Department of Policy and Planning Sciences

Discussion Paper Series

No.1391

混合ガウスモデルを用いたデータ駆動型企業分類の提案と検証

(Proposal and Validation of Data-Driven Company Classification Using Gaussian Mixture Model)

by

寺田海渡、鱒 涼稀、竹田俊彦、高野祐一、岡田幸彦
(Kaito TERADA, Ryoki MOTAI, Toshihiko TAKEDA, Yuichi TAKANO and Yukihiko OKADA)

May 13, 2024

UNIVERSITY OF TSUKUBA
Tsukuba, Ibaraki 305-8573
JAPAN

混合ガウスモデルを用いたデータ駆動型企業分類の提案と検証

Proposal and Validation of Data-Driven Company Classification Using Gaussian Mixture Model

筑波大学大学院サービス工学プログラム　寺田海渡
筑波大学大学院社会工学学位プログラム　鐘　涼稀
水戸信用金庫　竹田俊彦
筑波大学システム情報系／人工知能科学センター　高野祐一
筑波大学システム情報系／人工知能科学センター　岡田幸彦

〈論文要旨〉

日本の中小企業は，国の経済成長のために不可欠であり，今後の更なる成長のために中小企業の労働生産性向上が求められている。労働生産性をはじめとする財務指標から企業の状態を適切に評価し，それに基づいた意思決定を行うためには，正確な財務分析が重要である。財務分析の際には，同業他社との財務比率の比較が行われるが，現在用いられている産業分類は，静的であり，その分類に企業が属するかどうかという情報しか与えない。近年では，技術の発展により様々な産業形態の企業が生まれつつある。また，特定の産業にとどまらず，複数の産業特性を持つ企業も存在する。このように，時間経過とともに変化していく企業や市場について，既存の産業分類ではその産業構造を適切に捉えられないことが指摘されている。本研究では，既存の産業分類に加えて，財務比率から得られるデータ駆動型の企業分類を用いることで，既存の産業分類の課題を改善できるのかを検証した。具体的には，労働生産性の増減に関わる財務比率を L1 正則化ロジスティック回帰モデルによって選択し，その財務比率を用いたクラスタリングによって企業分類を得て，これを用いて財務分析と次年度の労働生産性増減予測モデルの構築を行った。この結果，既存の産業分類では不可能であった特定の財務比率の時系列的推移が類似した企業間での比較と，業種の枠組みではサンプルが少ないためにモデルが構築不可能であった企業も含めた重みづけモデルが構築可能となった。

〈キーワード〉

中小企業，地域金融機関，産業分類，機械学習，ソフトクラスタリング，アカウンティング・インフォマティクス

1. 問題設定

1.1. 社会的背景

中小企業は日本の企業数の 99.7%を占め，日本全体の雇用の約 70%を抱えている（中小企業庁 2022)。これらの企業は，地域経済の活性化に寄与し，多様な雇用機会を提供しており，国の経済成長に不可欠な存在である。しかし，これらの中小企業は，大規模企業と比較して労働生産性が半分以下であることが指摘されている。中小企業庁（2019）は,「我が国全体の付加価値額を引き上げるためには，大企業だけでなく，中小企業の労働生産性も向上させることが重要であるといえる」と述べている。

企業は事業活動により生み出した「付加価値」を基に，人件費などの諸費用を賄い，利益を得ている。そして，付加価値額を労働投入量で除した「労働投入量 1 単位当たり付加価値額」のことを労働生産性という。労働投入量としては，従業員数や，従業員数に労働時間を乗算したものなど用途に応じて様々なものが用いられる[1]。労働生産性は労働の効率性を測る尺度であり，労働生産性が高いほど，投入された労働力が効率的に利用されていると言える。将来的に人口減少が見込まれる中，我が国経済のさらなる成長のためには，中小企業の労働生産性を高めることが重要である（中小企業庁 2022)。

このような指標によって表される企業の状態を適切に評価し，それに基づいて意思決定を行うためには，正確な財務分析が不可欠である。財務分析は，投資家や金融機関の意思決定，経営状態の把握，融資の可否判断など，多くの重要なビジネスプロセスに影響を及ぼす。また，企業の財務状況や業績，成長潜在力を把握するうえで，財務分析は重要な手段となる。これには，財務諸表の詳細な分析や，財務比率の計算，業界平均との比較などが含まれる。特に，業界平均との比較は，企業が自身の業績を業界全体の状況と照らし合わせるために重要である。この比較には，効果的な産業分類が不可欠となる。

しかし，現在広く用いられている産業分類にはいくつかの課題が存在していることが知られている。例えば，海外で広く用いられている SIC コード[2]に基づいた産業分類に関していうと，複数の産業特性を持つコングロマリット企業に対してこの分類は適していないことが指摘されている（Amit and Livnat 1990)。さらに，Hoberg and Phillips（2016）は，既存の産業分類は静的であり，変化する産業構造を捉えられないと述べている。企業や市場は時間の経過とともに進化しているため，産業分類はそのダイナミックな側面を捉える必要がある（Fang et al. 2013)。 また，Kile and Phillips（2009）は，既存の産業分類がハイテク・サービス部門の多様な活動を十分にカバーしていない可能性を指摘しており，このような新興の産業に対して対応できない点も課題である。

日本で一般的に用いられている日本標準産業分類においても，上記と同様の課題が考えられる。既存の産業分類方法では，企業の多様性や産業の動的な特性を適切に反映できていない。これは，収益性や生産性などの指標を比較分析することによる企業価値の評価と意思決定にとって大きな問題となる。

1.2. 学術的背景

[1] 日本生産性本部 HP『生産性とは』（https://www.jpc-net.jp/movement/productivity.html，閲覧日 2023 年 12 月 29 日）

[2] Standard industrial classification code の略。1937 年に米国で制定された，政府機関が産業を分類するために使用する 4 桁のコードのこと。

財務分析では，様々な財務比率を用いて収益性，安全性，成長性など多方面から企業の評価を行う。この分析プロセスにおいて，産業別の比較基準は重要な役割を果たす。なぜなら，企業の財務状態や業績を正確に理解し，評価するためには，それぞれの業種の特徴や動向に即した分析が求められるからである。しかし，この産業ごとの比較分析の有効性は，使用される産業分類の性質と直接的に関連している。

財務分析における企業分類の重要性にも関わらず，既存の分類方法が直面している課題は無視できない。Mensah（1984）や Clarke（1989）の研究は，産業別モデルの有用性を示しつつも，分類の精度に依存するその限界を浮き彫りにしている。特に，Clarke（1989）は，SIC コードに基づく分類が，より細かいレベルでの有用性に欠けることを指摘しており，これは産業分類が市場の実態を反映していないことを示唆している。

さらに，Bhojraj et al.（2003）は，異なる産業分類基準を比較し，会計研究における産業識別力の不足を指摘している。Fairfield et al.（2009）の研究は，成長性予測における産業別モデルの有用性を示しながらも，収益性予測においてはその限界を示している。これらの研究結果は，産業分類が一貫して財務分析のニーズを満たしていないことを示唆している。

財務分析の伝統的なアプローチでは，産業分類に基づいて企業を比較することが一般的であるが，最近の研究はこの枠組みを超えた方法を探求している。Chen et al.（2022）は，機械学習技術を活用して，産業分類に依存せずに財務データから企業の収益性を予測する手法を提案している。この研究は，XBRL[3]から得られる膨大で詳細な財務諸表をデータとして機械学習モデルを構築し，先行研究よりも高い予測精度を得ている。このようなデータ駆動型のアプローチは，市場の変動や産業の新たな動向に迅速に対応できる柔軟性を持ち，従来の産業分類に基づく方法では捉えきれない企業間の微妙な違いを明らかにすることができる。

このアプローチは，産業分類が提供する一般的な枠組みを超えて，個々の企業の独自性や市場環境の変化に即した分析を行うことの重要性を示唆している。従来の産業分類に頼ることなく，個別企業の財務データに基づいて直接分析を行うことで，より詳細で個別化された企業評価が可能になり，投資家や経営者にとってより有益な情報を提供することが期待される。

このような最新の研究動向は，財務分析における産業分類の役割を再考するきっかけを提供している。産業分類が持つ伝統的な枠組みの限界を認識し，データ駆動で適応性の高い分析手法の開発が求められている。これにより，財務分析は，企業の真の財務状態をより正確に反映し，投資や経営の意思決定においてより有効な支援を提供することができるようになると考えられる。

1.3. 本研究の目的

本研究の目的は，既存の産業分類だけでなく，分析目的に即したデータ駆動型の企業分類から得られる情報を分析に用いることにより，先行研究において指摘されている既存の産業分類の課題を改善できるのかを検証することである。本研究は，水戸信用金庫との共同研究であり，共同研究契約および秘密保持契約を厳守して実施された。水戸信用金庫が保有する取引先事業者の財務諸表データは，取引先事業者の機

[3] 各種事業報告用の情報を作成・流通・利用できるように標準化された XML ベースのコンピュータ言語のこと。XBRL Japan『XBRL とは』（https://www.xbrl.or.jp/modules/pico1/index.php?content_id=9，閲覧日 2024 年 1 月 10 日）

密データである。そのため，当該データについて企業が特定できないかたちで秘匿化加工を行い，かつ，秘匿化加工された分析用データは外部に持ち出さず，水戸信用金庫本店内の定められた部屋でのみデータ解析とモデルの構築を行った。信用金庫には取引先中小企業の生産性向上のための身近な相談相手としての役割が求められている（信金中央金庫，2018）。それゆえ，取引先中小企業の生産性について分析し，向上の要因を探求することは，地域経済を活性化させるうえで信用金庫にとっても重要である。本研究では，中小・地域金融機関の実際の取引先企業の財務諸表データを用いて，来年度の労働生産性の方向性に影響を与える財務指標を特定し，その指標を基に企業分類を行う。このようにして得られた企業分類は，企業間比較によって労働生産性の向上を図るための戦略を考える際に有効な分析基盤となることが期待される。この新たな企業分類アプローチにより，本研究でいうところの「次年度の労働生産性の方向性」のような分析者が注目する特定の目的に即したより深い分析が可能になると考えられる。さらに，この手法は伝統的な産業分類では見過ごされがちである，異なる産業分類に属する企業同士についても，横断的な視点を提供する。結果として，このデータ駆動型の企業分類は，労働生産性の方向性を探求する分析において，既存の産業分類よりも柔軟で詳細な洞察を提供する可能性がある。

本研究では，提案手法によって得られた新分類に関して解釈を行い，分類を利用した財務分析の方法論について提案する。また，新分類を用いて「次年度の労働生産性の方向性」を予測するモデルを構築し，既存の方法論や産業分類モデルとの精度の比較も行う。

本論文の残りの部分は以下のように構成されている。まず，第 2 節で先行研究についてのレビューを行う。次に，第 3 節で研究方法について述べ，第 4 節でその結果を示す。そして第 5 節で結果に対する考察を行い，第 6 節で研究の結果と今後の課題についてまとめる。

2. 先行研究

2.1. 会計データと伝統的な産業分類の関連

会計データを用いて，統計的なグルーピングを行った研究として，Gupta and Huefner（1972）がある。Gupta and Huefner（1972）は，財務比率について，ある業種はほかの業種に比べて特定の財務比率が高く，ある業種は中程度であるというように，財務比率で業種の属性を説明できるのではないかと考えた。そこで，The Statistics of Income[4]のデータより，各業種における財務比率の代表値を計算し，その比率によって業種の属性を説明できるのかというマクロレベルでの財務比率の有用性を検証した。具体的には，産業経済学の専門家の知識をもとにいくつかの財務比率によって産業を分類し，その結果と財務比率による業種クラスタリングの結果を比較することで，財務比率に産業特性を説明する能力があるのかを確かめた。検証の結果，業界の資産の専門性が表れる固定資産回転率が業種間で最も顕著な差異を示し，より集計的な比率である流動資産回転率や総資産回転率は業種間の差異を表す上で一般的には役に立たないことがわかった。この結果より，財務比率は，それ自体では定量化が極めて困難な産業特性の集合に対応し，産業特性基準の代用品として役立つ可能性が示唆された。そして，この論文では，会計データのグルーピングにおける分類基準をどのようにするかについて更なる研究が必要とされている。

会計データによる分類に限らず，企業を分類する際の基準に関しては多くの研究が行われている。伝統

[4] IRS，SOI Tax Stats – Statistics of Income（https://www.irs.gov/statistics/soi-tax-stats-statistics-of-income，閲覧日 2023 年 12 月 29 日）

的に使用されてきた企業分類として，主要な製品や製造工程をもとに企業を集約した標準産業分類（SIC）がある。しかし，製品の多様化や，サービスの重要性の増大，技術や企業構成の変化により，SICの有用性が疑問視されている（Clarke 1989）。Fama and French（1997）は，企業の 4 桁の SIC コードから，共通のリスク特性を持つものを 48 の産業グループに再編成した。この分類は大きな影響力を持ち，様々な分野の学術的研究で用いられている（Brennan et al. 2004; Chan et al. 2004; Flannery and Rangan 2006）。

SIC と異なるアプローチで開発された産業分類として，世界産業分類（GICS）がある。GICS は S&P とMSCI（モルガン・スタンレー・キャピタル・インターナショナル）が共同開発した産業分類であり，企業の事業特性だけでなく，何がその企業の主要事業であるかについての投資家の認識に関する情報にも基づいた投資家向けの産業分類である。Chen et al.（2007）は，SIC に基づいて Fama and French（1997）が再編した FF 産業分類と，GICS 分類を比較することにより，産業分類に対する学術的分野と実務家のアプローチ間の格差の解消を試みた。また，上記の分類を，リターン相関に基づいた統計的クラスタ分析によって得られた分類とも比較することにより，銘柄分類における GICS 分類と FF 分類の有用性を示した。

2.2. 新たな視点に基づく企業分類の開発

先行研究で指摘されている伝統的な産業分類の課題に対し，新たな視点に基づいた企業分類によって解決を試みる研究が行われている。

Fan and Lang（2000）は，既存の産業分類では異なる産業間の関連性が不明であるという課題に対して，産業関連表を用いて関連性の指標を定義することにより，産業間の関連を表そうとした。また，Dalziel（2007）は，モジュール性理論を活用し，システムベースの産業分類を提案した。産業を表す垂直方向での分類と，その産業において統合的な役割をするのか補完的な役割をするのかを表す水平方向の分類の 2 方向の企業構造を構築することで，既存の産業分類よりも複雑な分析が可能になることを主張した。近年では，製品説明からテキスト分析の技術を用いて産業分類を行ったもの（Hoberg and Phillips 2016）や，ウェブでの同時検索に基づいて企業間の関連性を説明したもの （Lee et al. 2015）など，革新的な方法が提案されている。

さらに，XBRL の導入によりデジタル化された大規模な財務諸表を効率的に扱うことが可能となった。これに伴い，財務データから企業の分類を得ようとする研究も行われている。Chong and Zhu（2012）は，企業と企業が使用する GAAP タクソノミ[5]の要素を二部構成のソーシャルネットワークとしてモデル化し，それに対してスペクトルクラスタリングを適用することにより，企業のクラスタを特定するために財務データを利用することの実現可能性を示した。Yang et al.（2019）は，グラフ類似性アルゴリズムとスペクトルクラスタリングを組み合わせて，財務諸表構造に基づく産業分類を提案した。この研究において提案された産業分類法による分類は，既存の産業分類よりもグループ内の財務比率分散が小さくなり，企業の事業性の同質性を識別するうえで既存の産業分類を凌駕することを実証した。財務諸表は定期的に作成されるため，このような財務諸表ベースのアプローチは業界のダイナミクスに関する洞察をもたらす可

[5] GAAP（Generally accepted accounting principles）は，財務諸表の作成にあたり，その基準となる会計原則である。タクソノミとは，情報やデータなどを階層構造で整理したものを指す言葉であり，例えば「借方・貸方」や「勘定科目」のような属性を定義するためのものである。金融庁『EDINET タクソノミの概要説明』（https://www.fsa.go.jp/search/20130821/1b_1.pdf，閲覧日 2024 年 1 月 10 日）

能性がある。

上に述べたような先行研究では，一つの企業に対して一つの産業分類への割り当てを前提としている。このような分類に対して，Fang et al.（2013）は次のように述べている。「既存の企業分類は，2 つの企業が同じ業界にあるか，異なる業界にあるかという二項関係を前提としており，類似性の程度を測定していない。同じ業界内の企業間の類似度は千差万別であり，最も類似している企業を比較対象として選択したい。効果的な業界分類アプローチは，業界を特定できるだけでなく，業界内の差異を測定できる，つまり業界内の異質性を捕捉できるものではなくてはならない。」この考えのもと，Fang et al.（2013）では，10K フォームの文章に対してソフトクラスタリング手法である LDA（Latent Dirichlet Allocation）[6]を適用して企業分類を行った。LDA は各グループへの割り当てをハードに行うのではなく，どの程度そのグループに属するのかという確率によってソフトに行うため，その確率分布からハードな分類よりも詳細な企業間類似度を提供する。このほかにも，Fuzzy クラスタリングによってソフトな企業分類を試みた研究（Lenard et al. 2000）も存在する。

以上のような先行研究を踏まえて，本研究では中小企業の財務諸表から計算された，「次年度の労働生産性の方向性」に関連する財務比率を基に，ハードクラスタリングとソフトクラスタリングによる企業分類を行い，両者を比較する。なお，ハードクラスタリングには様々な分野で広く用いられている k-means を利用し，ソフトクラスタリングには実装面や評価指標の観点から，確率密度を基にした手法である混合ガウスモデルを利用する。

2.3. 財務諸表を用いた予測

財務諸表を用いて，企業の将来の収益性を予測しようとした古典的な研究として，Ou and Penman（1989）がある。Ou and Penman（1989）は，財務諸表をもとに将来の収益性を予測するロジスティック回帰モデルを構築し，その結果に基づいた投資戦略のリターンについて分析を行った。これによって，財務諸表を分析することで，株価には反映されていない企業の本質的価値が発見できることを明らかにしようとした。この研究の結果は，財務諸表には企業価値を把握するうえで重要な情報が含まれている可能性が示唆されるものであった。

Ou and Penman（1989）は，モデルの係数の推定に十分なデータがある場合，業種別モデルや企業別モデルによってその精度が改善する可能性があると述べている。しかし，学術的背景でも述べたように，既存の産業分類を用いた業種別モデルの有用性については，その識別力に疑問が投げかけられている（Clarke 1989; Bhojraj et al. 2003; Fairfield et al. 2009）。

近年では，機械学習モデルの発展により，大規模なデータを用いた分析が行われている。Chen et al.（2022）は XBRL から得られる大規模な財務諸表データを用いて，将来の収益性を予測する機械学習モデルを構築し，Ou and Penman（1989）よりも高い予測精度を達成した。この結果より，大規模データと機械学習モデルによる予測は会計データに対しても有効であることがうかがえる。ほかにも，中小企業の財務諸表データと機械学習モデルを用いて，長期借入金の増減を予測した研究も存在する（鱒ほか 2022; 鱒

[6] 潜在的ディリクレ配分法とも呼ばれる，トピックモデルの一種。文書中の単語がどの潜在トピックによって生成されたかを示す潜在変数を導入することで，文書に含まれるトピックを推定する。

ほか 2023)。

本研究で企業分類のために用いる混合ガウスモデルは，訓練データを用いて分布の推定を行った後，未知のデータに対して分布への当てはめが可能となる。よって，学習に用いていない未知のデータに対しての精度を見ることでモデルの汎化性能を評価する機械学習の方法論においても，この分類は適用可能である。そこで，本研究では次年度の労働生産性の方向性を目的変数として，①すべての企業のデータを用いたモデル，②業種別の企業データを用いたモデル，③提案手法による分類から得られる情報を用いたモデルを比較することにより，将来の労働生産性予測における既存の産業分類の有用性と提案手法による分類の有用性についても検証する。なお，それぞれの実験において，近年その高い精度から様々な分野で利用されている Light GBM による予測を行う。

3. 研究方法

3.1. 使用データ

本研究で用いるデータは，共同研究先である水戸信用金庫が保有する取引先事業者の財務諸表データである。使用するデータの期間は 2016 年 1 月から 2023 年 3 月までであり，サンプル数の合計は 41767 件である。決算年度ごとのサンプル数と，財務諸表データが含む項目の内訳については表 1 に示すとおりである。ただし，t 年度とは，t 年 4 月 1 日から t+1 年 3 月 31 日までの期間を意味する。そのため，2015 年度のサンプル数はほかの年度と比較して少なくなっている。

表 1　年度ごとのサンプル数と財務諸表データ項目の内訳

決算年度	個数
2015	1303
2016	4909
2017	5337
2018	5860
2019	6233
2020	6469
2021	6305
2022	5351
計	41767

項目	個数
属性（顧客番号等）	16
B/S（貸借対照表）	95
P/L（損益計算書）	59
CF（キャッシュフロー計算書）	40
変動計算書	36
財務比率	87
経常収支比率	22
脚注 （従業員数等）	22
計	377

財務諸表データのうち B/S（貸借対照表），P/L（損益計算書），CF（キャッシュフロー計算書）の項目は，中小・地域金融機関において電子データとして保管されている可能性が最も高く，信頼性の高い情報である（鐘ほか 2022)。この考えのもと，本研究では，表 1 に示した財務諸表データ項目のうち，B/S（貸借対照表），P/L（損益計算書），CF（キャッシュフロー計算書）と，それらを用いて計算が可能である財務比率の計 281 項目を説明変数の候補として用いる。

また，今回使用するデータには事業者ごとに業種が登録されている。業種の区分には大分類と小分類が

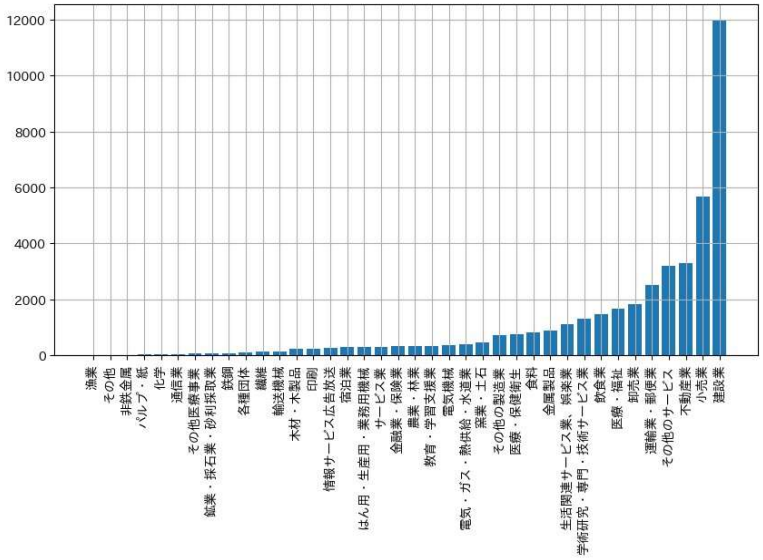
存在し，今回は，業種として大分類を使用する。業種ごとのサンプル数は表 2 に，サンプル数の分布を図 1 に示す。

