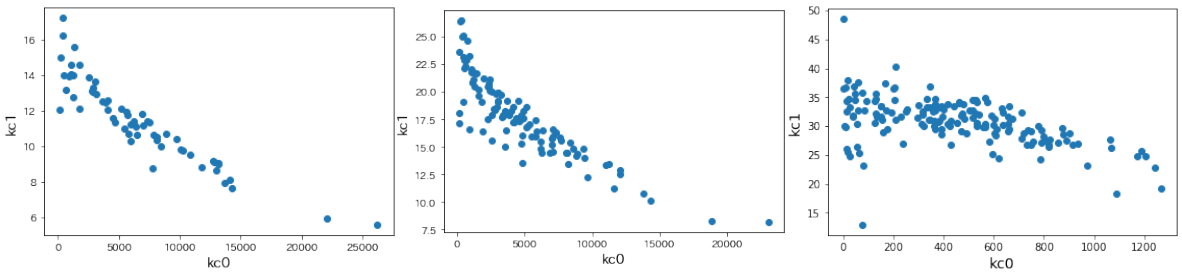
(a)クラスタ数64　(b)クラスタ数125　(b)クラスタ数216

Fig. 4.20 世帯データの遍在度と売り上げ(NMFで両方の情報を集約)

4.5 ガソリン抜きの世帯ごとの購買データ

世帯ごとデータにおいて、ガソリンが全体の購買金額でもその割合を大きく占めるため、比較優位の計算や、NMFによるクラスタリングへの影響が大きい。よって、データからガソリンを除いて再度計算を行った。

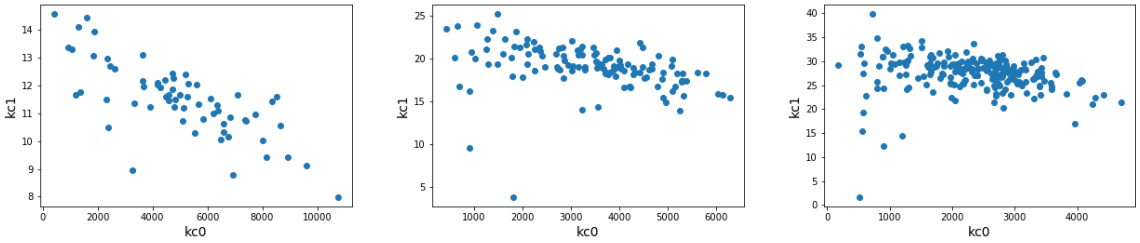
図(4.21)にデータからガソリンを除きRFMでユーザーの情報集約を行いユーザーの複雑度マップを算出した結果を示す。 ランク数64と125では石油を除く前に存在した左下の枝分かれ部分がなくなり左上から右下にかけてクラスタが分布する形となった。ランク数216では分類が細かすぎるためか、多様度が低く購入している商品の平均遍在度が低いニッチなクラスタが確認できる。



(a)ランク数64 (b)ランク数125 (c)ランク数215

Fig. 4.21 世帯データ(ガソリン抜き)のユーザーの複雑度マップ(RFM)

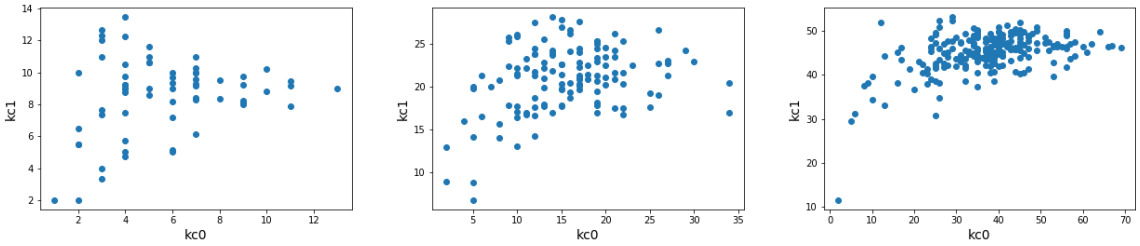
図(4.22)にデータからガソリンを除き、NMFで得たクラスタに基づき、ユーザーの情報集約を行い、ユーザーの複雑度マップを算出した結果を示す。ガソリンを購入しているユーザーが属していた、突出して多様度の高いクラスタがなくなったのが確認できる。また左下に購買商品の遍在度平均が突出して低いニッチな購買傾向を持つクラスタが分布する事が確認できる。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

