統計数理(2018)
第 66 巻 第 2 号 279–294

特集「サービス科学の今」
[原著論文]

集約的シンボリックデータのカイ2乗統計量を用いた非類似度とその不動産情報データへの適用

清水 信夫[1]・中野 純司[1]・山本 由和[2]

(受付 2017 年 11 月 9 日；改訂 2018 年 9 月 18 日；採択 9 月 19 日)

要　　旨

近年，サービス科学においては連続変数とカテゴリー変数が混在している大量のデータが得られることが多い．そしてそれらの個体データはいくつかの自然なグループに分かれる場合がある．そのとき，個々の個体データそのものではなく，その集合であるグループに対する推論および解析に興味があることがある．われわれは，そのようなグループを表すためにいくつかの記述統計量の集合をデータと考え，それを集約的シンボリックデータ(Aggregated Symbolic Data，ASD)と呼ぶ．ここでは，連続変数とカテゴリー変数がともに含まれる場合に，2 次以下のモーメントに関する統計量を ASD と考える．また，連続変数をカテゴリー化することによりすべての変数について同様の基準によるカイ 2 乗統計量を考えた上で，それらの和として ASD 間の非類似度を構成する手法を提案する．そして，この方法を東京都区部の不動産情報データに適用し，各区ごとのデータの集合を考え，それらの ASD を計算する．さらに各 ASD の値から区の間の非類似度を求め，各区の階層的クラスタリングおよび多次元尺度構成法による分析を行う．

キーワード：Burt 行列，カイ 2 乗統計量，階層的クラスタリング，多次元尺度構成法，ビッグデータ．

1. はじめに

近年，サービス産業においては Web システムを用いたデータの収集が多用されており，その活動の詳細なデータが計算機上に連続的に蓄積されるようになっている．それらのデータは連続変数とカテゴリー変数が混在した多次元データであることが多い．また，それらのデータ数は非常に多く，いわゆる“ビッグデータ”の代表例となっている．

このようなデータの全体像を見るためには，個体データに着目するこれまでの方法ではその個数や変数の多さのために計算が困難な場合がある．ただし，そのようなときには個体データは意味のある自然な比較的少数のグループに分かれることが多い．したがって，個体データそのものではなく，個体がまとめられたグループに関心を向けた手法が必要である．その方法の一つとして Diday (1988)はシンボリックデータ解析を提唱した．

シンボリックデータ解析においては，データとして各連続変数ごとに 1 つの値ではなく，ある値を中心としてばらつきをもつデータ(区間データや分布値データ)などで表されるものが考

[1] 統計数理研究所：〒 190–8562 東京都立川市緑町 10–3
[2] 徳島文理大学 理工学部：〒 769–2193 香川県さぬき市志度 1314–1

表 1. 東京都区部の不動産情報データ(一部).

No.	区	賃料 (万円)	面積 (m^2)	物件種別	構造種別	···	管理形態
1	荒川区	8.25	26.83	マンション	鉄筋コンクリート	···	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
557	北区	19.80	89.84	一戸建て	木造	···	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
4588	港区	22.30	71.28	マンション	鉄筋コンクリート	···	巡回管理
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
17512	中央区	21.20	75.46	マンション	鉄骨鉄筋	···	巡回管理
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
498088	足立区	6.40	33.34	アパート	軽量鉄骨	···	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮
714202	新宿区	16.40	55.64	マンション	鉄骨鉄筋	···	常駐管理
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮

えられ，それらに従来の各種多変量解析手法を拡張する研究が Bock and Diday (2000)，Billard and Diday (2006)，Diday and Noirhomme-Fraiture (2008)などにまとめられている．それら以外にも，シンボリックデータに対するクラスタリングに関しては Verde (2004)や Irpino and Verde (2006)など，多次元尺度構成法に関しても Denœux and Masson (2000)や Groenen et al. (2006)などの研究がある．これまでのシンボリックデータ解析においては，データは最初から区間のような形で与えられている場合が多く，そこではグループ内の複数の変数間の関係は無視される．例えば，2 つの連続変数間の相関関係は考慮されない．

しかしながら，現代のビッグデータにおいては、元の個体データは保持されている．超大量データの場合は移動や計算に困難を伴うが，どうしても必要ならばグループに関するいかなる記述統計量も計算することは可能である．そこで，グループにおける多次元データの情報を可能な限り簡潔な形で持つために，複数の記述統計量を考えることにし，それを集約的シンボリックデータ(Aggregated symbolic data，ASD)と呼ぶこととする．

われわれはそのようなビッグデータとして，表 1 で示される大規模な不動産情報データを持っている．このデータはいくつかのグループに自然に分けることができ，ASD の適用が有効なデータと考えられる．

本論文では，グループ内の個体データのそれぞれの変数および複数の変数の組み合わせに関して 2 次までのモーメントに関する統計量を用い，それをグループを表す ASD と考える．第 2 節で連続変数とカテゴリー変数が混在する多次元データにおける ASD を定義し，その意味を考える．第 3 節では ASD 間の非類似度をカイ 2 乗統計量を用いて表す手法を提案する．第 4

節では，提案した手法を表 1 で示される東京都区部における不動産情報データに適用し，23 区ごとのデータを 23 個の ASD として考え，その相互間の非類似度を算出して階層的クラスタリングや多次元尺度構成法を行った結果について考察する．第 5 節では，本研究に関するまとめについて述べる．

2. 集約的シンボリックデータ

ここでは p 個の連続変数および q 個のカテゴリー変数からなる n 個の個体データが与えられている場合を考える．それらの個体データは G 個の自然な意味のあるグループに分かれると仮定する．グループ g に含まれる個体データの数を $n^{(g)}$ とし，グループ g の個体 i の連続変数 l の値を $x_{il}^{(g)}$ と書く．カテゴリー変数 k は m_k 個のカテゴリー値を取るとすると，それは m_k 個のダミー変数で表すことができる．すなわちグループ g の個体 i のカテゴリー変数 k が j 番目のカテゴリー値を取るとき $x_{ij}^{(g,k)}$ は 1，それ以外は 0 とする．するとグループ g のすべての個体データは

$$X^{(g)} = \begin{bmatrix} x_{11}^{(g)} & \cdots & x_{1p}^{(g)} & x_{11}^{(g,1)} & \cdots & x_{1m_1}^{(g,1)} & \cdots & x_{11}^{(g,q)} & \cdots & x_{1m_q}^{(g,q)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{n^{(g)}1}^{(g)} & \cdots & x_{n^{(g)}p}^{(g)} & x_{n^{(g)}1}^{(g,1)} & \cdots & x_{n^{(g)}m_1}^{(g,1)} & & x_{n^{(g)}1}^{(g,q)} & \cdots & x_{n^{(g)}m_q}^{(g,q)} \end{bmatrix} \tag{2.1}$$

