統計研究彙報　第83号　2026年3月(113～136)

ワッサースタイン距離などの分布間距離の公的統計への活用

上田　聖†

Application of Distributional Distances such as the Wasserstein Distance to Official Statistics

Ueda Sei

近年、画像処理や人工知能分野において分布間の比較が重要性を増しており、最適輸送理論に基づくワッサースタイン距離も容易に計算可能となっている。本稿では、公的統計に対する分布間距離の応用例を二点に分けて示す。第一に、経済統計における集中度指標としてワッサースタイン距離を導入し、従来のジニ係数やハーフィンダル指数では考慮されなかった地域間の空間的距離を反映する新しい集積度指標を提案する。具体例として、東京都と福岡県に分散する産業と、東京都と神奈川県に分散する産業を比較し、従来指標では同一と評価される状況において、後者の方を高い集積度を示す指標としている。第二に、地理的距離のみならず人口分布や産業分布など多様な分布間距離を利用し、空間統計学におけるバリオグラム線形の枠組みを応用することで、様々な分布空間における空間的自己相関の検出を SSDSE（Standardized Statistical Data Set for Education）のデータを用いて行った。また、モランＩ統計との関係性を整理した。

キーワード：ワッサースタイン距離、集中度、集積度、バリオグラム線形、SSDSE、モランＩ統計、
マハラノビス距離

In recent years, the comparison of probability distributions has gained increasing importance in fields such as image processing and artificial intelligence, with the Wasserstein distance based on optimal transport theory now being computationally feasible. This study demonstrates two applications of distributional distances to official statistics. First, we introduce the Wasserstein distance as a concentration measure in economic statistics and propose a new agglomeration index that incorporates spatial distances between regions, an aspect overlooked by conventional measures such as the Gini coefficient and the Herfindahl index. For example, industries distributed between Tokyo and Fukuoka are compared with those distributed between Tokyo and Kanagawa. While conventional indices evaluate both cases as equivalent, the proposed index indicates stronger agglomeration in the latter. Second, extending beyond geographic distance, we employ various distributional distances, including population and industrial distributions, and apply the framework of the linear variogram in spatial statistics. Using data from the SSDSE (Standardized Statistical Data Set for Education), we detect spatial autocorrelation in diverse distributional spaces and clarify its relationship with Moran’s I statistic.

Keywords: Wasserstein distance, concentration, agglomeration, linear variogram, SSDSE, Moran’s I statistic,
Mahalanobis distance

† 独立行政法人統計センター　Email:sueda2@nstac.go.jp

1　はじめに

人口減少社会に入り、少ない資源を効果的に活用するため、行政では EBPM の推進が、公的統計改革と一体として進められており、総務省も自治体に対する EBPM 支援事業[1]を実施し、地方自治体においても EBPM の推進が図られている。このような状況において、自治体間の成功事例の横展開が行われるとともに、類似する他の自治体の成功事例を参考にすることも行われている。

「似ている」ことを定量的に示す指標として分布間の距離がある。緯度・経度が似ている近隣自治体というものもあれば、性・年齢階級別の人口の構成比や従業者数の産業別構成比が似ているものもある。これらの相違を１つの数値で表したものが分布間の距離である。分布間の距離は経済分析にも多く活用されている。例えば経済分析でよく目にするジニ係数は、その代表的なものの１つで、完全に平等な分布と現状の配分の分布の距離を数量化している。

分布間の距離の計測技術は、画像処理や AI 研究に欠かせない技術として、近年、飛躍的に発展し、様々な計測方法や計算技術が示されている。解析的に計算できるユークリッド距離に加え、解析的な算式で示せない最適輸送問題から派生したワッサースタイン距離なども、統計解析ソフトで簡単に計算できるようになっている（佐藤(2023)）。

本稿の目的は、この分布間の距離を公的統計に活用するいくつかの方法を示すことである。特にワッサースタイン距離を活用した研究は進展途上であり、この距離の活用を含めた指標の提示や計算結果を示す。

本稿の構成は、まず、第２節で分布間距離とその特性を概説する。第３節では、分布間距離を経済統計に活用する方法として、ジニ係数やハーフィンダル指数などの集中度・集積度の指標と分布間距離との関係を整理したうえで、この分野へのワッサースタイン距離を適用した計算結果と評価を述べる。第４節では、分布間距離を活用した、経済統計の空間的自己相関（地域性、人口構造依存性、産業構造依存性）の判断指標の１つとその計算結果を示し、既存の空間的自己相関に関する統計量（モラン I 統計量）との関係についても整理する。最後に今後の課題などについて述べる。

2　地理的距離と分布間距離

本節では、以降で応用する分布間の距離について、その主要なものと基本的な特性を紹介する。

2.1　地理的距離（座標系ユークリッド距離）

地理的距離とは、地点iの緯度x_i、経度y_iの情報に基づき地球上の 2 地点(x_1, y_1)，(x_2, y_2)の物理的距離を定量化したもので、通常、地点$i(x_i, y_i)$と地点$j(x_j, y_j)$のユークリッド距離$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$で計算される。座標系の回転や平行移動に対して不変である特徴を持つ。本稿では、都道府県や県庁所在市の位置を代表する座標として、各都道府県や県庁所在市の行政区域の面積重心を用いる。

2.2　分布間距離

2.2.1　情報量系距離

確率分布間の差異を定量化する指標として、情報理論に基づく Kullback-Leibler ダイバージェンス（KL-ダイバージェンス:$D_{KL}(*|*)$）をベースに対称性を持たせた Jensen-Shannon 距離（JS 距離）がある。JS 距離は、確率（密度）関数$P(X)$、$Q(X)$の平均関数 $M(X) = (P(X) + Q(X))/2$を用いて、

$$d_{JS}(P|Q) = \sqrt{\frac{1}{2}D_{KL}(P|M) + \frac{1}{2}D_{KL}(Q|M)} \tag{2.1}$$

で定義される。

2.2.2　ノルム系距離

確率分布間の差異を示す別指標として、L_p 距離がある。確率（密度）関数$P(X)$、$Q(X)$に対して

[1] 例えば、総務省統計局実施の「EBPM ブートキャンプ」（https://www.stat.go.jp/dstart/research/）など

（離散分布の場合）　$d_{Lp}(P,Q) = (\sum |P(X) - Q(X)|^p)^{\frac{1}{p}}$　(2.2)

（連続分布の場合）　$d_{Lp}(P,Q) = (\int |P(X) - Q(X)|^p \, dX)^{\frac{1}{p}}$　(2.3)

で定義され、特に p=1 の場合はマンハッタン距離、p=2 の場合はユークリッド距離と呼ばれる。

2.2.3　最適輸送距離（Wasserstein 距離）

ワッサースタイン距離（Wasserstein distance）は、2 つの確率分布の差異を「確率質量を移動させるコスト」に基づいて定義する距離であり、最適輸送問題の解として与えられる。直感的には、ある分布を「砂山」とみなし、もう一方の分布に一致させるために砂を移し替える際に必要な最小の輸送コストを距離とするものである。

まず、ワッサースタイン距離をわかりやすい離散型で紹介する。

2 つの確率分布PとQが、離散的な点集合$X = \{x_1, x_2, \cdots, x_n\}$上で、それぞれの確率が$P = (p_1, p_2, \cdots, p_n), \sum_i p_i = 1$，$Q = (q_1, q_2, \cdots, q_n), \sum_i q_i = 1$で与えられたとする。ここで、「$x_i$にある$p_i$の確率質量のうち、$x_j$に輸送する確率質量」を$\gamma_{ij}$とすると、「輸送計画」は、すべての$x_i$と$x_j$の組み合わせ（n×n 組）に対する$\gamma_{ij}$で構成され、n×n 行列 $\Gamma(P,Q) = (\gamma_{ij})$で表される。また、$x_i$から運び出される確率質量 $\sum_j \gamma_{ij}$はp_iに等しく、x_jに持ち込まれる$\sum_i \gamma_{ij}$はq_jに等しく計画することが求められるため、$\Gamma(P,Q)$の i 行の総和はp_iに等しく、j 列の総和はq_jに等しい制約条件が課される。

さらに、x_iからx_jに単位量を輸送するコスト（直感的には、x_iとx_jとの間の距離）を$d(i,j)$とすると、ある 1 つの輸送計画$\Gamma(P,Q)$の輸送コストは、$\sum_i \sum_j \gamma_{ij} d(i,j)$で示され、無数にある輸送計画$\Gamma(P,Q)$の集合を$\Gamma_{all}$とすると、この中からコストが最小となるものがワッサースタイン距離として定義される。改めて整理すると次のようになる。

離散点集合$X = \{x_1, x_2, \cdots, x_n\}$上で、$P, Q$の確率分布が$P = (p_1, p_2, \cdots, p_n), \sum_i p_i = 1$，$Q = (q_1, q_2, \cdots, q_n), \sum_i q_i = 1$で与えられ、$x_i$から$x_j$へ 1 単位の確率質量を輸送するコストを$d(i,j)$としたとき、コスト$d(i,j)$に付随する分布$P$と$Q$間のワッサースタイン-p 距離 $W_p(P,Q)$は以下のように定義[2]される。

$$W_p(P,Q) = \min_{\Gamma(P,Q) \in \Gamma_{all}} \left(\sum_i \sum_j \gamma_{ij} d(i,j)^p\right)^{\frac{1}{p}} \quad (2.4)$$

$\sum_j \gamma_{ij} = p_i$, $\sum_i \gamma_{ij} = q_j$, $\Gamma(P,Q) = (\gamma_{ij})$, $\gamma_{ij} \geq 0$, Γ_{all}:左の条件を満たす全ての$\Gamma(P,Q)$の集合

ワッサースタイン距離は、L_1・L_2 距離や JS 距離などとは異なる特性を持つ。特に、年齢分布や所得分布を階級化して比較する場合、Rubner et al. (2000)によると、情報量系距離（JS 距離など）やノルム系距離は、階級数が増えるほど差異が過剰に強調され感覚と乖離する傾向がある一方、ワッサースタイン距離は、分布形状の感覚的な差異を安定的に測定できるとされており、この特徴は、公的統計の分析においても特に注目に値する。

図 1 は、2020 年国勢調査に基づく北海道における江別市、南富良野町、音威子府村[3]および北海道全体の 1 歳階級別および 10 歳階級別の人口構成比を示したグラフである。また、図 2 の横軸は北海道の全市町村に対する「各歳階級」の人口構成比における北海道全体の分布との距離、縦軸は「10 歳階級」で計算した同じ距離を表している[4]。この結果から、L_2 距離は階級の取り方に影響を受けやすいのに対し、ワッサースタイン-1 距離は影響をほとんど受けない。このように、ワッサースタイン距離は、ノルム系距離では過剰に反応する分布形状の微小な差異に影響されず、ある程度平準化して数量化可能であることが確認される。

---

[2] ワッサースタイン距離は連続型確率変数にも定義されるが、本稿では p=1 の離散型のみを対象として議論を行う。

[3] 比較を分かりやすくするため、北海道全体の分布から最も乖離が大きい市・町・村をそれぞれ 1 つずつ選定した。

[4] 図 1 及び図 2 に示すワッサースタイン-1 距離のコストは、階級を 1 つ移動させるには、その年齢幅分の年齢差が必要と設定した。また、L_2 距離は構成比の差を 2 乗して計算するため、階級数が n 倍になると、距離はおよそ$1/\sqrt{n}$倍に縮小する。そこで、L_2 距離については縦軸と横軸の縮尺を揃えるため、距離に階級数の平方根$\sqrt{n}$を乗じて調整した。

図 1 各歳階級別および 10 歳階級別の北海道全体、江別市、南富良野町および音威子府村の人口分布

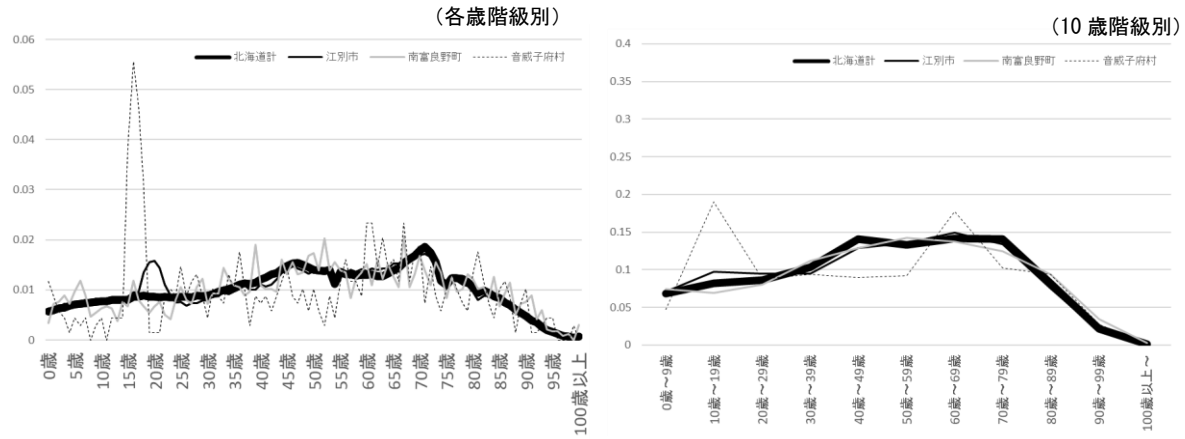
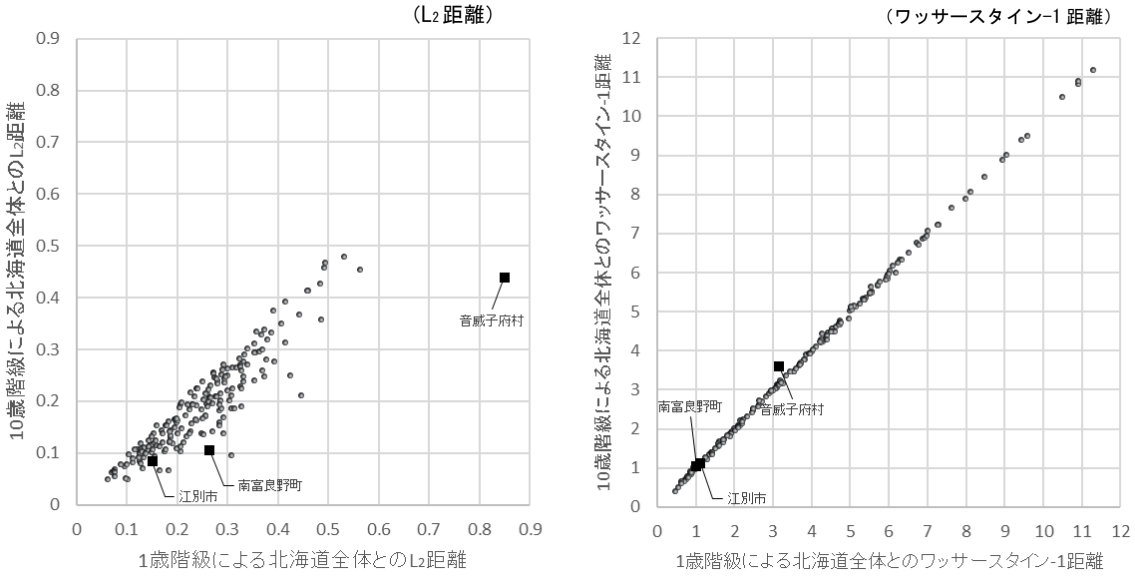