Fig. 4.22 世帯データ(ガソリン抜き)のユーザーの複雑度マップ(NMFでユーザーのみ情報集約)

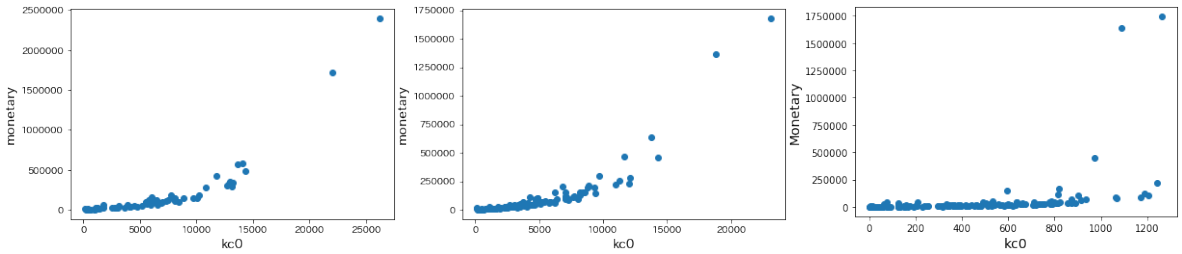
図(4.23)にデータからガソリンを除き、NMFで得たクラスタに基づき、ユーザーと商品両方の情報集約を行い、算出したユーザーの複雑度マップを示す。両方の情報集約した場合に見られる、右下に分布するクラスタが存在しない複雑度マップが得られた。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

Fig. 4.23 世帯データ(ガソリン抜き)の複雑度マップ(NMFで両方の情報を集約

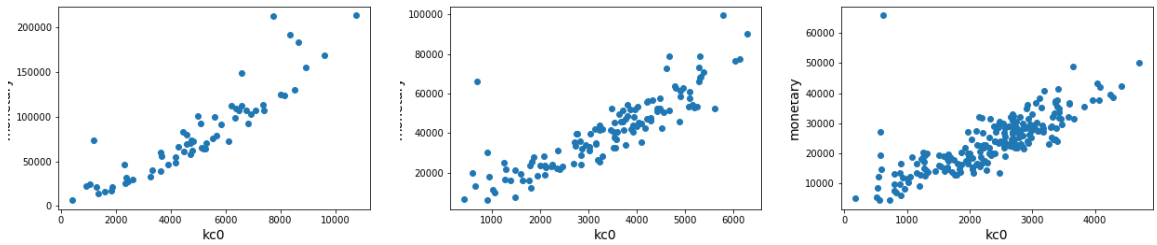
図(4.24)にデータからガソリンを除き、RFMでユーザーの情報集約を行い算出した、ユーザーの多様度と購買額の散布図を示す。多様度と購買額の間に正の相関が確認できる。ランク数を216にすると2つのセグメントに売り上げの多くが集まる結果となった。



(a)ランク数64 (b)ランク数125 (c)ランク数216

Fig. 4.24 世帯データ(ガソリン抜き)の多様度と購買額(RFM)

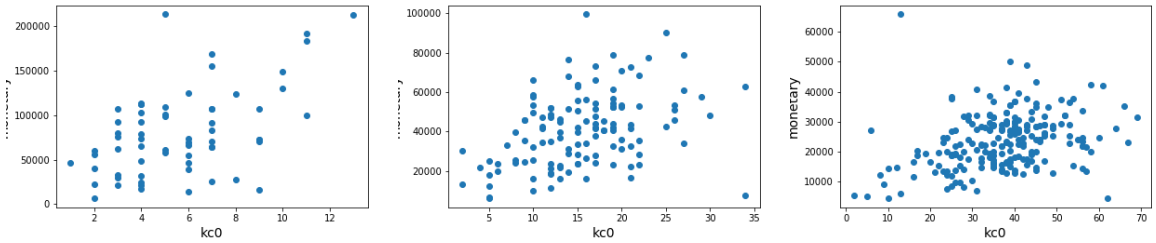
図(4.25)にデータからガソリンを除き、NMFで得たクラスタに基づき、ユーザーの情報集約を行い 算出したユーザーの多様度と購買額の散布図を示す。多様度と購買額の間に強い正の相関が確認できる。これは買う商品が多様なクラスタほど、購買額も高くなり、また突出して特殊な購買傾向を持つようなクラスタはごくわずかで、どのクラスタ数も似た購買傾向を持つことを示していると考えられる。