と表すことができる．$X^{(g)}$ の最初の p 列からなる部分行列 $X_1^{(g)}$ は p 個の連続変数に対応する．また部分行列

$$X_2^{(g,k)} = \begin{bmatrix} x_{11}^{(g,k)} & \cdots & x_{1m_k}^{(g,k)} \\ \vdots & & \vdots \\ x_{n^{(g)}1}^{(g,k)} & \cdots & x_{n^{(g)}m_k}^{(g,k)} \end{bmatrix} \tag{2.2}$$

はカテゴリー変数 k のダミー変数からなる行列であり，q 個のカテゴリー変数に対応するのは $X_2^{(g)} = [X_2^{(g,1)} \cdots X_2^{(g,q)}]$ である．これらを用いると $X^{(g)} = [X_1^{(g)}\ X_2^{(g)}]$ である．

個体数 $n^{(g)}$ が非常に大きいとき，$X^{(g)}$ を保持し続けるのは記憶領域の制約などのために困難を伴う．またそのような状況で $X^{(g)}$ をそのまま用いた詳細な計算は長い時間がかかり，データの全体像を捉えるという面での意義も乏しい．そこでこのグループを表すいくつかの記述統計量を考え，それを用いてグループに対する統計的推論を行うことにする．

ひとつの連続変数データを集約する簡単な方法は標本平均と標本分散を用いることである．さらに詳しい情報として尖度，歪度を用いることもある．複数の連続変数に関しては標本相関係数も用いられる．これらはモーメントを表す記述統計量である．各グループにおけるデータの情報をモーメントに関する記述統計量で表す場合，低次のモーメントによる統計量だけでは多くの情報が抜け落ちてしまうし，高次のモーメントによる統計量を多く用いると，情報の脱落は少なくなるものの保持すべき値が多くなり扱いが難しくなる．そこで，われわれは，2 次以下のモーメントにより表される記述統計量の集合を考え，これを集約的シンボリックデータ (Aggregated symbolic data，ASD) と呼ぶことにする．

まず，重要な情報としてグループ g の個体数 $n^{(g)}$ が考えられる．これはデータの値の 0 乗 (=1) の合計と考えると 0 次のモーメントと言える．

次に 1 次のモーメントは各変数ごとの和に対応する．これは $X^{(g)}$ に関しては

$$\mathbf{1}'_{n^{(g)}} X^{(g)}/n^{(g)} = [\bar{x}_1^{(g)}, \ldots, \bar{x}_p^{(g)}, \hat{p}_1^{(g,1)}, \ldots, \hat{p}_{m_1}^{(g,1)}, \ldots, \hat{p}_1^{(g,q)}, \ldots, \hat{p}_{m_q}^{(g,q)}] \tag{2.3}$$

と同じ情報である．ここで $\mathbf{1}'_{n^{(g)}}$ はすべての成分が 1 である $n^{(g)}$ 次元横ベクトルを表す．明らかにこれらは各連続変数の平均および各カテゴリー変数の周辺分布のパラメータである．

さらに 2 次のモーメントは(2.1)式より

$$X^{(g)'}X^{(g)} = \begin{bmatrix} X_1^{(g)'}X_1^{(g)} & X_1^{(g)'}X_2^{(g)} \\ X_2^{(g)'}X_1^{(g)} & X_2^{(g)'}X_2^{(g)} \end{bmatrix} \equiv \begin{bmatrix} S_{11}^{(g)} & S_{12}^{(g)} \\ S_{21}^{(g)} & S_{22}^{(g)} \end{bmatrix} \tag{2.4}$$

を考えることになる．$S_{11}^{(g)}$ は連続変数データの積和行列である．$X_2^{(g,k_1)'}X_2^{(g,k_2)} = S^{(g,k_1k_2)}$ がカテゴリー変数 k_1 とカテゴリー変数 k_2 に対する分割表となることを考えると，$S_{22}^{(g)}$ はそれらを部分行列とする Burt 行列である．$S_{12}^{(g)}$ は連続変数とカテゴリー変数に関する 2 次のモーメントを表す $p \times (m_1 + \cdots + m_q)$ 行列であるが，その第 k 部分行列 $X_1^{(g)'}X_2^{(g,k)}$ の (l, j) 成分はカテゴリー変数 k の値がカテゴリー値 j を取る個体における連続変数 l の合計である．

われわれはこれらの記述統計量で各グループの特徴が表されていると考え，これを用いてグループの関係を調べることにする．

3. ASD 間の非類似度

ここで考えているデータは連続変数とカテゴリー変数の両方を含む．これらを統一的に考える方法の一つとして，連続変数をカテゴリー変数に変換する．そして 2 つのグループのカテゴリー変数の分布の差を考えるためにカイ 2 乗統計量を用いることにする．

3.1 2 つのカテゴリー変数の組み合わせに関する非類似度

まず，2 つの ASD g_1, g_2 における異なる 2 つのカテゴリー変数 (k_1, k_2) の組により形成される分割表の間の非類似度を考える．2 つのグループの分割表は $S^{(g_1,k_1k_2)} = [s_{j_1j_2}^{(g_1,k_1k_2)}]$，$S^{(g_2,k_1k_2)} = [s_{j_1j_2}^{(g_2,k_1k_2)}]$ であり，$j_a = 1, \ldots, m_{k_a}$ $(a = 1, 2)$ である．2 つの ASD が同じ性質を持つ場合，それぞれの分割表の各セルの出現確率は等しい．この仮定が正しい場合，分割表のセル (j_1, j_2) の出現個数の期待値の推定量は

$$\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})} = \frac{s_{j_1j_2}^{(g_1,k_1k_2)} + s_{j_1j_2}^{(g_2,k_1k_2)}}{n^{(g_1)} + n^{(g_2)}} n^{(g_a)} \quad (a = 1, 2)$$

と考えられる．一方，2 つの ASD が異なる場合にはそれぞれの分割表の各セルにおける出現個数 $s_{j_1j_2}^{(g_a,k_1,k_2)}$ と $\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})}$ よりカイ 2 乗統計量を求めることができ，その総和を非類似度と考える．この場合，$\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})}$ が分母になるので，これが 0 になるときは無視しなくてはならない．すなわち $s_{j_1j_2}^{(g_1,k_1k_2)} = s_{j_1j_2}^{(g_2,k_1k_2)} = 0$ となるセルは無視してカイ 2 乗統計量を考える．これより

$$\chi^{2(g_1g_2,k_1k_2)} = \sum_{\substack{a=1}}^{2} \sum_{j_1=1}^{m_{k_1}} \sum_{j_2=1}^{m_{k_2}} \frac{\left\{ s_{j_1j_2}^{(g_a,k_1,k_2)} - \widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})} \right\}^2}{\widehat{E(s_{j_1j_2}^{(g_a,k_1k_2)})}} \tag{3.1}$$
$$s_{j_1j_2}^{(g_1,k_1k_2)} + s_{j_1j_2}^{(g_2,k_1k_2)} \geq 1$$