表 2　業種ごとのサンプル数

業種（大分類）	サンプル数
建設業	11951
学術研究・専門・技術サービス業	1304
その他の製造業	720
その他のサービス	3192
医療・保健衛生	764
小売業	5675
繊維	133
飲食業	1479
医療・福祉	1671
運輸業・郵便業	2506
不動産業	3289
生活関連サービス業,娯楽業	1113
教育・学習支援業	345
金融業・保険業	327
卸売業	1826
窯業・土石	461
食料	831
印刷	233
宿泊業	290
各種団体	98
情報サービス広告放送	253
電気機械	348
はん用・生産用・業務用機械	304
その他医療事業	62
金属製品	881
輸送機械	135
木材・木製品	221
サービス業	304
農業・林業	327
電気・ガス・熱供給・水道業	406
鉄鋼	76
通信業	52
漁業	5
パルプ・紙	45
鉱業・採石業・砂利採取業	73
化学	49
非鉄金属	11
その他[7]	7

[7] その他は，業種コードに一致する業種が存在しなかったものである．

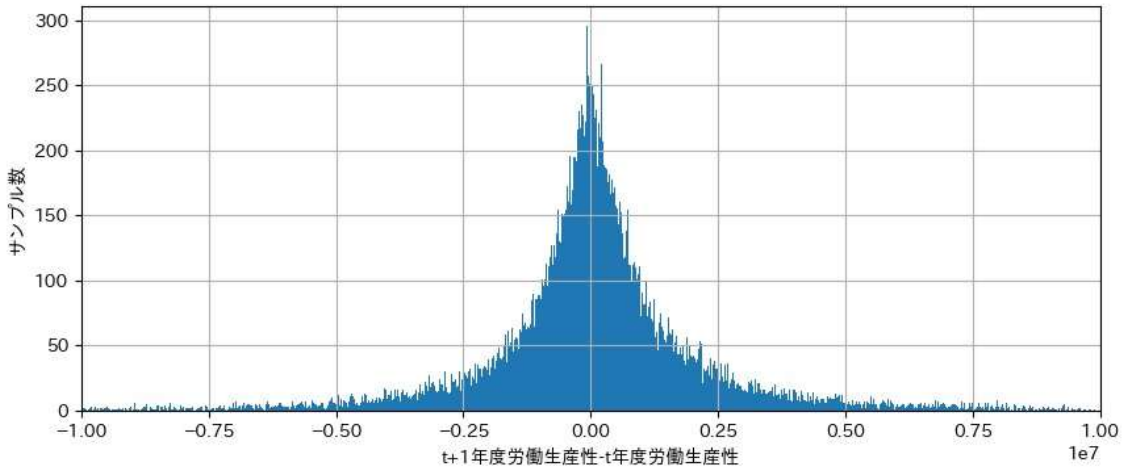
図 1　業種ごとのサンプル数分布



3.2.　前処理

本研究では，中小・地域金融機関の取引先事業者の財務諸表データを用いて，企業の労働生産性向上のための分析に有用な企業分類を得ることが目的である。そこで，目的変数を「t+1 年度の労働生産性－t 年度の労働生産性」が 0 以上である事業者が 1，0 未満である事業者を 0 とする二値の変数とし，これを t 年度の財務諸表データのうち財務比率項目を説明変数とする分類モデルで予測するうえで，説明力があるとされる 10 個の財務比率をもとに企業分類を行う。また，予測面での企業分類の有用性検証では，t 年度の財務諸表データのうち 3.1 節で述べた項目を説明変数の候補とする。なお，目的変数を「t+1 年度の労働生産性－t 年度の労働生産性」や「t+1 年度の労働生産性/t 年度の労働生産性」としてモデルを構築することも可能であるが，前者は事業者の規模によってスケールが大きく異なってしまう点，後者は労働生産性が負の値をとり得る指標である点から，本研究では二値で表される労働生産性の方向性を目的変数とした。参考として二値に変換する前の目的変数の分布を図 2 に示す。

図 2　「t+1 年度の労働生産性-t 年度の労働生産性」の分布



注）横軸のスケールが 1e7 であることに注意されたい。-10000000 から 10000000 の間を 1000 分割したヒストグラムである。基本統計量は，最小値 -262862600，25%点 -774610.9，中央値 30239.50，75%点 917263.0，最大値 200012400，平均値 146647.3，標準偏差 6269388 である。

まず，目的変数を作成するために，事業者の各年度財務諸表データに対して労働生産性を算出する必要がある。労働生産性は「労働投入量 1 単位当たりの付加価値額」として表され，労働者 1 人あたり，あるいは労働時間 1 時間当たりでどれだけ成果を生み出したかを示すものである。本研究では，利用できるデータの観点から，各事業者の期末従業員数を労働投入量とする。また，付加価値額については，売上高から売上原価を差し引いたものによって近似したものを用いる。つまり，本研究では「(売上高－売上原価）/期末従業員数」で労働生産性を計算し，これをもとに目的変数を作成する。目的変数の作成にあたって，売上高，売上原価，従業員数のいずれかが欠損しているサンプルを除外した。これによってサンプル数は 41767 から 36244 となった。本研究では，t 年度の財務諸表データから，t 年度から t+1 年度にかけての労働生産性の方向性を予測するモデルを構築するため，モデル構築のために連続した 2 年度分の財務諸表データが必要である。したがって，連続した 2 年度分の財務諸表データをもたない事業者を除外した。以上の操作によって，年度ごとのサンプル数は表 3 のようになった。また，各年度における前処理後の業種別サンプル数の分布は付録 A に記載する。

表 3　年度ごとのサンプル数（前処理前後）

決算年度	個数（前処理前）	個数（前処理後）	除外数
2015	1303	1048	255
2016	4909	4100	809
2017	5337	4448	889
2018	5860	4813	1047
2019	6233	5182	1051
2020	6469	5319	1150
2021	6305	4525	1780
2022	5351	0	5351
計	41767	29435	12332

また，説明変数に対する前処理として，次のような操作を行った。

1. すべての値が欠損している説明変数を削除
2. 値が無限大もしくは負の無限大に発散しているデータを，同項目内の最大値もしくは最小値で補完

以上のような操作の結果，本研究に用いる説明変数の内訳は表 4　説明変数の内訳（前処理後）表 4 のようになった。

表 4　説明変数の内訳（前処理後）

項目	個数
B/S	80
P/L	45
CF	39
財務比率	76
t 年度労働生産性	1
計	241

3.3.　全体概要

目的でも述べた通り，本研究では，得られた新分類に関して「①財務分析」と，「②予測モデルの構築」という大きく分けて 2 つの利用方法を想定する。ここでは，提案する企業分類を得る方法と，それぞれの利用方法における新分類の有用性の検証方法について概要を示す。なお，以下での目的変数とは前節でも述べた通り，次年度の労働生産性の方向性を指す。

まず，提案する企業分類を得る方法について述べる。

1. 取引先事業者の財務諸表データにおける財務比率項目内の説明変数から，L1 正則化ロジスティック回帰モデルを用いて，目的変数の予測に対して説明力のある財務比率を 10 個選択する。

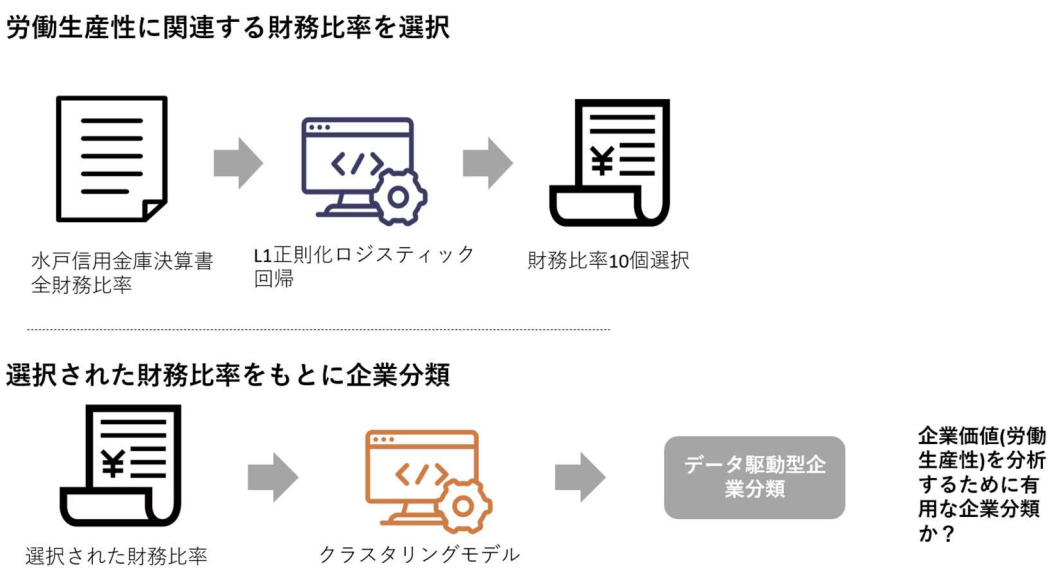
2. 選択した財務比率をもとに，k-means および GMM で取引先事業者の財務諸表データをクラスタリングする。

次に，それぞれの検証方法について概要を述べる。

① 財務分析における新分類の有用性検証（図 3）

取引先事業者の全年度財務諸表データを用いて上記の方法で企業分類を行う。その後，得られた各分類に対して解釈を行い，これを利用した企業間比較の方法論を提案する。そして，実際の中小・地域金融機関の取引先事業者の財務諸表データを用いて財務分析を行い，有用性を検証する。

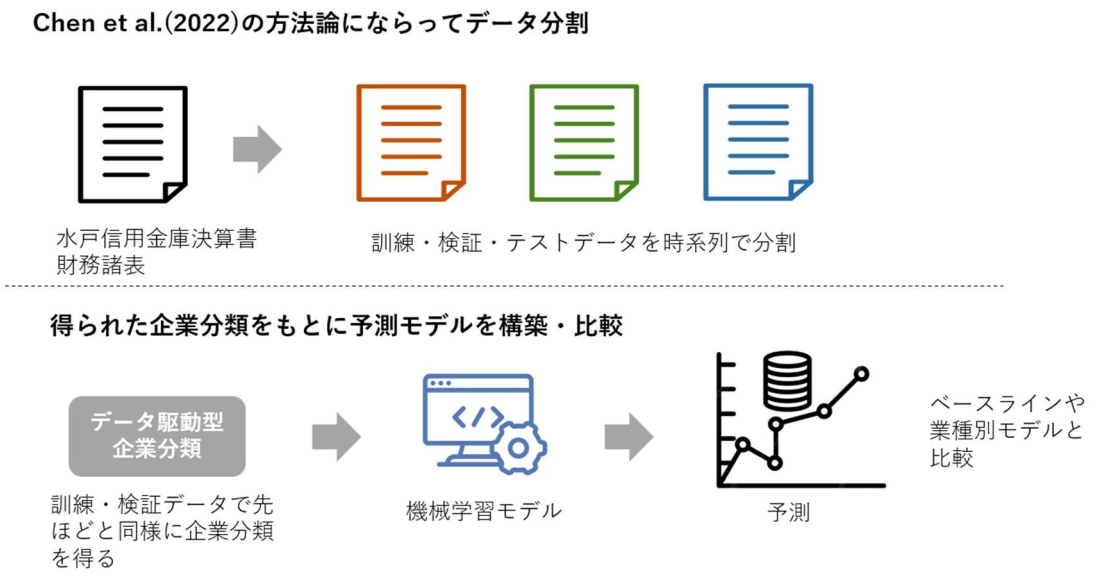
図 3 財務分析における新分類の有用性検証概要図



② 予測モデルの構築における新分類の有用性検証（図 4）

まず，Chen et al.（2022）にならって，取引先事業者の財務諸表データを年度ごとに訓練データ，検証データ，テストデータに分割する。その後，訓練データと検証データに相当する年度の財務諸表データを用いて上記の方法で企業分類を行う。訓練データと検証データを用いて学習させた GMM でテストデータに対しても企業分類を行い，それを用いて次年度の労働生産性の方向性を予測するモデルを構築する。そして，予測精度の結果を，全企業のデータをそのまま用いたモデルと既存の産業分類を用いたモデルと比較することで有用性を検証する。

図 4　予測モデルの構築における新分類の有用性検証概要図



3.4.　L1 正則化ロジスティック回帰による変数選択

本研究では，「将来の労働生産性向上のための分析」という目的に対して有用な企業分類を得るため，1 年後の労働生産性の方向性に関係していると考えられる財務比率を用いて企業の分類を行う。目的変数に関連する変数選択を行った研究として Ou and Penman（1989）がある。Ou and Penman（1989）では，会計の教科書で一般的に用いられる 68 個の財務比率から，ステップワイズロジット法によって 1 年後の収益の増減の予測に関係している財務比率を選択した。しかし，ステップワイズ法による変数選択には，変数間に多重共線性の問題がある場合に必ずしも安定な推定量が得られるわけではないという問題がある（Bangchang 2015; Yanke et al. 2022; Leiby and Ahner 2023）。

このような問題を踏まえ，今回は，変数選択のために L1 正則化ロジスティック回帰（Tibshirani 1983）を用いる。L1 正則化ロジスティック回帰は数値データに対して線形回帰を行う手法の一つである。L1 正則化ロジスティック回帰ではモデルの過学習を抑えるため，誤差関数に正則化項として L1 ノルムを組み込み，この正則化項による制約を満たす条件のもとで誤差関数を最小化する。例えば，目的変数yとm次元の説明変数ベクトル$\boldsymbol{x}$に関するn組のデータセット$D = \{\boldsymbol{X}, \boldsymbol{y}\} = \{(\boldsymbol{x_1}, y_1), \dots, (\boldsymbol{x_n}, y_n)\}$が与えられたとき，$\boldsymbol{x_i}$に対する予測値$\hat{y_i}$は式 1）で与えられる。

$$\hat{\boldsymbol{y_i}} = \boldsymbol{\beta_0} + \boldsymbol{x_i^T \beta} \tag{1}$$

ここで，β_0は切片，$\boldsymbol{\beta} = (\beta_1, \dots, \beta_m)$は回帰係数ベクトルである。L1 正則化ロジスティック回帰では，次の式 2 を解くことを考える。

$$\min_{\beta_0, \beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \boldsymbol{x_i^T \beta} \right)^2 \right\} \; subject\ to \sum_{j=1}^{m} |\beta_j| \leq t \tag{2}$$

ここで，tは正則化の度合いを決定するパラメータである。この正則化の作用により，変数間に多重共線性がある場合でも従来の線形回帰より安定して係数の推定を行うことができることが知られている。ま

た，L1 正則化ロジスティック回帰では，L1 ノルムによる制約境界の形状上，最適解が係数の軸上で得られることが多いため，モデルの係数が 0 になりやすく，変数選択に用いることができる。

なお，本研究において選択する財務比率の数は，クラスタリングにおいて結果の解釈を行う関係から，クラスタの解釈が容易である範囲内の 10 個とした。

3.5. k-means による企業分類

クラスタリングの手法において最もよく知られた手法として k-means がある。k-means は教師なし学習の非階層型クラスタリング手法の一つであり，シンプルなアルゴリズムと計算効率の高さから，多くの場面で用いられている。本研究では，L1 正則化ロジスティック回帰によって得られた 10 個の財務比率をもとに企業分類を行う方法の一つとして，この k-means を用いる。k-means によるデータの分割のアルゴリズムは以下のとおりである。

k-means によって$\boldsymbol{x_i}$で表される n 個のデータ点を，中心が$\boldsymbol{\mu_k}$で表されるm個のクラスタに分割するとき，式 3 のような目的関数の最小化を行う。

$$J = \sum_{i=1}^{n}\sum_{k=1}^{m} q_{ik}|\boldsymbol{x}_i - \boldsymbol{\mu}_k|^2 \tag{3}$$

ここで，q_{ik}は，データ点x_iがクラスタkに所属するとき 1，そうでないときに 0 となるような関数である。目的関数の最小化は以下のようなステップで行われる。

1. 初期化：各データをランダムにクラスタに割り当てる。割り当てた各データの平均をクラスタの中心$\boldsymbol{\mu_k}$（$k = 1,2,\dots,m$）とする。
2. q_{ik}について目的変数を最小化：求めたμ_kを固定し，q_{ik}についてJを最小化する。
3. μ_{ik}について目的変数を最小化：2. で求めたq_{ik}を固定し，μ_kについてJを最小化する。
4. 2. 3. を収束するまで繰り返す。

このようなアルゴリズムでデータのクラスタリングを行うのだが，k-means は 1. の初期化の結果によって結果が変わることが知られており，クラスタの数も分析者が決定する必要がある。そこで，今回はクラスタリングの際に異なる 100 パターンの初期値を試し，次に示す 3 つの評価指標を総合的に考慮して，もっとも良い分割が行えていると考えられる初期値とクラスタ数を採用する。

- シルエット係数

  クラスタリングの性能を評価する指標の 1 つであり，クラスタ内のデータ同士は近く，異なるクラスタとの距離は遠いほど良い分類であるとする指標である。n個のサンプルが与えられたとき，サンプルiが属するクラスタの他のサンプルまでの平均距離を$a(i)$，サンプルiに最も近い別のクラスタに属するサンプルまでの平均距離を$b(i)$とすると，シルエット係数は式 4 のように計算される。

$$\bar{S} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}\right) \tag{4}$$

  シルエット係数は$[-1,1]$の範囲となり，値が 1 に近いほど良い分類とされる。また，値が負になると所属クラスタの判別が間違っている可能性がある。

- Calinski Harabasz 基準

  クラスタリングの性能を評価する指標の 1 つであり，クラスタ内のデータの分散が小さく，クラス

タ間の分散が大きいほど良い分類であるとする指標である。

クラスタ数がk，全サンプル数がnであるとき，Calinski Harbasz 基準は，クラスタ内分散を$WCSS$（$Within-Cluster\ Sum\ of\ Squares$），クラスタ間分散を$BCSS$（Between-Cluster Sum of Squares）として式 5 のように計算される。

$$CH = \frac{\mathrm{BCSS} / (\mathrm{k}-1)}{\mathrm{WCSS} / (\mathrm{n}-\mathrm{k})} \tag{5}$$

この指標は，値が大きいほど良い分類とする指標である。

- Davies Bouldin 基準

クラスタリングの性能を評価する指標の 1 つであり，クラスタ内距離とクラスタ間距離の比に基づいた指標である。

クラスタi内のデータ間距離の平均をS_i，クラスタiとクラスタjの間の距離を$M_{i,j}$としたとき，2 つの異なるクラスタに対して式 6 のように$R_{i,j}$が計算される。

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \tag{6}$$

これは，2 つのクラスタに注目してどれだけうまく分類が行えているかを測る指標である。そして，クラスタiに関して，$R_{i,j}$が最も大きいものをD_iとする（式 7）。

$$D_i = \max_{j \neq i} R_{i,j} \tag{7}$$

最後に，k個のクラスタに対してD_iの平均をとることで，Davies Bouldin 基準は計算される（式 8）。

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_{i,j} \tag{8}$$

この指標は値が小さいほど良い分類とする指標である。

3.6. GMM による企業分類

本研究では，各サンプルを 1 つのクラスタへ割り当てる k-means に加えて，各サンプルに対して全クラスタへの所属確率を提供する GMM（混合ガウスモデル）による企業分類も行う。k-means のような手法がハードクラスタリングと呼ばれるのに対し，サンプルに対して複数クラスタへの所属を許容する GMM のようなクラスタリング手法はソフトクラスタリングと呼ばれる。既存の産業分類に対する課題でも挙げられているように，企業の多様化が進む現代では，ハードな分類では説明しきれない企業が存在する。その課題についての対応策として，ソフトクラスタリングが有効であるのかを，ハードな分類と比較しながら検証する。

GMM は，与えられたデータセットを複数のガウス分布の重ね合わせで表現し，その確率分布からデータの分類を行うクラスタリング手法の 1 つである。平均が$\boldsymbol{\mu_k}(k = 1,2,\dots,K)$，共分散行列が$\sum_k(k = 1,2,\dots,K)$で与えられるガウス分布を，混合係数$\pi_k$によって重ねた混合ガウス分布から，サンプル$\boldsymbol{x}$が得られる確率は式 9 のような確率密度関数で表される。

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}) \tag{9}$$

GMM で混合ガウス分布を推定する際には，与えられたサンプル集合$\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_N}\}$に対して混合ガウス分布を当てはめ，そのパラメータについて式 10 の尤度関数の最大化が行われる。

$$p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \sum) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}) \tag{10}$$

なお，計算の便宜上，式の対数をとった対数尤度（式 11）が最大化される。

$$\ln p(\boldsymbol{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \sum) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}}) \right\} \tag{11}$$

対数尤度の最大化には，EM アルゴリズムが適用される。以下に EM アルゴリズムによる対数尤度最大化のステップを示す。

1. 初期化：平均$\boldsymbol{\mu_k}$，分散行列$\sum_k$，そして混合係数π_kを初期化し，対数尤度の初期値を決定する。
2. E ステップ：現在のパラメータ値を使って，式 12 の負担率を計算する。

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \sum_{\boldsymbol{k}})} \tag{12}$$