図 2 北海道全体と北海道各市・区・町・村の各歳階級別および 10 歳階級別の人口分布に基づく距離の関係



3　集中度・集積度の指標

本節では、集中度を測定する主要な指標であるジニ係数、ハーフィンダル指数、Theil 指数を分布間距離に基づいて再解釈し、さらにワッサースタイン距離を用いた新たな集中度・集積度指標を提案する。

3.1　ジニ係数と分布間の差異としての解釈

所得や資産などの分布における集中度や不平等度を定量化する代表的な指標としてジニ係数(G)がある。所得などのn個の非負の観測値$x_1, x_2, \cdots, x_n$に対し、これを昇順に並べ変えたものを$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$、それぞれのシェアを$s_i = x_i / \sum_j x_j$、$s_{(i)} = x_{(i)} / \sum_j x_{(j)}$とすると、ジニ係数($G$)は、実際の観測値から得られるローレンツ曲線（$L(i) = \sum_{s_{(j)} \le s_{(i)}} s_{(j)}$）と、完全平等の観測値つまり$x_{(1)} = x_{(2)} = \cdots = x_{(n)}$、$s_{(i)} = 1/n$から得られる完全平等線$L(i) = i/n$で囲まれる面積を 2 倍した値として定義され、0 から 1 の範囲を取り、0 は完全平等、1 は完全不平等（所得等の独占）を示す。

ジニ係数は、「観測されたシェアの分布」と「完全平等のシェアの分布」の確率分布関数の分布間距離として解釈でき、距離の公理を満たし、L_1 距離の性質を有していることが知られている。

また、ジニ係数は、観測値ペア間の平均的な格差

$$G = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} |s_i - s_j| \tag{3.1}$$

としても計算できる。

3.2　ハーフィンダル指数(HHI)と分布間の差異としての解釈

市場の集中度を表す代表的な指標としてハーフィンダル指数（Herfindahl-Hirschman Index; *HHI*）が広く用いられている。これは企業の独占度を把握するために用いられ、例えば特定の製品を生産する企業の出荷額などの非負の観測値 $x_1, x_2, \cdots, x_n$、シェアを $s_i = x_i / \sum_j x_j$とすると、

$$HHI = \sum_{i=1}^{n} {s_i}^2 = \sum_{i=1}^{n} \left(\frac{x_i}{\sum_j x_j} \right)^2 \tag{3.2}$$

で定義される。この式は、ゼロ分布（無シェア状態）と現状のシェア分布との L_2 距離の 2 乗と解釈することもできる。

3.3　Theil 指数と分布間の差異としての解釈

Theil 指数(T)は情報理論から導出される集中度指標であり、所得などの非負の観測値$x_1, x_2, \cdots, x_n$と、平均値 $\mu = (\sum_i x_i)/n$を用いて、

$$T = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i}{\mu} \ln \frac{x_i}{\mu} \tag{3.3}$$

で定義される。すべての観測値が等しいとき $T = 0$ となり、不平等（独り占め・富の集中）になるほど Tは増加する。Theil 指数は、観測値をシェアとした$s_i = x_i / \sum_j x_j = x_i/(n\mu)$の分布と完全平等の$s_i = 1/N$の分布との KL ダイバージェンスに対応し、

$$T = \sum_{i=1}^{n} s_i \ln \frac{s_i}{(1/n)} \tag{3.4}$$

と表現できる。Theil 指数の理論上の最大値は ln(n)であり、これは 1 つの観測値が全資源を独占する場合に達成される。

3.4　ワッサースタイン-1 距離による集中度指標の提案

ジニ係数、*HHI*、Theil 指数はいずれも「完全平等分布」（または「零分布」）からの乖離を測る指標である。ここでは、同様の考え方に基づき、ワッサースタイン-1 距離を用いて「均等分布との距離」を集中度として定義することを提案する。

3.4.1　ワッサースタイン-1 距離の可処分所得の集中度指標への活用

まず、可処分所得分布を例に取る。n 世帯の非負の可処分所得$x_1, x_2, \cdots, x_n$を、全世帯が平均所得 $\mu = (\sum_i x_i)/n$を持つ分布に移し替えるとする。この時、各世帯間の距離をすべて 1 と見なすと、ワッサースタイン-1 距離は次式で表される。

$$W = \frac{1}{2} \sum_{i=1}^{n} \frac{1}{n} |x_i - \mu| \tag{3.5}$$

これは「平均より多い世帯から少ない世帯へ再配分する金額の合計」を示し、古くから用いられている散布度となる。最大値は、1 世帯のみが全所得を独占する場合で$[(n-1)/n]\mu$、最小値は、全世帯が同一所得の場合で 0 となる。したがって、[0,1]区間に正規化した指標を次のように定義できる。

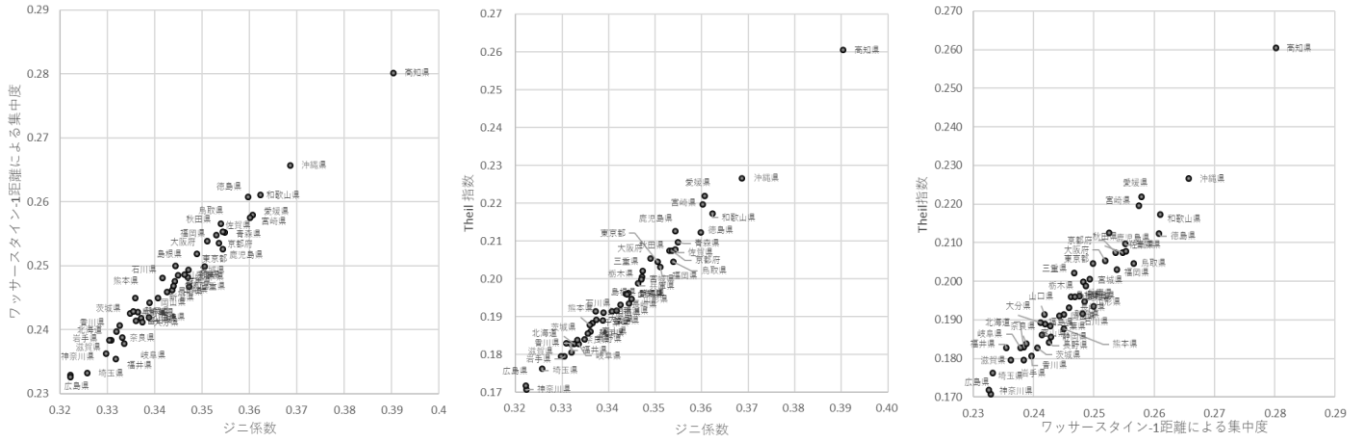
$$W' = \frac{1}{2\mu(n-1)} \sum_{i=1}^{n} \frac{1}{n} |x_i - \mu| \tag{3.6}$$

このW'は、「完全平等を達成するために再配分が必要な所得の割合」を示す直感的な指標である。

図 3 は、2019 年全国家計構造調査の結果に基づき、都道府県ごとにジニ係数、ワッサースタイン-1

距離（(3.6) 式によるもの)、および Theil 指数を算出しプロットした結果である。使用データは、世帯員 2 人以上の世帯について都道府県別に可処分所得を 40 階級に区分された公表統計表とし、各階級の代表値を中央値として計算した。この結果から、ジニ係数が 0.35、Theil 指数が 0.20 のとき、ワッサースタイン-1 距離は概ね 0.25 に対応しており、「総所得の 25%を再配分すれば均等化が達成される」と直感的に理解できる。このように、ジニ係数や Theil 指数に加えて、計算が容易で解釈の明確なワッサースタイン-1 距離を用いることで、集中度指標としての理解をより深化させることができる。

図 3　都道府県別ジニ係数、Theil 指数、ワッサースタイン-1 距離による集中度指標の比較図



3.4.2　ワッサースタイン-1 距離による地理的空間を考慮した産業などの集中度指標への活用

従来、産業や人口の空間的集中度には、自治体別シェアに基づくジニ係数やハーフィンダル指数が用いられてきた。これらは分布の偏りを把握する上で有用だが、地理的な位置関係を考慮しないという限界がある。例えば、産業立地が東京都と福岡県に集中する場合と、東京都と神奈川県に集中する場合とでは、ジニ係数やハーフィンダル指数は同一の集中度となる。しかし、実際には、東京都と神奈川県の方が地理的に近く、より高い集積度と考えるのが自然である。

さらにもう一歩踏み込むと、集積に対する人の感覚や経済的な効果として、東京都・神奈川県のように近い場合は集積していると捉える一方、一定距離以上離れた場合に違いが生じないことも考えられる。例えば、東京都・大阪府の 2 か所に集中している場合と東京都・福岡県の 2 か所に集中している場合の違いを過度に考慮することで逆に人の感覚にあわなくなったり、経済的な考察をする際に逆に大きなノイズとなったりする恐れもある。

したがって、考察すべき指標の特性として、

1)自治体間の距離が離れると集積関係を必ず弱くする厳密な距離関係を持たせる

2)距離が離れると集積関係は弱くなるが、一定距離以上離れるとそれ以上集積関係は限界に達しそれ以上弱くならない関係を持たせる

といった対応が考えられ、個別に対処できる指標が望ましい。

この課題等に対し、ワッサースタイン-1 距離を用いて地理的距離を組み込んだ集中度指標を提案する。

今、i,jを自治体、$d(i,j)$を自治体間の距離または近接関係（例：同一市区町村なら 0、隣接関係なら 1、隣接していないが同一都道府県内なら 2、それ以外を 3 とする）とする。例えば、集中度を計測する指標を上記 1)として計算する場合は距離を選び、指標を 2)の性質に類似するものとして計算する場合、県外に出れば集積関係は限界に達しどれだけ離れても集積関係は弱くならない場合はこの近接関係を選択する。Pを何らかの観測された自治体毎のシェア分布、Qを均等分布とすると、前述の式(2.4)によるPとQの間のワッサースタイン-1 距離を集中度指標とする。この最適輸送解 $\Gamma = (\gamma_{ij})$は、R や Python を用いて数値的に求めることができる。このとき、東京都と福岡県の双方に集中した分布は、東京都・神奈川県に集中する分布よりも全体への配分コストが低く、より低い集積度として評価される。

このアプローチは、産業の立地パターンや人口分布の空間的偏在を測定する際に、地理的距離を直接考慮できるという点で、従来指標に比べて優位性を持つ。一方で、地理的な距離を導入することにより、北海道や沖縄県に集積している産業は高い集積度として、逆に日本の地理的重心に位置する石川県などに集積している産業は低い集積度として評価される傾向が生じる。

この指標を評価するため、「2021 年経済センサス-活動調査」の都道府県別産業小分類別従業者数のデータx_{ij}（i:都道府県, j:産業（小分類 635 区分及びその上位区分を含め 767 区分））を用い、産業ごとに以下の 6 つの指標[ｉ）～vi)]を計算する。

ｉ）ハーフィンダル指数（**HHI**）　　ii）ジニ係数（**Gini**）　　iii）Theil 指数（**Theil**）

iv）実距離によるワッサースタイン-1 距離（**W(実距離)**）：コスト行列を、都道府県の重心の地理的緯度経度からユークリッド距離で計算された距離とする（距離が離れると集積関係を必ず弱くする関係を持たせ、厳密な空間的な関係を考慮している）

v）隣接考慮によるワッサースタイン-1 距離（**W(隣接考慮)**）：コスト行列を、同一都道府県間を 0、隣接都道府県間を 1、それ以外の都道府県間を 2 とする（隣接関係を考慮し、それ以上の距離の乖離は同一と見なし集積関係を弱くしない。空間的な粗い関係を考慮している）

vi）同一コストによるワッサースタイン-1 距離（**W(同一コスト)**）：コスト行列を、同一都道府県間の距離を 0、他の都道府県間の距離を 1 とする（この指標は距離を全く考慮せず、都道府県間の確率質量の差異を計測する指標となり、ジニ係数に近い性質を持つ）

さらに、均等分布の考え方として、①各都道府県で一様に 1／47 を割り当てる場合、②全産業計における都道府県別従業者数構成比を基準とする場合、の 2 種類が考えられることから、合計 12 の指標を計算・比較する。

（テストデータによる確認）

ワッサースタイン距離の特性を理解するため、表 1 に示す 4 つのテストデータ（Test1 は沖縄県に集中、Test2 は石川県に集中、Test3 は東京都 50%・福岡県 50%、Test4 は東京都 50%・神奈川県 50%と仮定したもの）に対し、前述の 12 種類の集中度指標を算出した。

Test1 と Test2 の比較では、HHI、Gini、Theil はいずれもほぼ同値を示し、分布②を用いた際に生じるわずかな差は、産業計のシェアの違いによるものである。なお、分布②で HHI が 1 を超えるのは L_2 距離の定義により沖縄県または石川県以外の県における差（マイナス値）の平方値を加算したためである。