(a)ランク数64 (b)ランク数125 (c)ランク数216

Fig. 4.25 世帯データ(ガソリン抜き)の多様度と購買額(NMFでユーザーのみ情報集約)

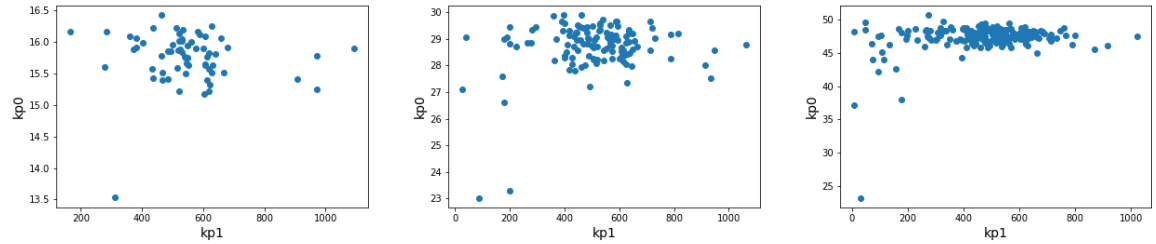
図(4.26)にデータからガソリンを除き、NMFで得たクラスタに基づき、ユーザーと商品の情報集約を行い算出したユーザーの多様度と購買額の散布図を示す。ユーザーのみ情報集約を行った場合と比べて、多様度が高くても購買額は高くないクラスタが存在し、相関が弱くなったことが確認できる。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

Fig. 4.26 世帯データ(ガソリン抜き)の多様度と購買額(NMFで両方の情報集約)

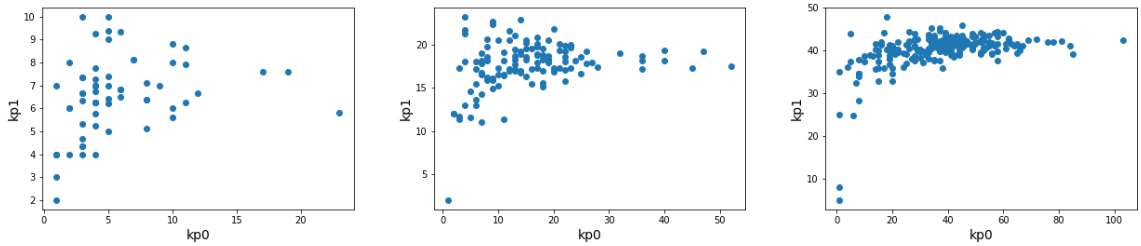
図(4.27)にデータからガソリンを除き、NMFで得たクラスタに基づき、商品の情報集約を行い商品の複雑度マップを算出した結果を示す。一部左下に属すクラスタを除いて、ほぼ横並びであった。つまり遍在度は変わるが、商品クラスタに対して優位なユーザーの平均多様度はほんとど変わらないという事である。



(a)クラスタ数64　(b)クラスタ数125　(b)クラスタ数216

Fig. 4.27 世帯データ(ガソリン抜き)の商品複雑度マップ(NMFで商品のみ情報集約)

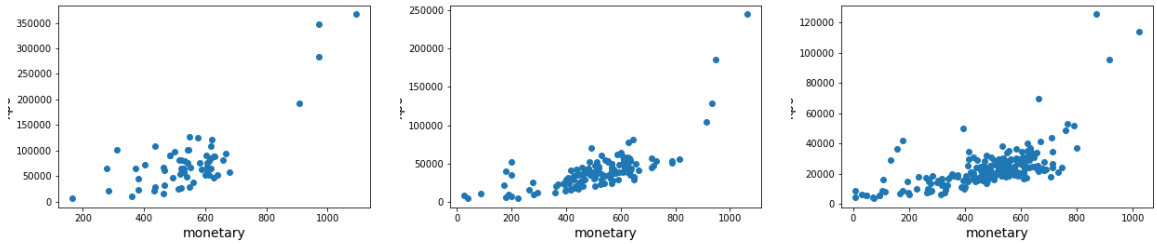
図(4.28)にデータからガソリンを除きNMFで商品とユーザーの情報集約を行い、商品の複雑度マップを算出した結果を示す。両側の情報を集約するとやはり右下にクラスタが分布しない結果となった。