3. M ステップ：現在の負担率を使って，次式でパラメータ値を再計算する。

$$\boldsymbol{\mu_k^{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \boldsymbol{x_n} \tag{13}$$

$$\sum_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})(\boldsymbol{x}_n - \boldsymbol{\mu}_k^{new})^T \tag{14}$$

$$\pi_k^{new} = \frac{N_k}{N} \tag{15}$$

ただし，

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{16}$$

4. 対数尤度（式 11）を計算し，パラメータ値の変化または対数尤度の変化を見て収束性を確認し，収束基準を満たしていなければステップ 2 に戻る。

GMM も k-means と同様，初期化の結果によって結果が変わることが知られており，クラスタの数も分析者が決定する必要がある。そこで，k-means で分類を得る際と同じように 100 パターンの初期値を試し，先述の 3 つの評価指標と AIC（式 17），BIC（式 18）を総合的に考慮して，もっともよい分割が行えていると考えられる初期値とクラスタ数を採用する。

$$AIC = -2 \ln L + 2k \tag{17}$$

$$BIC = -2 \ln L + k \ln n \tag{18}$$

ここで，Lはモデルの尤度，kはパラメータ数，nはデータ数である。AIC と BIC は情報量基準と呼ばれるモデルの当てはまりの良さを表す指標であり，小さいほど良いモデルであるとされる。

3.7. 財務分析における新分類の有用性検証方法

k-means や GMM を用いて得られた新分類を用いて，財務分析を行うことを考える。財務分析では，収益性や安全性，生産性に関する指標を他社と比較することによって，分析対象企業の状態を把握する。このとき，分析対象企業との比較に用いる他社は，分析対象企業とより実態の近い企業であることが望ましい。ハードクラスタリングである k-means を用いた場合の企業分類を財務分析に用いる場合には，従来の産業分類と同様に，分析対象企業と同じ企業群に分類された企業が実態の近い企業であると考えられ，これを用いることで分析が可能であるだろう。一方，ソフトクラスタリングである GMM は，各企業に対してクラスタへの所属確率を与える。最も所属確率の高いクラスタに各企業を割り当てることで，ハードクラスタリングと同様の分析方法も可能であるが，所属確率を用いることで従来の産業分類とは異なる類似企業の選択が可能となりそうである。

LDA によって企業のソフトクラスタリングを行った研究である Fang et al.（2013）では，ソフトクラスタリングの特性を生かして，各企業のクラスタ所属確率分布を基にした類似企業測定スキームを提案した。その具体的な方法を以下に示す。

1. LDA で得られた企業F_1，F_2のクラスタ所属確率分布を基に KL-ダイバージェンスを計算する（式 19)。ここで，$F_j(T_i)$は企業jのクラスタi所属確率を表す。

$$D_{KL}\ (F_1||F_2) = \sum_{T_i} F_1(T_i) \times log \frac{F_1(T_i)}{F_2(T_i)} \tag{19}$$

2. 企業同士の分類的差異を式 20 のように定義する。

$$D_{genre}(F_1, F_2) = D_{KL}(F_1||F_2) + \mathrm{D}_{KL}(F_2||F_1) \tag{20}$$

3. 企業規模を株価×発行株数と定義する（式 21)。

$$market\ cap = share\ price \times number\ of\ outstanding\ shares \tag{21}$$

4. 企業規模の差異を式 22 のように定義する。

$$D_{scale}(F_1, F_2) = \log_{10} \max\left(\frac{market\ cap_{F_1}}{market\ cap_{F_2}}, \frac{market\ cap_{F_2}}{market\ cap_{F_1}}\right) \tag{22}$$

5. 2.3.で得られた分類的差異，企業規模差異をもとに企業差異を式 23 のように定義する。

$$D_{business}(F_1, F_2) = D_{genre}(F_1, F_2) + D_{scale}(F_1, F_2) \tag{23}$$

Fang et al.（2013）は，この方法により従来のハードな分類よりもより実態の近い企業が選択できると主張している。本研究では，GMM による分類において類似企業を選択する際に，Fang et al.（2013）の方法を少し変更して用いる。Fang et al.（2013）では企業規模を株価×発行株数としたが，今回用いるデータには，個人事業主や株式会社ではない事業者も含まれるため，これをそのまま用いることはできない。中小企業基本法では，資本金と従業員数によって中小企業者を定義しているが，地域金融機関の融資実務では，売上高によって大まかな企業規模の把握が行われることが多い[8]。そこで，本研究では，売上高によって企業規模を定義することとする。

また，今回は財務比率を基にしたクラスタリングを行うため，年度によって所属クラスタが大きく変わる場合があることが想定される。そのため，単年度でのクラスタ所属確率ではなく，複数年度のクラスタ所属確率から類似企業を選択することによって，より効果的な比較分析が可能になると考えられる。この

[8] 水戸信用金庫職員の方へのヒヤリングより

考えのもと，本研究では，単年度で分類的差異を計算する場合と，複数年度で分類的差異を計算する場合の 2 つで企業間比較を行う。複数年度で分類的差異を計算する場合の類似企業測定スキームは以下のとおりである。

I. 2015,2016, …,2021年度のうち 2 つの企業に共通してデータが存在する年度を$year$とする。GMM で得られた$year$年度の企業F_1，F_2のクラスタ所属確率分布を基に KL-ダイバージェンスを計算する（式 24)。ここで，$F_j^{year}(T_i)$は$year$年度の企業jのクラスタi所属確率を表す。

$$\mathrm{D_{KL}}\left(F_1^{year} \,||\, F_2^{year}\right) = \sum_{T_i} F_1^{year}(T_i) \times log \frac{F_1^{year}(T_i)}{F_2^{year}(T_i)} \tag{24}$$

II. 企業同士の分類的差異を式 25 のように定義する。ここで，Yは 2 つの企業に共通してデータが存在する年度の数である。

$$D_{genre}(F_1, F_2) = \frac{1}{Y} \sum_{year} D_{KL}\left(F_1^{year} || F_2^{year}\right) + D_{KL}\left(F_2^{year} || F_1^{year}\right) \tag{25}$$

III. 企業規模を$year$のうち最も新しいデータの売上高とする（式 26)。

$$market\ cap = sales \tag{26}$$

IV. 4.と同様

V. 5.と同様

以上のようなスキームで，GMM を用いた場合の類似企業を選択する。そして，選択された類似企業との比較分析を通じて，提案手法が財務分析において有用であるかを 5 章の考察にて論じる。

3.8. 予測モデルの構築における新分類の有用性検証方法

予測モデルでは，既知のデータから目的変数を予測するための重要なパターンを学習し，未知のデータに対して予測を行う。このようなモデルの性能を適切に評価するために，データは次のような 3 つに分割される。

- 訓練データ：モデルの学習に用いるためのデータ
- 検証データ：モデルのハイパーパラメータ調整や，モデル選択に用いるためのデータ
- テストデータ：モデルの構築には用いず，モデルの最終的な性能を評価するためのデータ

k-means や GMM のような教師なし学習の場合，データには事前に定義されたラベルやカテゴリが存在せず，学習に用いるデータ内のパターンや構造を学習し，それに基づいて結果を示す。つまり，これらの結果はデータセット固有のものとなる。そのため，学習に用いていない未知のデータに対して，単純にその結果を適用することは困難である。しかし，学習に用いたデータから得られた分類軸は，未知データに対する分類の基準として機能する可能性がある。

今回，GMM の実装に用いた sklearn.mixture.GaussianMixture[9]では，既知データから推定した混合ガウス分布への未知データの当てはめを簡単に行うことができる。本研究では，これを用いて，GMM によって得られた企業分類から予測モデルを構築する。そして，その予測精度を，①すべての企業のデータを用い

[9] 本研究では python 3.8.17 を用いており，主なパッケージのバージョンは numpy 1.24.3，pandas 2.0.3，scikit-learn 1.3.0 である．

たモデル，②業種別の企業データを用いたモデルと比較することにより，予測モデル構築における新分類の有用性検証を行う。

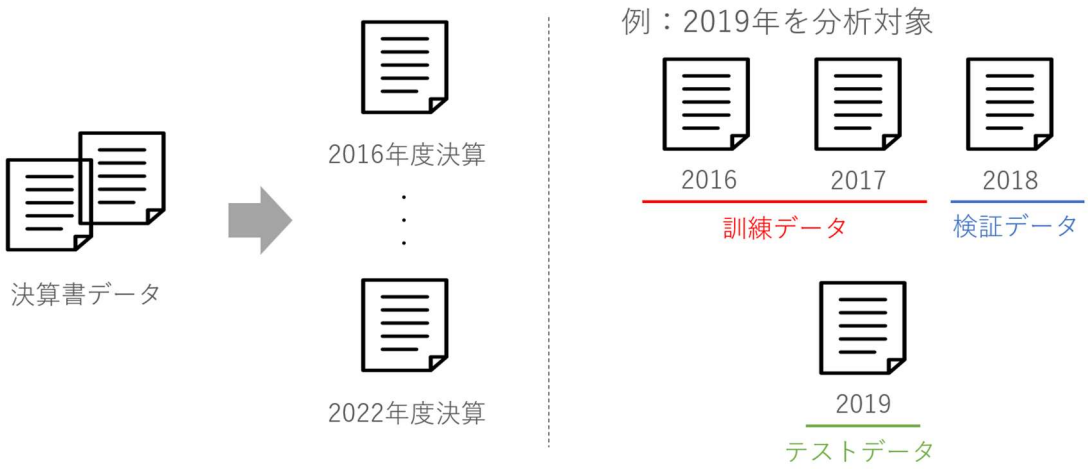
3.8.1. データ分割

前節でも述べたが，予測モデリングでは，データは通常，訓練データ，検証データ，テストデータに分割され，一般的にデータの分割は交差検証によって行われる。

Chen et al.（2022）は，この方法では，過去のイベント（例えば，2017 年の目的変数増減）を用いて将来のデータ（例えば，2019 年の目的変数の増減）から推定されたモデルを評価し，パラメータ調整を行うことが考えられ，時系列的な予測タスクとは矛盾することを懸念している。そこで，Chen et al.（2022）では訓練データ，検証データ，テストデータを時系列で分割し，訓練データと検証データを徐々に時間的に次へシフトさせながら，各データの年数を一定に保つローリングサンプル分割法を使用している。2019 年から 2022 年までのテスト期間の各年度について，モデルは 2 年前と 3 年前のデータで訓練され，1 年前のデータでパラメータ調整のための検証が行われる。これによりモデル構築に最新の情報を利用できるという利点がある。

本研究でのデータ分割も Chen et al.（2022）と同様の方法を採用する。具体的な分割方法の例を図 5 に示す。

図 5　データ分割の方法



3.8.2. 予測モデル

ここでは，①すべての企業のデータを用いたモデル，②業種別の企業データを用いたモデル，③提案手法による分類から得られる情報を用いたモデルについてそれぞれ構築方法を述べる。なお，すべてのモデルに共通することとして，目的変数は「次年度の労働生産性の方向性」，説明変数は 3.2 節表 3 に示した通りであり，アルゴリズムは LightGBM である。

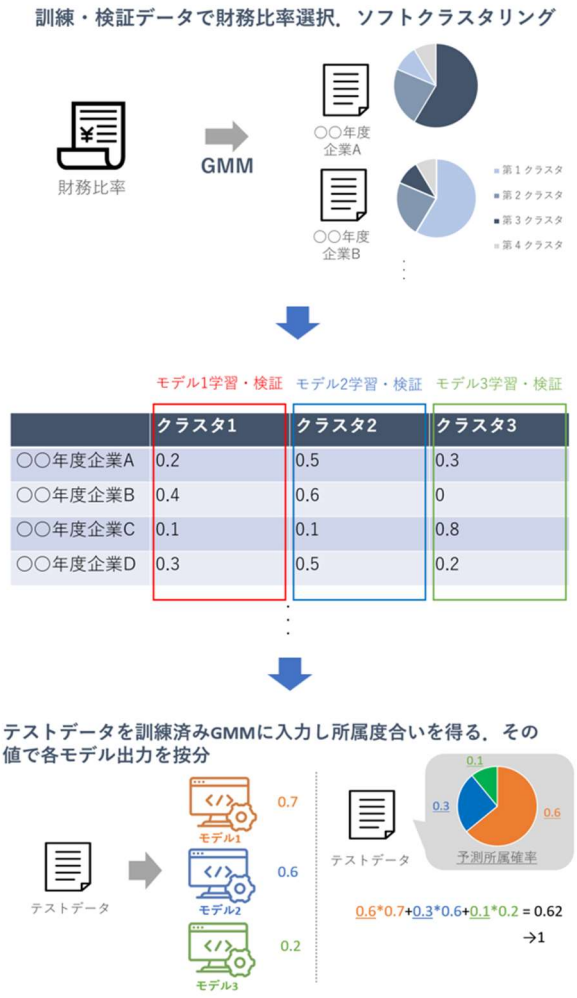
- 全ての企業のデータを用いたモデル

  I. データ分割に示した方法でデータを分割する。

II. LightGBM によって予測を行う。

- 業種別の企業データを用いたモデル
  I. 3.1 節に示したような業種の大分類区分に基づいてデータを業種別に分ける
  II. データ分割に示した方法でデータを分割する。
  III. LightGBM によって予測を行う。
- 提案手法による分類から得られる情報を用いたモデル（図 6）
  I. データ分割に示した方法でデータを分割する。
  II. 訓練データ・検証データを用いて 3.3 節で述べた方法により，企業分類を得る。
  III. 各クラスタ毎に，クラスタ所属確率でデータに対して重みづけ[10]を行ったモデルを構築する。
  IV. 2 で学習させた GMM により，テストデータに対してクラスタ所属確率を求める。
  V. クラスタ所属確率で各クラスタ別モデルの出力を按分し，最終的な出力とする。

図 6　提案手法による分類から得られる情報を用いたモデル：例



[10] 重みづけには，lightgbm.LBGMClassifier の引数である weight を用いて行う．

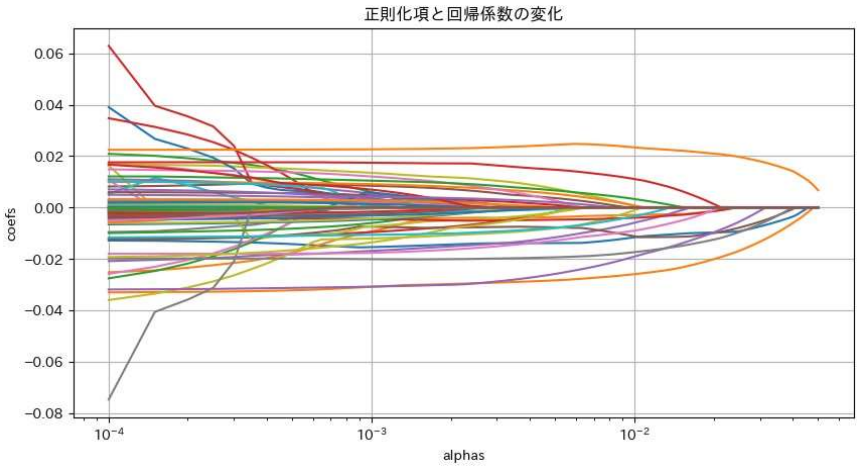
これらのモデルを，テストデータに対する ACC（Accuracy）[11]と AUC（Area Under the ROC Curve）[12]によって評価・比較する。また，今回用いる LightGBM は，目的変数が 0 か 1 で表される二値の場合，1 となる確率を出力とする。よって，ACC の算出の際には確率を二値に変換する閾値を決定する必要がある。本研究では，この閾値決定のために Youden Index（Youden 1950）を用いた。Youden Index は「真陽性率－偽陽性率」で求められ，これを最大にする閾値を用いて確率値を二値に変換した。

4. 結果

4.1. 財務分析のための企業分類結果

前処理後の全年度データにおける財務比率項目内の説明変数から，L1 正則化ロジスティック回帰モデルを用いて，目的変数の予測に対して説明力があるとされるものを 10 個選択した。具体的には，sklearn.linear_model.Lasso を用いて L1 正則化ロジスティック回帰モデルを構築し，正則化項の係数を決定するハイパーパラメータである alpha を[0.0001, 0.05]の範囲で 1000 分割して結果の出力を行い，係数が 0 でない値が 9 個以下となる直前の alpha を採用して変数選択を行った。その他のハイパーパラメータはデフォルトのものを使用した。以下に，alpha の変化による説明変数係数の推移（図 7）と，選択された財務比率（表 5）について示す。表 5 における項目は，企業のどの側面を分析するために用いられる比率であるかを表している。

図 7　L1 正則化ロジスティック回帰による全年度データを用いた財務比率選択



[11] 予測モデルの目的変数に対する正解率を表し，予測結果と正解ラベルが等しいサンプル数/全サンプル数で計算可能。
[12] 縦軸に真陽性率，横軸に偽陽性率をとり，様々な閾値に対して点をプロットして描いた曲線である ROC 曲線の曲線下部面積のこと。閾値の恣意的な設定に影響を受けない指標（Mandrekar 2010）として広く利用されている。

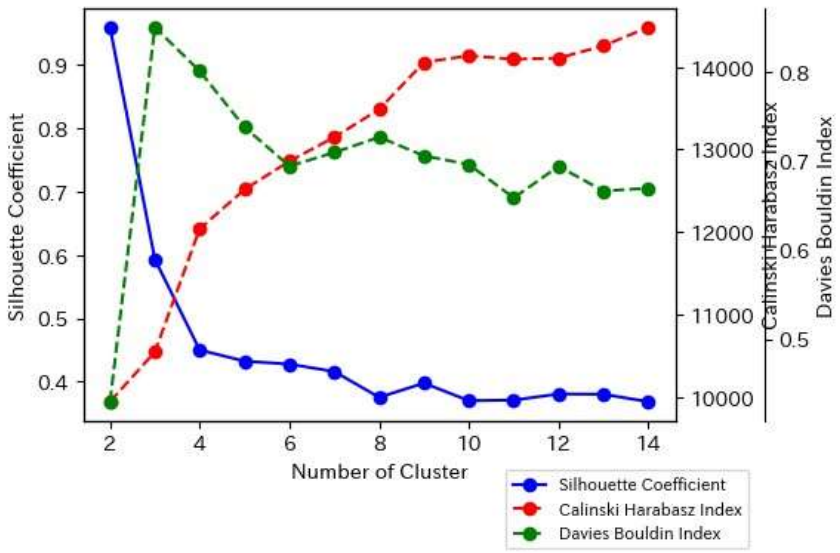
表 5　L1 正則化ロジスティック回帰によって選択された財務比率（全年度データ）

項目	財務比率
収益性	総資本営業利益率
収益性	総資本経常利益率
収益性	売上高営業利益率
収益性	経常損益比率
生産性	従業員一人当たり経常利益
生産性	従業員一人当たり人件費
成長性	営業利益増加率
成長性	経常利益増加率
安全性	負債増加率
安全性	債務超過解消年数

次に，選択された財務比率を用いて k-means と GMM によるクラスタリングを行った。これらの手法における初期値依存性とクラスタ数決定のため，複数の評価指標を用いて企業分類に用いるモデルを決定した。

まず，k-means による企業分類について述べる。k-means の実装には，sklearn.cluster.KMaeans を使用し，各クラスタ数において，ハイパーパラメータである random_state の値を[0,99]の範囲の 100 通りで実行したときの評価指標を見ることで，初期値とクラスタ数を決定した。なお，その他のハイパーパラメータはデフォルトのものを利用した。図 8 に，クラスタ数の変化に対する評価指標の推移を示す。

図 8　クラスタ数と評価指標（k-means による全年度データ分類）



シルエット係数と Davies Bouldin 基準によると，クラスタ数が 2 のときが最もよい分割ができている。

しかし，Calinski Harabasz 基準は，探索範囲においてはクラスタ数の増加に伴ってよいスコアとなっている。この場合，これら 3 つの指標だけでクラスタ数を決定することが難しいため，それぞれのクラスタ数における具体的な分類結果についても見てみる。以下にクラスタ数 2（表 6），3（表 7），4（表 8），5（表 9）のときのクラスタ中心の値と，各クラスタに所属するサンプルの割合について示す。

表 6　k-means による分類（クラスタ数 2）

財務比率	クラスタ 0	クラスタ 1
総資本営業利益率	-3.22	9.35
総資本経常利益率	-0.68	13.14
売上高営業利益率	-3.64	31.61
経常損益比率	93.29	156.09
従業員一人当たり経常利益	319845.40	76076384.33
従業員一人当たり人件費	3819923.56	19568713.89
売上総利益増加率	7.39	23.89
営業利益増加率	-197.37	67.94
負債増加率	17.32	10.14
債務超過解消年数	357.21	0.00
所属サンプル割合	99.92	0.08

表 7　k-means による分類（クラスタ数 3）

財務比率	クラスタ 0	クラスタ 1	クラスタ 2
総資本営業利益率	-4.35	9.35	5.74
総資本経常利益率	-1.72	13.14	7.59
売上高営業利益率	-4.71	31.61	4.90
経常損益比率	92.34	156.09	100.87
従業員一人当たり経常利益	34382.62	76076384.33	2597186.58
従業員一人当たり人件費	3255745.62	19568713.89	8320775.33
売上総利益増加率	5.94	23.89	18.95
営業利益増加率	-215.73	67.94	-50.88
負債増加率	17.02	10.14	19.66
債務超過解消年数	382.72	0.00	153.66
所属サンプル割合	88.75	0.08	11.16

表 8　k-means による分類（クラスタ数 4）

財務比率	クラスタ 0	クラスタ 1	クラスタ 2	クラスタ 3
総資本営業利益率	-5.87	9.08	4.03	-10.09
総資本経常利益率	-3.03	13.03	5.72	-4.31
売上高営業利益率	-6.05	32.59	2.88	-2.14
経常損益比率	91.31	158.09	98.66	91.02
従業員一人当たり経常利益	-85249.94	78145875.81	1398600.99	2191084.27
従業員一人当たり人件費	2822323.94	13758004.27	6173454.18	36267861.00
売上総利益増加率	5.62	23.57	12.00	22.08
営業利益増加率	-235.70	114.37	-93.86	-187.33
負債増加率	17.76	10.56	15.87	39.15
債務超過解消年数	419.16	0.00	189.98	315.01
所属サンプル割合	72.72	0.08	26.90	0.30