表 1　テストデータによる各指標の比較

	均等分布：全都道府県=1/47（①の分布）					
	HHI	Gini	Theil	W(実距離)	W(隣接考慮)	W(同一コスト)
Test1（沖縄県に100%集積）	1.000	0.979	3.850	12.631	1.915	0.979
Test2（石川県に100%集積）	1.000	0.979	3.850	4.002	1.894	0.979
Test3（東京都に50%、福岡県に50%集積）	0.500	0.957	3.157	2.763	1.745	0.957
Test4（東京都に50%、神奈川県に50%集積）	0.500	0.957	3.157	4.634	1.830	0.957

	均等分布：産業計の分布（②の分布）					
	HHI	Gini	Theil	W(実距離)	W(隣接考慮)	W(同一コスト)
Test1（沖縄県に100%集積）	1.034	0.989	4.579	13.595	1.960	0.990
Test2（石川県に100%集積）	1.036	0.991	4.663	3.592	1.950	0.991
Test3（東京都に50%、福岡県に50%集積）	0.352	0.860	1.831	2.877	1.410	0.798
Test4（東京都に50%、神奈川県に50%集積）	0.331	0.829	1.617	3.260	1.449	0.777

一方、W (実距離) は本テストデータの場合、理論上、全確率質量を最も距離のある北海道・沖縄県間で移送する対応が最大となることから [0, 22.194]の値を取る。また、W (実距離) は沖縄県、石川県の地理的配置の違いから全国に再配分するコストを反映し、沖縄県集中（Test1）の方が大きな値を示した。W (隣接考慮) は隣接県間移動をコスト 1、非隣接県間の移送コストを一定（＝2）とする粗く緩やかな地理関係を反映しているため、日本のどこに集積されているか、その配置に敏感に反応しにくく、理論上[0,2]の範囲をとる指標である。

次に、Test3 と Test4 の比較では、分布①の場合、HHI、Gini、Theil はいずれも同値だが、分布②の場合、均等分布の神奈川県のシェアが福岡県より大きいため Test4 の値がやや小さい。一方、空間的距離を考慮する W(実距離)および W(隣接考慮)は、東京都―神奈川県間の近接性を反映して、Test4 でより大きな値を示しており、これにより、ワッサースタイン-1 距離に基づく W(実距離)および W(隣接考慮)の指標は、地理的な集約と分散を明確に区別でき、定義で記述した特性のとおり W(隣接考慮)と比較して W(実距離)は両者の地理的近接性の差を最も細かく数量化しており、地域性を反映した集中度評価指標として最も情報量が大きい指標であることが確認された。

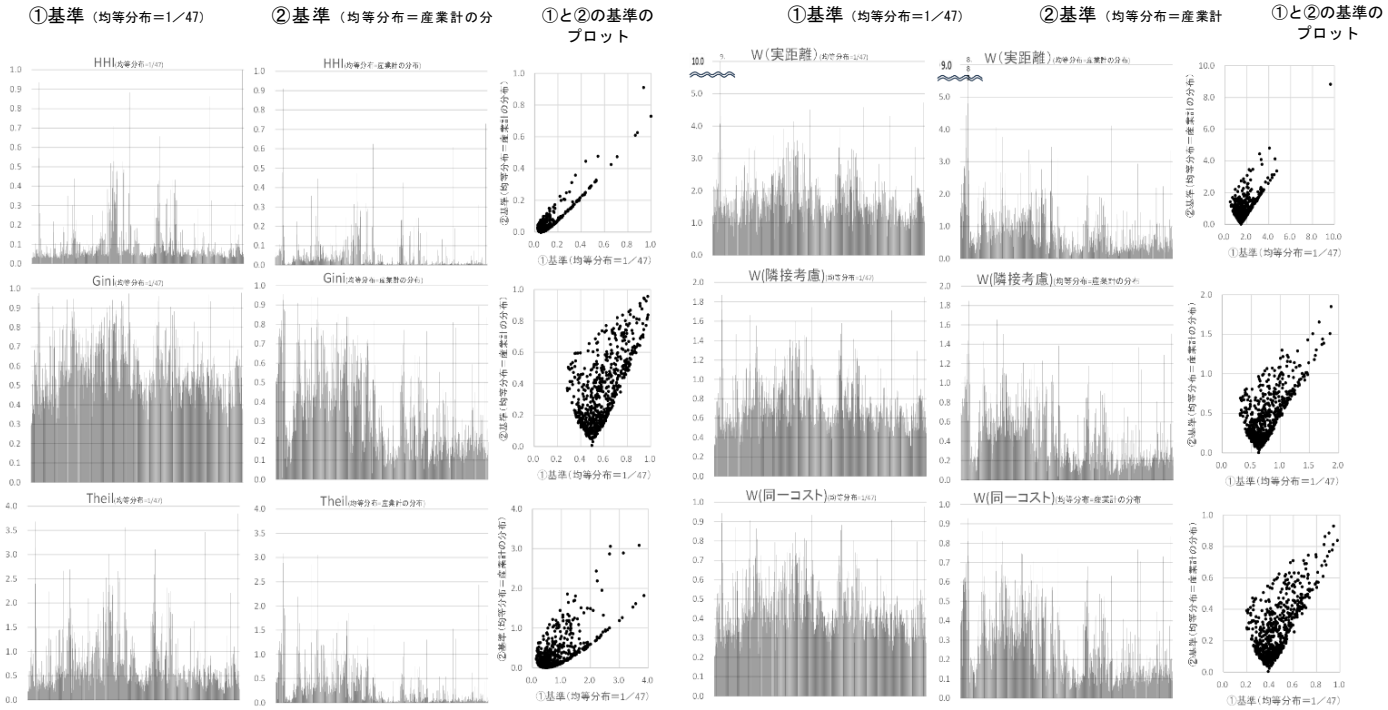
（経済センサス-活動調査の結果による確認）

実際の産業データの適用結果を示す。図 4 は、「2021 年経済センサス - 活動調査」に基づく都道府県別産業別従業者数構成比（全 767 産業区分のうち「管理・補助的経済活動を行う事業所」を除外）に対して、日本標準産業分類順に、2 種類の均等分布（①全都道府県均等、②産業計の構成比）を用いた 12 の集中度の指標を示している。

まず、①と②の違いを示すプロット図の挙動は、$W_{(実距離)}$、$W_{(隣接考慮)}$および $W_{(同一コスト)}$は類似している。また、算式構造の類似性から、$W_{(同一コスト)}$は Gini と類似した挙動を示し、HHI は Theil と類似した挙動を示している。

次に、均等分布②を用い、縦軸と横軸に異なる指標を配置し、産業ごとの指標値をプロットした結果を図 5 に示す。図 5①では、Gini と Theil の値が概ね曲線上で 1 対 1 の関係を示している。また、図 5④に示す Gini と $W_{(同一コスト)}$ の関係は、ほぼ線形であり、$W_{(同一コスト)} \fallingdotseq 0.75 \times Gini$ の関係が確認された。これは、ジニ係数に 0.75 を乗じた値が、概ねの「均等分布へ再配分される確率質量」となることを意味する。

図 4　令和 3 年経済センサス-活動調査に適用した産業別集中度指標の結果



さらに、$W_{(同一コスト)}$・$W_{(実距離)}$・$W_{(隣接考慮)}$の比較（図 5⑥～⑧）を行うことにより、空間的特性が数値に現れることを確認する。一例として、東京都の従業者数がシェア 55%を占める、集積度の高い情報通信業について紹介する。

図 6 は図 5⑥および⑦のうち、情報通信業に属する産業のデータを抜き出したものである。この図から、電気通信に付帯するサービス業は、他の産業と比べて W(同一コスト)の値に対して、$W_{(実距離)}$が高く出ていることが確認でき、何らかの空間的特性が存在しているものと推察される。一方、$W_{(隣接考慮)}$は他の産業と同様の傾向を示していることから、電気通信に付帯するサービス業は「隣接県での集中」ではないことも確認できる。

図 7 及び表 2 は情報通信業における各指標の結果と、シェア 1 位～3 位の都道府県におけるシェアの値を示している。実際に、電気通信に付帯するサービス業は大阪府がシェア 1 位であり、他産業とは異なる特徴的な結果となっている。これは、全産業平均がある程度東京都に集中している一方で、同産業が大阪府や福岡県に分散しているためである。その結果、情報通信業における他産業と比較して、当該産業は均等分布からの乖離が大きくなっており、$W_{(実距離)}$が空間的な特徴を数値として表している事例となる。