を (k_1, k_2) の組による分割表における ASD 間の非類似度と考えることができる．これを $k_1 < k_2$ なる全ての (k_1, k_2) に関して考え，その総和をとったものが Burt 行列における ASD 間の非類似度

$$d_{(cc)}^{(g_1g_2)} = \sum_{k_1=1}^{q-1} \sum_{k_2=k_1+1}^{q} \chi^{2(g_1g_2,k_1k_2)} \tag{3.2}$$

と考えられる．なお，Burt 行列の対角成分の違いは考えていないことに注意する．それは Burt 行列の対角成分は各カテゴリー変数の周辺分布を表し，その情報は Burt 行列の非対角成分にある分割表の列和，行和として含まれているからである．

3.2　2 つの連続変数の組み合わせに関する非類似度

2 つの連続変数 l_1, l_2 のデータは 2 次元平面上にプロットできる．その平面を格子状に分割し，それぞれの格子をカテゴリー値と考える．2 次元平面 $(-\infty,\infty)\times(-\infty,\infty)$ を各次元ごとに N 個ずつの区間に分割するとし，その各区間の境界値 $h_j^{(l)}$ $(j=0,1,\ldots,N)$ について

$$-\infty = h_0^{(l)} < h_1^{(l)} < \cdots < h_{N-1}^{(l)} < h_N^{(l)} = \infty$$

とする．われわれは ASD のみを保持していると考えるので，各格子内に何個のデータがあるかの情報は持っていない．持っているのはグループ g に含まれる個体数 $n^{(g)}$ および標本平均 $\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)} = \begin{bmatrix} \hat{\mu}_{l_1}^{(g)} \\ \hat{\mu}_{l_2}^{(g)} \end{bmatrix}$ と標本分散共分散行列 $\hat{\Sigma}_{l_1l_2}^{(g)} = \begin{bmatrix} \hat{\sigma}_{l_1l_1}^{(g)} & \hat{\sigma}_{l_1l_2}^{(g)} \\ \hat{\sigma}_{l_2l_1}^{(g)} & \hat{\sigma}_{l_2l_2}^{(g)} \end{bmatrix}$ である．したがって，これらを用いて各セルの個体数を推定することを考える．

これだけの情報からだと，連続変数 l_1, l_2 の実現値 $\boldsymbol{x}_{l_1l_2} = [x_{l_1}, x_{l_2}]'$ の同時分布は標本平均が $\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)}$，標本分散共分散行列が $\hat{\Sigma}_{l_1l_2}^{(g)}$ である 2 変量正規分布に従うと仮定するのが自然である．なお，連続変数が従う確率分布について非対称性などのより複雑な状況を考えるには，3 次以上のモーメントにあたる情報が必要となるので，ここでの ASD を用いる限り考慮することはできない．たとえば，各個体の連続変数の値の分布の非対称性が強い場合は，変数変換を行うなどして非対称性を可能な限り弱め，対称性が担保されているとみなせる変数にした方がよい．そのような変換により，ある程度は外れ値の影響を軽減することができる．

この 2 変量正規分布の密度関数を $\varphi(\boldsymbol{x}_{l_1l_2}|\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)}, \hat{\Sigma}_{l_1l_2}^{(g)})$ と書くと，ASD g において領域であるカテゴリー $[h_{j_1}^{(l_1)}, h_{j_1+1}^{(l_1)}] \times [h_{j_2}^{(l_2)}, h_{j_2+1}^{(l_2)}]$ における出現確率は

$$\hat{p}_{j_1j_2}^{(g,l_1l_2)} = \iint_{[h_{j_1}^{(l_1)}, h_{j_1+1}^{(l_1)}]\times[h_{j_2}^{(l_2)}, h_{j_2+1}^{(l_2)}]} \varphi\left(\boldsymbol{x}_{l_1l_2}|\,\hat{\boldsymbol{\mu}}_{l_1l_2}^{(g)}, \hat{\Sigma}_{l_1l_2}^{(g)}\right) d\boldsymbol{x}_{l_1l_2}$$

となる．これより，異なる 2 つの連続変数 (l_1, l_2) の組による $N\times N$ 分割表をそれぞれ $S^{(g_1,l_1,l_2)} \simeq [\hat{p}_{j_1j_2}^{(g_1,l_1l_2)} n^{(g_1)}]$，$S^{(g_2,l_1,l_2)} \simeq [\hat{p}_{j_1j_2}^{(g_2,l_1l_2)} n^{(g_2)}]$ と近似できる．

2 つの ASD が同じ性質を持つ場合，それぞれの分割表の各セルの出現確率は等しいと仮定できる．これが正しい場合，セル (j_1, j_2) の出現個数の期待値の推定量は

$$\widehat{E(s_{j_1j_2}^{(g_a,l_1l_2)})} = \frac{\hat{p}_{j_1j_2}^{(g_1,l_1l_2)} n^{(g_1)} + \hat{p}_{j_1j_2}^{(g_2,l_1l_2)} n^{(g_2)}}{n^{(g_1)} + n^{(g_2)}} n^{(g_a)} \quad (a=1,2)$$

と書ける．カテゴリー変数同士の組み合わせの場合と同様の基準でカイ 2 乗統計量を考える場合，$\widehat{E(s_{j_1j_2}^{(g_a,l_1l_2)})}$ で割ることになるので，これが 0 もしくは極端に小さな値となるときは無視しなくてはならない．そこで，$\hat{p}_{j_1j_2}^{(g_1,l_1l_2)} n^{(g_1)} + \hat{p}_{j_1j_2}^{(g_2,l_1l_2)} n^{(g_2)} < 1$ となるセルは無視してカイ 2 乗統計量を考える．これより

$$\chi^{2(g_1g_2,l_1l_2)} \simeq \sum_{a=1}^{2}\sum_{j_1=1}^{N}\sum_{j_2=1}^{N}{}_{\hat{p}^{(g_1,l_1l_2)}_{j_1j_2}n^{(g_1)}+\hat{p}^{(g_2,l_1l_2)}_{j_1j_2}n^{(g_2)}\geq 1} \frac{\left\{\hat{p}^{(g_a,l_1l_2)}_{j_1j_2}n^{(g_a)} - E(\widehat{s^{(g_a,l_1l_2)}_{j_1j_2}})\right\}^2}{E(\widehat{s^{(g_a,l_1l_2)}_{j_1j_2}})} \tag{3.3}$$

を非類似度と考えることができる．これを $l_1 < l_2$ なる全ての (l_1, l_2) に関して考え，その総和をとったものが連続変数に関する ASD 間の非類似度

$$d^{(g_1g_2)}_{(rr)} = \sum_{l_1=1}^{p-1}\sum_{l_2=l_1+1}^{p}\chi^{2(g_1g_2,l_1l_2)} \tag{3.4}$$