表 9　k-means による分類（クラスタ数 5）

財務比率	クラスタ 0	クラスタ 1	クラスタ 2	クラスタ 3	クラスタ 4
総資本営業利益率	-5.62	1.50	8.13	17.08	-12.19
総資本経常利益率	-2.65	3.03	13.53	20.24	-5.93
売上高営業利益率	-6.11	1.12	34.47	18.31	-3.21
経常損益比率	91.39	96.54	164.32	124.49	89.09
従業員一人当たり経常利益	-30953.43	682947.41	89384618.18	14226560.38	1014099.36
従業員一人当たり人件費	2675620.69	5941932.92	12144597.20	7523899.72	37596578.63
売上総利益増加率	6.74	6.89	18.91	65.09	22.56
営業利益増加率	-233.24	-125.98	54.78	47.80	-202.70
負債増加率	17.94	15.10	4.53	37.36	42.61
債務超過解消年数	414.28	241.76	0.00	35.66	346.09
所属サンプル割合	67.81	30.89	0.06	0.96	0.28

表 6 より，シルエット係数と Davies Harabasz 基準が最もよい値となったクラスタ数 2 のときの分類は，ほとんど全てのサンプルが所属する大きなクラスタとそれ以外というような分類となっている。この場合，大半の企業に対して類似企業を絞り込むことができず，財務分析に利用できる分類として適していない。クラスタ数を増やしていくと，3 つのクラスタに分けたときに大きなクラスタが 2 つに分割された。また，クラスタ数 4 のときには小さなクラスタがもう 1 つ形成された。その後も同様に，今回の評価指標に基づいた初期値決定方法においては，クラスタ数の増加につれてほとんどの場合小さなクラスタが形成され，稀に大きなクラスタの分割が行われた。小さなクラスタが多数形成されることによる財務分析利用面での効用は，少数企業を対象にする場合以外はほとんどないため，最初に大きなクラスタが分割された，3 つのクラスタによる分類を採用することとする。

次に，GMM による企業分類について述べる。GMM の実装には，sklearn.mixture.GaussianMixture を使用し，k-means と同様の方法で初期値とクラスタ数を決定した。また，sklearn.mixture.GaussianMixture では，covariance_type 引数によって混合ガウス分布における各ガウス分布の形状（共分散行列）に制約を課すことができる。選択できる共分散行列の制約には次の 4 つがある。

- full：それぞれのガウス分布が独自の共分散行列を持つ。これによって，各分布の形状はそれぞれ異なったものになる。
- tied：すべてのガウス分布が共通の共分散行列を持つ。これによって，各分布の形状はすべて同じになる。
- diag：それぞれのガウス分布が独自の対角共分散行列を持つ。これによって，各分布の形状は変数軸と並行な軸を持つ n 次元楕円となる。
- spherical：それぞれのガウス分布が独自の単位共分散行列を持つ。これによって，各分布の形状は n 次元球となる。

これによってクラスタリング結果が大きく変わることが予想されるため，すべての場合について実験を行った。以下に，それぞれの制約を課したときのクラスタ数の変化に対する評価指標の推移を示す。

図 9　クラスタ数と評価指標（GMM_full による全年度データ分類）

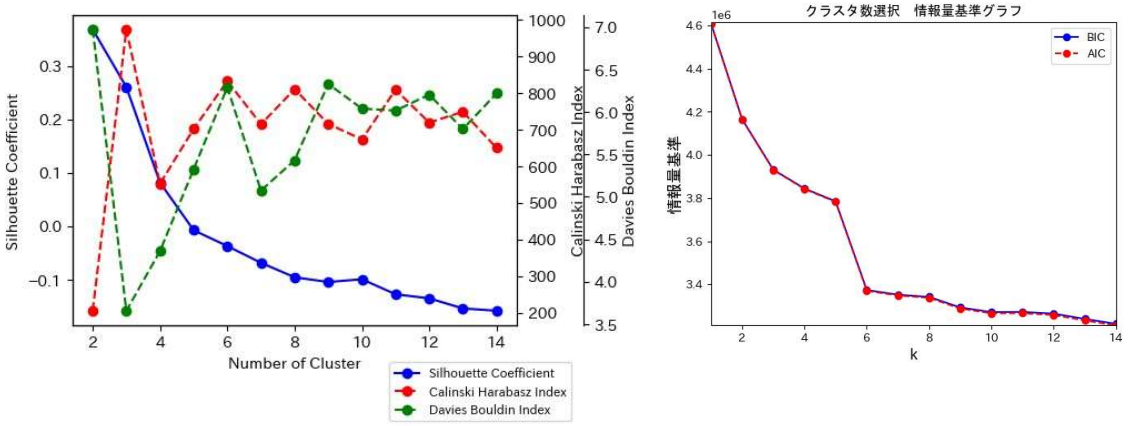


図 10　クラスタ数と評価指標（GMM_tied による全年度データ分類）

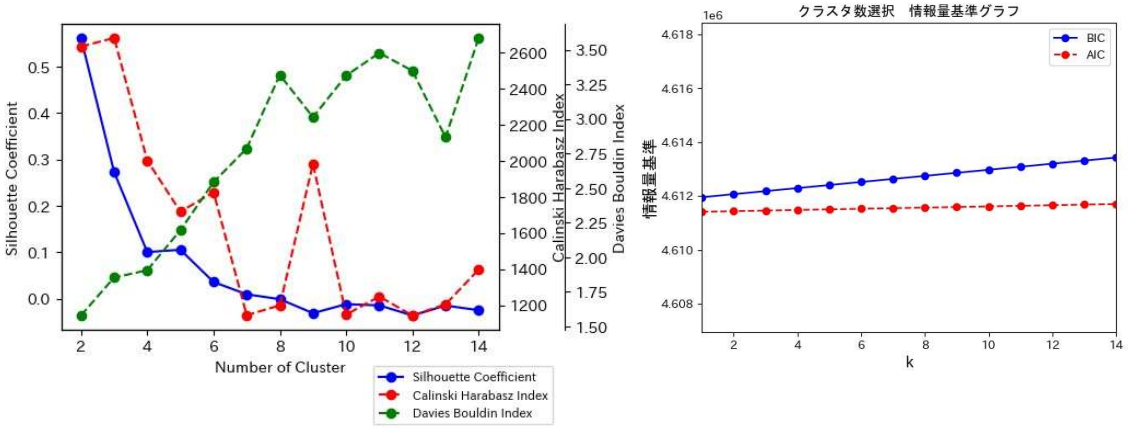


図 11　クラスタ数と評価指標（GMM_diag による全年度データ分類）

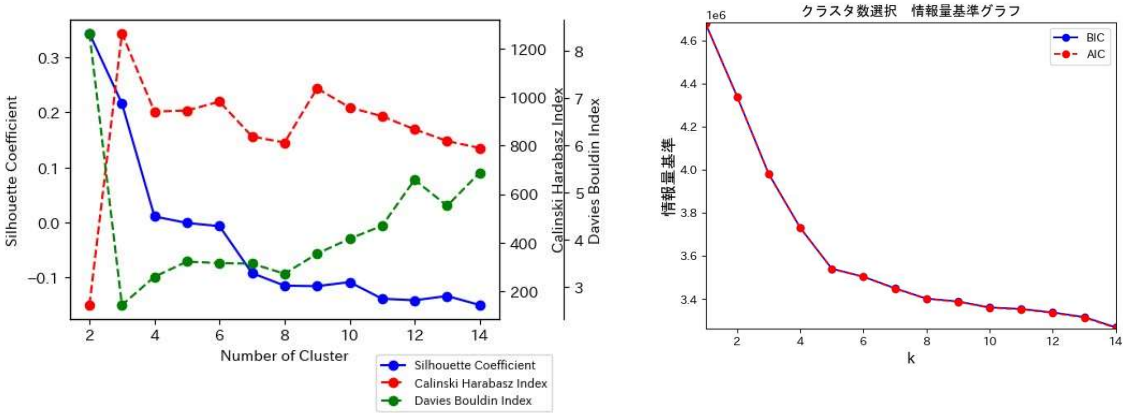
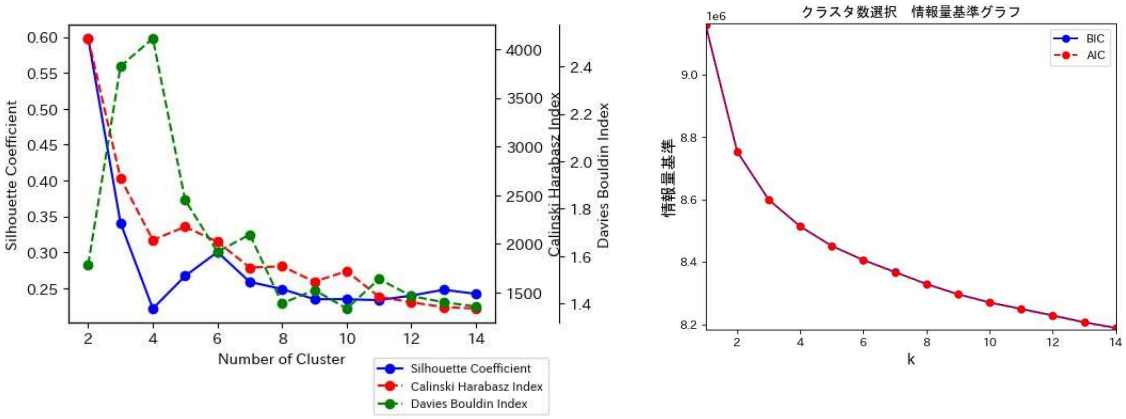


図 12　クラスタ数と評価指標（GMM_spherical による全年度データ分類）



AIC と BIC は，full 制約を課したときのモデルにおいて最も良い値を示している。一方で，シルエット係数と Calinski Harabasz 基準，Davies Bouldin 基準は spherical の制約を課したときのモデルにおいて最も良い値を示している。この結果より，full か spherical を制約としたときのモデルを企業分類に使用することとする。

図 9 より，full 制約を課したときのクラスタ数の候補として，シルエット係数と Calinski Harabasz 基準，Davies Bouldin 基準が総合的に良い結果を示している 3 か，AIC と BIC が大きく減少している 6 が挙げられる。ここで，クラスタ数が 6 のときのシルエットスコアを見ると，値が負になっている。この場合，所属クラスタの判別が間違っている可能性が示唆される。このことから，full 制約を課したときのクラスタ数は 3 が最適であると判断した。

また，図 12 より spherical 制約を課したときには AIC と BIC からクラスタ数を決定するのは難しいため，シルエット係数と Calinski Harabasz 基準，Davies Bouldin 基準を参考に決定する。クラスタ数の候補として，これら 3 つの指標が総合的に良い値を示している 2 もしくは 6 がクラスタ数の候補として挙げられる。

以上より，「full 制約，クラスタ数 3（表 10)」，「spherical 制約，クラスタ数 2（表 11)」，「spherical 制約，クラスタ数 6（表 12)」の 3 つの場合について具体的なクラスタリング結果を確認し，最終的な GMM による企業分類を得る方法を決定する。以下にそれぞれの場合のクラスタ中心の値と，単一のガウス分布で表される各クラスタの混合ガウス分布における混合係数を示す。

表 10　GMM_full による分類（クラスタ数 3）

財務比率	クラスタ 0	クラスタ 1	クラスタ 2
総資本営業利益率	-0.21	1.38	-21.38
総資本経常利益率	1.25	1.49	-11.16
売上高営業利益率	-0.32	0.99	-22.88
経常損益比率	101.21	79.99	83.89
従業員一人当たり経常利益	282996.70	146156.08	1105845.97
従業員一人当たり人件費	3796273.75	3684622.20	4192522.98
売上総利益増加率	6.10	-1.62	25.85
営業利益増加率	-231.28	-45.83	-296.68
負債増加率	12.99	-0.83	60.82
債務超過解消年数	372.75	241.66	470.84
混合係数値（%）	59.71	24.28	16.01

表 11　GMM_spherical による分類（クラスタ数 2）

財務比率	クラスタ 0	クラスタ 1
総資本営業利益率	-3.33	-2.41
総資本経常利益率	-0.67	-0.67
売上高営業利益率	-3.11	-7.02
経常損益比率	92.76	97.32
従業員一人当たり経常利益	127878.39	2114038.38
従業員一人当たり人件費	3303923.56	7443519.29
売上総利益増加率	6.78	11.64
営業利益増加率	-203.44	-154.21
負債増加率	16.64	21.94
債務超過解消年数	364.76	303.34
混合係数値（%）	87.18	12.82

表 12　GMM_spherical による分類（クラスタ数 6）

財務比率	クラスタ 0	クラスタ 1	クラスタ 2	クラスタ 3	クラスタ 4	クラスタ 5
総資本営業利益率	-0.11	-5.85	1.17	16.48	-4.03	-35.38
総資本経常利益率	1.80	-1.46	2.50	19.88	-2.67	-33.49
売上高営業利益率	-0.07	-4.92	0.71	9.60	-6.96	-32.49
経常損益比率	94.08	90.33	95.72	106.63	97.19	77.71
従業員一人当たり経常利益	146013.18	29436.13	330384.09	2440768.99	3026495.50	-1788338.48
従業員一人当たり人件費	3442095.07	1739340.17	5464781.05	3734268.24	9224367.98	3371326.64
売上総利益増加率	8.24	10.35	2.81	38.88	15.94	-31.37
営業利益増加率	-163.13	-232.64	-136.02	4.73	-165.67	-578.92
負債増加率	14.79	18.67	14.32	14.75	24.40	25.82
債務超過解消年数	297.03	431.37	250.33	24.21	318.14	954.02
混合係数値（%）	27.13	26.74	21.94	9.03	6.51	8.65

それぞれのクラスタ中心と混合係数値を確認すると，spherical 制約を課したクラスタ数 6 の分類が，クラスタ中心値が特徴づいている点と分布間のサンプルの偏りが小さい点で，今回の企業分類に最適であると判断した。以降は，GMM による企業分類としてこれを用いる。

ここまでの実験結果をまとめると，評価指標や分類結果から，企業分類を得るためにクラスタ数 3 の k-means（表 7）と spherical 制約を課したクラスタ数 6 の GMM（表 12）を採用することとなった。ここからはこれらを用いて，いくつかの企業に対して類似企業を選択した結果について示す。なお，各クラスタの解釈や類似企業を用いた比較分析については次章の考察で行う。

まず，k-means による企業分類を用いた類似企業選択結果を示す。例示のため，分析対象の企業として，水戸信用金庫の業種分類において建設業（一般建築）に属するサンプル A と，小売業（機械工具小売業）に属するサンプル B を設定する。k-means による分類の場合，類似企業選択の際に参考にできる事柄としては，同じ企業分類に属するかどうかということである。今回の k-means を用いた企業分類では，サンプル A はクラスタ 0（所属サンプル割合 88.75%）に所属し，サンプル B はクラスタ 2（所属サンプル割合 11.16%）に所属する。k-means による分類から得られる情報では，これ以上類似企業の絞り込みはできないため，今回は同じクラスタ内で売上高が近いサンプルを比較対象として選択した。この結果，サンプル A の類似企業として，建設業（給排水設備）の企業が選択され，サンプル B の類似企業として，その他のサービス業（人材派遣業）の企業が選択された。比較分析は次章の考察で行うとして，ここでは代表的な財務比率を示すのみとする（表 13，表 14）。

表 13　サンプル A と類似企業の代表的な財務比率（k-means）

財務比率	サンプル A	類似企業
総資本経常利益率	1.86	0.26
売上高経常利益率	1.96	0.32
一人当たり人件費	3501295.25	4412490.25
自己資本比率	62.75	49.54
流動比率	244.69	469.22
売上高増加率	-25.94	13.41
総資本回転率[13]	0.95	0.81

[13] 財務諸表データには存在しなかったため，B/S 項目の売上高を B/S 項目の資産合計で除算することで算出した。

表 14　サンプル B と類似企業の代表的な財務比率（k-means）

財務比率	サンプル B	類似企業
総資本経常利益率	0.54	5.11
売上高経常利益率	0.31	3.95
一人当たり人件費	12964759.50	205468.25
自己資本比率	8.89	10.8
流動比率	205.53	206.49
売上高増加率	47.97	118.07
総資本回転率	1.71	1.29

次に，GMM による企業分類を用いた類似企業選択結果を示す。分析対象とする企業は k-means の場合と同じとする。3.6 節に示した方法により，クラスタ所属確率分布と売上高規模から類似企業を選択する。また，クラスタ所属確率分布は単年度の類似性を測る場合と複数年度の類似性を測る場合の両方について実験を行った。まず，単年度の類似性を測った場合の結果について述べる。単年度のクラスタ所属確率類似性と，売上規模で類似企業を選択した結果，サンプル A の類似企業として，建設業（一般建築）の企業が選択され，サンプル B の類似企業として，建設業（土木工事）の企業が選択された。代表的な財務比率を表 15，表 16 に示す。

表 15　サンプル A と類似企業の代表的な財務比率（GMM）

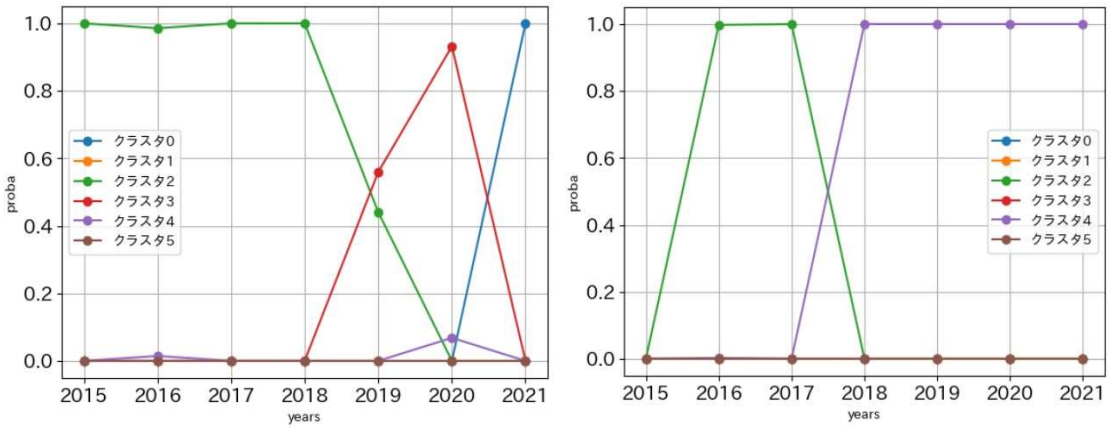
財務比率	サンプル A	類似企業
総資本経常利益率	1.86	0.92
売上高経常利益率	1.96	3.6
一人当たり人件費	3501295.25	3236829.20
自己資本比率	62.75	26.68
流動比率	244.69	395.60
売上高増加率	-25.94	-16.56
総資本回転率	0.95	0.26

表 16　サンプル B と類似企業の代表的な財務比率（GMM）

財務比率	サンプル B	類似企業
総資本経常利益率	0.54	-4.09
売上高経常利益率	0.31	-3.29
一人当たり人件費	12964759.50	10218569.82
自己資本比率	8.89	11.18
流動比率	205.53	201.40
売上高増加率	47.97	-11.44
総資本回転率	1.71	1.24

次に，複数年度の類似性を測った結果について述べる。財務諸表より計算される財務比率は年度によって大きく変わることが予想される。そこで，クラスタ所属確率の時系列的な遷移が似ている企業を選択することにより，より適切な比較分析が可能となるのではないかと考えた。ここでサンプル A とサンプル B のクラスタ所属確率の時系列的な遷移を可視化したものを図 13 に示す。

図 13　クラスタ所属確率の時系列遷移（左：サンプル A，右：サンプル B）



3.7 節の方法論で，複数年度にわたるクラスタ所属確率の類似性を計算することにより，図 13 に示したような遷移の仕方が似ている企業を定量的に選択した。その結果，サンプル A の類似企業として，建設業（電気設備）の企業が選択され，サンプル B の類似企業として，不動産業（不動産仲介業）の企業が選択された。図 14 にサンプル A とその類似企業のクラスタ所属確率遷移を，図 15 にサンプル B とその類似企業のクラスタ所属確率遷移を示す。また，各企業の代表的な財務比率についても表 17，表 18 に示す。ただし，サンプル A の類似企業については 2021 年度の財務諸表データが存在しなかったため，2020 年の財務諸表から計算された財務比率を掲載する。

図 14　クラスタ所属確率の時系列遷移（左：サンプル A，右：A の類似企業）

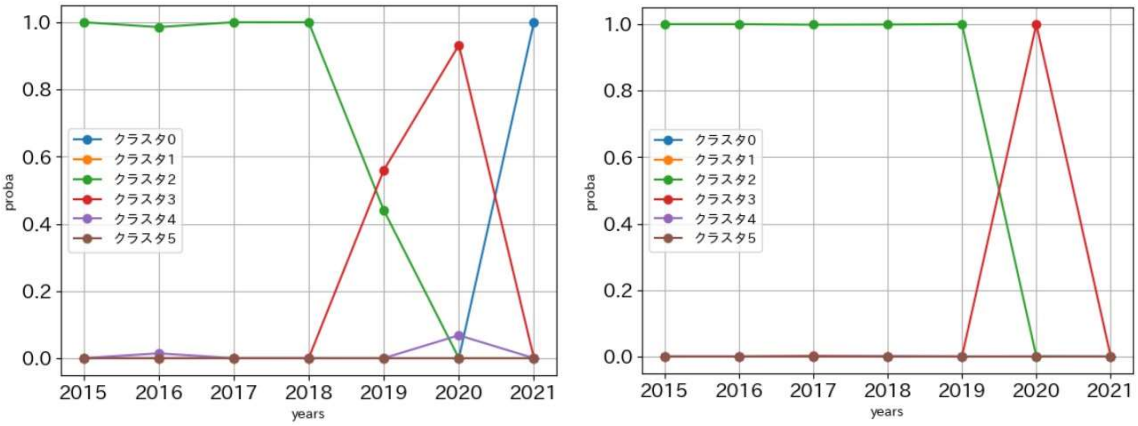


図 15　クラスタ所属確率の時系列遷移（左：サンプル B，右：B の類似企業）

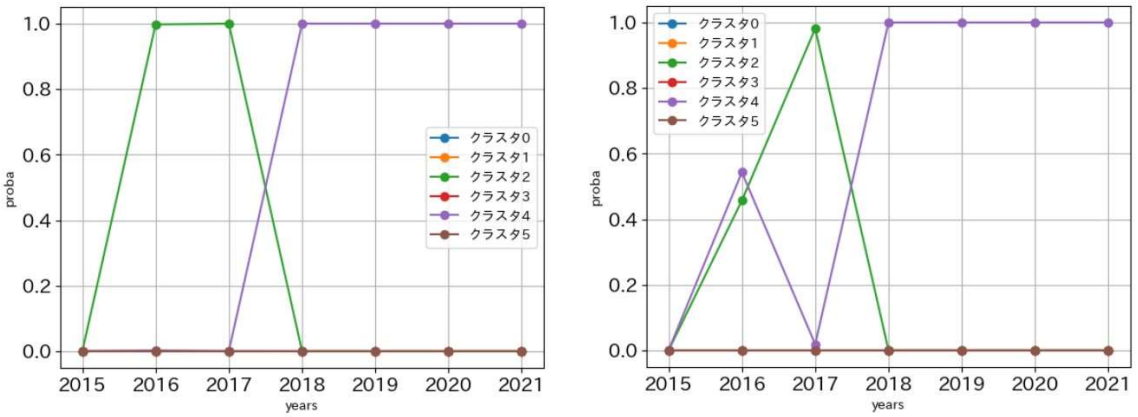


表 17　サンプル A と類似企業の代表的な財務比率（GMM 複数年度）

財務比率	サンプル A（2020 年度）	類似企業
総資本経常利益率	14.28	6.13
売上高経常利益率	12.78	10.25
一人当たり人件費	7171037.50	6624750.35
自己資本比率	54.49	51.15
流動比率	247.79	532.15
売上高増加率	-51.33	-26.64
総資本回転率	1.12	0.59

表 18　サンプル B と類似企業の代表的な財務比率（GMM 複数年度）

財務比率	サンプル B	類似企業
総資本経常利益率	0.54	11.59
売上高経常利益率	0.31	10.35
一人当たり人件費	12964759.50	8706670.63
自己資本比率	8.89	49.52
流動比率	205.53	899.69
売上高増加率	47.97	28.21
総資本回転率	1.71	1.12

以上のようにして，提案手法で企業分類を行った場合に比較対象となる類似企業を得た。次章の考察で，基礎統計量をもとに各クラスタに対する解釈を行い，実際の財務分析における有用性について考察する。

4.2.　予測モデルの構築における新分類の有用性検証

財務諸表データから「次年度の労働生産性の方向性」を予測するモデルを構築するうえで提案手法が有用か検証を行った結果について述べる。なお，LightGBM のアルゴリズム実装には，lightgbm.LGBMClassirier[14]を用いており，検証データによるハイパーパラメータの決定には optuna[15]を用いた。

まず，ベースラインとなる全ての企業データを用いたモデルの結果を表 19 に示す。以降は，訓練データに対する ACC と AUC をそれぞれ TRAIN_ACC，TRAIN_AUC とし，テストデータに対する ACC と AUC をそれぞれ TEST_ACC，TEST_AUC とする。

表 19　全企業データを用いたモデルの予測精度

テストデータ年度	TRAIN_ACC	TRAIN_AUC	TEST_ACC	TEST_AUC
2018	0.638	0.687	0.590	0.627
2019	0.603	0.648	0.560	0.602
2020	0.676	0.731	0.580	0.613
2021	0.633	0.678	0.598	0.637

次に，業種別モデルにおいて，TEST_AUC がベースラインモデルよりも向上したものについて表 20 に示す。なお，全業種別モデルの結果は付録 A に記載する。

[14] Lightgbm 4.0.0 を使用.