図５　各指標の相関関係

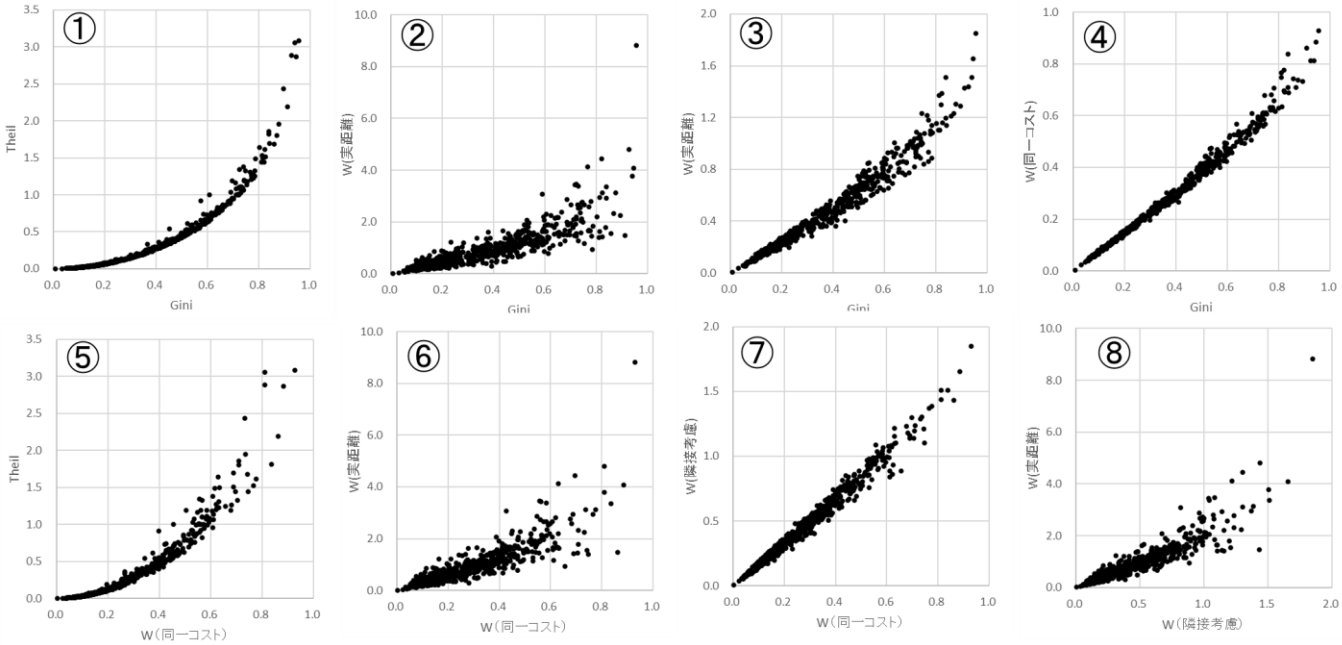


図６　情報通信業における W（同一コスト）と W（隣接考慮）、W（実距離）の関係

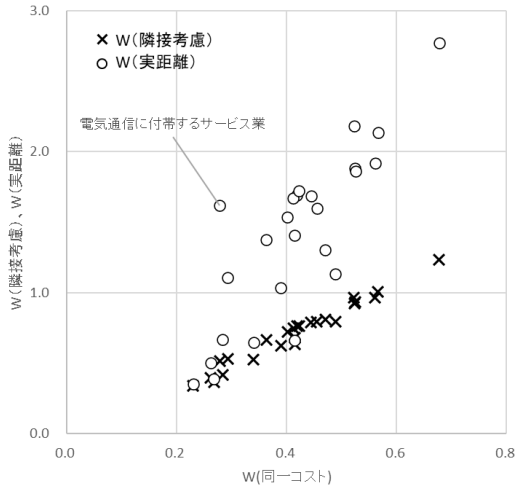


図７　情報通信業におけるシェア１位都道府県のシェアと W 指標の関係

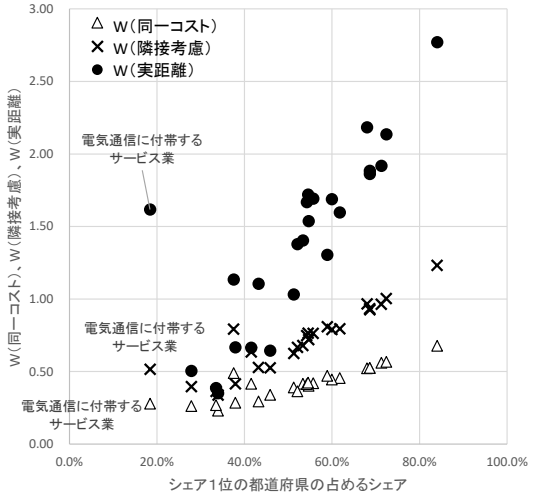


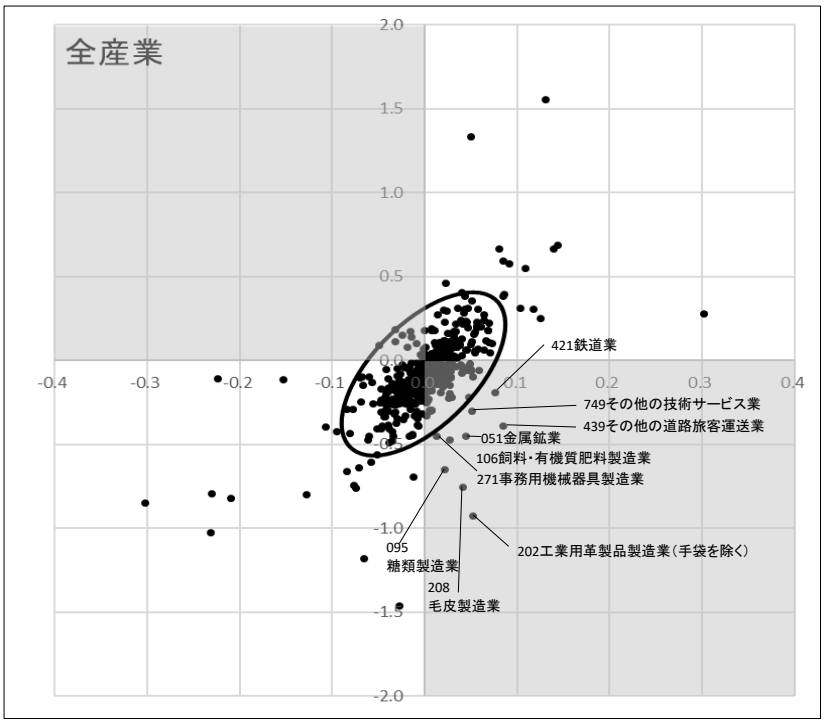
表２　情報通信業の集積指標

	Gini	W(同一コスト)	W(隣接考慮)	W(実距離)	シェア1位	シェア2位	シェア3位
情報通信業	0.501	0.402	0.720	1.537	東京都 0.55	大阪府 0.09	神奈川県 0.06
情報通信業（通信業，放送業，映像・音声・文字情報制作業）	0.481	0.390	0.625	1.032	東京都 0.51	大阪府 0.12	愛知県 0.04
通信業	0.497	0.415	0.634	0.664	東京都 0.42	大阪府 0.21	福岡県 0.06
固定電気通信業	0.594	0.489	0.792	1.135	東京都 0.38	大阪府 0.35	福岡県 0.04
移動電気通信業	0.691	0.561	0.963	1.918	東京都 0.71	大阪府 0.07	愛知県 0.05
電気通信に附帯するサービス業	0.379	0.278	0.515	1.617	大阪府 0.18	東京都 0.13	福岡県 0.11
放送業	0.322	0.230	0.338	0.351	東京都 0.34	大阪府 0.07	愛知県 0.06
公共放送業（有線放送業を除く）	0.454	0.340	0.525	0.644	東京都 0.46	大阪府 0.06	北海道 0.05
民間放送業（有線放送業を除く）	0.373	0.268	0.362	0.387	東京都 0.33	大阪府 0.10	愛知県 0.06
有線放送業	0.335	0.262	0.396	0.504	東京都 0.28	愛知県 0.07	千葉県 0.06
映像・音声・文字情報制作業	0.539	0.455	0.794	1.597	東京都 0.62	大阪府 0.07	愛知県 0.04
映像情報制作・配給業	0.612	0.524	0.925	1.884	東京都 0.69	大阪府 0.07	愛知県 0.03
音声情報制作業	0.745	0.678	1.232	2.771	東京都 0.84	大阪府 0.05	神奈川県 0.02
新聞業	0.382	0.284	0.415	0.667	東京都 0.38	大阪府 0.10	北海道 0.07
出版業	0.638	0.566	1.003	2.135	東京都 0.72	大阪府 0.05	愛知県 0.03
広告制作業	0.531	0.415	0.679	1.404	東京都 0.53	大阪府 0.12	愛知県 0.06
映像・音声・文字情報制作に附帯するサービス業	0.518	0.444	0.789	1.688	東京都 0.60	大阪府 0.07	神奈川県 0.03
情報通信業（情報サービス業，インターネット附随サービス業）	0.521	0.419	0.763	1.692	東京都 0.56	大阪府 0.08	神奈川県 0.08
情報サービス業	0.512	0.412	0.748	1.668	東京都 0.54	大阪府 0.09	神奈川県 0.08
ソフトウェア業	0.525	0.423	0.764	1.720	東京都 0.55	神奈川県 0.09	大阪府 0.09
情報処理・提供サービス業	0.463	0.363	0.666	1.378	東京都 0.52	大阪府 0.08	神奈川県 0.05
情報処理サービス業	0.387	0.293	0.528	1.106	東京都 0.43	大阪府 0.09	神奈川県 0.04
情報提供サービス業	0.562	0.471	0.808	1.305	東京都 0.59	福岡県 0.06	大阪府 0.05
その他の情報処理・提供サービス業	0.652	0.522	0.966	2.184	東京都 0.68	大阪府 0.08	神奈川県 0.06
インターネット附随サービス業	0.627	0.525	0.933	1.862	東京都 0.69	大阪府 0.07	福岡県 0.03

次に、指標の時間的変化から空間的に特徴的な産業を $W_{(同一コスト)}$ および $W_{(実距離)}$ を用いて確認する。図 8 は、2012 年および 2021 年の「経済センサス-活動調査」に基づき、産業計の都道府県シェアを均等分布とした場合における $W_{(同一コスト)}$ および $W_{(実距離)}$ の変化を示している。横軸は $W_{(同一コスト)}$ の変化量（2021 年－2012 年)、縦軸は $W_{(実距離)}$ の変化量である。

楕円は、両指標の変化の分布を二変量正規分布と仮定した場合の 95%信頼区間を示し、この外側に位置する産業は、変動が統計的に有意（上位 5 %）と見なされるものである。

図 8　産業別 $W_{(同一コスト)}$ 及び $W_{(実距離)}$ の時間的変化



総じて、$W_{(同一コスト)}$ と $W_{(実距離)}$ は同方向に変動しており、両者の変化量の符号が異なり、かつ 95%信頼区間ラインを超える産業は、空間的変動に特徴を有するものと考えられる。これらの産業の状況を図 9 に示す。図 9 は、楕円外に位置する 9 産業について、都道府県別シェア（2012－2021 年）およびその変化を示している。全体的には横軸目盛り“13”と表示されている東京都への集約が進む傾向がみられる一方で、これらの産業には、①集積度（$W_{(同一コスト)}$）が上昇する一方で集中先が複数地域に分散、②上位都道府県以外のシェアの増加、③北海道や九州・沖縄地域のシェア減少、のいずれかの傾向がみられた。このように、ワッサースタイン距離を用いることで、地理的関係を反映した集中・集積の動態分析が可能であることが確認された。

4　公的統計の様々な空間的特性の把握

本節では、分布間の距離を活用し、都道府県など自治体別に作成される公的統計データを対象として、①地理的配置、②性・年齢階級別人口構成比、③産業別従業者構成比などから構成される多元的距離空間において、公的統計の空間的自己相関の有無を判定する手法及び計算結果を提示する。

公的統計では、人口、産業、消費、物価、就業状況、生活時間などを把握する各種統計調査において、都道府県別結果が作成されている。例えば、消費に関する統計では、総額の消費支出のみならず、数百に及ぶ品目区分ごとの消費支出額も作成されている。本節では、そのうち、平均値が算出され、正規分布を仮定できる統計調査の結果を対象として考察を行う。

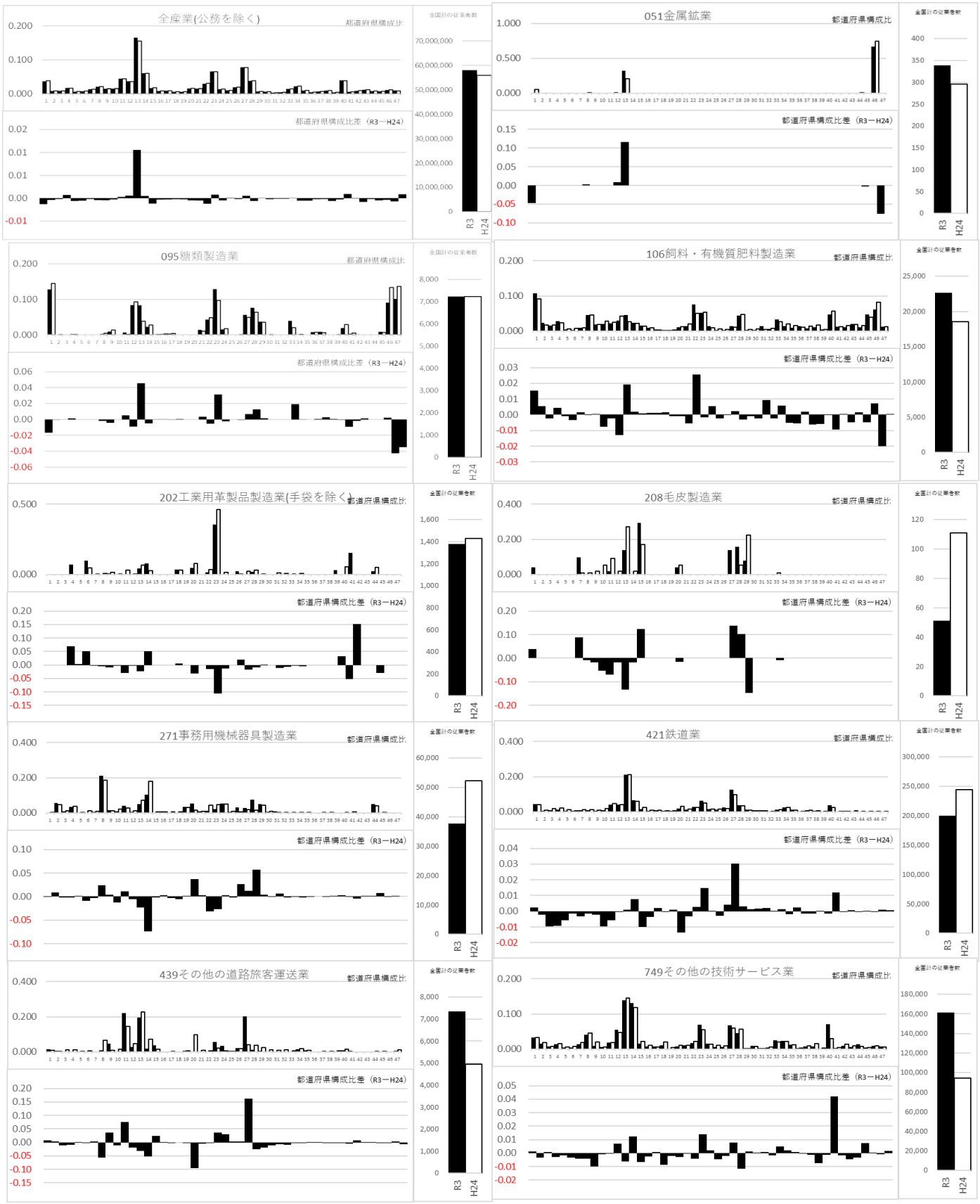
自治体 i 毎に観測される観察値 X_i について、次を仮定する：

$$\vec{X} = (X_1, X_2, \cdots, X_N)^T \sim N(\mu\vec{1}, \kappa\Sigma)$$
$$\vec{1} = (1,1,\cdots,1)^T,\ \mu, \kappa > 0$$

平均を $\mu\vec{1}$ とするのは、$\vec{X}$ が 1 標本であり、一定の条件のもとで分析を進めるためである。

これら統計調査の結果において、「地域的な類似性」や「性・年齢階級別人口構成、産業別従業者構成比が近接する地域での結果の類似性」といった多元的空間における空間的自己相関の存在を事前に把握できれば、その後の結果分析において有用である。従来、空間的自己相関の判定には、観測点間の隣接情報などから構築した重み行列を用いるモランⅠ統計量やギアリーC 統計量が用いられてきた。しかし、県庁所在市のように隣接情報の設定に判断を要する場合や、地理的距離以外（例えば性・年齢別人口構成比空間、産業別従業者構成比空間）における近接関係を扱う場合には、新たに重み行列を設定する必要が生じる。

図９　時間的変化で特異な変動を示した産業の状況



そこで本稿では、隣接関係の設定を必要としない空間的自己回帰モデルのうち、最も単純なモデルの一つである「バリオグラム線形モデル」[5]を用い、尤度統計量に基づいて都道府県別統計や県庁所在市別統計の特性を評価する方法を提示する。

4.1　モデルの構築（バリオグラム線形モデルに基づくモデル構築）

地理的空間においては、距離が近いほど事物の性質が似るという傾向、すなわち空間的自己相関が観察されることがある（例：関東では納豆の消費額が高く、西日本では低い傾向）。同様の自己相関は、性・年齢別人口構成比や産業別従業者構成比などの分布空間でも観察される。

観測点（例えば自治体）iおよびjの状態や構造（例えば、地理的配置（緯度・経度）、性・年齢別人口構成比、産業別従業者構成比、気象情報など）をベクトル$\vec{s_i}$と$\vec{s_j}$とする。このとき、$\vec{s_i}$と$\vec{s_j}$から判断して両観測点が「近接」している場合、観測値X_iとX_jも近い値をとるかどうかを把握することを考える。この関係性を評価するため、観測値X_iおよびX_jに対して、空間統計学における「バリオグラム線形モデル」を適用し、ナゲット効果[6]を 0 とするモデルのフィッティングを行う。

ここで、「バリオグラム線形モデル」でナゲットを 0 とするモデルは、$d(\vec{s_i},\vec{s_j})$を自治体 i と j の間の距離（$\vec{s_i},\vec{s_j}$の設定に応じ，地理上の距離、性・年齢階級別人口構成比間の距離、産業別従業者構成比間の距離、気象条件の距離などが該当し、具体的には、ノルム系の距離、ワッサースタイン距離、ダイバージェンス距離などで計算される）とすると、観測値の差の分散が両者の距離に比例するという以下の構造を持つモデルである。

$$Var(X_i - X_j) = C \times d(\vec{s_i},\vec{s_j}) \qquad (4.1)$$
$$C > 0：比例定数$$

このモデルは、端的に i と j の状態や構造（$\vec{s_i}$と$\vec{s_j}$）が類似していれば観測値X_iとX_jも近い値を取る（差$|X_i - X_j|$が小さくなる可能性が高くなる）構造を示している。この構造に対し、基準となる i=0 点を決めると、$Var(X_i - X_0) = C \times d(\vec{s_i},\vec{s_0})$，　$Var(X_j - X_0) = C \times d(\vec{s_j},\vec{s_0})$，　$Var(X_i - X_j) = C \times d(\vec{s_i},\vec{s_j})$となり、$X_i - X_j = X_i - X_0 + X_0 - X_j$として変形すると、

$$Cov(X_i - X_0, X_j - X_0) = C \times \frac{d(\vec{s_i},\vec{s_0}) + d(\vec{s_j},\vec{s_0}) - d(\vec{s_i},\vec{s_j})}{2} \qquad (4.2)$$