と考えられる．

残った問題は各区間の境界値 $h^{(l)}_j$ $(j = 0, 1, \ldots, N)$ の定め方である．非類似度はすべてのグループのペアに対して計算しなければならないので，統一性のためにもこの境界値は同一のものを利用するのがよい．そこでわれわれはすべてのデータに関する連続変数 l に対する平均と分散を考え，それを用いた正規分布の確率が同じになるように境界値を取ることにする．すなわち分割数 N に関して

$$\hat{\mu}_l = \frac{1}{n}\sum_{g=1}^{G} n^{(g)}\hat{\mu}^{(g)}_l$$

$$\hat{\sigma}_{ll} = \frac{1}{n}\sum_{g=1}^{G} n^{(g)}\hat{\sigma}^{(g)}_{ll} + \frac{1}{n}\sum_{g=1}^{G} n^{(g)}(\hat{\mu}^{(g)}_l - \hat{\mu}_l)^2$$

を用いた 1 次元正規分布 $\varphi(x_l|\hat{\mu}_l, \hat{\sigma}_{ll})$ が各区間でそれぞれ $1/N$ ずつの確率を持つように $h^{(l)}_j$ を定める．すなわち $h^{(l)}_j = \hat{\mu}_l + \hat{\sigma}_{ll}\boldsymbol{\Phi}^{-1}(j/N)$ となるように取る．ただし $\boldsymbol{\Phi}(x_l)$ は標準正規分布の分布関数である。N の値については，小さくしすぎると分布の特徴がとらえられず，一方で大きくしすぎると各セル内のデータ個数が少なくなり，セル数の増加に伴い計算時間の増大につながる．そのため，適当な範囲内でいくつかの場合に対する結果を求め，その中で適当なものを選べばよい．

3.3 連続変数とカテゴリー変数の組み合わせの場合

連続変数 l とカテゴリー変数 k のペアを考える．連続変数に関しては前節と同様にカテゴリー化する．このペアの場合の 2 次モーメントは(2.4)式の $S^{(g)}_{12}$ に対応するが，この中にはカテゴリー変数 k の各カテゴリー値が j_2 となる個体における連続変数 l の標本分散に対応する値は含まれない．そのため，この場合の標本分散に関しては全て同一の値，すなわち $\hat{\sigma}^{(g)}_{ll}$ を使用せざるを得ないことに注意する．

ここで保持する情報からだと，カテゴリー変数 k のカテゴリー値が j_2 である場合の連続変数 l の実現値 $x^{(g,k)}_{j_2l}$ の分布は，標本平均が $\hat{\mu}^{(g,k)}_{j_2l}$，標本分散が $\hat{\sigma}^{(g)}_{ll}$ である 1 変量正規分布に従うと仮定するのが自然であり，その密度関数を $\varphi(x^{(g,k)}_{j_2l}|\hat{\mu}^{(g,k)}_{j_2l}, \hat{\sigma}^{(g)}_{ll})$ と書く．すると ASD g において区間であるカテゴリー $[h^{(l)}_{j_1}, h^{(l)}_{j_1+1}]$ における出現確率は

$$\hat{p}^{(g,lk)}_{j_1j_2} = \int_{h^{(l)}_{j_1}}^{h^{(l)}_{j_1+1}} \varphi\left(x^{(g,k)}_{j_2l}\middle|\hat{\mu}^{(g,k)}_{j_2l}, \hat{\sigma}^{(g)}_{ll}\right) dx^{(g,k)}_{j_2l}$$

となる．ただし，$h^{(l)}_{j_1}$ は前節と同じものである．これより，連続変数とカテゴリー変数 (l, k) の組による分割表をそれぞれ $S^{(g_1,lk)} \simeq [\hat{p}^{(g_1,lk)}_{j_1j_2}n^{(g_1,k)}_{j_2}]$，$S^{(g_2,lk)} \simeq [\hat{p}^{(g_2,lk)}_{j_1j_2}n^{(g_2,k)}_{j_2}]$ で近似できる．

2 つの ASD が同じ性質を持つ場合，それぞれの分割表の各セルの出現確率は等しい．これが正しい場合，セル (j_1, j_2) の出現個数の期待値の推定量は

$$\widehat{E(s_{j_1j_2}^{(g_a,lk)})} = \frac{\hat{p}_{j_1j_2}^{(g_1,lk)} n_{j_2}^{(g_1,k)} + \hat{p}_{j_1j_2}^{(g_2,lk)} n_{j_2}^{(g_2,k)}}{n_{j_2}^{(g_1,k)} + n_{j_2}^{(g_2,k)}} n_{j_2}^{(g_a,k)}$$

と考えられる．前の 2 つの節と同様に，カイ 2 乗統計量を考える場合，$\widehat{E(s_{j_1j_2}^{(g_a,lk)})}$ で割ることになるので，これが 0 もしくは極端に小さな値となるときは無視しなくてはならない．カテゴリー変数同士の組み合わせの場合と同様の基準で考えるため，$n_{j_2}^{(g_1,k)}$ と $n_{j_2}^{(g_2,k)}$ のうち少なくとも 1 つが 0 である場合，およびそれらがいずれも正であっても $\hat{p}_{j_1j_2}^{(g_1,lk)} n_{j_2}^{(g_1,k)} + \hat{p}_{j_1j_2}^{(g_2,lk)} n_{j_2}^{(g_2,k)} < 1$ となる場合のセルは無視してカイ 2 乗統計量を考える．これより

$$\chi^{2(g_1g_2,lk)} \simeq \sum_{\substack{a=1}}^{2} \sum_{j_1=1}^{N} \sum_{j_2=1}^{m_k} \frac{\left\{ \hat{p}_{j_1j_2}^{(g_a,lk)} n_{j_2}^{(g_a,lk)} - \widehat{E(s_{j_1j_2}^{(g_a,lk)})} \right\}^2}{\widehat{E(s_{j_1j_2}^{(g_a,lk)})}} \tag{3.5}$$

$$\hat{p}_{j_1j_2}^{(g_1,lk)} n_{j_2}^{(g_1,k)} + \hat{p}_{j_1j_2}^{(g_2,lk)} n_{j_2}^{(g_2,k)} \geq 1,$$
$$n_{j_2}^{(g_1,k)} n_{j_2}^{(g_2,k)} > 0$$

を非類似度と考えることができる．これを全ての (l, k) に関して考え，その総和をとったものが 2 つのグループの連続変数とカテゴリー変数間の非類似度

$$d_{(rc)}^{(g_1g_2)} = \sum_{l=1}^{p} \sum_{k=1}^{q} \chi^{2(g_1g_2,lk)} \tag{3.6}$$