[15] ベイズ最適化によりハイパーパラメータを自動で最適化するフレームワーク．https://www.preferred.jp/ja/projects/optuna/

表 20　業種別モデルの予測精度

予測データ年度	業種名	TRAIN_ACC	TRAIN_AUC	TEST_ACC	TEST_AUC
2018	建設業	0.914	0.975	0.603	0.640
2019	小売業	0.890	0.946	0.576	0.605
2019	食料	0.930	0.981	0.560	0.611
2019	不動産業	0.893	0.946	0.613	0.673
2019	建設業	0.880	0.950	0.621	0.673
2019	医療・福祉	0.827	0.897	0.597	0.609
2020	飲食業	0.903	0.967	0.609	0.643
2020	不動産業	0.944	0.986	0.581	0.630
2020	金属製品	0.860	0.933	0.603	0.651
2020	建設業	0.884	0.946	0.610	0.658
2020	医療・福祉	0.854	0.920	0.601	0.620
2020	印刷	0.906	0.951	0.649	0.638
2021	農業・林業	0.744	0.809	0.564	0.643
2021	金属製品	0.894	0.963	0.634	0.646
2021	教育・学習支援業	0.736	0.811	0.640	0.649
2021	電気機械	0.724	0.803	0.618	0.672
2021	建設業	0.830	0.909	0.602	0.661
2021	情報サービス広告放送	0.881	0.908	0.538	0.683

最後に，提案手法による分類から得られる情報を用いたモデルの結果を表 21 に示す。具体的には，クラスタの所属確率を重みとして設定し，クラスタ毎にモデルを構築した。最終的な出力には，各モデルの重みを所属確率で按分したものを用いた。なお，クラスタリングにおけるクラスタ数は，前節と同様に評価指標を参考に決定した。その結果，テストデータが 2018，2019，2020，2021 年度のデータであるモデルに対し，クラスタ数はそれぞれ 5，6，6，6 個とした。その詳細は付録 A に記載する。

表 21　提案手法による分類から得られる情報を用いたモデルの予測精度

テストデータ年度	クラスタ番号	混合係数値	TRAIN_ACC	TRAIN_AUC	TEST_ACC	TEST_AUC
2018	0	26.06	0.624	0.669	0.587	0.625
	1	23.35	0.627	0.669	0.589	0.624
	2	21.25	0.609	0.650	0.587	0.623
	3	2.55	0.601	0.644	0.583	0.624
	4	26.78	0.648	0.701	0.590	**0.632**
最終的な出力			0.625	0.664	0.588	0.622
2019	0	7.67	0.613	0.658	0.584	**0.616**
	1	23.43	0.622	0.668	0.587	**0.623**
	2	27.93	0.610	0.664	0.600	**0.619**
	3	6.44	0.602	0.642	0.572	**0.603**
	4	26.22	0.604	0.637	0.566	0.594
	5	8.31	0.603	0.645	0.570	**0.606**
最終的な出力			0.609	0.646	0.576	**0.606**
2020	0	6.21	0.661	0.717	0.580	**0.618**
	1	23.22	0.713	0.775	0.586	**0.624**
	2	8.15	0.659	0.721	0.579	**0.616**
	3	27.92	0.672	0.734	0.578	**0.617**
	4	26.68	0.647	0.700	0.583	**0.618**
	5	7.82	0.687	0.743	0.582	**0.617**
最終的な出力			0.674	0.730	0.584	**0.619**
2021	0	27.35	0.633	0.680	0.601	0.636
	1	26.60	0.680	0.736	0.604	**0.643**
	2	22.05	0.655	0.709	0.596	0.633
	3	8.50	0.640	0.694	0.597	0.630
	4	8.38	0.660	0.713	0.602	**0.639**
	5	7.11	0.651	0.700	0.600	**0.639**
最終的な出力			0.654	0.704	0.600	0.635

注）ベースラインモデルの TEST_AUC を超えたものが太字になっている。

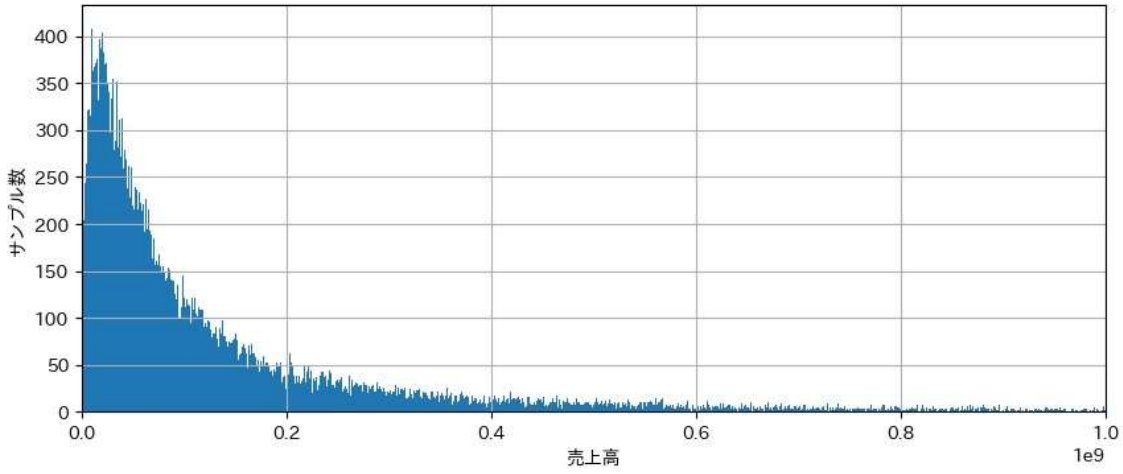
以上の結果から，業種別モデルによって大きく予測精度が向上する業種が存在することが分かった。また，今回の提案手法を用いた実験では，最終的な予測精度がベースラインを大きく超えることはなかったが，いくつかのクラスタ別重みづけモデルにおいて予測精度が向上しているものが見られた。これらの結果についても，次章の考察で詳しく議論する。

5. 考察

クラスタリング結果について

今回は，L1 正則化ロジスティック回帰によって目的変数の予測に有用である財務比率を選択した。その結果としては，目的変数の計算に用いた売上高や従業員数に関連するものが多く選択された印象であった。これを用いて，k-means と GMM による企業のクラスタリングを行ったところ，k-means では評価指標が良い値を示すも，類似した企業群を得るには粒度が大きすぎる分類と，小さすぎる分類に分かれた。これは，今回用いた中小企業財務諸表データの特性が原因であると考えられる。中小企業法で定義される中小企業には，零細企業から一般的にイメージされる中小企業まで様々な規模のものが含まれる。今回用いたデータにもそのような特性が顕著に表れている。図 16 は，今回用いたデータにおける売上高の分布である。

図 16　使用データの売上高分布



注）横軸のスケールが 1e9 であることに注意されたい。0 から 1000000000 の間を 1000 分割したヒストグラムである。基本統計量は，最小値 596，25%点 31533900，中央値 75874120，75%点 192174300，最大値 536212000000，平均値 404085800，標準偏差 622123000 である。

売上高は，あるスケールにサンプルが集中していて裾の長い分布となっている。つまり，使用したデータは主にデータが密集した部分と外れ値から成る。k-means は，データ点とクラスタ中心の距離に基づいたクラスタリングを行うため，外れ値が存在する場合に均質にクラスタリングできない場合がある。今回のデータはまさにそのようなデータであり，外れ値でない部分がかなり密になっているため，クラスタ数 2

の k-means が評価指標の観点では最も良い値になったのだと考えられる。これに対して GMM が比較的均質にクラスタリングできた要因としては，GMM が確率分布に基づいたクラスタリング手法であることが挙げられる。データが密になっている部分に対して，特に密になっている部分とそうではない部分を確率分布によって表現した結果，比較的均質なクラスタリングが可能となったことが考えられる。しかし，k-means と GMM ともに初期値によってかなり結果が変わってしまった。このような結果から，中小企業データのクラスタリングにおける，外れ値対処と適切な分類軸選択の必要性が示唆された。

今回得られたクラスタについての解釈

クラスタの基礎統計量を用いて，k-means と GMM による分類結果の各クラスタについての解釈を行った。解釈に用いた箱ひげ図については付録 B に記載する。

まず，k-means で形成されたクラスタについての解釈を行う。k-means では非常に大きなクラスタが 1 つと小さいクラスタが 1 つ，そして非常に小さいクラスタが 1 つというような分類となっている。非常に小さいクラスタ 1 に分類されることがほとんどないことを考慮し，クラスタ 0 と 2 のみに対して解釈を行う。非常に大きなクラスタ 1 に関しては，財務比率の分散が大きくほとんどすべての指標でクラスタを特徴づけるのが困難であったが，２つのクラスタに大きく違いが出ていたのは一人当たり生産性指標である。クラスタ 1 よりもクラスタ 2 の方が良い値を示していた。今回の k-means の分類の粒度では，この程度の解釈が限界であった。簡単ではあるが各クラスタを次のように命名した。

- クラスタ 0 ：低生産性企業群
- クラスタ 1 ：特徴づけるための十分なサンプルなし
- クラスタ 2 ：高生産性企業群

次に，GMM で形成されたクラスタについての解釈を行う。

収益性指標である売上高営業利益率や，売上高経常利益率などをみると，クラスタ 0，１，２はそれほど違いはないが，クラスタ 3 が特に良い値となっており，クラスタ 4 もそれに続いて良い値となっている。対して，クラスタ 5 は全体的に見てもかなり悪い値となっている。

効率性指標は，分散が大きいためクラスタを特徴づけるためには用いなかった。

一人当たり生産性指標である労働生産性は，クラスタ 4，３，２の順で比較的良い値となっていて，クラスタ 1，５が比較的悪い値となっている。今回の分類は労働生産性向上のための分析に用いることが目的であるので，良い値を示したクラスタへの遷移について詳しく分析することが重要となるだろう。また，従業員一人当たり人件費と従業員一人当たり売上高がクラスタ 0，１，２間で大きく違う結果となり，最も良い値を示したのはクラスタ 2，４である。収益性指標が同程度であるこれらを特徴づけるためにはこの指標を用いるのが適していると考えられる。

安全性指標と C/F 指標に関しては，効率性指標と同様に分散が大きいためクラスタの特徴づけには用いない。

成長性指標は，四分位範囲ではクラスタ 3 が良い値を示し，クラスタ 5 が悪い値を示したがこれも分散が大きいため参考程度に用いることとする。

以上のことから，クラスタを次のように命名した。

- クラスタ 0：収益性・生産性中堅企業群
- クラスタ 1 ：低生産性企業群
- クラスタ 2 ：中堅収益・高生産性企業群
- クラスタ 3 ：高収益性企業群
- クラスタ 4 ：高収益・高生産企業群
- クラスタ 5 ：危険企業群

得られた企業分類の財務分析における有用性について

4.1 で分析例として挙げたサンプル A とサンプル B について，類似企業を k-means で選択した場合と，GMM で単年度の類似性から選択した場合，そして GMM で複数年度の類似性から選択した場合のそれぞれの結果から，財務分析によく用いられる項目を使用して簡易的な財務分析表を作成した（表 22，表 23，表 24）。これを用いることで，統計モデルによって選択された，次年度の労働生産性の方向性に関連がある財務比率を基準として，近い企業と比較分析することができる。

表 22　k-means による分類から選択した類似企業の比較分析

指標	財務比率	サンプル A（0）	類似企業（0）	サンプル B（2）	類似企業（2）
収益性	売上高営業利益率	2.11	-1.03	0.80	4.12
	売上高経常利益率	1.96	0.32	0.31	3.95
	売上高総利益率	51.46	33.76	10.44	17.45
	売上高金融損益比率	0.23	0.46	0.98	0.17
	総資本営業利益率	2.01	-0.84	1.38	5.33
	総資本経常利益率	1.86	0.26	0.54	5.11
	自己資本当期利益率	0.37	0.20	4.76	35.46
効率性	売上債権回転期間	25.75	96.72	31.54	40.56
	棚卸資産回転期間	18.09	283.38	123.00	0.00
	総資本回転期間	384.67	449.99	212.88	282.53
1 人当たり生産性	労働生産性	4567273.25	2996246.00	34748820.00	562191.40
	従業員一人当たり人件費	3501295.25	4412490.25	12964760.00	205468.20
	従業員一人当たり売上高	8874395	8874861.75	332731100.00	3220099.00
安全性	固定比率	63.29	16.72	54.29	647.09
	固定長期適合率	52.7	10.29	8.99	81.83
	当座比率	225.59	146.30	57.07	204.75
	流動比率	244.69	469.22	205.53	206.49
	自己資本比率	62.75	49.54	8.89	10.80
C/F 指標	債務償還年数	9.07	0.00	22.41	9.64
成長性	売上総利益増加率	-21.14	36.97	-9.88	128.01
	売上高増加率	-25.94	13.41	47.97	118.07
	経常利益増加率	-88.61	-103.56	-54.45	287.89

注）() 内の数字は分類されたクラスタ番号を指す。

表 23　GMM による分類の単年度類似性から選択した類似企業の比較分析

指標	財務比率	サンプル A	類似企業	サンプル B	類似企業
収益性	売上高営業利益率	2.11	-16.27	0.80	-3.95
	売上高経常利益率	1.96	3.60	0.31	-3.29
	売上高総利益率	51.46	48.94	10.44	16.64
	売上高金融損益比率	0.23	0.97	0.98	1.10
	総資本営業利益率	2.01	-4.18	1.38	-4.91
	総資本経常利益率	1.86	0.92	0.54	-4.09
	自己資本当期利益率	0.37	3.47	4.76	-21.31
効率性	売上債権回転期間	25.75	50.49	31.54	25.59
	棚卸資産回転期間	18.09	1000.49	123.00	0.03
	総資本回転期間	384.67	1420.84	212.88	293.66
1 人当たり生産性	労働生産性	4567273.25	3468809.80	34748820.00	3182634.00
	従業員一人当たり人件費	3501295.25	3236829.20	12964760.00	10218570.00
	従業員一人当たり売上高	8874395	7087147.00	332731100.00	19123770.00
安全性	固定比率	63.29	71.22	54.29	569.57
	固定長期適合率	52.7	23.89	8.99	77.69
	当座比率	225.59	51.37	57.07	194.08
	流動比率	244.69	395.60	205.53	201.40
	自己資本比率	62.75	26.68	8.89	11.18
C/F 指標	債務償還年数	9.07	0.00	22.41	5.61
成長性	売上総利益増加率	-21.14	6.90	-9.88	-4.66
	売上高増加率	-25.94	-16.56	47.97	-11.44
	経常利益増加率	-88.61	-335.68	-54.45	-999.99

注）() 内の数字は分類されたクラスタ番号を指す。複数数字が記載されているものは，ソフトクラスタリングにおいて単一のクラスタ所属確率が極端に高くならなかったものである。

表 24　GMM による分類の複数年度の類似性から選択した類似企業の比較分析（サンプル A）

指標	財務比率	2018_A(2)	2019_A(3,2)	2020_A(3)	2018_類似(2)	2019_類似(2)	2020_類似(3)
収益性	売上高営業利益率	1.75	5.71	8.72	-4.78	0.82	10.71
	売上高経常利益率	1.45	5.71	12.78	-4.90	1.03	10.25
	売上高総利益率	63.38	27.24	48.33	23.47	25.75	41.08
	売上高金融損益比率	0.30	0.00	0.14	0.66	0.51	0.69
	総資本営業利益率	1.55	11.66	9.73	-3.32	0.69	6.31
	総資本経常利益率	1.28	11.66	14.28	-3.41	0.87	6.03
	自己資本当期利益率	4.08	24.82	21.28	-7.19	1.78	11.80
効率性	売上債権回転期間	15.69	25.74	31.46	103.17	89.85	92.86
	棚卸資産回転期間	123.43	0.20	0.25	11.87	17.45	20.79
	総資本回転期間	412.16	178.83	326.84	525.19	434.19	619.77
1 人当たり生産性	労働生産性	11251560.00	8945780.00	11583636.00	6853474.00	9288035.00	10323577.05
	従業員一人当たり人件費	7523292.00	5255770.00	7171037.50	6812240.00	7130664.00	6624750.35
	従業員一人当たり売上高	17750390.00	32836250.00	23967620.00	29193460.00	36056290.00	25127131.15
安全性	固定比率	99.10	76.55	61.88	57.26	68.10	51.70
	固定長期適合率	47.56	76.55	46.03	32.88	38.67	30.69
	当座比率	90.78	114.27	247.50	388.75	375.68	501.46
	流動比率	148.33	114.45	247.79	416.64	403.38	532.15
	自己資本比率	23.01	38.14	54.49	47.37	47.24	51.15
C/F 指標	債務償還年数	13.33	3.03	2.11	999.99	23.28	4.00
成長性	売上総利益増加率	24.79	-20.49	-13.67	-37.40	35.52	16.99
	売上高増加率	28.04	84.98	-51.33	-46.49	23.50	-26.64
	経常利益増加率	10.66	628.31	8.82	-179.62	-126.05	626.09

注）() 内の数字は分類されたクラスタ番号を指す。複数数字が記載されているものは，ソフトクラスタリングにおいて単一のクラスタ所属確率が極端に高くならなかったものである。

表 25　GMM による分類の複数年度の類似性から選択した類似企業の比較分析（サンプル B）

指標	財務比率	2019_B(4)	2020_B(4)	2021_B(4)	2019_類似(4)	2020_類似(4)	2021_類似(4)
収益性	売上高営業利益率	0.98	1.79	0.80	14.08	9.34	10.39
	売上高経常利益率	0.55	1.02	0.31	13.64	11.04	10.35
	売上高総利益率	10.96	17.14	10.44	59.78	61.18	50.72
	売上高金融損益比率	0.66	0.94	0.98	0.51	1.28	0.86
	総資本営業利益率	2.11	2.16	1.38	21.03	5.54	11.64
	総資本経常利益率	1.18	1.23	0.54	20.38	6.55	11.59
	自己資本当期利益率	8.70	10.73	4.76	31.13	17.88	17.18
効率性	売上債権回転期間	38.15	38.19	31.54	0.00	0.00	0.00
	棚卸資産回転期間	87.68	204.78	123.00	188.91	345.32	172.48
	総資本回転期間	169.49	303.25	212.88	244.36	615.06	325.94
1 人当たり生産性	労働生産性	33019980.00	38560560.00	34748820.00	24753050.00	20278930.00	21557880.00
	従業員一人当たり人件費	10904340.00	13654300.00	12964760.00	9031714.00	8666040.00	8706670.00
	従業員一人当たり売上高	301066700.00	224858900.00	332731100.00	41404600.00	33145730.00	42496410.00
安全性	固定比率	54.61	52.44	54.29	23.31	17.95	15.08
	固定長期適合率	13.52	10.30	8.99	22.75	10.08	8.32
	当座比率	64.26	35.47	57.07	22.63	40.05	313.51
	流動比率	163.59	172.79	205.53	169.17	188.45	899.69
	自己資本比率	10.49	8.80	8.89	46.10	27.86	49.52
C/F 指標	債務償還年数	8.56	6.12	22.41	0.00	1.12	0.00
成長性	売上総利益増加率	5.33	16.77	-9.88	9.53	-18.07	6.30
	売上高増加率	-24.83	-25.31	47.97	0.05	-19.94	28.21
	経常利益増加率	-36.09	39.11	-54.45	208.75	-35.18	20.17