を得る。共分散行列は半正値対称行列となる必要があるが、Schoenberg(1937)及びそれ以降の研究により、$d(\vec{y_i},\vec{y_j})$がユークリッド距離の場合、任意の観測点 i のいかなる状態や構造$\vec{y_i}$であっても、共分散行列は半正値対称行列となること、また、それ以外の L_1 距離、ワッサースタイン-1 距離や Jensen-Shannon 距離などの場合、半正値対称行列になることは一般には保証されないことが示されている。そのため、ユークリッド距離以外の距離関数で共分散行列を構成した場合に、半正値行列となることを数値計算で確認する必要がある。

また、上記 (4.2) で定義した共分散行列は、基準点 (0 自治体) をどの自治体に設定するか依存する。そこで、すべての自治体を公平に扱うため、基準点（0 自治体）をすべての自治体に順次設定し、それぞれの共分散行列を求め、その平均を取ることが行われる。この対応に沿って、バリオグラム線形モデル（ナゲット＝0）に基づく共分散行列を以下のように計算する。

$$\Sigma = [\sigma_{ij}] = [Cov(X_i, X_j)] = [\sum_{k=1}^{N} \frac{1}{N} Cov(X_i - X_k, X_j - X_k)] = [\sum_{k=1}^{N} \frac{1}{N} \left(\frac{d(\vec{y_i},\vec{y_k}) + d(\vec{y_j},\vec{y_k}) - d(\vec{y_i},\vec{y_j})}{2} \right)] \qquad (4.3)$$

この行列Σは、すべての行和・列和が同値となり、その値はΣの最大固有値と一致する（補論参照）。

実際の計算では、自治体の状態や構造を示す$\vec{s_i}$の指標によりΣの各要素の水準は大きく異なる。しかし、共分散行列にスカラーλを乗じ、そのλを最尤推定するため、事前に共分散行列 Σ をスカラー倍しても問題ない。そこで、地理的距離、性・年齢別人口構成比に基づく距離、産業別従業者構成比に基づく距離における共分散行列の数値を比較しやすくするため、Σ を行和（＝列和＝最大固有値）で割ってリ

[5] バリオグラム線形モデル：観測値間の差の分散が観測点間距離に比例するという仮定に基づく。距離が近い地点ほど観測値が類似するという空間的自己相関を表現するモデルの１つ。

[6] ナゲット効果とは、距離がゼロに近づいても観測値が完全に一致せず、ある程度のばらつき（誤差）が残る現象を表す。本研究では簡略化のためナゲット＝0 と仮定する。

スケーリングした共分散行列（行和＝列和＝最大固有値＝１となる）を用いる。

$$\Sigma \leftarrow \frac{\Sigma}{\sum_{i=1}^{N}\sigma_{ij}} \tag{4.4}$$

この共分散行列Σをスカラー倍した$\lambda\Sigma$を実際の観測値の共分散行列と仮定し、

$$\vec{X} = (X_1, X_2, \cdots, X_N)^T \sim N(\mu\vec{1}, \lambda\Sigma),\ \text{パラメータ}\mu\text{、}\lambda\text{は最尤推定量を代入}$$

に対する当てはまりの良さを対数尤度によって評価する。このとき、観測値$\vec{X}$に対する平行移動及びスカラー倍はμ、λの最尤推定量に吸収され、結果として対数尤度に影響を与えない。そこで、

$$Z_i = \frac{X_i - \bar{\mu}}{\sqrt{\overline{\sigma^2}}}\quad \left(\bar{\mu} = \frac{1}{N}\sum_{i=1}^{N} X_i\ ,\ \overline{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{\mu})^2\right) \tag{4.5}$$

と標準化した、$\vec{Z} = (Z_1, Z_2, \cdots, Z_N)^T$に対して、$\vec{Z} \sim N(\nu\vec{1}, \lambda\Sigma)$を仮定し、空間的自己相関「ありモデル」と「なしモデル」の対数尤度差を空間的自己相関の有無の判断指標とする。ここで、空間的自己相関「なしモデル」は、共分散行列が$\mathrm{Diag}(\Sigma)$または単位行列Ｉの場合とする。なお、$\vec{Z}$は標準化されており、関心事は$\vec{Z}$の共分散行列$\lambda\Sigma$への適合度にあるため、平均について$\nu = 0$とするモデル、Σを相関行列化した$\mathrm{Corr}(\Sigma)$に置き換えたモデルも評価対象に含め、評価モデルを表３のとおりとする。

表３　検討モデル

	空間的自己相関「ありモデル」	
Model(1): $\vec{Z} \sim N(\nu\vec{1}, \lambda\Sigma)$	Model(2): $\vec{Z} \sim N(\vec{0}, \lambda\Sigma)$	Model(3): $\vec{Z} \sim N(\vec{0}, \lambda\mathrm{Corr}(\Sigma))$
	空間的自己相関「なしモデル」	
Model(4): $\vec{Z} \sim N(\nu\vec{1}, \lambda\mathrm{Diag}(\Sigma))$	Model(5): $\vec{Z} \sim N(\vec{0}, \lambda\mathrm{Diag}(\Sigma))$	Model(6): $\vec{Z} \sim N(\vec{0}, \lambda I)$

さらに、同一データセット$X = (X_{ij}$, i: 自治体, j: データ項目)でモデルの統一化を図るため、以下の手順を採用する。

１）i 自治体の状態・構造のデータ$\overrightarrow{s^k_{\ i}}$[$k$: 地理的配置，人口, 産業など]とする。このとき、全ての自治体の組合せに対して距離関数$d_l\left(\overrightarrow{s^k_{\ i1}}, \overrightarrow{s^k_{\ i2}}\right)$[i1 i2:自治体，l:L1,L2,JS,ワッサースタイン-1]を用い、式(4.3)及び(4.4)により、kとlの全ての組合せについて共分散行列Σ_{kl}を作成する。

２）各データ項目について、全てのΣ_{kl}に対し Model(1)～(6)の AIC を計算[7]する。

３）Model(1)～(3)の中で最も多く支持されたモデルを Model(Q1)、Model1(4)～(6)の中で最も多く支持されたモデルを Model(Q2)とし、これらをデータセットXにおける共通モデルとする。

４）モデルを Model(Q1)及び Model(Q2)に固定した上で、状態・構造の情報 k も固定し、Σ_{kl}の距離関数の種類 l ごとに全データ項目の AIC を計算し、l の中で最も多く支持された距離関数の種類 L を、状態・構造の情報 k に対応する距離関数として採択する。

５）データセットX全体において Model(Q1)、Model(Q2)、距離関数 L を採用し、状態・構造 k ごとに得られる共分散行列Σ_{kL}を用いて、各データ項目について Model(Q1)と Model(Q2)の対数尤度差を計算し、状態・構造 k 別に各データ項目の空間的自己相関の有無の指標とする。

4.2　試算結果

試算は、（独）統計センターが提供する教育用データセット SSDSE（詳細は統計センターHP 参照）を以下のとおり一部加工して使用した。

１）家計データ（SSDSE-C）：県庁所在市別の１世帯当たりの食料費目の平均消費支出額C_i^j (i:県庁所在市,j:品目)を、各品目の食料支出に占める構成比$X_i^j = C_i^j / \sum_j C_i^j$に加工した情報

２）行動データ（SSDSE-D）：都道府県別の行動データ（このうち行動率のみ使用）

あわせて、都道府県・県庁所在市 i の状態や構造ベクトル$\vec{s_i}$を次のとおり設定した。

A）緯度・経度（地理的重心）

B）性・年齢階級別構成比（2020 年国勢調査：性・１歳階級別の人口構成比）

[7] 共分散行列Σ_{kl}の組み合わせの数にデータ項目数 M と Model の数（＝6）を乗じた数の AIC が計算される。

C）産業別従業者数構成比（2021 年経済センサス：産業小分類別従業者数構成比）
D）気象情報（SSDSE-F の県庁所在市別年間 6 指標（平均気温、平均現地気圧、平均風速、日照時間の合計、降水量の合計、雪日数を、47 市の単純平均を引いて単純標準偏差で割った値）

これらの情報から、表 4 に示す選択可能な距離関数を設定してそれぞれ共分散行列Σを計算した[8]。

この共分散行列Σに基づき、各データセット（家計データ、行動データ）について 4.1 節で前述した AIC の計算によりモデル統一化を行い（結果は表 5）、これに基づき、家計データ及び行動データについて、空間的自己相関性の有無の指標を計算した。

表 4　状態・構造情報と距離関数の適用一覧

状態・構造の情報	対象地域	Σを作成するために使用する距離関数	1）家計データ	2）行動データ
A）緯度・経度情報	都道府県	L_2 距離	－	○
	県庁所在市	L_2 距離	○	－
B）性・年齢階級別人口構成比	都道府県	L_1・L_2 距離／ワッサースタイン-1 距離[9]／JS 距離	－	○
	県庁所在市	L_1・L_2 距離／ワッサースタイン-1 距離 ／JS 距離	○	－
C）産業別従業者数構成比	都道府県	L_1・L_2 距離／JS 距離	－	○
	県庁所在市	L_1・L_2 距離／JS 距離	○	－
D）気象情報	県庁所在市	L_1・L_2 距離	○	－

表 5　データセット毎の採択されたモデルおよびΣの距離関数

データセット	採択モデル（対数尤度差計算モデル）	Σの距離関数			
		地理的配置	性・年齢階級別分布	産業別従業者分布	気象情報
家計データ	Model3－Model6	L_2 距離	JS 距離	L_1 距離	L_2 距離
行動データ	Model2－Model6	L_2 距離	L_2 距離	L_1 距離	－

＜家計データの結果＞

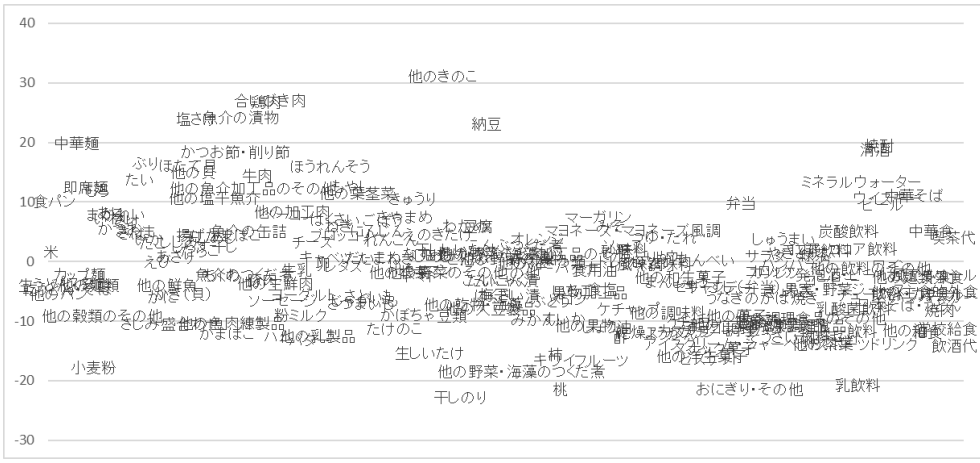
図 10-1～10-4 は、家計データについて、共分散行列Σを構成した 4 種類の状態・構造（「緯度・経度情報」、「性・年齢階級別人口構成比」、「産業別従業者数構成比」、「気象情報」）ごとに、以下の内容を示したものである。

① 対数尤度差をデータセットにおける品目格納順にプロットした図
② 対数尤度差上位 20 項目の結果
③ Σの元となった距離行列に基づき多次元尺度法で県庁所在市を 2 次元配置し、最大領域制約付きボロノイ分割（領域内の最大距離を L_1 距離で制約）でエリアを構成し、対数尤度差上位 2 品目の同差を該当するボロノイ図のエリアにヒートマップとして表示した図

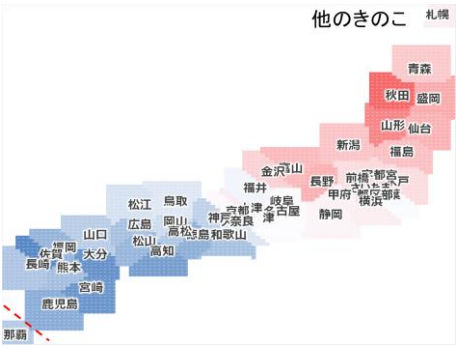
対数尤度差の高い品目をヒートマップで確認すると、4 種類の状態・構造に基づく空間配置に対して、それぞれ、偏りが生じている品目が存在していることが分かり、対数尤度差によって空間的自己相関性のあるものが抽出されていることが確認される。特に家計データは、緯度・経度情報やこれに関係のある気象情報の共分散行列に対し高い対数尤度差を持つ品目が複数観察され、家計データは地理的要因との結びつきが強いことを示唆する結果となった。

[8] L1 距離、ワッサースタイン-1 距離及び JS 距離で作成したΣの半正値性は数値的に検証し、固有値が非負であることを確認している。
[9] 性・年齢階級別人口構成比に対するワッサースタイン-1 距離は、年齢を 1 歳あたりの移動コストを 1 とし、男性⇔女性の移動コストを、一案として「移動後に得られる女性の分布を元の女性分布に戻すために要する最小輸送コスト」とし、全国計の男性の年齢階級別構成比と全国計の女性の年齢階級別構成比の間のワッサースタイン-1 距離（男性⇔女性の移動コスト＝3.2 歳）で評価した。

図 10-1　地理的配置（L_2 距離）における対数尤度差と上位項目のヒートマップ



	品目名	対数尤度差
1	他のきのこ	31.10
2	合いびき肉	27.06
3	鶏肉	26.87
4	魚介の漬物	24.19
5	塩さけ	23.92
6	納豆	23.19
7	中華麺	19.97
8	焼酎	19.46
9	清酒	18.86
10	かつお節・削り節	18.43
11	ぶり	16.50
12	ほたて貝	16.10
13	ほうれんそう	16.05
14	他の貝	15.00
15	牛肉	14.56
16	たい	13.76
17	ミネラルウォーター	13.43
18	もやし	12.58
19	即席麺	12.46
20	他の魚介加工品のその他	12.36