(a)NMF商品のみ (b)NMF両方 (b)NMF両方

Fig. 4.28 世帯データ(ガソリン抜き)の商品複雑度マップ(NMFで両方の情報集約)

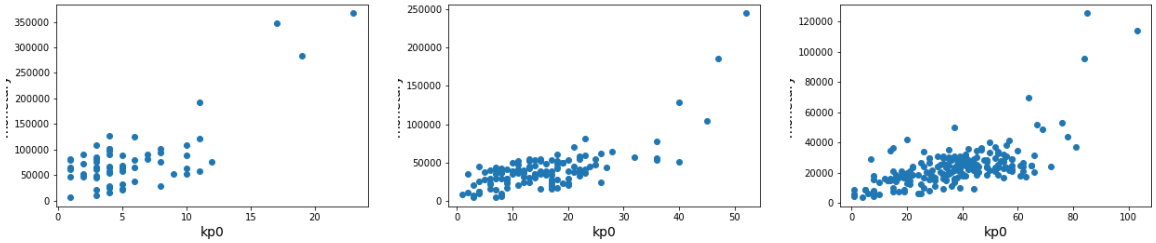
図(4.29)にデータからガソリンを除き、NMFで得たクラスタに基づき、商品の情報集約を行い算出した商品の遍在度と売り上げの散布図を示す。遍在度と売り上げに正の相関が見られた。多くのユーザーに優位に購入されている、人気な商品ほど売り上げが多いと考えられる。またB2Bのデータと異なり、ニッチだが売り上げの高い商品クラスタは存在しなかった。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.29 世帯データ(ガソリン抜き)の遍在度と売り上げ(NMFで商品のみ情報集約)

図(4.30)にデータからガソリンを除き、NMFで得たクラスタに基づき、商品とユーザーの両方を情報集約を行い算出した商品の遍在度と売り上げの散布図を示す。商品のみの時と大きな差異は見られなかった。



(a)クラスタ数64　(b)クラスタ数125　(c)クラスタ数216

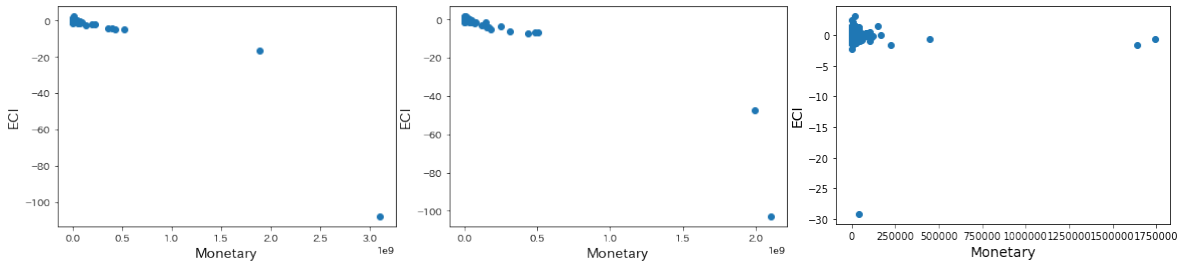
Fig. 4.30 世帯データ(ガソリン抜き)の遍在度と売り上げ(NMFで両方の情報集約)

4.6 購買データの経済複雑性指標

経済複雑性指標も複雑度マップと同様に、NMFとRFMのセグメントを行ってから計算を行い、どのような傾向や違いが見られるか、売り上げとどのような関係があるかを検討した。また2年分のデータから、1年目の経済複雑性指標と2年目の売り上げを比較する事で将来の優良顧客や人気商品を予測することが可能か検討した。 NMFでユーザーと商品の両方を情報集約した複雑度マップが全て似通った分布となったため、本稿ではユーザーと商品の両方を情報集約した場合の複雑性指標については省略する。