注）() 内の数字は分類されたクラスタ番号を指す。複数数字が記載されているものは，ソフトクラスタリングにおいて単一のクラスタ所属確率が極端に高くならなかったものである。

表 22 や表 23 のような単年度での類似性から類似企業を選択したものについては，財務比率は年度によって大きく変化することがあるため，比較に適した企業であるのか，またどこに注目して分析すれば良いのかが分かりにくい。対して，表 24 のように時系列的な企業分類の遷移が類似している企業を選択すれば，先述のクラスタの解釈を用いて，意味のある分析をすることが可能になりそうである。

まず，サンプル A に関しての分析についてである。サンプル A とその類似企業との比較分析において参考にするクラスタはクラスタ 2,3 である。クラスタ 2 は中堅収益・高生産性企業群，クラスタ 3 は高収益企業群であった。どちらの企業も年々収益性指標が向上しており，2020 年には高収益企業群に分類されている。2019 年度のサンプル A に注目してみると，従業員一人当たりの人件費が低く，一人当たり売上高がかなり高くなっている。また，労働生産性が前年より下がり，売上総利益増加率がマイナスであることから，原価にかなりの金額をかけて高い売上を得たことが予想される。その結果翌年には高収益企業群に分類されている。対して，類似企業は収益性指標とともに労働生産性も向上していっており，売上総利益率も向上している。しかし，売上高増加率が 2020 年においてマイナスになっていることから，類似企業は売上高の増加よりも売上原価を減少させることによって労働生産性を向上させていることが読み取れる。

次に，サンプル B に関しての分析についてである。サンプル B とその類似企業との比較分析において参考にするクラスタはクラスタ 4 である。クラスタ 4 は高収益・高生産企業群であり，今回の分類では最も優良とされる分類である。どちらの企業も 2019 年度から 2021 年度までクラスタ 4 に分類されており，労働生産性や従業員一人当たり売上高は継続して高い値となっている。しかし，売上高総利益率を見てみると，サンプル B が 10%程度なのに対し，類似企業は 50%超となっており，かなり違いが出ている。また，サンプル B は売上高増加率が増えると売上総利益増加率が減少しているが，類似企業はどちらも同じような動きをしている。このことから，収益性・生産性指標では類似していると見られる企業であっても，売上高がそのまま利益に直結する場合と，原価がかかる場合と様々なパターンがあった。

以上のことから，本研究で提案した企業分類を用いた財務分析では，労働生産性に関連する収益性・生産性指標が類似している企業であっても，業務形態によって向上の要因が様々であることがわかった。業務形態が近い企業で比較したい場合には，今回提案した企業分類に加えて，既存の産業分類を併用することが考えられる。それによって，ある業務形態のなかで収益性・生産性指標が類似している企業間での比較が可能となり，労働生産性向上の要因が見える可能性がある。

予測モデルについて

「次年度の労働生産性の方向性」を予測するモデルを構築するにあたって，全データをそのまま用いた場合と，業種別モデルを用いた場合，提案手法による企業分類から得られる情報を用いた場合の 3 つで実験を行ったところ，企業を分類することによって予測精度が向上する場合があることが確かめられた。業種別モデルでいうと，建設業や不動産業がより大幅な精度向上が見られた。これらは，今回用いたデータの中でもサンプル数が多い業種であるため，業種別にモデルを構築することによってノイズとなっていたほかの業種が排除され，精度が良くなったということが要因の一つとして考えられる。もしくは，建設業や不動産業という業種の財務諸表データには，他の業種よりも 1 年先の目的変数の変化が顕著に表れてい

るということも考えられるだろう。また，本研究の提案手法である企業分類を用いた予測でも精度が向上したものがあることから，財務データを基にした企業分類が予測モデルの構築において効果的である可能性も示された。

今回，データ分割の方法として，時系列に基づいた方法論を採用した。この結果，年度によって予測の精度に差が出ることがわかった。2020 年ごろから猛威を振るった新型コロナウイルスによる影響により，2020 年以降の予測は精度が低下すると予想されたが，予測精度への影響は想像していたよりは小さかった。今回は，１年後の目的変数を予測するモデルであったが，目的変数の変化がどの時期の財務諸表に現れるかについても調査する必要があるだろう。

6. まとめ

本研究では，地域金融機関が保有する中小企業財務諸表データを用いたクラスタリング手法により，分析目的に応じたデータ駆動型の企業分類が得られるかについて検証した。その結果，既存の産業分類に依存しない，特定の財務比率から特徴づけられた企業分類を得ることができた。また，実験を通して，中小企業データから既存のクラスタリングモデルで有用な分類を得るためには，外れ値を適切に処理する必要性が示唆された。なお，中小企業データの外れ値や欠損値に言及している研究としては高橋（2015）が詳しい。今回は，外れ値に対する処理は特に行わず，評価指標をもとに企業のクラスタリングを行った。

クラスタリングには一般的に良く用いられるハードクラスタリング手法である k-means と，企業分類間の関係性を表すことができる可能性がある手法として，ソフトクラスタリング手法である GMM を用いた。アルゴリズムの性質から，今回用いたデータでは GMM を用いたクラスタリングによって有用な分類が得られた。この分類を用いた財務分析では，労働生産性に関連する収益性・生産性指標が類似している企業であっても，業務形態によって向上の要因が様々であることがわかった。 業務形態でさらに絞り込むには，既存の産業分類と併用するか，もしくは売上原価や固定資産回転率などの業務形態の特徴が出やすい指標を分類基準として用いることが考えられる。また，既存の産業形態や提案手法による企業分類を用いた予測モデルの構築では，企業を何らかの基準で分割することによって予測モデルの精度を向上することができる可能性が示された。業種別モデルで用いたのは，大分類の業種であり，その中にも様々な形態の企業が存在する。業種という静的なものではなく，目的変数に対する説明力が高い財務諸表データを分類することができれば，業種別ではサンプルが少数になってしまったような業種に対する精度も改善できるかもしれない。まずは，予測精度の高い建設業や不動産業などの特定の業種に対して，その理由を分析する必要性がある。

今回用いた方法論は，変数選択からクラスタリングまで全てデータに基づいたデータ依存が高い手法となっている。このような手法のメリットとしては，通常の分析では得られないような知見を得ることができる可能性があることが挙げられる。対してこのような手法のデメリットは，機械学習モデルや統計モデルによる結果が必ずしも人間の感覚と一致しないことがあることである。そのため，結果に対して後付けで解釈を行う際に，伝統的なパターンや一般論ではうまく説明できない場合がある。データ駆動型手法のメリットを残しつつ，デメリットを軽減する方法として，ドメイン知識を用いた変数選択ののちに，機械的な手法を適用することが考えられる。今後も発展が見込まれる情報技術の適用について，投資や融資の

ように信用が重視される会計の分野においては，機械的なモデルの出力が妥当であるようなプロセスの開発が重要となるだろう。

参考文献

高橋淳一. 2015.「財務諸表データに対する外れ値処理と信用リスク評価モデリング」

信金中央金庫. 2018.『信金中金月報 第 17 巻 第 9 号（通巻 554 号）』

中小企業庁. 2019.『2019 年度中小企業白書』

中小企業庁. 2022.『2022 年度中小企業白書』

鱒涼稀・秦涼太・今倉暁・櫻井鉄也・岡田幸彦. 2022.「財務諸表データを用いた資金ニーズの見過ごしチェック AI の開発」.『Department of Policy and Planning Sciences Discussion Paper Series』1383.

鱒涼稀・竹田俊彦・今倉暁・櫻井鉄也・岡田幸彦. 2023.「リレーションシップバンキング機能の向上を目的とした中小企業の資金ニーズ判別法とその活用の提案」.『Department of Policy and Planning Sciences Discussion Paper Series』1385.

Amit, R. and Livnat, J. 1990. GROUPING OF CONGLOMERATES BY THEIR SEGMENTS' ECONOMIC ATTRIBUTES: TOWARDS A MORE MEANINGFUL RATIO ANALYSIS. *Journal of Business Finance & Accounting* 17(1): 85-100.

Bangchang, K. N. 2015. A Comparison of Variable Selection by Tabu Search and Stepwise Regression with Multicollinearity Problem. *Journal of Statistical Science and Application* 3(1-2): 16-24.

Bhojraj, S., M. C, Lee and K. D, Oler. 2003. What's my line? A comparison of industry classification schemes for capital market research. *Journal of Accounting Research* 41(5): 745-774.

Brennan, M. J., W. A, Wang and Y, Xia. 2004. Estimation and test of a simple model of intertemporal capital asset pricing. *The Journal of Finance* 59(4): 1743-1776.

Chan, L. K., J, Lakonishok and B, Swaminathan. 2007. Industry classifications and return comovement. *Financial Analysts Journal* 63(6): 56-70.

Chen, S., T, Gao, Y, He and Y, Jin. 2019. Predicting the stock price movement by social media analysis. *Journal of Data Analysis and Information Processing* 7(4): 295-305.

Chen, X., H. Y, Cho, Y, Dou and B, Lev. 2022. Predicting future earnings changes using machine learning and detailed financial data. *Journal of Accounting Research* 60(2): 467-515.

Chong, D. and H. H, Zhu. 2012, December. Firm clustering based on financial statements. In *22nd Workshop on Information Technology and Information Systems (WITS'12)*.

Clarke, R. N. 1989. SICs as delineators of economic markets. *Journal of Business*: 17-31.

Dalziel, M. 2007. A systems-based approach to industry classification. *Research Policy* 36(10): 1559-1574.

Fairfield, P. M., S, Ramnath and L. T, Yohn. 2009. Do industry-level analyses improve forecasts of financial performance? *Journal*

of Accounting Research 47(1): 147-178.

Fama, E. F., and R. K, French. 1997. Industry costs of equity. *Journal of financial economics* 43(2): 153-193.

Fan, J. P., and H. L, Lang. 2000. The measurement of relatedness: An application to corporate diversification. *The Journal of Business* 73(4): 629-660.

Fang, F., K, Dutta and A, Datta. 2013. LDA-based industry classification. *International Conference on Information Systems (ICIS 2013): Reshaping Society Through Information Systems Design* 3 (2013): 2500-2509.

Flannery, M. J., and K. P, Rangan. 2006. Partial adjustment toward target capital structures. *Journal of financial economics* 79(3): 469-506.

Gupta, M. C., and J. R, Huefner. 1972. A cluster analysis study of financial ratios and industry characteristics. *Journal of Accounting Research* 77-95.

Hoberg, G., and G, Phillips. 2016. Text-based network industries and endogenous product differentiation. *Journal of Political Economy* 124(5): 1423-1465.

Kile, O., and E. C, Phillips. 2009. Using industry classification codes to sample high-technology firms: Analysis and recommendations. *Journal of Accounting, Auditing & Finance* 24(1): 35-58.

Lee, C. M., P, Ma and C. C, Wang. 2015. Search-based peer firms: Aggregating investor perceptions through internet co-searches. *Journal of Financial Economics* 116(2): 410-431.

Leiby, B. D., and D. K, Ahner. 2023. Multicollinearity applied stepwise stochastic imputation: a large dataset imputation through correlation-based regression. *Journal of Big Data* 10(1): 1-20.

Lenard, M. J., P, Alam and D, Booth. 2000. An analysis of fuzzy clustering and a hybrid model for the auditor's going concern assessment. *Decision Sciences* 31(4): 861-884.

Mandrekar, J. N. 2010. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology* 5 (9): 1315-1316.

Mensah, M. Y. 1984. An examination of the stationarity of multivariate bankruptcy prediction models: A methodological study. *Journal of accounting research* 380-395.

Ou, J. A., and H. S, Penman. 1989. Financial statement analysis and the prediction of stock returns. *Journal of accounting and economics* 11(4): 295-329.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58(1): 267-288.

Yang, S. Y., F. C, Liu, X, Zhu and D. C, Yen. 2019. A graph mining approach to identify financial reporting patterns: an empirical examination of industry classifications. *Decision Sciences* 50(4): 847-876.

Yanke, A., E. N, Zendrato and A. M, Soleh. 2022. Handling Multicollinearity Problems in Indonesia's Economic Growth Regression Modeling Based on Endogenous Economic Growth Theory: Penanganan Masalah Multikolinieritas pada Pemodelan Pertumbuhan Ekonomi Indonesia Berdasarkan Teori Pertumbuhan Ekonomi Endo. *Indonesian Journal of Statistics and Its*

Applications 6(2): 228–244.

Youden, J. W. 1950. Index for rating diagnostic tests. *Cancer* 3(1): 32-35.

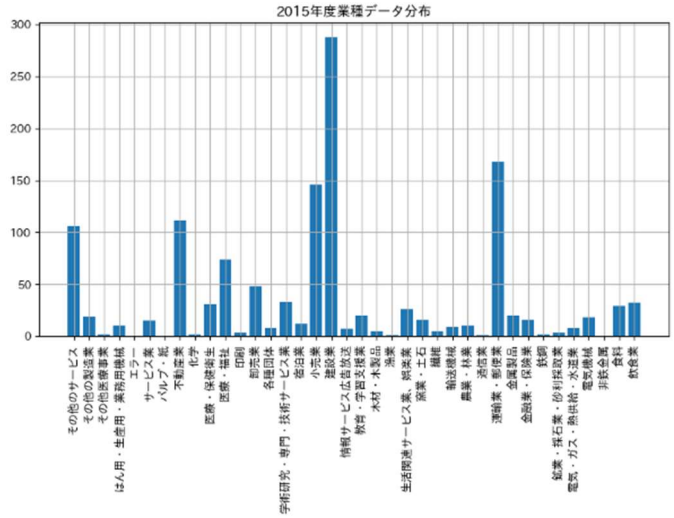
（付記）本研究は，JSPS 科研費（23K22166）による研究成果の一部である。

付録

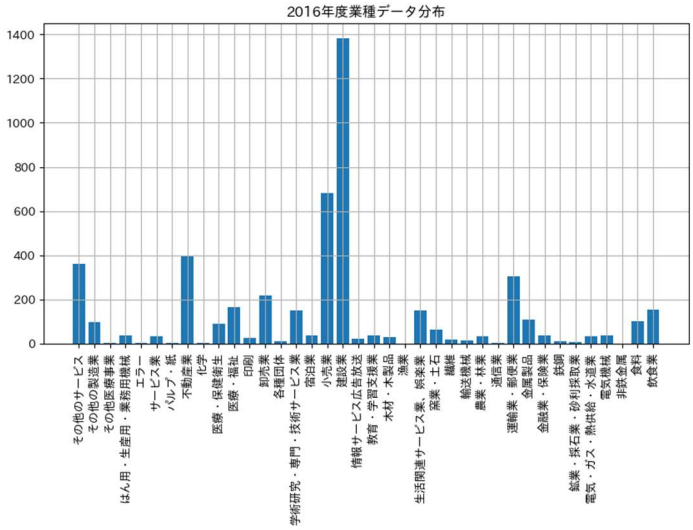
付録 A

ここでは，予測モデルの構築における新分類の有用性検証において行った実験の詳細を記載する。まず，以下に年度別の業種別サンプル分布を示す。

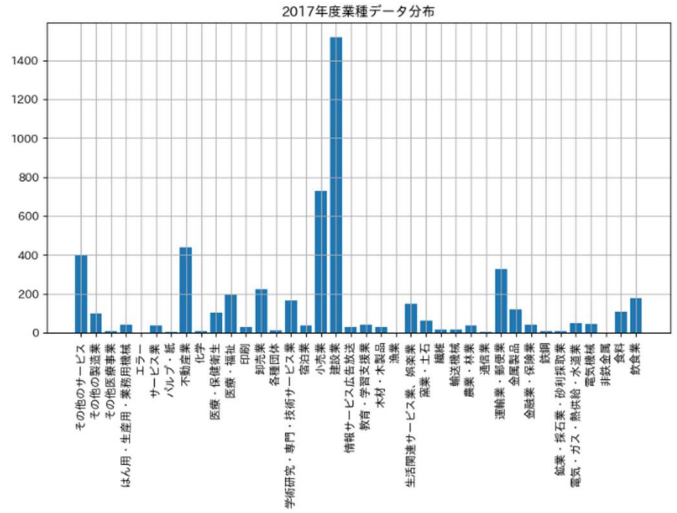
付録図 1　2015 年度業種別サンプル数分布



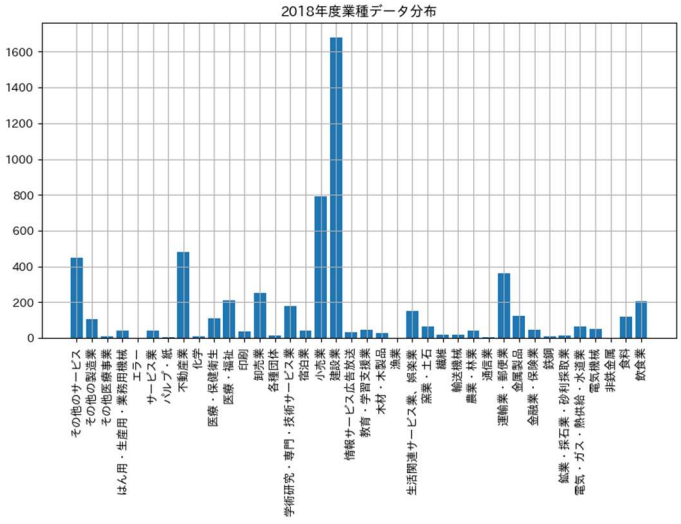
付録図 2　2016 年度業種別サンプル数分布


2016年度業種データ分布
1400
1200
1000
800
600
400
200
0
その他のサービス
その他の製造業
その他医療事業
はん用・生産用・業務用機械
ゴム
サービス業
パルプ・紙
不動産業
化学
医療・保健衛生
医療・福祉
印刷
船舶業
各種団体
学術研究・専門・技術サービス業
宿泊業
小売業
建設業
情報サービス広告放送
教育・学習支援業
木材・木製品
漁業
生活関連サービス業、娯楽業
窯業・土石
繊維
輸送機械
農業・林業
通信業
運輸業・郵便業
金属製品
金融業・保険業
鉄鋼
鉱業・採石業・砂利採取業
電気・ガス・熱供給・水道業
電気機械
非鉄金属
食料
飲食業