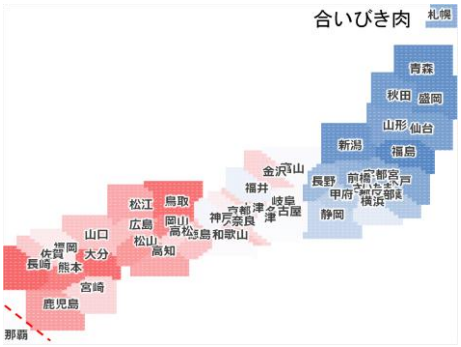
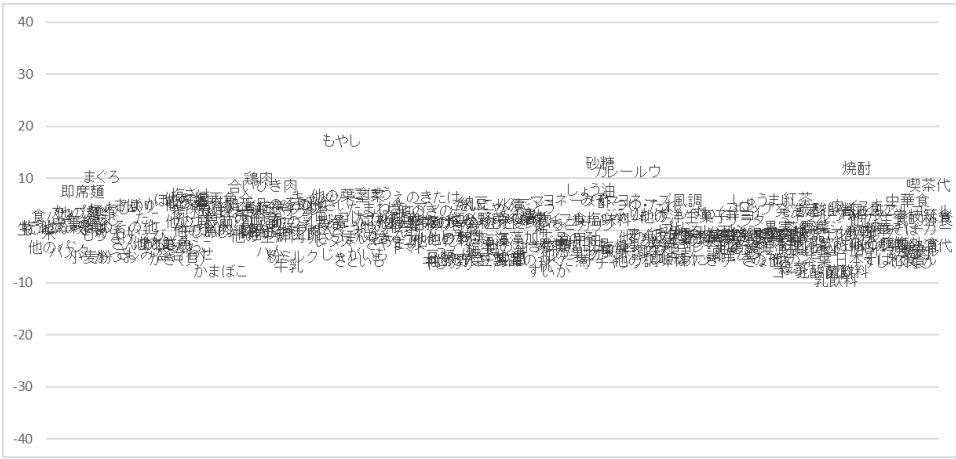
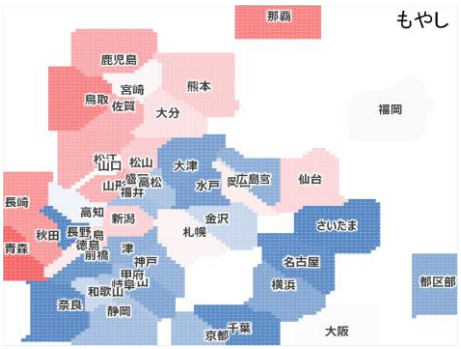


図 10-2　性・年齢階級別人口構成比による配置（JS 距離）における対数尤度差と上位項目のヒートマップ



	品目名	対数尤度差
1	もやし	17.31
2	砂糖	12.90
3	焼酎	12.10
4	カレールウ	11.41
5	まぐろ	10.29
6	鶏肉	10.01
7	喫茶代	8.84
8	合いびき肉	8.70
9	しょう油	7.79
10	即席麺	7.50
11	ごぼう	7.03
12	塩さけ	6.89
13	他の葉茎菜	6.75
14	えのきたけ	6.54
15	キャベツ	6.25
16	紅茶	6.22
17	他の塩干魚介	6.11
18	みそ	5.96
19	中華食	5.92
20	しゅうまい	5.81



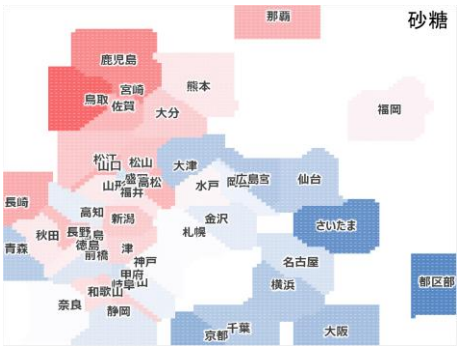
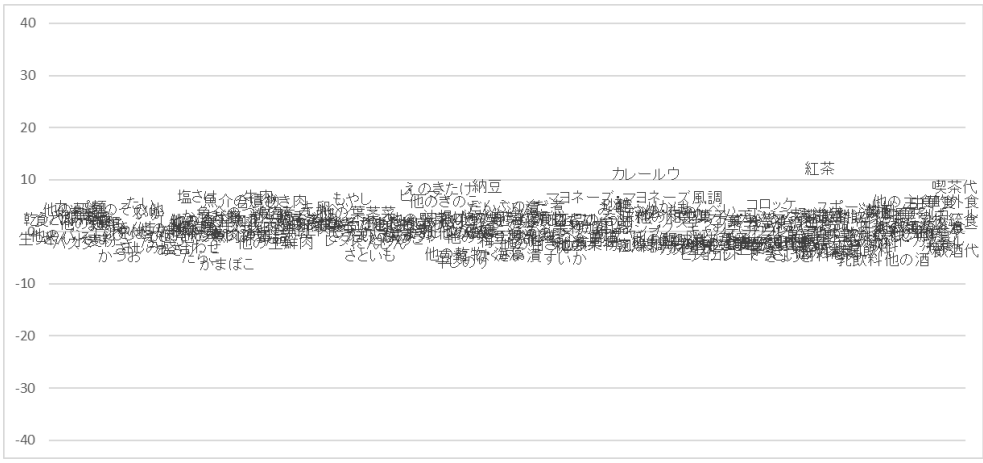
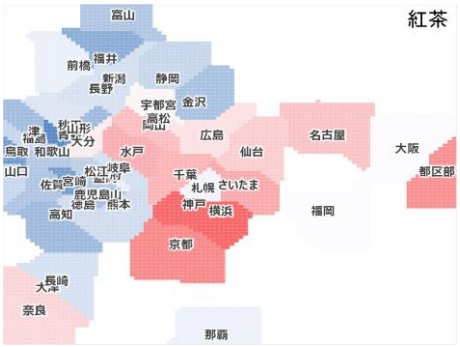


図 10-3　産業別従業者構成比による配置(L_1 距離)における対数尤度差と上位項目のヒートマップ



	品目名	対数尤度差
1	紅茶	12.26
2	カレールウ	11.24
3	納豆	8.91
4	喫茶代	8.64
5	えのきたけ	8.41
6	ピーマン	7.17
7	牛肉	6.87
8	塩さけ	6.76
9	もやし	6.37
10	魚介の漬物	6.08
11	他の主食的外食	5.93
12	他のきのこ	5.91
13	合いびき肉	5.83
14	中華食	5.81
15	たい	5.65
16	コロッケ	5.37
17	酢	5.20
18	砂糖	5.07
19	こんぶつくだ煮	5.00
20	ふりかけ	4.83



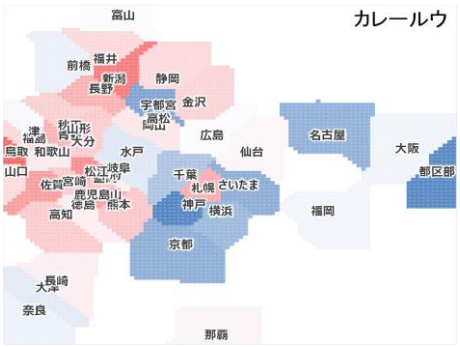
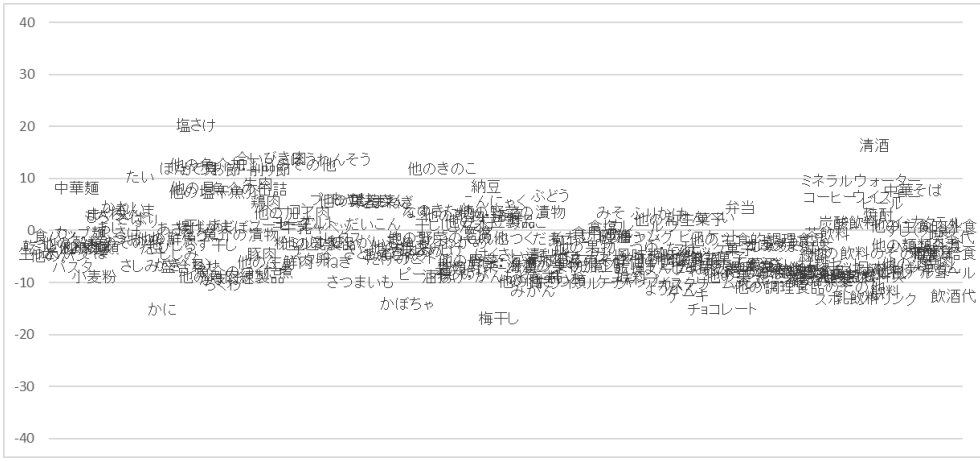
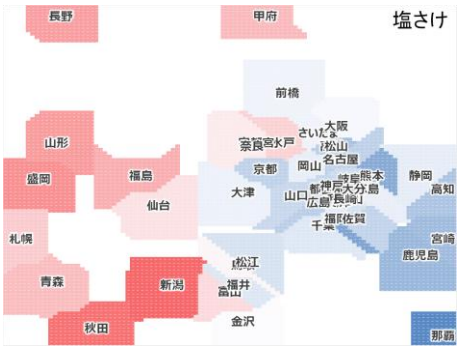
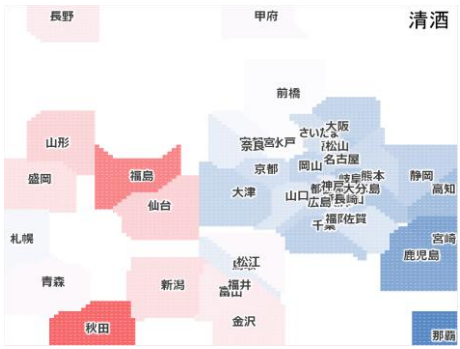


図 10-4　気象情報による配置(L_2 距離)における対数尤度差と上位項目のヒートマップ



	品目名	対数尤度差
1	塩さけ	20.33
2	清酒	16.41
3	合いびき肉	13.97
4	ほうれんそう	13.67
5	他の魚介加工品のそ	12.67
6	他のきのこ	11.98
7	ほたて貝	11.97
8	かつお節・削り節	11.76
9	たい	10.46
10	ミネラルウォーター	9.69
11	牛肉	9.00
12	納豆	8.59
13	中華麺	8.24
14	他の貝	8.18
15	魚介の缶詰	8.12
16	中華そば	7.82
17	他の塩干魚介	7.42
18	ぶどう	6.70
19	ウイスキー	6.48
20	コーヒー	6.48





＜行動データの結果＞

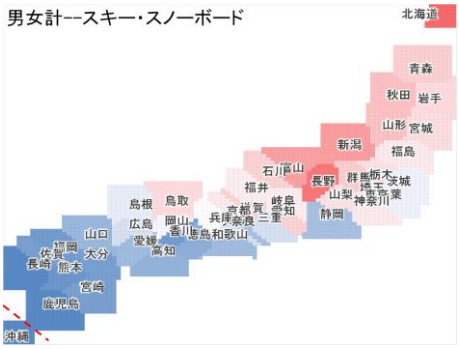
行動データにも家計データと同様の計算を行った。

図 11-1～11-3 は、行動データについて、「緯度・経度情報」、「性・年齢階級別人口構成比」、「産業別従業者数構成比」に基づく共分散行列Σごとに、①対数尤度差の上位項目（男女計、男性、女性）、②男女計で上位２項目のヒートマップの結果を示したものである。

結果として、家計データの場合と同様、状態・構造に応じて構成したΣにより、空間的自己相関の強い項目が抽出されることが確認できた。とりわけ行動データでは、地理的配置よりも、性・年齢階級別人口構成比や産業別従業者構成比に基づくΣにおいて、高い対数尤度差が多く観察され、行動データが人口・産業構成と相関関係が強いことを示唆する結果となった。

図 11-1　地理的配置における対数尤度差上位項目とヒートマップ

男女計			男性			女性		
	品目名	対数尤度差		品目名	対数尤度差		品目名	対数尤度差
1	スキー・スノーボード	25.08	1	スキー・スノーボード	24.69	1	スキー・スノーボード	17.33
2	つり	15.84	2	つり	17.79	2	行楽(日帰り)	5.89
3	登山・ハイキング	6.71	3	登山・ハイキング	10.96	3	絵画・彫刻の制作	3.37
4	行楽(日帰り)	5.05	4	写真の撮影・プリント	4.53	4	スマートフォン・家庭用ゲーム機などによるゲーム	2.53
5	ウォーキング・軽い体操	3.79	5	観光旅行	4.06	5	スポーツの総数	1.25
6	スポーツの総数	3.65	6	パソコンなどの情報処理	3.08	6	登山・ハイキング	0.82
7	ゴルフ(練習場を含む)	3.44	7	ウォーキング・軽い体操	2.37	7	旅行・行楽の総数	0.64
8	観光旅行	2.54	8	行楽(日帰り)	2.03	8	観光旅行	0.55
9	写真の撮影・プリント	1.31	9	趣味・娯楽の総数	1.71	9	映画館での映画鑑賞	-0.43
10	絵画・彫刻の制作	1.08	10	スポーツの総数	1.05	10	ウォーキング・軽い体操	-0.53
11	スマートフォン・家庭用ゲーム機などによるゲーム	1.05	11	商業実務・ビジネス関係	1.01	11	邦楽(民謡、日本古来の音楽を含む)	-1.25
12	高齢者を対象とした活動	0.84	12	ゴルフ(練習場を含む)	0.89	12	趣味としての読書（マンガを除く）	-1.48
13	旅行・行楽の総数	0.69	13	商業実務・ビジネス関係(総数)	0.62	13	編み物・手芸	-1.82
14	趣味・娯楽の総数	0.04	14	趣味としての料理・菓子作り	0.32	14	楽器の演奏	-2.12
15	邦楽(民謡、日本古来の音楽を含む)	-0.82	15	旅行・行楽の総数	0.20	15	写真の撮影・プリント	-2.14
16	自然や環境を守るための活動	-0.95	16	グラウンドゴルフ	0.18	16	ヨガ	-2.24
17	パソコンなどの情報処理	-1.03	17	旅行(１泊２日以上)	-0.06	17	趣味・娯楽の総数	-2.71
18	まちづくりのための活動	-1.11	18	国内	-0.16	18	映画館以外での映画鑑賞（テレビ・DVD・パソコンなど）	-2.76
19	趣味としての読書（マンガを除く）	-1.23	19	学習・自己啓発・訓練の総数	-0.19	19	高齢者を対象とした活動	-3.12
20	グラウンドゴルフ	-1.52	20	まちづくりのための活動	-1.02	20	まちづくりのための活動	-3.13