4.7 B2Bのデータの複雑性指標

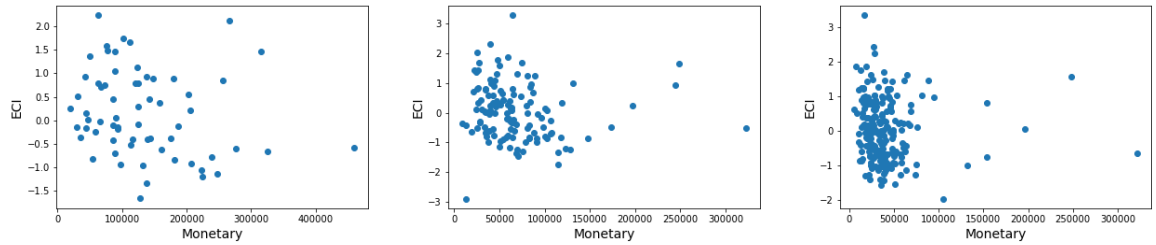
図(4.31)にRFMで情報集約したユーザーから算出した一年目の複雑性指標と購買額の散布図を示す。クラスタ数64と125の場合は、売り上げとECIの間に負の相関が見られる。クラスタ数を216にした場合は、負の相関がなくなったのが確認できる。これは複雑度マップと同様でユーザーを細かく分類しすぎたために、全体の傾向から大きく外れたクラスタが出来たからであると考えられる。



(a)ランク数64 (b)ランク数125 (c)ランク数216

Fig. 4.31 1年目のユーザーの複雑性指標と購買額(RFM)

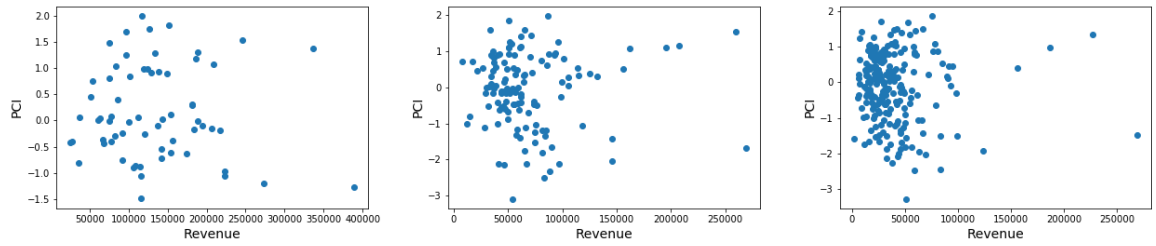
図(4.32)にNMFで得たクラスタに基づき、ユーザーの情報集約し算出した1年目のユーザーの複雑性指標と購買額の散布図を示す。 複雑性指標が負の側に購買額の高いユーザーが多い事が分かる。クラスタ数を増やすほど傾向が顕著になっていくが、複雑性指標が正になるようなニッチな商品を優位に買っているユーザーはクラスタ数を増やした時に、新たに別のクラスタに分類されることが多いためであると考えられる。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.32 1年目のユーザーの複雑性指標と購買額(NMFでユーザーのみ情報集約)

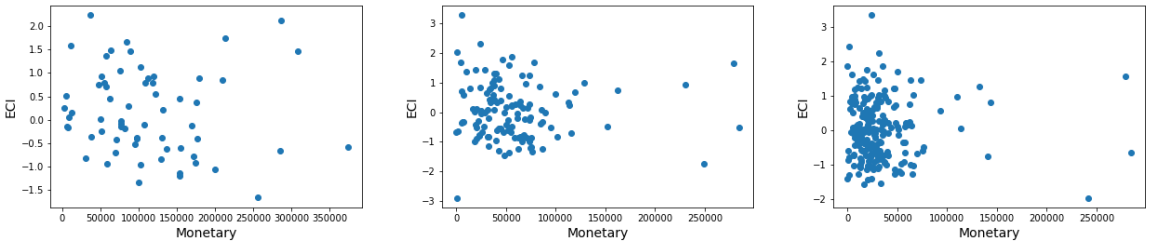
図(4.33)にNMFで得たクラスタに基き、商品の情報集約し算出した、1年目の商品の複雑性指標と売り上げの散布図を示す。 商品の複雑性指標の正負に関わらず、両側に売り上げの高い商品クラスタが分布していることがわかる。複雑性指標が負になるような遍在度の高い人気商品クラスタと複雑性指標が正になるようなニッチな商品クラスタを多く購入しているユーザーが存在するため、このような結果になると考えられる。