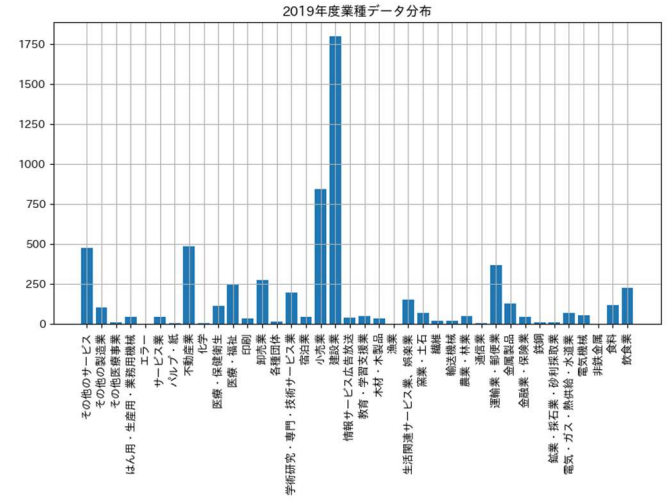
付録図 3　2017 年度業種別サンプル数分布



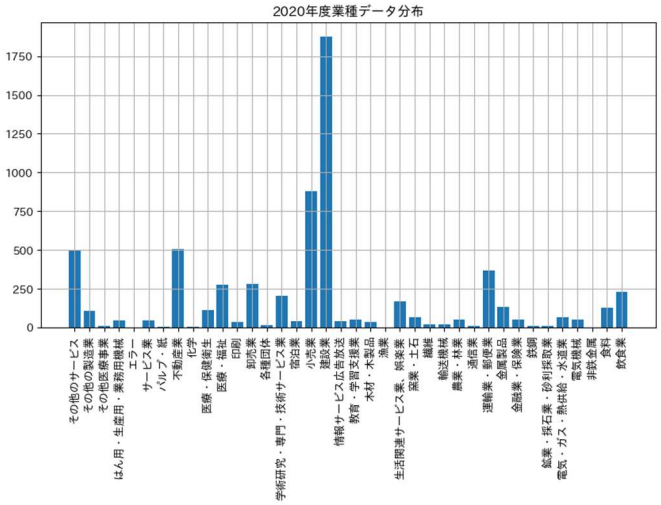
付録図 4　2018 年度業種別サンプル数分布



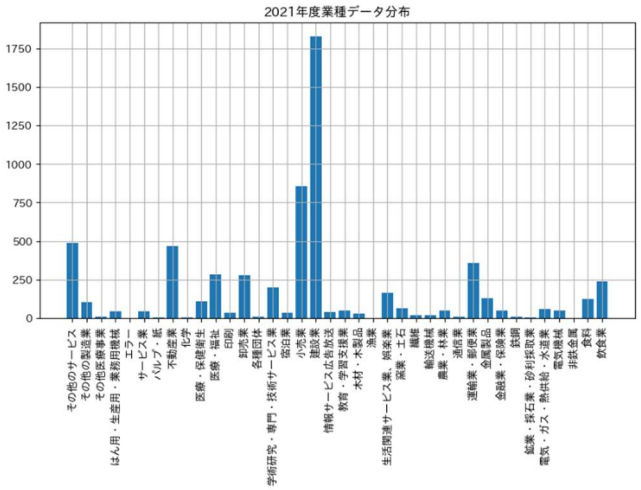
付録図 5　2019 年度業種別サンプル数分布



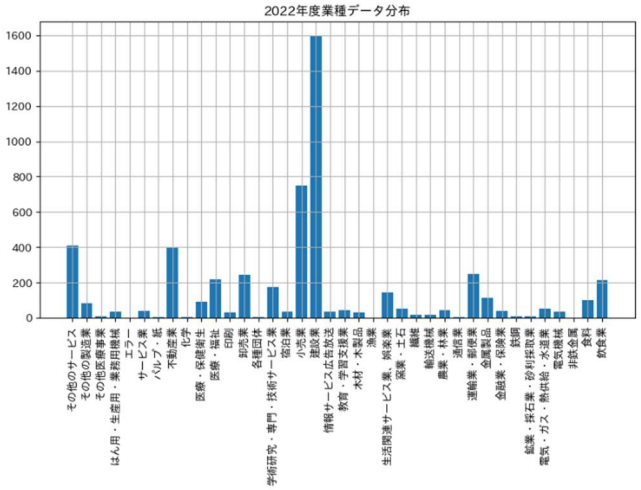
付録図 6　2020 年度業種別サンプル数分布



付録図 7　2021 年度業種別サンプル数分布



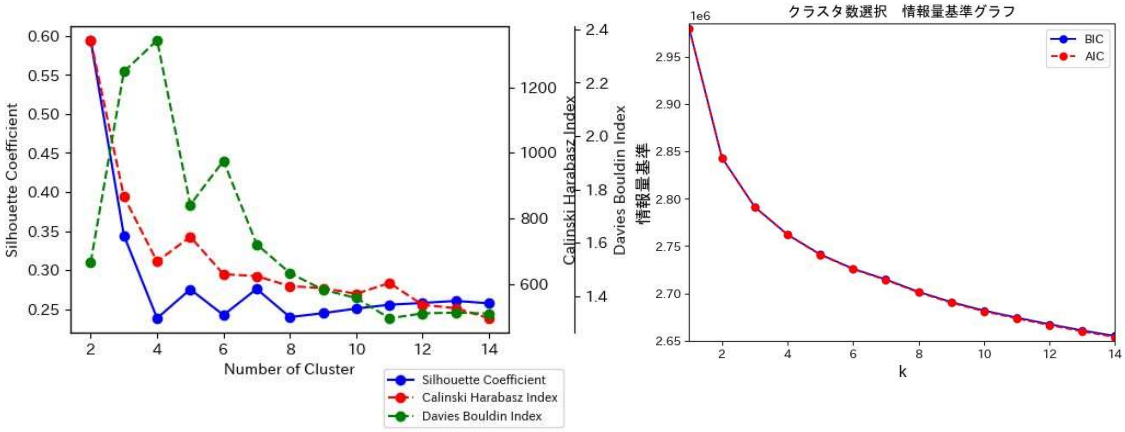
付録図 8　2021 年度業種別サンプル数分布



次に，テストデータの年度ごとに L1 正則化ロジスティック回帰によって選択された財務比率と，それを用いて GMM で分類したときの評価指標を示す。

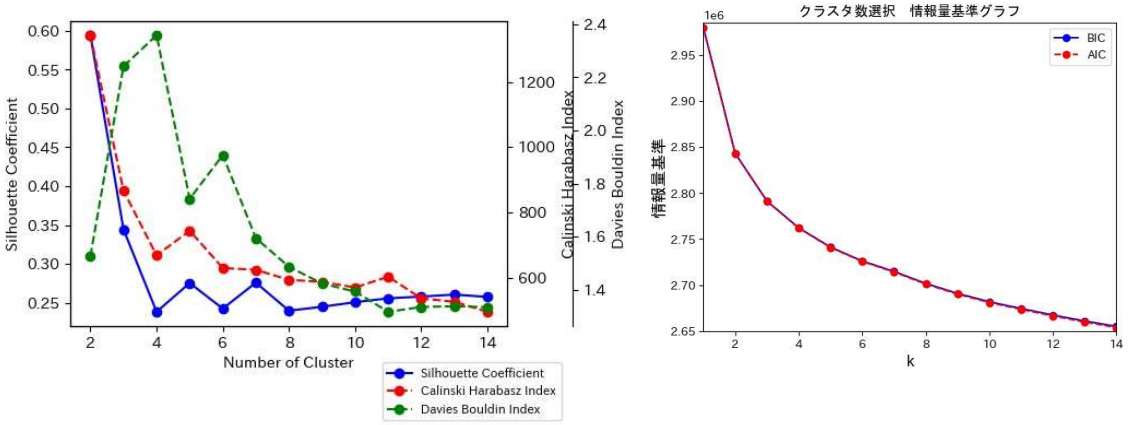
2018 年度のデータをテストデータとした場合には，総資本営業利益率，売上高営業利益率，売上高経常利益率，経常損益比率， 従業員一人当たり経常利益，従業員一人当たり人件費，営業利益増加率，経常利益増加率，負債増加率，債務超過解消年数の 10 個の財務比率が選択された。そしてこれを用いて GMM でサンプルの分類を行ったところ，評価指標は付録図 9 のようになり，これをもとにクラスタ数は 5 とした。

付録図 9　2018 年度データ予測モデル評価指標



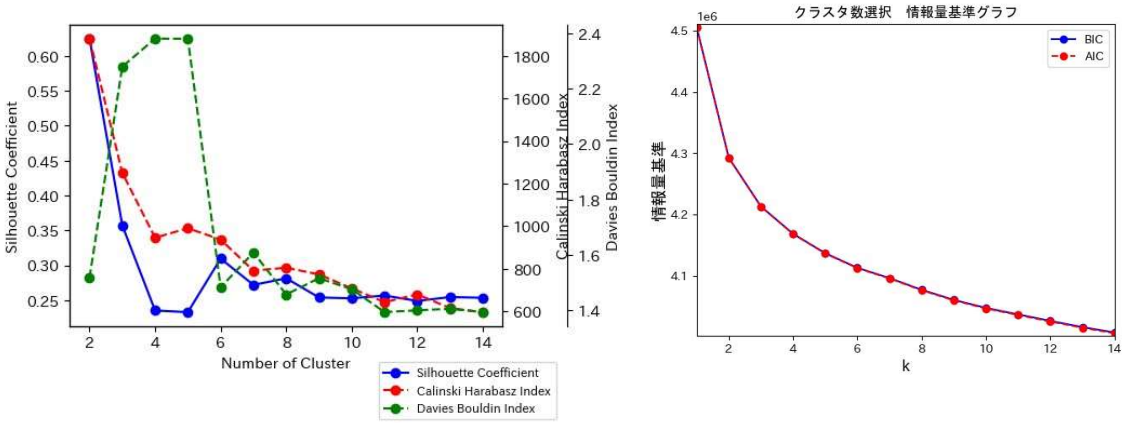
2019 年度のデータをテストデータとした場合には，総資本営業利益率，売上高営業利益率，売上高人件費比率，経常損益比率，従業員一人当たり経常利益，従業員一人当たり人件費，営業利益増加率，経常利益増加率，負債増加率，損益分岐点比率， 債務超過解消年数の 11 個の財務比率が選択された。この場合では，L1 正則化ロジスティック回帰の正則化項係数パラメータの探索幅の関係で，財務比率が 11 個となった。そしてこれを用いて GMM でサンプルの分類を行ったところ，評価指標は付録図 10 のようになり，これをもとにクラスタ数は 6 とした。

付録図 10　2019 年度データ予測モデル評価指標



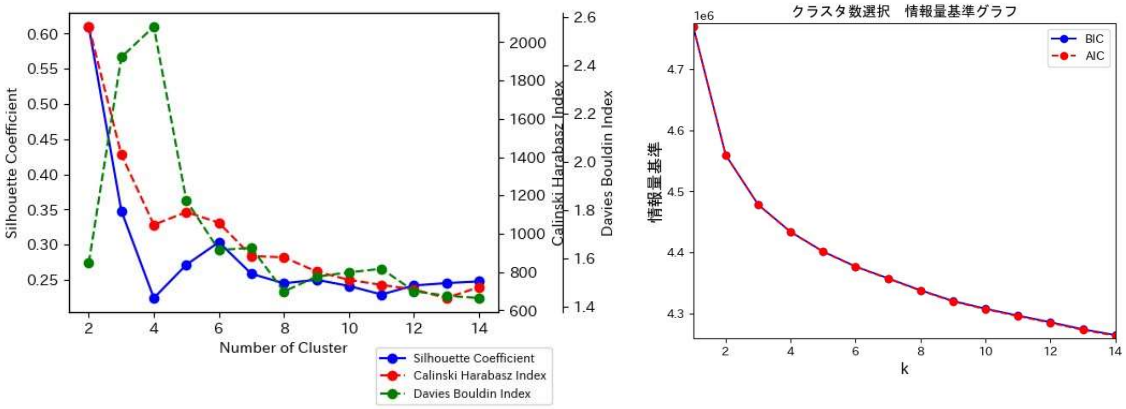
2020 年度のデータをテストデータとした場合には，総資本営業利益率，売上高営業利益率，売上高経常利益率，経常損益比率，従業員一人当たり経常利益，従業員一人当たり人件費，営業利益増加率，負債増加率，損益分岐点比率，債務超過解消年数の 10 個の財務比率が選択された。そしてこれを用いて GMM でサンプルの分類を行ったところ，評価指標は付録図 11 のようになり，これをもとにクラスタ数は 6 とした。

付録図 11　2020 年度データ予測モデル評価指標



2021 年度のデータをテストデータとした場合には，総資本営業利益率，売上高総利益率，売上高営業利益率，経常損益比率，従業員一人当たり経常利益，従業員一人当たり人件費，営業利益増加率，負債増加率，債務超過解消年数，D/EBITDA の 10 個の財務比率が選択された。そしてこれを用いて GMM でサンプルの分類を行ったところ，評価指標は付録図 12 のようになり，これをもとにクラスタ数は 6 とした。

付録図 12　2021 年度データ予測モデル評価指標



次に，業種別モデルによる予測精度を掲載する。本文でも述べたが，業種別にモデルを構築することにより，全サンプルを用いた場合よりも精度が向上したものが見られた。また，nan となっているものは，データ数が不足しているために評価を行うことができなかったものである。

付録表 1　業種別モデル予測精度

テストデータ年度	業種	TRAIN_ACC	TRAIN_AUC	TEST_ACC	TEST_AUC
2018	その他の製造業	0.853	0.931	0.483	0.605
2019	その他の製造業	0.728	0.806	0.448	0.512
2020	その他の製造業	0.767	0.835	0.583	0.590
2021	その他の製造業	0.918	0.959	0.481	0.527
2018	医療・保健衛生	0.798	0.877	0.520	0.557
2019	医療・保健衛生	0.734	0.826	0.530	0.546
2020	医療・保健衛生	0.759	0.815	0.588	0.589
2021	医療・保健衛生	0.758	0.838	0.429	0.561
2018	その他のサービス	0.871	0.943	0.522	0.535
2019	その他のサービス	0.787	0.876	0.559	0.566
2020	その他のサービス	0.892	0.956	0.516	0.553
2021	その他のサービス	0.905	0.970	0.583	0.592
2018	小売業	0.974	0.995	0.541	0.549
2019	小売業	0.890	0.946	0.576	0.605
2020	小売業	0.944	0.984	0.559	0.584
2021	小売業	0.870	0.940	0.566	0.589
2018	農業・林業	0.421	0.500	0.541	0.500
2019	農業・林業	0.934	0.968	0.561	0.557
2020	農業・林業	0.739	0.738	0.477	0.488
2021	農業・林業	0.744	0.809	0.564	0.643
2018	食料	0.919	0.904	0.495	0.467
2019	食料	0.930	0.981	0.560	0.611

2020	食料	0.898	0.967	0.474	0.552
2021	食料	0.912	0.970	0.567	0.555
2018	飲食業	0.881	0.936	0.539	0.571
2019	飲食業	0.901	0.962	0.607	0.595
2020	飲食業	0.903	0.967	0.609	0.643
2021	飲食業	0.758	0.853	0.482	0.493
2018	不動産業	0.855	0.934	0.559	0.563
2019	不動産業	0.893	0.946	0.613	0.673
2020	不動産業	0.944	0.986	0.581	0.630
2021	不動産業	0.971	0.995	0.598	0.610
2018	金属製品	0.764	0.794	0.508	0.603
2019	金属製品	0.886	0.947	0.553	0.598
2020	金属製品	0.860	0.933	0.603	0.651
2021	金属製品	0.894	0.963	0.634	0.646
2018	金融業・保険業	0.500	0.500	0.500	0.500
2019	金融業・保険業	0.545	0.500	0.286	0.500
2020	金融業・保険業	0.533	0.500	0.600	0.500
2021	金融業・保険業	0.400	0.500	0.200	0.500
2018	窯業・土石	0.784	0.812	0.467	0.486
2019	窯業・土石	0.814	0.875	0.410	0.409
2020	窯業・土石	0.849	0.912	0.476	0.467
2021	窯業・土石	0.934	0.989	0.566	0.622
2018	卸売業	0.728	0.817	0.475	0.487
2019	卸売業	0.918	0.969	0.514	0.519
2020	卸売業	0.868	0.930	0.579	0.581
2021	卸売業	0.762	0.842	0.475	0.496
2018	はん用・生産用・業務用機械	0.689	0.702	0.342	0.338
2019	はん用・生産用・業務用機械	0.733	0.773	0.581	0.513
2020	はん用・生産用・業務用機械	0.711	0.787	0.500	0.578
2021	はん用・生産用・業務用機械	0.728	0.764	0.529	0.580
2018	教育・学習支援業	0.636	0.500	0.583	0.500
2019	教育・学習支援業	0.553	0.500	0.414	0.500
2020	教育・学習支援業	0.773	0.798	0.452	0.464
2021	教育・学習支援業	0.736	0.811	0.640	0.649
2018	電気機械	0.796	0.839	0.426	0.442
2019	電気機械	0.807	0.873	0.490	0.442
2020	電気機械	0.720	0.770	0.540	0.490
2021	電気機械	0.724	0.803	0.618	0.672

2018	運輸業・郵便業	0.910	0.956	0.544	0.584
2019	運輸業・郵便業	0.964	0.993	0.564	0.571
2020	運輸業・郵便業	0.852	0.925	0.581	0.595
2021	運輸業・郵便業	0.848	0.926	0.534	0.513
2018	建設業	0.914	0.975	0.603	0.640
2019	建設業	0.880	0.950	0.621	0.673
2020	建設業	0.884	0.946	0.610	0.658
2021	建設業	0.830	0.909	0.602	0.661
2018	学術研究・専門・技術サービス業	0.748	0.787	0.487	0.446
2019	学術研究・専門・技術サービス業	0.850	0.898	0.569	0.594
2020	学術研究・専門・技術サービス業	0.871	0.945	0.559	0.589
2021	学術研究・専門・技術サービス業	0.846	0.912	0.544	0.573
2018	繊維	0.524	0.500	0.500	0.500
2019	繊維	0.419	0.500	0.526	0.500
2020	繊維	0.484	0.500	0.600	0.500
2021	繊維	0.514	0.500	0.333	0.500
2018	鉱業・採石業・砂利採取業	0.583	0.500	0.455	0.500
2019	鉱業・採石業・砂利採取業	0.737	0.500	0.667	0.500
2020	鉱業・採石業・砂利採取業	0.619	0.500	0.500	0.500
2021	鉱業・採石業・砂利採取業	0.550	0.500	0.250	0.500
2018	医療・福祉	0.810	0.879	0.535	0.548
2019	医療・福祉	0.827	0.897	0.597	0.609
2020	医療・福祉	0.854	0.920	0.601	0.620
2021	医療・福祉	0.956	0.992	0.539	0.549
2018	木材・木製品	0.419	0.500	0.556	0.500
2019	木材・木製品	0.885	0.922	0.548	0.397
2020	木材・木製品	0.769	0.782	0.613	0.609
2021	木材・木製品	0.741	0.753	0.407	0.372
2019	各種団体	0.500	0.500	0.800	0.500
2020	各種団体	0.500	0.500	0.000	nan
2021	各種団体	0.778	0.500	0.250	0.500
2018	その他医療事業	0.375	0.500	0.857	0.500
2019	その他医療事業	0.357	0.500	0.571	0.500
2020	その他医療事業	0.600	0.500	0.625	0.500

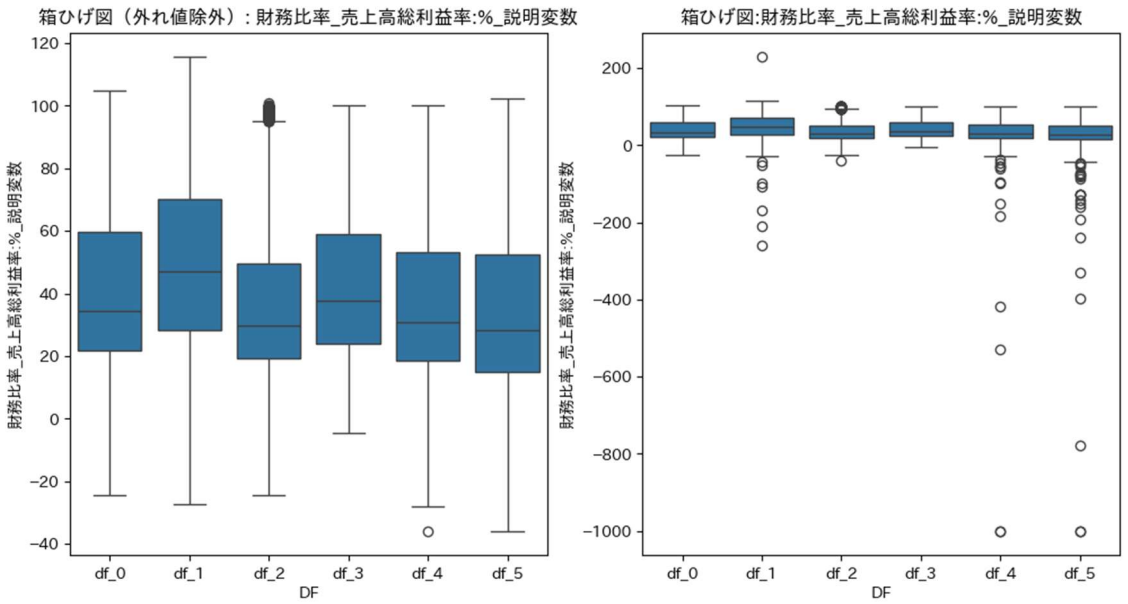
2021	その他医療事業	0.714	0.500	0.429	0.500
2018	宿泊業	0.800	0.803	0.500	0.489
2019	宿泊業	0.696	0.690	0.590	0.510
2020	宿泊業	0.905	0.946	0.622	0.612
2021	宿泊業	0.924	0.971	0.294	0.322
2018	情報サービス広告放送	0.481	0.500	0.444	0.500
2019	情報サービス広告放送	0.778	0.783	0.531	0.578
2020	情報サービス広告放送	0.765	0.829	0.594	0.573
2021	情報サービス広告放送	0.881	0.908	0.538	0.683
2018	印刷	0.464	0.500	0.559	0.500
2019	印刷	0.727	0.720	0.432	0.418
2020	印刷	0.906	0.951	0.649	0.638
2021	印刷	0.775	0.828	0.613	0.592
2019	化学	0.538	0.500	0.286	0.500
2020	化学	0.714	0.500	0.500	0.500
2021	化学	0.500	0.500	0.167	0.500
2018	輸送機械	0.500	0.500	0.333	0.500
2019	輸送機械	0.516	0.500	0.706	0.500
2020	輸送機械	0.441	0.500	0.222	0.500
2021	輸送機械	0.514	0.500	0.563	0.500
2018	生活関連サービス業，娯楽業	0.822	0.904	0.484	0.486
2019	生活関連サービス業，娯楽業	0.849	0.904	0.493	0.564
2020	生活関連サービス業，娯楽業	0.833	0.890	0.527	0.565
2021	生活関連サービス業，娯楽業	0.712	0.801	0.488	0.433
2018	鉄鋼	0.583	0.500	0.333	0.500
2019	鉄鋼	0.474	0.500	0.556	0.500
2020	鉄鋼	0.333	0.500	0.545	0.500
2021	鉄鋼	0.444	0.500	0.500	0.500
2019	通信業	0.625	0.500	0.250	0.500
2020	通信業	0.500	0.500	0.500	0.500
2021	通信業	0.500	0.500	0.333	0.500
2018	電気・ガス・熱供給・水道業	0.250	0.500	0.550	0.500

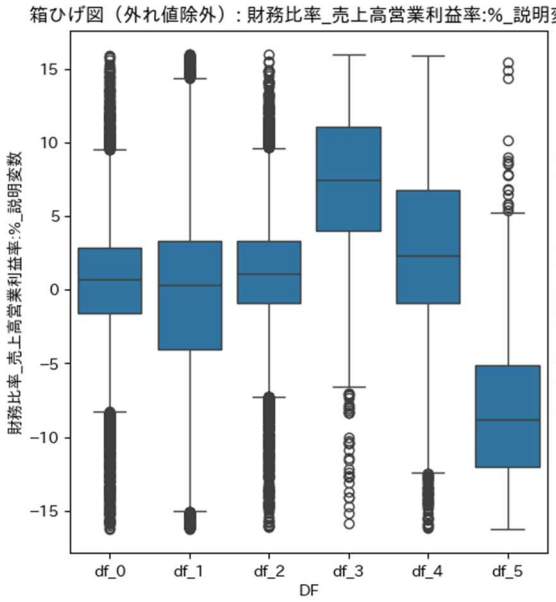
2019	電気・ガス・熱供給・水道業	0.433	0.500	0.500	0.500
2020	電気・ガス・熱供給・水道業	0.526	0.500	0.370	0.500
2021	電気・ガス・熱供給・水道業	0.773	0.774	0.478	0.492
2018	サービス業	0.553	0.500	0.472	0.500
2019	サービス業	0.860	0.892	0.474	0.530
2020	サービス業	0.692	0.700	0.462	0.448
2021	サービス業	0.743	0.768	0.486	0.434
2019	パルプ・紙	0.222	0.500	0.571	0.500
2020	パルプ・紙	0.538	0.500	0.714	0.500
2021	パルプ・紙	0.714	0.500	0.500	0.500