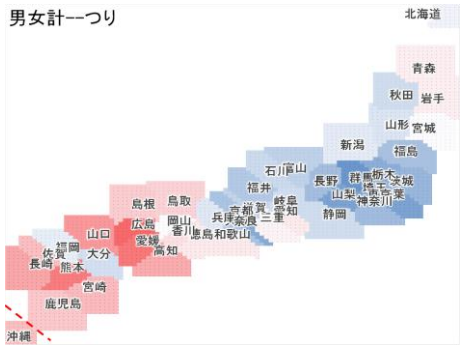
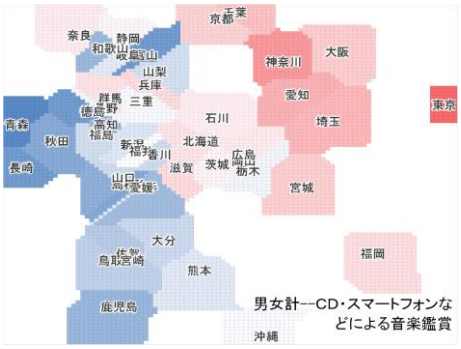


図 11-2　性・年齢階級別人口構成比の配置における対数尤度差上位項目とヒートマップ

	男女計		男性		女性	
	品目名	対数尤度差	品目名	対数尤度差	品目名	対数尤度差
1	ＣＤ・スマートフォンなどによる音楽鑑賞	37.58	ＣＤ・スマートフォンなどによる音楽鑑賞	32.24	ＣＤ・スマートフォンなどによる音楽鑑賞	37.92
2	外国語	33.81	外国語	29.18	映画館以外での映画鑑賞（テレビ・DVD・パソコンなど）	32.69
3	映画館以外での映画鑑賞（テレビ・DVD・パソコンなど）	32.15	英語	28.40	外国語	28.52
4	英語	31.94	スマートフォン・家庭用ゲーム機などによるゲーム	28.05	スマートフォン・家庭用ゲーム機などによるゲーム	26.41
5	スマートフォン・家庭用ゲーム機などによるゲーム	31.67	学習・自己啓発・訓練の総数	25.64	趣味としての読書（マンガを除く）	25.33
6	英語以外の外国語	27.79	商業実務・ビジネス関係	24.78	英語	24.98
7	学習・自己啓発・訓練の総数	27.44	商業実務・ビジネス関係(総数)	24.58	演芸・演劇・舞踊鑑賞(テレビ・スマートフォン・パソコンなどは除く)	24.55
8	商業実務・ビジネス関係	26.27	パソコンなどの情報処理	23.57	英語以外の外国語	24.39
9	商業実務・ビジネス関係(総数)	25.89	映画館以外での映画鑑賞（テレビ・DVD・パソコンなど）	23.35	スポーツの総数	23.64
10	趣味としての読書（マンガを除く）	25.31	趣味としての読書（マンガを除く）	22.07	学習・自己啓発・訓練の総数	23.44
11	マンガを読む	25.04	趣味・娯楽の総数	21.22	趣味・娯楽の総数	23.44
12	趣味・娯楽の総数	24.98	旅行(１泊２日以上)	19.98	ヨガ	22.98
13	演芸・演劇・舞踊鑑賞(テレビ・スマートフォン・パソコンなどは除く)	24.17	国内	19.91	マンガを読む	22.05
14	パソコンなどの情報処理	23.71	英語以外の外国語	19.83	写真の撮影・プリント	21.43
15	スポーツの総数	23.57	マンガを読む	19.32	楽器の演奏	21.35
16	写真の撮影・プリント	22.45	観光旅行	19.24	コンサートなどによるポピュラー音楽・歌謡曲鑑賞	20.56
17	コンサートなどによるポピュラー音楽・歌謡曲鑑賞	22.22	まちづくりのための活動	16.86	国内	19.27
18	楽器の演奏	22.21	映画館での映画鑑賞	16.79	旅行(１泊２日以上)	19.25
19	ヨガ	21.52	ボランティア活動の総数	16.52	サイクリング	18.60
20	サイクリング	20.80	演芸・演劇・舞踊鑑賞(テレビ・スマートフォン・パソコンなどは除く)	16.51	映画館での映画鑑賞	18.32



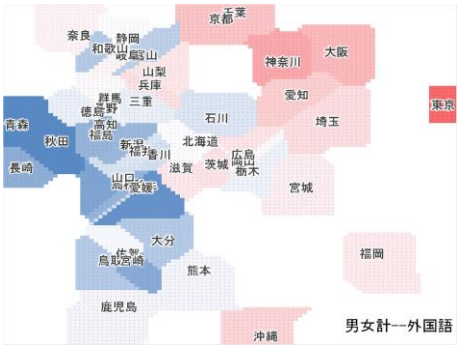
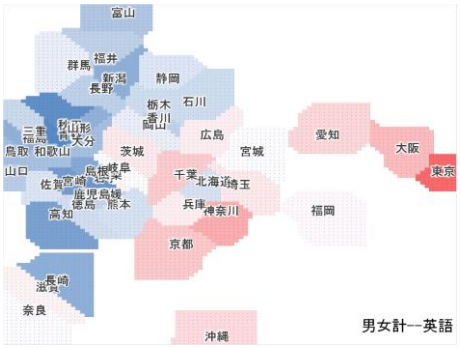
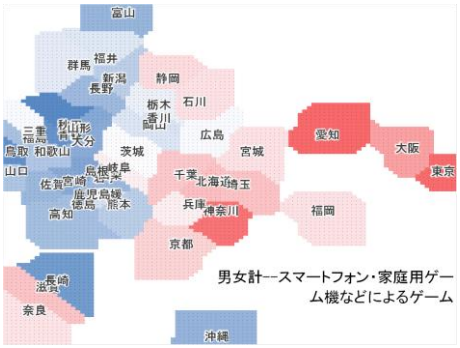


図 11-3　産業別従業者数構成比の配置における対数尤度差上位項目とヒートマップ

	男女計		男性		女性	
	品目名	対数尤度差	品目名	対数尤度差	品目名	対数尤度差
1	外国語	27.37	スマートフォン・家庭用ゲーム機などによるゲーム	24.98	ＣＤ・スマートフォンなどによる音楽鑑賞	26.45
2	英語	26.18	外国語	24.58	映画館以外での映画鑑賞（テレビ・DVD・パソコンなど）	25.26
3	ＣＤ・スマートフォンなどによる音楽鑑賞	26.11	英語	24.09	外国語	23.26
4	スマートフォン・家庭用ゲーム機などによるゲーム	25.86	ＣＤ・スマートフォンなどによる音楽鑑賞	22.95	演芸・演劇・舞踊鑑賞(テレビ・スマートフォン・パソコンなどは除く)	22.49
5	英語以外の外国語	25.29	商業実務・ビジネス関係(総数)	21.12	マンガを読む	22.22
6	映画館以外での映画鑑賞（テレビ・DVD・パソコンなど）	25.03	パソコンなどの情報処理	21.00	スマートフォン・家庭用ゲーム機などによるゲーム	22.15
7	商業実務・ビジネス関係(総数)	23.69	英語以外の外国語	20.15	英語以外の外国語	21.53
8	パソコンなどの情報処理	22.57	学習・自己啓発・訓練の総数	19.77	英語	21.25
9	マンガを読む	22.51	映画館以外での映画鑑賞（テレビ・DVD・パソコンなど）	19.76	趣味としての読書（マンガを除く）	20.71
10	学習・自己啓発・訓練の総数	21.79	商業実務・ビジネス関係	19.07	学習・自己啓発・訓練の総数	19.51
11	演芸・演劇・舞踊鑑賞(テレビ・スマートフォン・パソコンなどは除く)	21.55	ウォーキング・軽い体操	18.70	ヨガ	18.17
12	趣味としての読書（マンガを除く）	20.51	趣味としての読書（マンガを除く）	18.46	趣味・娯楽の総数	17.91
13	商業実務・ビジネス関係	20.48	趣味・娯楽の総数	18.09	サイクリング	17.90
14	ウォーキング・軽い体操	19.54	映画館での映画鑑賞	16.92	商業実務・ビジネス関係(総数)	17.62
15	趣味・娯楽の総数	19.53	マンガを読む	16.86	スポーツの総数	17.08
16	スポーツの総数	18.35	演芸・演劇・舞踊鑑賞(テレビ・スマートフォン・パソコンなどは除く)	15.37	写真の撮影・プリント	16.14
17	サイクリング	18.16	スポーツの総数	15.13	パソコンなどの情報処理	16.05
18	楽器の演奏	17.76	観光旅行	14.52	楽器の演奏	15.86
19	人文・社会・自然科学(歴史・経済・数学・生物など)	17.31	サイクリング	13.80	映画館での映画鑑賞	15.78
20	映画館での映画鑑賞	17.25	人文・社会・自然科学(歴史・経済・数学・生物など)	13.79	ウォーキング・軽い体操	15.61





4.3　既存の空間統計手法との比較

これまで、「緯度・経度情報」や「性・年齢階級別人口構成比」などに基づく自治体間の距離から、バリオグラム線形モデルを用いて共分散行列を作成し、当該空間における「空間的自己相関」の有無を確認してきた。一方で、「空間的自己相関」の判定には、モランⅠ統計による検定などが存在する。従来手法は、都道府県の観測値であれば隣接情報が得られるため、本稿の結果とモランⅠ統計の結果を比較することで、本稿手法のパフォーマンスを確認することができる。

モランⅠ統計は、観測地点 i,j 間の重み行列 $W=(w_{ij})$ を隣接行列、

$$w_{ij}=\begin{cases}1 & (i\neq j\text{かつ}i\text{と}j\text{が隣接している場合})\\ 0 & (i=j\text{又は上記以外})\end{cases}$$

と定義し、観測値 $\vec{X}=(X_1,X_2,\cdots,X_N)^T$ に対して

$$I=\frac{N}{\sum_i\sum_j w_{ij}}\times\frac{\sum_i\sum_j w_{ij}(X_i-\bar{X})(X_j-\bar{X})}{\sum_i(X_i-\bar{X})^2} \tag{4.6}$$

で与えられ、$Z_i=(X_i-\bar{\mu})/\sqrt{\overline{\sigma^2}}$, $\bar{\mu}=(\sum_{i=1}^n X_i)/N$, $\overline{\sigma^2}=(\sum_{i=1}^N(X_i-\bar{\mu})^2)/N$, $\vec{Z}=(Z_1,\cdots,Z_N)^T$ とすると

$$I=\frac{1}{\sum_i\sum_j w_{ij}}\times\sum_i\sum_j w_{ij}Z_iZ_j=\frac{\vec{Z}^TW\vec{Z}}{\vec{1}^TW\vec{1}} \tag{4.7}$$

となる。モランⅠ統計は $I\in[-1,1]$ となり 1 に近いほど正の空間的自己相関が強いことを示唆する。

モランⅠ統計は、W による二次形式であり、(適切な正定値条件の下で) W^{-1} によるマハラノビス距離と対応づけて理解できる。一方、本稿の評価指標である対数尤度差は共分散関数Σ（あるいは相関行列 $\mathrm{Corr}(\Sigma)$）に基づくマハラノビス距離の関数として、

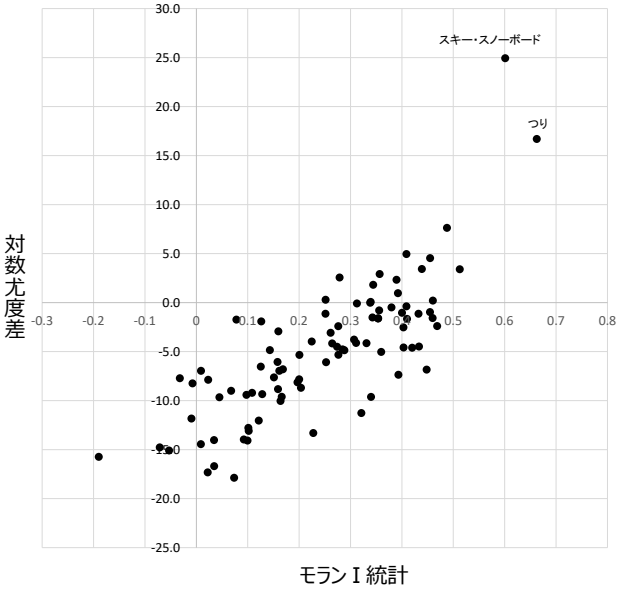
$$L_2-L_6=-\frac{N}{2}\ln\left(\frac{1}{N}\vec{Z}^T\Sigma^{-1}\vec{Z}\right)-\frac{1}{2}\ln(\det(\Sigma)) \tag{4.8}$$

$$L_3-L_6=-\frac{N}{2}\ln\left(\frac{1}{N}\vec{Z}^T\mathrm{Corr}(\Sigma)^{-1}\vec{Z}\right)-\frac{1}{2}\ln\left(\det\left(\mathrm{Corr}(\Sigma)\right)\right) \tag{4.9}$$

と表される。したがって、Σ・$\mathrm{Corr}(\Sigma)$ と W^{-1} の構造の違いが、モランⅠ統計と対数尤度差の相違として現れるが、同一データセットに対して、対数尤度差とモランⅠ統計に一定の相関関係が観察されれば、対数尤度差がモランⅠ統計と類似する効果を有することが示される。

図 12 は、都道府県別の「行動データ」について、横軸に隣接情報を重み行列としたモランⅠ統計、縦軸に前述の Model(2)と Model(6)の対数尤度差をとったプロット図であり、一定の相関関係がみられる。