(b)クラスタ数64 (b)クラスタ数125 (b)クラスタ数216

Fig. 4.33 1年目のユーザーの複雑性指標と購買額(NMFで商品のみ情報集約)

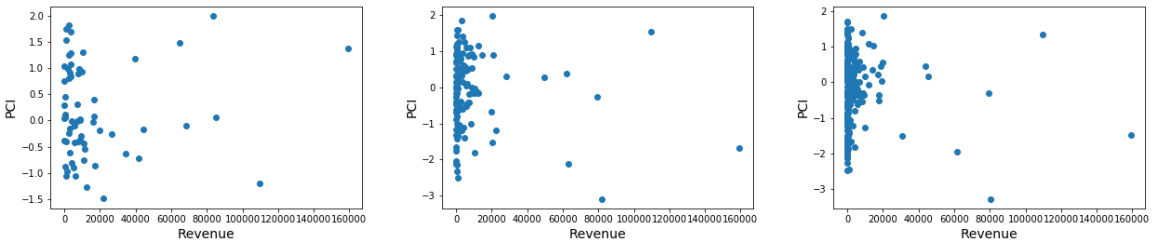
図(4.34)にNMFで得たクラスタに基づき、ユーザーの情報集約し、算出した1年目のユーザーの複雑性指標と2年目の購買額の散布図を示す。1年目と同様に、複雑性指標が正負どちらであっても購買額の高いクラスタが存在することがわかる。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.34 1年目のユーザーの複雑性指標と2年目の購買額(NMFでユーザーのみ情報集約)

図(4.34)にNMFで得たクラスタに基づき、商品の情報集約し算出した1年目の商品の複雑性指標と2年目の購買額の散布図を示す。こちらも複雑性指標の正負に関わらず売り上げの高いクラスタが存在することがわかる。



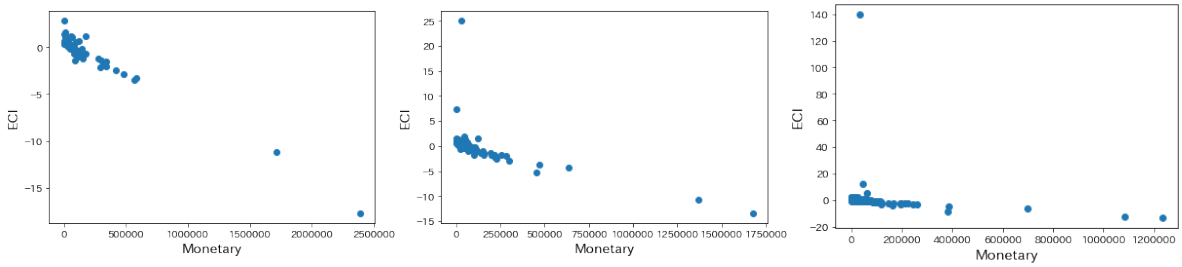
(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.35 1年目のユーザーの複雑性指標と2年目の購買額(NMFで商品のみ情報集約)

4.8 世帯ごと購買データの複雑性指標

世帯ごとデータでは石油を除いた結果と売り上げの図を示す。

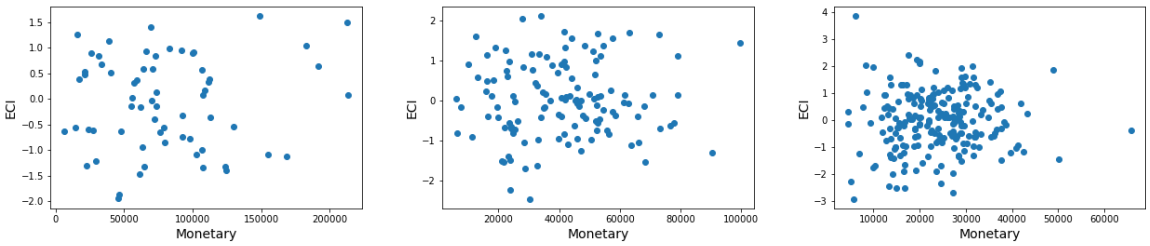
図(4.36)にRFMで情報集約したユーザーから、算出した一年目のユーザーの複雑性指標と売り上げの散布図を示す。ランク数64では購買額と複雑性指標の間に相関が見られる。ランク数125と216では他と比べて大きく正に外れている値も存在するがその数点を除けば、負の相関があることが確認できる。