と考えられる．

3.4　すべての変数を用いた非類似度

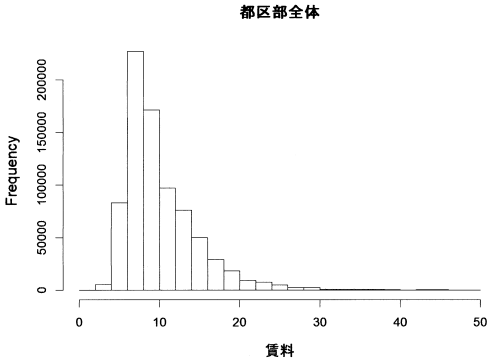
連続変数をカテゴリー変数化して考えることにより，(3.2)，(3.4)，(3.6)式は全てカテゴリー変数同士の組み合わせによる非類似度と考えられるので

$$d^{(g_1g_2)} = d_{(cc)}^{(g_1g_2)} + d_{(rr)}^{(g_1g_2)} + d_{(rc)}^{(g_1g_2)}$$

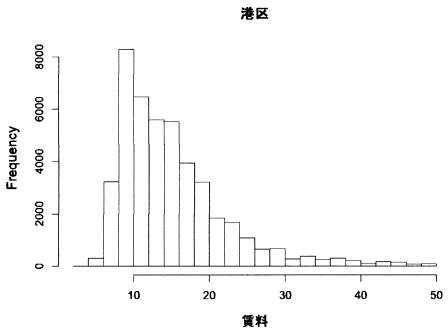
が ASD 間の全体のカイ 2 乗統計量に基づく非類似度となる．これは $X^{(g_1)'}X^{(g_1)}$ と $X^{(g_2)'}X^{(g_2)}$ の間の非類似度を計算したことになる．これらの行列はいずれも対称行列であり，対角成分よりも上側に位置する成分だけを考えていることに注意する．また，対角成分の情報は，対角成分よりも上側に位置する成分にほとんどが含まれていることより，陽には計算されていないことも注意しなければならない．

4.　不動産情報データへの適用

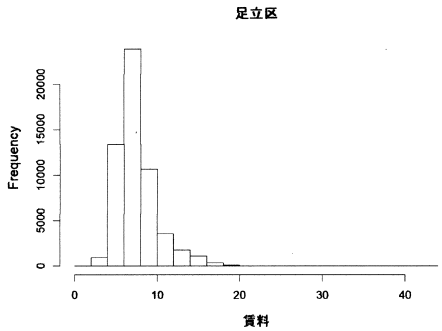
ここでは，各変数において欠測値が含まれる物件，書き間違いおよび外れ値と考えられる値が含まれる物件，他の変数への従属性が高いと考えられる変数をあらかじめ削除した後の東京都区部の不動産情報データにわれわれが提案した手法を適用し解析する．表 1 はそのデータの一部であるが，全体では 2 個の連続変数および 79 個のカテゴリー変数からなる，合わせて約 79 万件の賃貸物件情報である．なお，変数のうち「管理費」「礼金月数」「敷金月数」については元の数値がいずれも賃料を基準とする月数で表されているが，値が 0 である物件がいずれの変数でも全体の 14% 以上存在し，連続変数として考えるための適切な変数変換が難しいことか



(a)都区部全体



(b)港区



(c)足立区

図 1. 賃料のヒストグラム.

ら，ここでは各変数を少数のカテゴリー値をもつカテゴリー変数として考える．

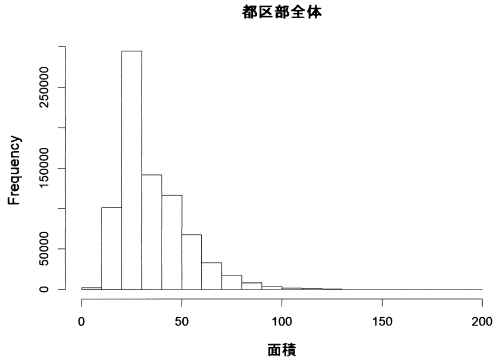
このデータにおける 2 個の連続変数「賃料」および「面積」に関し，例として都区部全体，港区，足立区それぞれの場合におけるヒストグラムを図 1(a)～(c)および図 2(a)～(c)に示す．

図 1 および図 2 より，両変数の値の分布には少なからず非対称性がみられる．そこで，このような各連続変数値に対し，各変数ごとに対数をとることを考える．対数変換後のヒストグラムの例を図 3(a)～(c)および図 4(a)～(c)に示す．

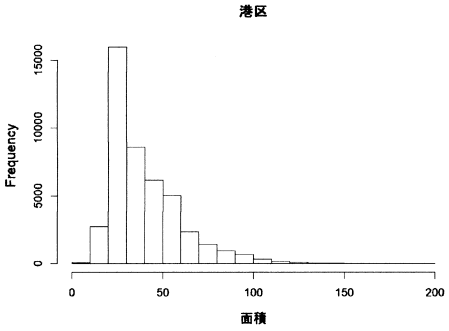
図 3 および図 4 より，対数変換後の各連続変数の値は，元の値の場合よりも非対称性が弱くなっている．そこで，以下ではこれらの値を連続変数の値として ASD を考えるものとする．

データを探索的に解析するにあたり，「区」というカテゴリー変数により各区ごとに 23 のグループに分けて考え，その ASD 間の非類似度を，連続変数をカテゴリー変数化するための分割数 N が 3 ～ 9 の場合においてそれぞれ計算した．そして，それらを用いてまず階層的クラスタリングを行う．ここでは最長距離法 (Defays, 1977)，最短距離法 (Sibson, 1973)，群平均法 (Sokal and Michener, 1958)の 3 種類の手法を用いる．なお，連続変数を含む部分の非類似度に関しては N の値により異なる値が得られることに注意する．

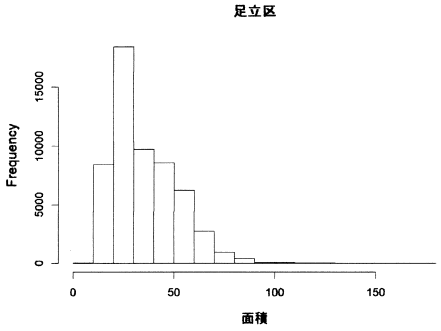
階層的クラスタリングにおいては，連続変数を含む部分のみの非類似度を用いた場合，全ての変数の組み合わせの非類似度を用いた場合についてのいずれでも，$N = 4$ 以下では N の値により大きな変化があったのに対し，$N = 5$ 以上ではどの N についても手法ごとの結果に大きな変化がみられなかった．そのため，ここでは $N = 5$ の場合を主として考えることにする．