図 12　モランⅠ統計と対数尤度差の関係



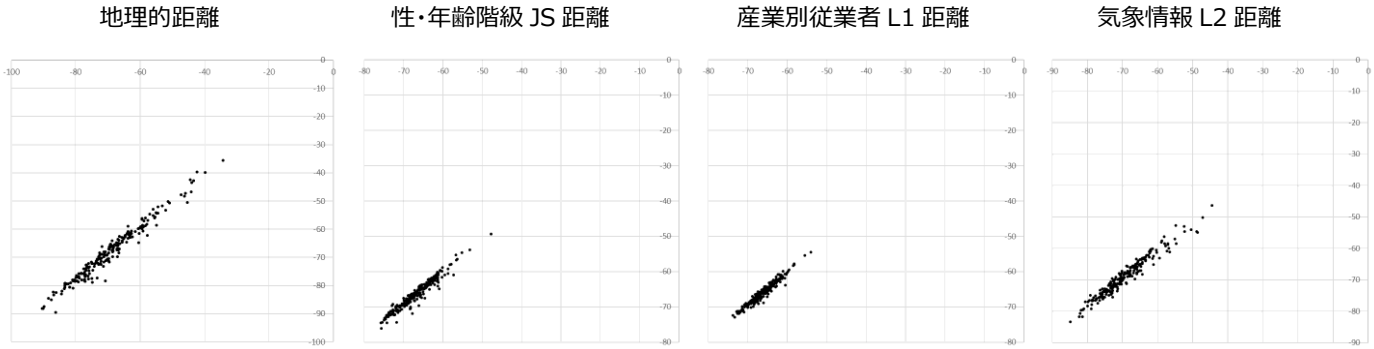
都道府県の場合には隣接関係を明確に定義できるが、県庁所在市を対象とする家計調査では、隣接関係の構築に一定の仮定を要する。また、「性・年齢階級別人口構成比」、「産業別従業者数構成比」、「気象情報」に基づく空間では、都道府県であっても妥当な隣接関係を構築することは容易ではない。距離の逆数やその二乗、距離二乗の負の指数などを重みとする実践的な方法も用いられているが、その根拠を理論的に提示することは難しい。それに対し、「観測値差の分散は観測地点の距離に比例する」という自然な仮定に基づき尤度関数により適合度を評価する本稿の方法は、実務上において一定のパフォーマンスを示す有効な手法と考えられる。

4.4　簡素化と検定統計量

最後に、簡素化及び数値判断の基準となる検定統計量を示す。

家計調査では Model(3)と Model(6)の対数尤度差を用いたが、本稿の目的はモデルの選択そのものではなく、様々な状態・構造から構成される空間における空間的自己相関の検出にあり、そのため、多変量正規分布が仮定できるデータに対し、実測データが空間的自己相関を有すると仮定した場合の発生しやすさと、有さないと仮定した場合の発生のしやすさの対数尤度比という分かりやすい指標を用いている。この数値は、指数変換したときに、空間的自己相関がない場合と比較して、実測データは何倍発生しやすい、という意味のある解釈しやすい数値となる。一方、4.2 における計算は統一モデルの選定のために AIC を計算するなどの煩雑化しており、したがって、4.1 節で示した AIC による統一モデルの選定作業を省略し、より簡潔に対応できるようにすることが望ましい。そこで、相関行列が選択された家計データについて、Model(2)と(3)の対数尤度の関係を確認した。図 13 は、家計調査における Model(2)と Model(3)のプロット図であり、高い連動性を示している。このことから、対数尤度の計算は相関行列を対象とすることなく、Model(2)に統一して統計量を考える。なお、距離関数の選択も省略したい場合、半正定値性が保証される L_2 を用いるのが安全である。

図 13　家計調査における Model(2)と Model(3)のプロット図



さらに、対数尤度差の分布は導出されているものの、ディガンマ関数を用いて表現されるなど、対数尤度差から直接検定を行うことは行政官としてハードルが高い。そこで、対数尤度差のうち、$\vec{Z}$に依存して変動するマハラノビス距離の二乗 $\vec{Z}^T\Sigma^{-1}\vec{Z}$に着目する。

モデル(2)に統一した対数尤度差は(4.8)で表され、$Y=\vec{Z}^T\Sigma^{-1}\vec{Z}$とすると(4.8)式は、Y に対する 1 対 1 対応の単調減少関数であることが分かる。このため対数尤度差の値から直接検定量を求めるのではなく、対数尤度差と 1 対 1 対応する$\vec{Z}^T\Sigma^{-1}\vec{Z}$の値から検定統計量を求めることとする。

具体的には、帰無仮説$\vec{Z}\sim N(\vec{0},I)$の下で、マハラノビス距離の 2 乗$\vec{Z}^T\Sigma^{-1}\vec{Z}$は固有値による加重$\chi^2_{\nu=47}$となり、自由度が比較的大きい場合には正規分布近似によって良好な結果が得られる。そこで、帰無仮説「与えられたΣの下で空間的自己相関が存在しない」を採用し、観測値$\vec{X}=(X_1,X_2,\cdots,X_N)^T$を

$$Z_i=\frac{X_i-\bar{\mu}}{\sqrt{\overline{\sigma^2}}}\quad\left(\bar{\mu}=\frac{1}{N}\sum_{i=1}^{N}X_i\ ,\ \overline{\sigma^2}=\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{\mu})^2\right)$$

で標準化して、$S=(s_{ij})=\Sigma^{-1}$とすると、帰無仮説の下で$E(\vec{Z}^T\Sigma^{-1}\vec{Z})=tr(\Sigma^{-1})=\sum_i s_{ii}$、$Var(\vec{Z}^T\Sigma^{-1}\overrightarrow{Z})=2tr(\Sigma^{-2})=2\sum_i\sum_j s_{ij}^2$が成り立つ。対数尤度差は、マハラノビス距離の二乗が小さいほど大きくなることから、次の標準化統計量

$$Q=-\frac{\vec{Z}^T\Sigma^{-1}\vec{Z}-\sum_i s_{ii}}{\sqrt{2\sum_i\sum_j s_{ij}^2}}\sim N(0,1)\qquad(4.10)$$

として、$Q>1.96$（上位 5%点）であれば、当該Σを構成した空間において空間的自己相関が有意と判断できる。表 6 は、Model(2)及び Model(6)に固定し、AIC に基づいて採択された距離関数を用いた共分散行列によって、データ項目ごとにQを計算し、そのうちQ>1.96 を満たす項目数を集計したものであり、各空間で多くの項目に空間的自己相関の傾向が認められることを示している。

表６ Ｑに基づき空間的自己相関が有意と認められた項目数

（家計データ）

Σの生成方法		上位 5%棄却域該当の項目数
分布情報	距離関数	
緯度･経度情報	L_2	44/212
性･年齢階級別構成比	JS	11/212
産業別従業者数構成比	L_1	2/212
気象情報	L_2	24/212

（行動データ）

Σの生成方法		上位 5%棄却域該当の項目数
分布データ	距離関数	
緯度･経度情報	L_2	8/91
性･年齢階級別構成比	L_2	43/91
産業別従業者数構成比	L_1	40/91

なお、4.3 までは対数尤度差で評価を行い、この場では検定統計量としてマハラノビス距離から標準化検定統計量を提示している。両者の使い分けは、空間的自己相関がない場合に比べて何倍発生しやすいかという観点から計算される 4.3 までの対数尤度差で各品目を網羅的に俯瞰し、個別の品目について、統計有意性を確認したい場合に 4.4 で示す検定統計量を用いることを想定している。

そして、前述のとおり$L_2 - L_6$とQは 1 対 1 の関係を持つため、双方の統計量の順序関係は同じものとなる。

まとめると、正規分布に従うと仮定される自治体別の観測値$\vec{X} = (X_1, X_2, \cdots, X_N)^T$について、地理的空間や性・年齢階級別構成比などで構成される多元的空間における空間的自己相関の有無を、次の手順で判定する。

1）標準化：$\vec{X}$を単純平均と単純な標準偏差で標準化し、$\vec{Z}$を作成する(式(4.5))。

2）共分散行列の構成：各観測点（自治体）の地理的配置、性・年齢階級分布、産業別従業者分布などに基づく距離を用い、式(4.3)及び(4.4)により共分散行列Σを作成する。

3）距離関数を決定する（簡便に行う場合、半正定値性が保証される L_2 距離を採択)。

4）対数尤度差の比較：式(4.8)により空間的自己相関ありモデル(Model(2))と無相関モデル(Model(6))との対数尤度差を算出し、別項目間で比較する。また、対数尤度差を指数変換することで、現状の実測データが無相関の場合と比べて何倍の発生確率があるかを確認する。

5）検定：４）に加え、検定統計量が必要な場合、対数尤度差の結果ではなく、対数尤度差と 1 対 1 に対応する$\vec{Z}^T\Sigma^{-1}\vec{Z}$に着目し、帰無仮説「空間的自己相関はない（$\vec{Z} \sim N(\vec{0}, I)$)」の下で、式(4.10)の検定統計量を計算し、値が 1.96（上位 5%点）を超えるかどうかで有意性を判断する。

以上により、各空間における空間的自己相関の有無に基づき指標の特性を把握し、以降の分析・解釈の参考とする。

5　まとめと今後の課題

本稿では、第 3 節においてワッサースタイン距離を用いて、空間的分布を考慮した新たな集中度指標を提案した。本指標の特性については、①既存のジニ係数や Theil 指数などの集中度指標と比較・整理し、②テストデータにより、距離を組み込んだ評価が直観と整合的であることを示し、③情報通信業の実証分析において、「電気通信に付帯するサービス業」に関して、ワッサースタイン-1 距離に基づく指標により空間的特徴の存在が数値に現れることを例証し、③経済センサス－活動調査（2012 年および 2021 年）の結果を用いて、各産業の差分二変量に対する空間情報を加味した集中度と、空間情報を加味しない集中度で異なる動向を示す事例を提示し、ワッサースタイン-1 距離に基づく集中度指標の特性の一例を示した。

今回は指標の提案及び指標の特性の一面を紹介したものであるが、今後は、提案した指標の特性を生かし経済効果測定や集積経済モデルへの適用などの検討の余地があると考える。

第 4 節においては、公的統計に内在する「空間的自己相関」の特性を解明することを目的とし、地理的距離、人口構成、産業構成、気象条件などの多様な分布に基づき自治体間の多元的距離空間を構築し、

その上で「空間的自己相関」の有無を抽出する方法を検討し、実際に SSDSE などの公的データを用いた試算においては、食料消費支出、行動率に関して、地理的配置・性・年齢別人口構成比・産業別従業者数構成比に基づく多元的距離空間で「空間的自己相関」が相当程度内在していることが確認された。これにより、統計指標がどの空間で自己相関を持つかを事前に把握することが可能となり、分析に資する補助情報を提供し得る。

もっとも、課題も残存する。第一に、距離関数の選択によっては共分散行列が半正定値性を必ずしも満たさない場合がある。第二に、本稿における分析対象は正規分布を仮定したものであるが、現実の統計には正規分布の仮定を満たさない場合も少なくない。これらの場合の対応方法については本稿では言及していない。さらに、第 4 節で十分に適用機会を見出せなかったワッサースタイン距離の活用も検討課題である。ワッサースタイン距離は移動コストの設定により他の距離関数を凌駕する適合度を持つモデルを構築できる可能性も存在していると考えている。

統計行政においては、自治体単位の統計作成や分析は行われているが、「空間」を前提とした集積分析や空間的自己相関の分析はあまり行われていない。しかしながら、地理的配置、年齢階級別人口分布に基づく距離空間、産業別従業者構成比などに基づく多元的距離空間を用い、自己と類似した状況にある自治体との比較により対象指標を分析することを統計行政に普及させることは、公的統計データを用いた EBPM の発展に資するものと期待される。

【謝辞】

本稿について、丁寧な査読をしていただき、多くの有益なコメントをしていただいた 2 名の匿名の査読者、及び丁寧に文章を確認いただいた事務局の担当者名に対して、深く感謝を申し上げる。

参考文献

[1]佐藤竜馬(2023),『最適輸送の理論とアルゴリズム』講談社
[2]瀬谷創・堤盛人(2014),『空間統計学』朝倉書店
[3]内閣官房統計改革推進会議(2017),「統計改革推進会議最終取りまとめ」
[4]中村良平(2008)「都市・地域における経済集積の測度（上)」, 岡山大学経済学会雑誌第 39 巻 4 号, P99-121
[5]Curriero, F.C.(2005) ”On the Use of Non-Euclidean Isotropy in Geostatistics”, Johns Hopkins University, Dept. of Biostatistics Working Papers
[6]Matheron, G.(1963), ”Principles of geostatistics, Economic Geology”, Vol.58, No.8, P1246-1266
[7]Rubner, Y. , Tomasi C. , and Guibas L.J.(2000) “The Earth Mover's Distance as a Metric for Image Retrieval”, International Journal of Computer Vision, Vol. 40, No.2, 99-121,
[8]Schoenberg, I. J.(1937), ”On Certain Metric Spaces Arising from Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space”, Annals of Mathematics, Vol 38, No.4, P787-793.

補論　行和と列和が等しくなること

以下の算式

$$\Sigma = [\sigma_{ij}] = [\sum_{k=1}^{N} \frac{1}{N}\left(\frac{d(\vec{s}_i, \overrightarrow{s_k}) + d(\vec{s}_j, \vec{s}_k) - d(\vec{s}_i, \vec{s}_j)}{2}\right)]$$

について、記述を簡素化するため、$d(\vec{s}_i, \vec{s}_j) = d_{ij}$として$[s_{ij}] = [\sum_{k=1}^{N}(d_{ik} + d_{jk} - d_{ij})]$の行和が等しいことについて証明する。

$$\sum_{k=1}^{N}(d_{ik} + d_{jk} - d_{ij}) = \sum_{k=1}^{N} d_{ik} + \sum_{k=1}^{N} d_{jk} - \sum_{k=1}^{N} d_{ij}$$

$$= \sum_{k=1}^{N} d_{ik} + \sum_{k=1}^{N} d_{jk} - Nd_{ij}$$

i 行の行和は、

$$\sum_{j=1}^{N}\left(\sum_{k=1}^{N} d_{ik} + \sum_{k=1}^{N} d_{jk} - Nd_{ij}\right) = \sum_{j=1}^{N}\sum_{k=1}^{N} d_{ik} + \sum_{j=1}^{N}\sum_{k=1}^{N} d_{jk} - \sum_{j=1}^{N} Nd_{ij}$$

$$= N\sum_{k=1}^{N} d_{ik} + \sum_{j=1}^{N}\sum_{k=1}^{N} d_{jk} - N\sum_{j=1}^{N} d_{ij}$$

ここで、第１項目と第３項目は同じ値になることから

$$= \sum_{j=1}^{N}\sum_{k=1}^{N} d_{jk}$$

この結果は、i に依存していないことが分かるので、任意の i 行の和は同じ値になる。また、S は対称行列であることから、同様に列和でも同じことがいえる。