(a)ランク数64 (b)ランク数125 (c)ランク数216

Fig. 4.36 1年目のユーザーの複雑性指標と購買額(RFM)

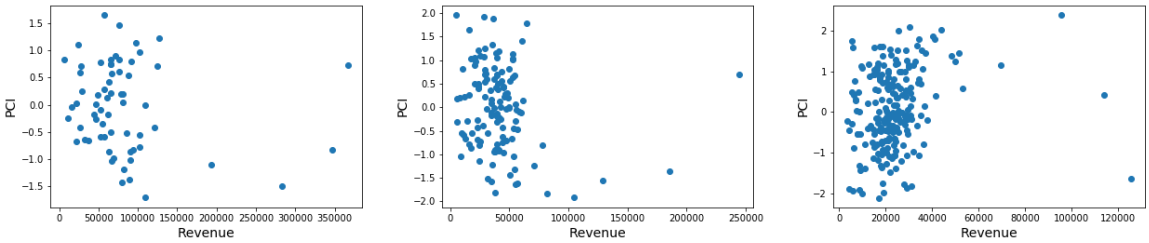
図(4.37)にNMFで得たクラスタに基づき、ユーザーの情報集約し算出した1年目のユーザーの複雑性指標と購買額の散布図を示す。 経済複雑性指標の値と購買額が無相関なのが確認できる。クラスタ数が変わっても大きな変化は見られなかった。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.37 1年目のユーザーの複雑性指標と購買額(NMFでユーザーのみ情報集約)

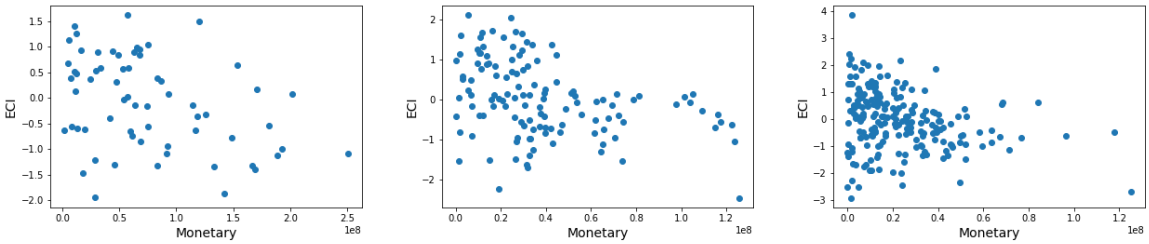
図(4.38)にNMFで得たクラスタに基づき、商品の情報集約し、算出した1年目の商品の複雑性指標と売り上げの散布図を示す。 商品の複雑性指標が負の側に売り上げの多い商品が集まっている。クラスタ数が216の時は大きく変化したが、これは固有値計算の過程でで正負が反転したものと考えられる。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.38 1年目の商品の複雑性指標と売り上げ(NMFで商品のみ情報集約)

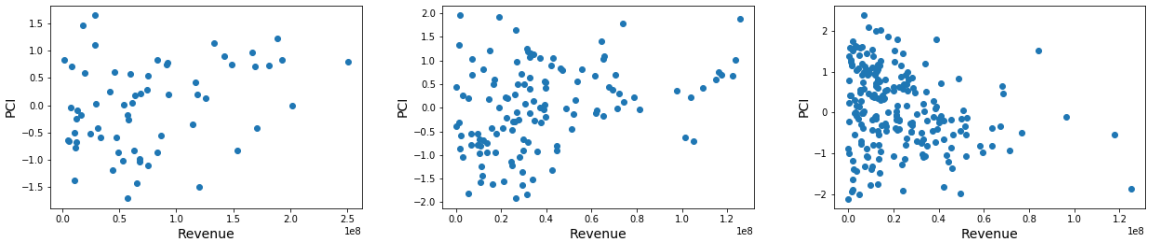
図(4.39)にNMFで得たクラスタに基づき、ユーザーの情報集約し算出した1年目のユーザーの複雑性指標と2年目の購買額の散布図を示す。 クラスタ数64では特に関係性は見られない。しかしクラスタ数125と216ではユーザーの複雑性指標の値が低い側に購買額の高いユーザーが集まっているのが確認できる。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.39 1年目のユーザーの複雑性指標と2年目の購買額(NMFでユーザーのみ情報集約)