(a)都区部全体



(b)港区



(c)足立区

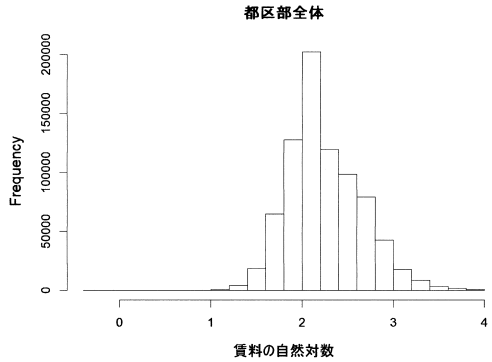
図 2．面積のヒストグラム．

この場合の，全ての変数の組み合わせの非類似度にそれぞれの手法を適用した結果が図 5(a)～(c)である．

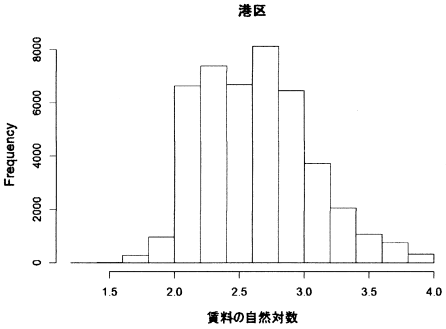
次に，$N = 5$ の場合の全ての変数の組み合わせの非類似度に多次元尺度構成法を適用する．その結果を図 6 に示す．$1 \leq g_1 < g_2 \leq 23$ に対し，図 6 における各 ASD 間のユークリッド距離 $\hat{d}^{(g_1 g_2)}$ と元々の非類似度行列における非類似度 $d^{(g_1 g_2)}$ の相関係数の 2 乗値(決定係数)は $R^2 = 0.996$ となり，非類似度行列による布置が高い精度で保持されていることがわかる．

階層的クラスタリングにおいてはクラスターをまとめるための基準により手法が特徴づけられる(Lance and Williams, 1966)ため，図 5(a)～(c)にも示されている通り，デンドログラム全体の形状については手法により多少の差異が見られる部分があるものの，どの手法でも共通する組み合わせが複数存在し，それらはそれぞれクラスターとみなせる．また図 6 における布置の全体的な形状からいくつかのクラスターが読み取れる．これらを合わせて考えると，概ね(1)中央区および港区，(2)足立区，(3)千代田区，新宿区，文京区，台東区，墨田区，江東区，品川区および渋谷区，(4)目黒区，大田区，豊島区および荒川区，(5)世田谷区，中野区，杉並区，北区，板橋区，練馬区，葛飾区および江戸川区，がそれぞれクラスターとみなせる．

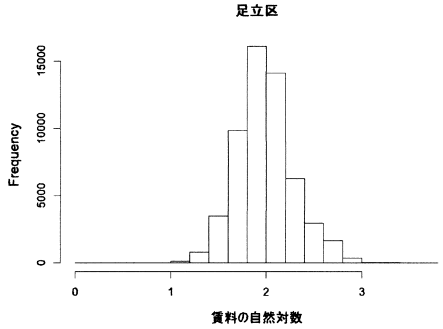
クラスター(1)およびクラスター(2)は他のクラスターと特に大きく異なる特徴があると考えられることから，後に詳細に考察する．それら以外のクラスターについて，クラスター(3)は東京都区部の中心から外側に向けて伸びる各鉄道路線の起点駅のうち特に乗降客の多い各駅(新宿駅，渋谷駅，上野駅など)が含まれる区およびその隣接区域に位置する各区が集まっている．



(a)都区部全体



(b)港区



(c)足立区

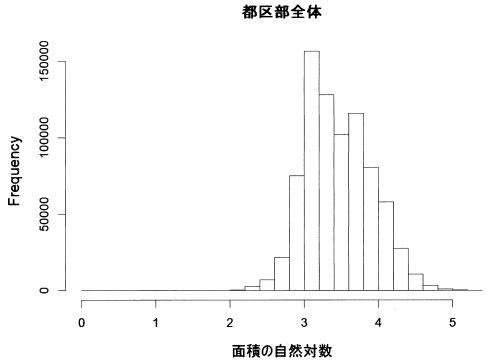
図 3. 賃料の対数のヒストグラム.

ただし，このクラスターには西武池袋線や東武東上線の起点駅として乗降客の多い池袋駅が含まれる豊島区が含まれていないのが興味深い．クラスター(5)は各区の地理的な位置より，東京都区部の中での外縁部およびその隣接区域に位置する区が集まっている．クラスター(4)はクラスター(3)とクラスター(5)の中間に位置する区の集合と考えられる．

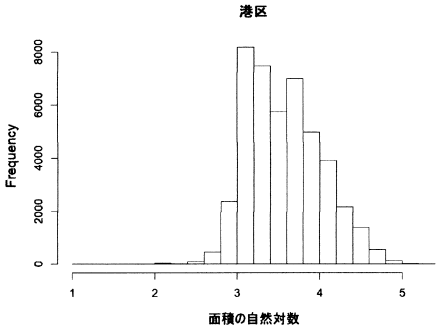
次に，中央区，港区，足立区の 3 区を選び，この不動産情報データにおけるいくつかの特徴的な 2 変数の組に関して考察する．

まず，カテゴリー変数のうち物件種別および構造種別の 2 つの組による分割表の各区ごとの違いを図示したものが図 7(a)～(c)である．これは物件数が多い組み合わせほど濃い色となるように表示した，いわゆるヒートマップである．また(b) (c)それぞれの図の下に，この 2 つの変数の組に関して中央区を基準とした各区への非類似度を記す．

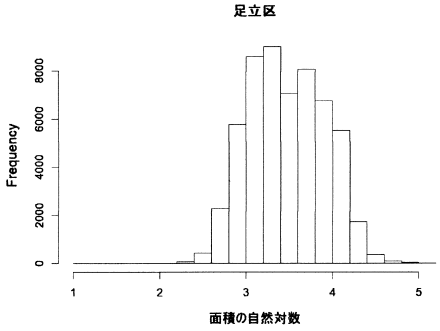
物件種別(当初のカテゴリー数は 5)においては，カテゴリー値 1 がマンション，カテゴリー値 2 がアパートであり，それ以外の種別を全てその他としてカテゴリー値 3 にまとめた．構造種別(当初のカテゴリー数は 10)においては，カテゴリー値 1 が鉄筋コンクリート造り，カテゴリー値 2 が鉄骨鉄筋造りなど，カテゴリー値 3 が鉄骨造りなど，カテゴリー値 4 が軽量鉄骨造りなど，カテゴリー値 5 が木造，カテゴリー値 6 がその他の 6 つのカテゴリーに集約した．この図より，中央区と港区における物件は鉄筋コンクリート造りもしくは鉄骨鉄筋造りなどのマンションが大半を占め，類似性が高いことがわかる．また，区の物件全体に占める鉄筋コンクリート造りの物件の割合は港区の方が中央区よりも高いことが読み取れる．一方で，足立区は