図(4.40)にNMFで得たクラスタに基づき、商品の情報集約し、算出した1年目の商品の複雑性指標と2年目の売り上げの散布図を示す。 商品の複雑性指標の高い商品が2年目は多く売れているが、この購買データにおいて、遍在度の高い商品である飲料や卵のような日常的に買う商品は各世帯で消費が大きく変動するようなことはないが、遍在度が低い雑誌や嗜好品は消費が大きく変動する事があるためだと考えられる。



(a)クラスタ数64 (b)クラスタ数125 (c)クラスタ数216

Fig. 4.40 1年目の商品の複雑性指標と2年目の売り上げ(NMFで商品のみ情報集約)

第5章　おわりに

本研究では、複雑度マップと経済複雑性指標を購買データに対して適用するため方法を提案した。実購買データとして二つのデータに対してこれらを適用して比較てし、以下の結果を得た。

- RFMによってクラスタリングした場合、国の複雑度マップと同様に左上から右下にかけてクラスタが分布するような結果が得られ、NMFによってクラスタリングした場合、左下に属す顕著にニッチなクラスタが存在する複雑度マップが得られる、全体の大まかな購買傾向を把握する事ができる。
- NMFで得たクラスタに基づき、ユーザーと商品の両方の情報集約を行ってしまうとデータに関わらず複雑度マップが似通った形になってしまうため適切ではない。
- RFMクラスタリングを施したユーザーの複雑性指標と売り上げの間には、2つのデータどちらでも負　の相関関係が確認できた。
- NMFで情報集約した場合の複雑性指標と売り上げの間には相関は無く、何らかの知見を得る事は難しいと考えらえる。

どちらの方法でも全体の大まかな購買傾向を把握出来るような複雑度マップを得る事が出来たが、複雑性指標は何らかの知見を得られるような結果は得られなかった。 全体に大きな影響を与えるユーザーや、商品といった外れ値は購買データにおいては常に存在すると考えられる。本研究ではデータから除外するという対処を行ったが、そういったユーザーや商品も重要な情報であるかもしれず、そういった外れ値を考慮した比較優位や複雑性指標の算出方法も今後の課題として挙げられる。

謝辞

本研究を行うにあたり、主指導教員である林幸雄教授からは多くの助言の言葉だけでなく、大学院での研究生活を行う上で手厚いサポートをして頂いて下さったことに深く感謝申し上げます。また副指導教員である金沢大学の寒河江雅彦教授からは経済学、統計学の視点からの有益な助言を賜りました事、感謝申し上げます。大学生活を有意義にしてくれた、研究室のメンバーにも感謝申し上げます。

References

[Sou] Source files - dunnhumby. https://www.dunnhumby.com/source-files/. (Accessed on 01/28/2021).

[UCI] Uci machine learning repository: Online retail ii data set. https://archive.ics.uci.edu/ml/datasets/Online+Retail+II. (Accessed on 01/29/2021).

[3] Balassa, B. (1965). Trade liberalisation and "revealed" comparative advantage 1. *The manchester school*, 33(2):99–123.

[4] Hausmann, R., Hidalgo, C. A., Bustos, S., Coscia, M., and Simoes, A. (2014). *The atlas of economic complexity: Mapping paths to prosperity*. Mit Press.

[5] Mealy, P., Farmer, J. D., and Teytelboym, A. (2018). A new interpretation of the economic complexity index. *Alexander, A New Interpretation of the Economic Complexity Index (February 4, 2018)*.

[6] Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

[7] Wei, J.-T., Lin, S.-Y., and Wu, H.-H. (2010). A review of the application of rfm model. *African Journal of Business Management*, 4(19):4199–4206.

[8] 澤田宏 (2012). 非負値行列因子分解 nmf の基礎とデータ/信号解析への応用. 信学誌, 95(9):829–833.