(a)都区部全体



(b)港区



(c)足立区

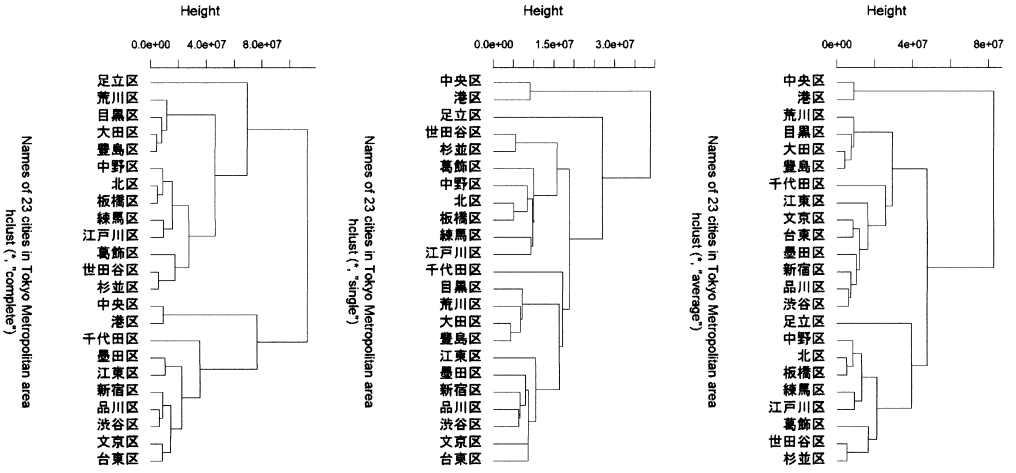
図 4. 面積の対数のヒストグラム.

鉄筋コンクリート造りもしくは鉄骨鉄筋造りなどのマンションの割合が中央区や港区と比べて低く，鉄骨造りのマンション，軽量鉄骨造りのアパート，木造アパートがそれぞれ一定の割合を占めている．すなわち，足立区においては物件の状況が中央区や港区とは大きく異なることがこの組み合わせからわかる．

さらに，中央区を基準とした各区への非類似度は，中央区と港区との間の値よりも中央区と足立区との間の値の方が極めて大きくなっており，足立区が中央区や港区と異なる状況に対応していると考えられる．

次に，連続変数に関して考える．表 2 は面積および賃料それぞれの対数に関し，各区ごとの各変数の平均および標準偏差，2 つの変数の相関係数，およびこの 2 つの変数の組に関して中央区を基準とした各区への非類似度をまとめたものである．これより，各変数の平均値は中央区および港区が足立区より高く，2 つの変数の相関係数も中央区と港区において足立区よりも大きいことがわかる．そして，この 2 つの変数に関する非類似度は，中央区と港区との間の値よりも中央区と足立区との間の値の方が極めて大きくなっており，ここでも足立区が中央区や港区と異なる状況に対応していると考えられる．

さらに，連続変数とカテゴリー変数の組に関しても考察する．表 3 はカテゴリー変数「物件種別」における各カテゴリー値ごとの件数と賃料の対数の平均，およびこの 2 つの変数の組に関して中央区を基準とした各区への非類似度を示したものである．この表より，3 区いずれにおいてもマンションの物件数の比率が高く，マンションの賃料の対数の平均が各区ごとの全体



(a)最長距離法 (b)最短距離法 (c)群平均法

図 5. 不動産情報データの集約的シンボリックデータの階層的クラスタリング($N = 5$).

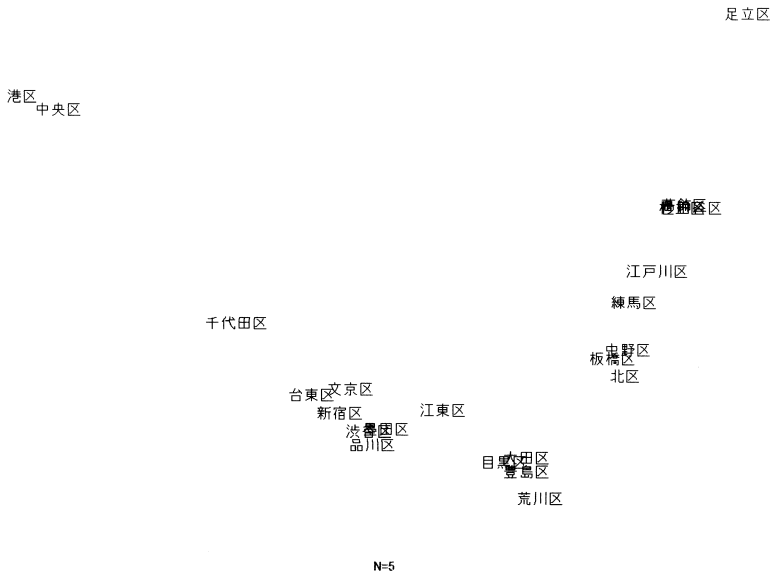
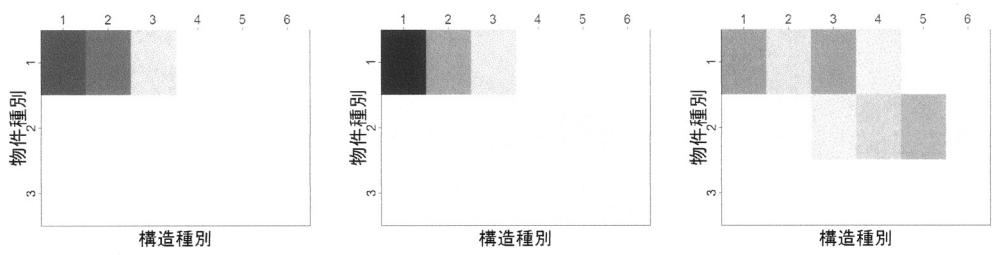


図 6. 不動産情報データの集約的シンボリックデータの多次元尺度構成法の布置($N = 5$).

の物件の賃料の対数の平均(表 2 参照)に近くなっている．各カテゴリー値ごとの賃料の対数の平均に注目すると，どのカテゴリーでも港区および中央区の値が足立区よりも高くなっている．そして，この 2 つの変数に関する非類似度も，中央区と港区との間の値よりも中央区と足立区との間の値の方が極めて大きくなっており，ここでも足立区が中央区や港区と異なる状況



-　　$\chi^{2(g_1g_2,k_1k_2)} = 2.32\times10^3$　　$\chi^{2(g_1g_3,k_1k_2)} = 3.47\times10^4$

(a) 中央区　　(b) 港区　　(c) 足立区

図 7. 物件種別と構造種別による分割表のヒートマップおよび中央区を基準とした非類似度（k_1：物件種別，k_2：構造種別，g_1：中央区，g_2：港区，g_3：足立区）．

表 2. 賃料と面積それぞれの対数に関する各区ごとの平均および分散，各区ごとの相関係数，および中央区を基準とした非類似度（$N = 5$）（l_1：賃料の対数，l_2：面積の対数，g_1：中央区，g_2：港区，g_3：足立区）．

区	賃料の対数 (万円)		面積の対数 (m^2)		相関係数	非類似度
	平均	標準偏差	平均	標準偏差		
中央区	2.492	0.3412	3.549	0.4045	0.9328	-
港区	2.629	0.4203	3.560	0.4557	0.9205	$\chi^{2(g_1g_2,l_1l_2)} = 5.128\times10^3$
足立区	1.990	0.2087	3.471	0.4365	0.8449	$\chi^{2(g_1g_3,l_1l_2)} = 5.796\times10^4$

表 3. 物件種別の各カテゴリー値ごとの件数，賃料の対数の平均（単位：万円）および中央区を基準とした非類似度（$N = 5$）（l：賃料の対数の平均，k：物件種別，g_1：中央区，g_2：港区，g_3：足立区）．

区	マンション		アパート		その他		非類似度
	件数	賃料の対数の平均	件数	賃料の対数の平均	件数	賃料の対数の平均	
中央区	29886	2.492	25	2.322	27	2.655	-
港区	43824	2.634	587	2.234	86	3.045	$\chi^{2(g_1g_2,lk)} = 1.275\times10^3$
足立区	37738	2.064	17408	1.817	707	2.303	$\chi^{2(g_1g_3,lk)} = 1.085\times10^4$

に対応していると考えられる．

なお，本論文で解析した不動産情報データについては，本節の最初に述べた通り，連続変数の数に対してカテゴリー変数の数が圧倒的に多いため，どの g_1 および g_2 についても，$d^{(g_1g_2)}_{(rr)}$ および $d^{(g_1g_2)}_{(rc)}$ よりも $d^{(g_1g_2)}_{(cc)}$ の値の方がずっと大きくなる．すなわち，カテゴリー変数同士の

みの非類似度から導出される構造が大きく影響している．

5. おわりに

本論文では，連続変数とカテゴリー変数が混在する多次元データ集合がいくつかのグループに分かれていると考え得る場合に，各グループを単位とする解析手法として集約的シンボリックデータ(ASD)解析法を提案し，その適用例として東京都区部の不動産情報データを解析した．そして，各 ASD 間の非類似度をカイ 2 乗統計量を用いて表す手法を提案し，これに対する階層的クラスタリングおよび多次元尺度構成法を，不動産情報データにおける各区ごとのグループを ASD と考えて適用し，いくつかの特徴的なクラスター構造を発見した．また特徴的ないくつの区に関して各変数の統計量および区の間の非類似度を計算し，区の間の差異について考察した．

連続変数とカテゴリー変数が混在する多次元データ場合については，連続変数を各グループ共通の少数の領域に離散化して考えることで，連続変数を含む組み合わせについても近似した分割表の組み合わせからなる Burt 行列を各 ASD ごとに考えることができ，全ての変数の組み合わせをカテゴリー変数同士の組み合わせとして統一的に考えることが可能になる．

異なる ASD 間の全体の非類似度については，データ集合の変数の数が多くなればなるほど大きな値となるため，冗長な変数が存在していると望ましい非類似度の値よりも大きな値が算出される．そのため，非類似度を計算するための適切な変数選択の手法の開発が今後の課題と考えられる．

謝　辞

本研究は，統計数理研究所共同研究(課題番号 27-共研-4204 および 28-共研-4105，研究代表者 清水信夫)，および科学研究費補助金基盤研究(C)（課題番号 26330054，研究代表者 中野純司)より支援を受けました．本論文で使用した不動産情報データは，(株)リクルート住まいカンパニーより提供を受けました．本論文の 2 名の査読者からは，内容に関して非常に有益なコメントを多く頂きました．深く感謝致します．

参 考 文 献

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis*: *Conceptual Statistics and Data Mining*, John Wiley & Sons Ltd, Chichester, UK.

Bock, H.-H. and Diday, E. (2000). *Analysis of Symbolic Data*: *Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin.

Defays, D. (1977). An efficient algorithm for a complete link method, *The Computer Journal*, **20**(4), 364–366, doi:10.1093/comjnl/20.4.364.

Denœux, T. and Masson, M. (2000). Multidimensional scaling of interval-valued dissimilarity data, *Pattern Recognition Letters*, **21**(1), 83–92.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis: The basic choices, In: Bock, H.-H. (ed.), *Classification and Related Methods of Data Analysis*, Proceedings of IFCS-87, Aachen, July 1987, North-Holland, Amsterdam, 673–684.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, John Wiley & Sons Ltd., Chichester, UK.

Groenen, P. J. F., Winsberg, S., Rodriguez, O. and Diday, E. (2006). I-Scal: Multidimensional scaling of interval dissimilarities, *Computational Statistics and Data Analysis*, **51**(1), 360–378.

Irpino, A. and Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data, *Data Science and Classification*, 185–192, Springer, Berlin.

Lance, G. N. and Williams, W. T. (1966). Computer programs for hierarchical polythetic classification ("similarity analyses"), *The Computer Journal*, **9**(1), 60–64.

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single-link cluster method, *The Computer Journal*, **16**(1), 30–34, doi:10.1093/comjnl/16.1.30.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships, *University of Kansas Science Bulletin*, **38**, 1409–1438.

Verde, R. (2004). Clustering methods in symbolic data analysis, *Data Science and Classification*, 299–317, Springer, Berlin.

Dissimilarity between Aggregated Symbolic Data Using Chi-squared Statistics and Its Application to Real Estate Data

Nobuo Shimizu[1], Junji Nakano[1] and Yoshikazu Yamamoto[2]

[1]Institute of the Statistical Mathemetics
[2]Faculty of Science and Engineering, Tokushima Bunri University

In recent service science research, we often have huge amount of individual data with both continuous and categorical variables. These data sets can sometimes be divided into rather small number of naturally defined groups. In such situations, we are interested in inference and analysis for these groups, not for individual data. For describing these groups, we consider a set of descriptive statistics, and call it "aggregated symbolic data" (ASD). We propose to use up to second moments descriptive statistics for both continuous and categorical variables as ASD, and define a dissimilarity as the sum of chi-squared statistics among all variables including continuous variables. We apply our method to real estate data in Tokyo metropolitan area. We consider 23 cities in Tokyo as ASD and calculate dissimilarity among 23 ASDs, and investigate some characteristics relationships among ASDs by using hierarchical clustering and multidimensional scaling.

Key words: Big data, Burt matrix, chi-square statistics, hierarchical clustering, multidimensional scaling.