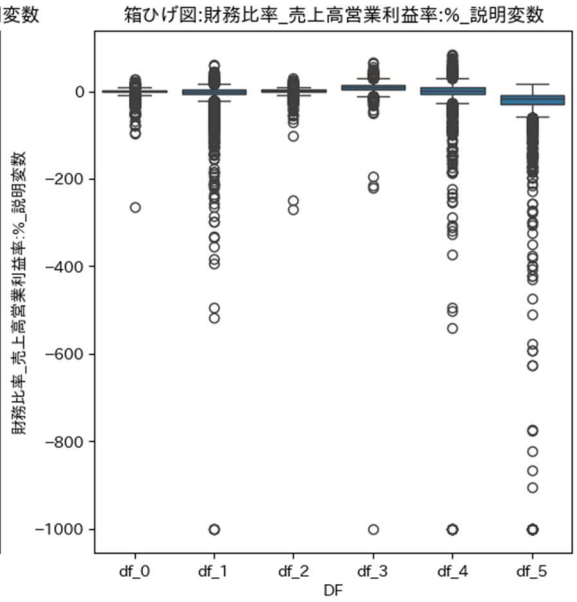
付録 B

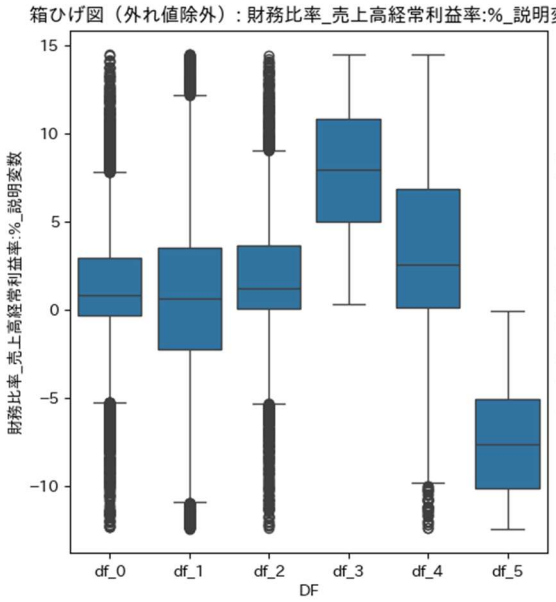
クラスタの性質を把握するために利用した箱ひげ図を以下に掲載する。左図が外れ値を除去したものであり，右図が外れ値除去前のものである。df_0 がクラスタ 0 を表しており，df_1…についても同様である。

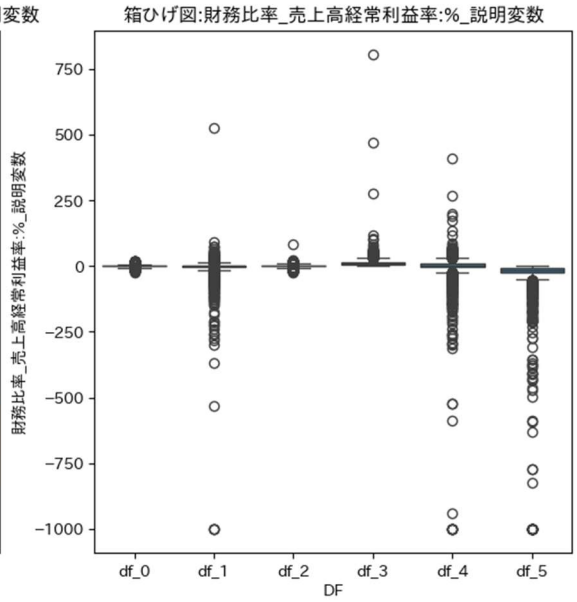
付録図 13　財務比率のクラスタ毎の箱ひげ図

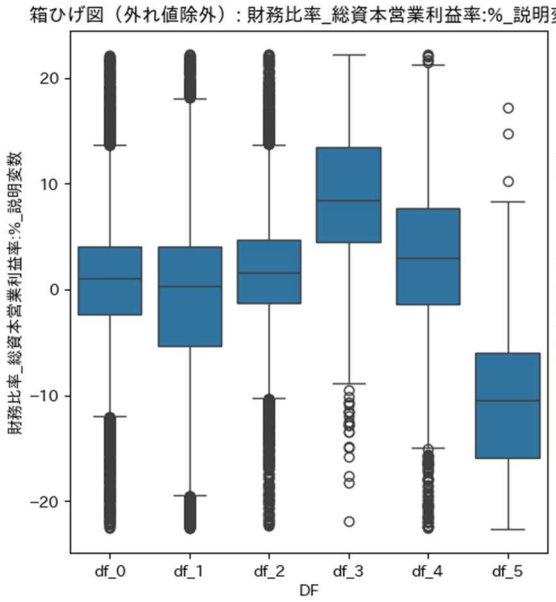


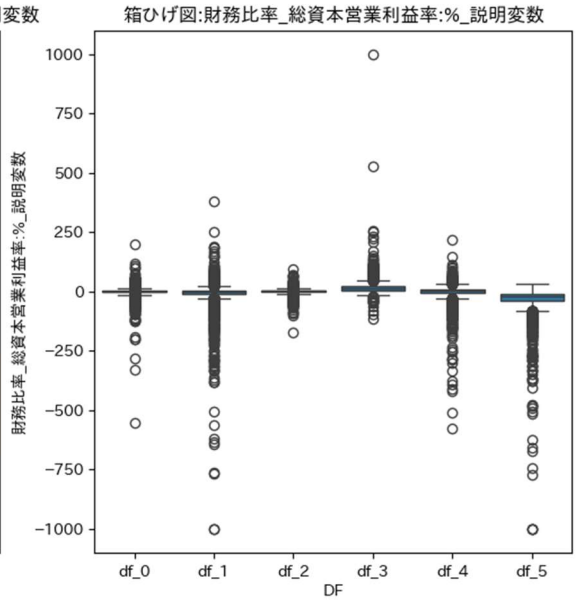

箱ひげ図（外れ値除外）：財務比率_売上高営業利益率:%_説明変数
財務比率_売上高営業利益率:%_説明変数
15
10
5
0
−5
−10
−15
df_0
df_1
df_2
df_3
df_4
df_5
DF

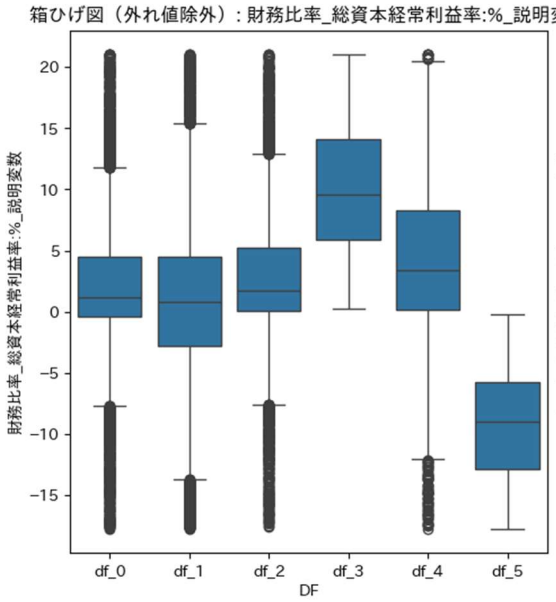

箱ひげ図:財務比率_売上高営業利益率:%_説明変数
財務比率_売上高営業利益率:%_説明変数
0
−200
−400
−600
−800
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF

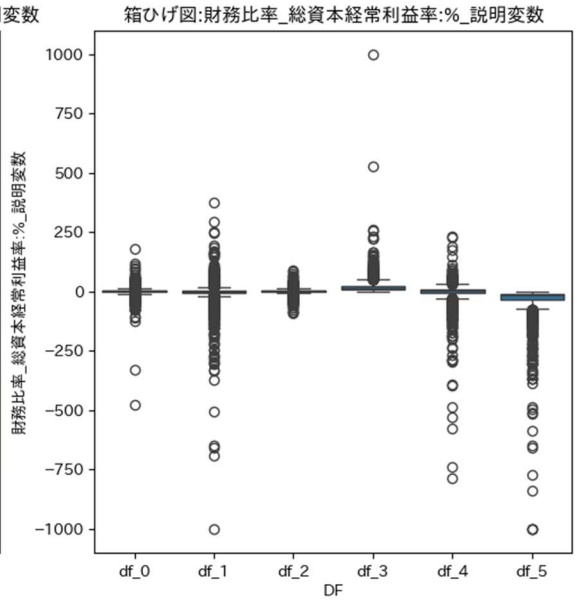

箱ひげ図（外れ値除外）：財務比率_売上高経常利益率:%_説明変数
財務比率_売上高経常利益率:%_説明変数
15
10
5
0
−5
−10
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_売上高経常利益率:%_説明変数
財務比率_売上高経常利益率:%_説明変数
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF

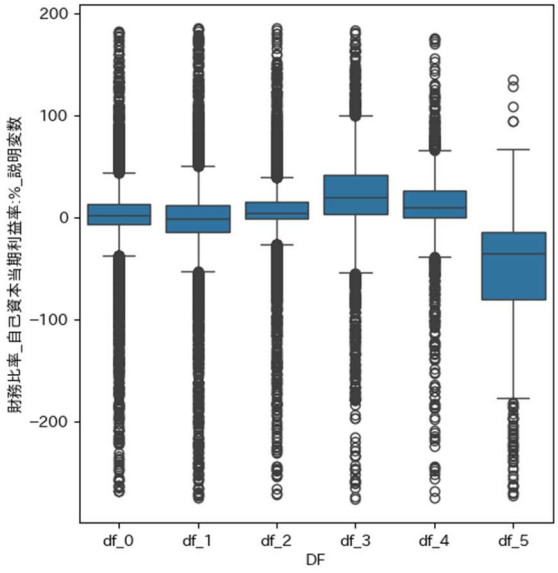

箱ひげ図（外れ値除外）：財務比率_総資本営業利益率:%_説明変数
財務比率_総資本営業利益率:%_説明変数
20
10
0
−10
−20
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_総資本営業利益率:%_説明変数
財務比率_総資本営業利益率:%_説明変数
1000
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF

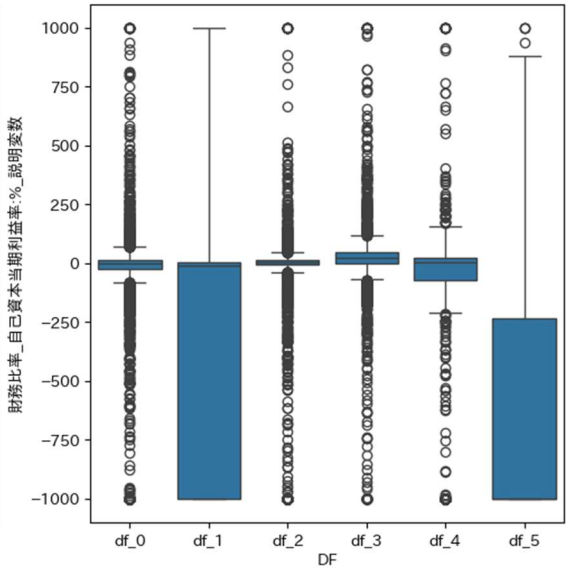

箱ひげ図（外れ値除外）：財務比率_総資本経常利益率:%_説明変数
財務比率_総資本経常利益率:%_説明変数
20
15
10
5
0
−5
−10
−15
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_総資本経常利益率:%_説明変数
財務比率_総資本経常利益率:%_説明変数
1000
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF

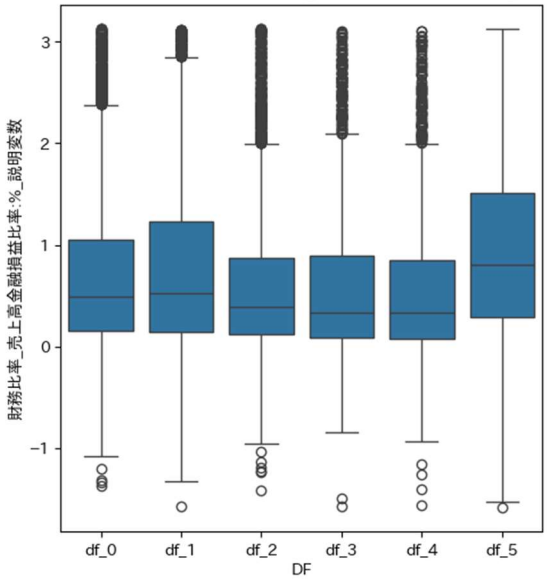
箱ひげ図（外れ値除外）：財務比率_自己資本当期利益率:%_説明変数



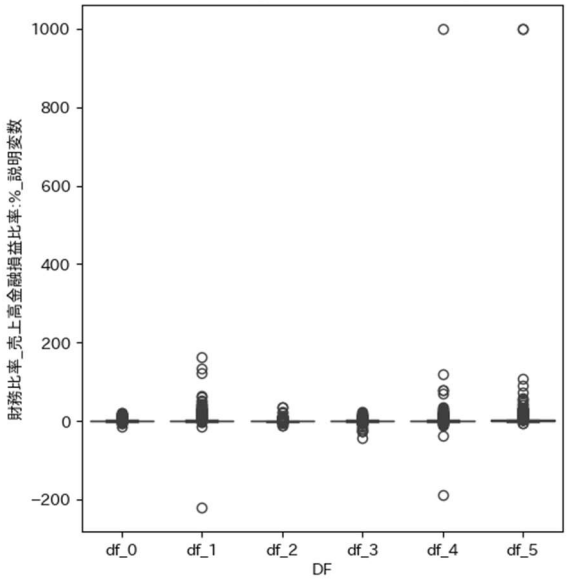
箱ひげ図:財務比率_自己資本当期利益率:%_説明変数

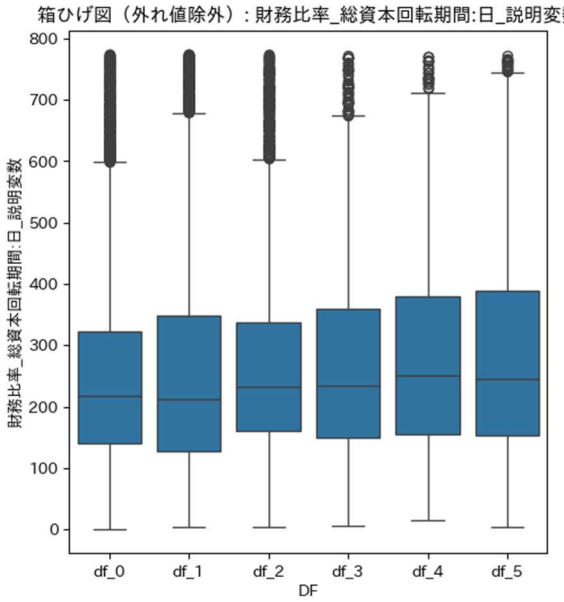


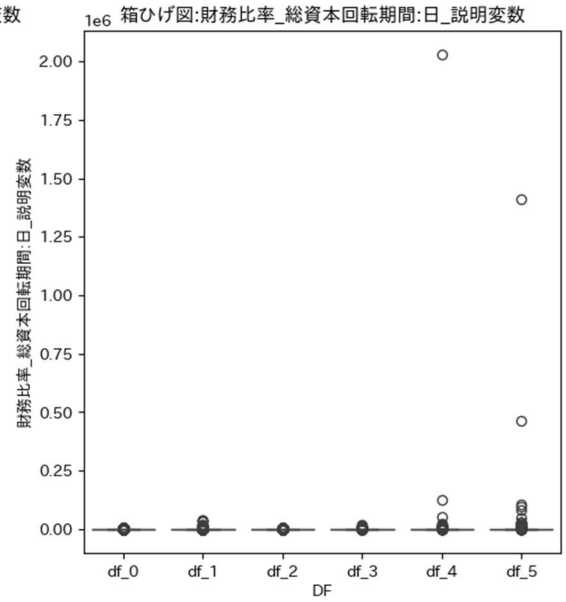
箱ひげ図（外れ値除外）：財務比率_売上高金融損益比率:%_説明変数

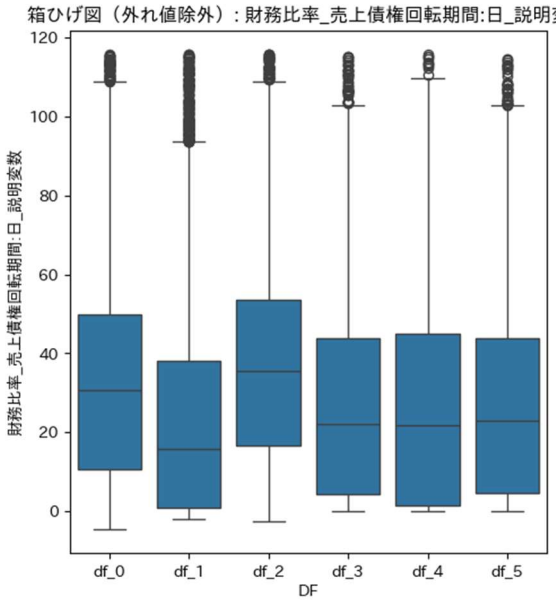


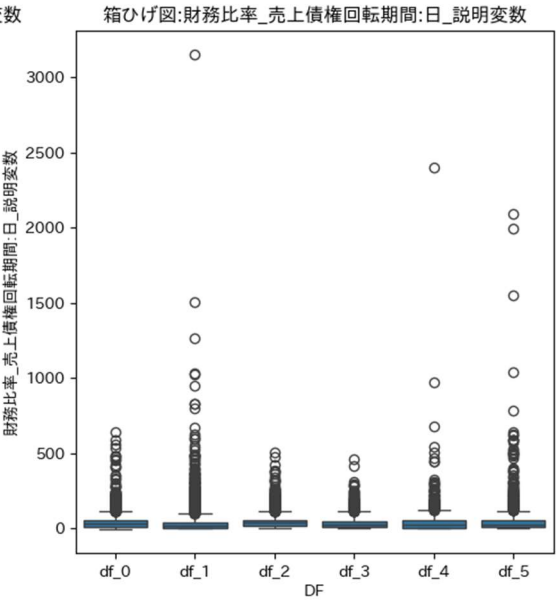
箱ひげ図:財務比率_売上高金融損益比率:%_説明変数

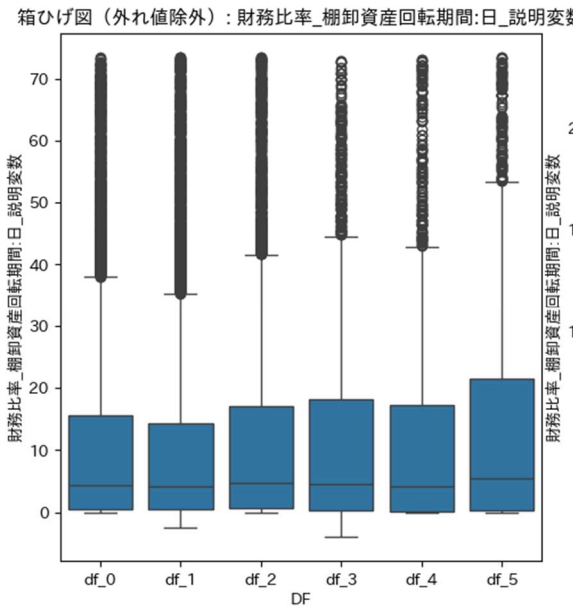


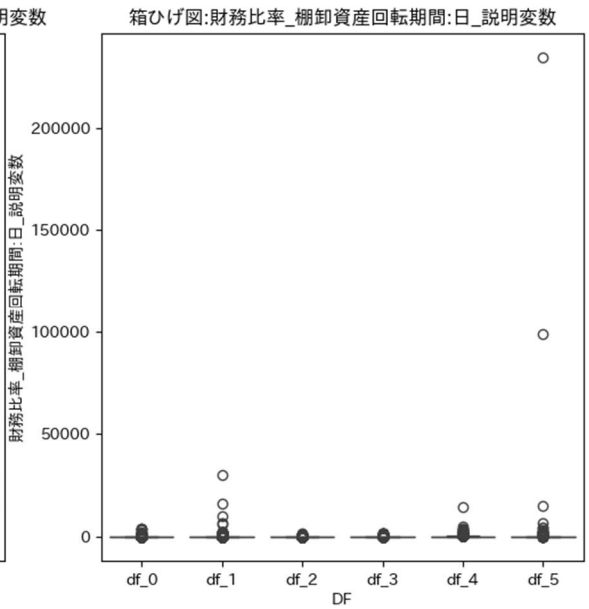

箱ひげ図（外れ値除外）：財務比率_総資本回転期間:日_説明変数
財務比率_総資本回転期間:日_説明変数
800
700
600
500
400
300
200
100
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

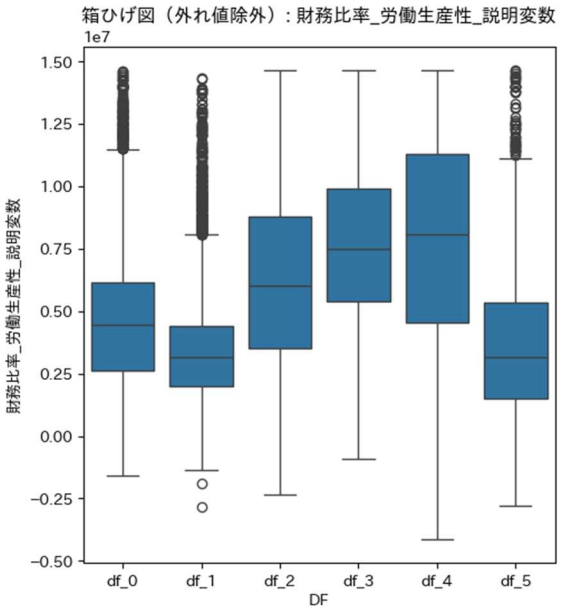

箱ひげ図:財務比率_総資本回転期間:日_説明変数
1e6
財務比率_総資本回転期間:日_説明変数
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
df_0
df_1
df_2
df_3
df_4
df_5
DF

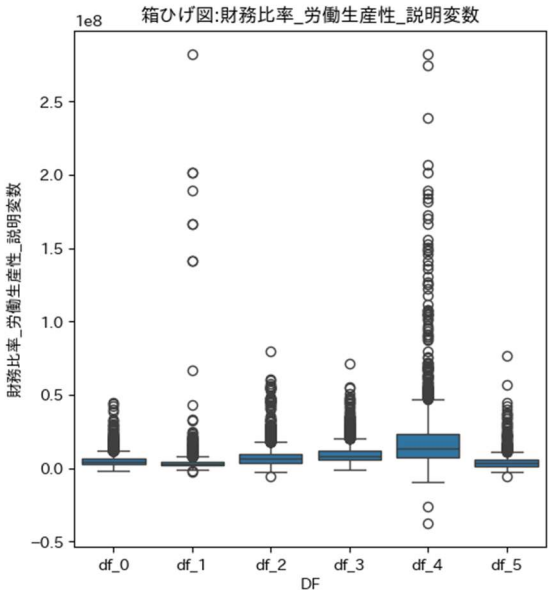

箱ひげ図（外れ値除外）：財務比率_売上債権回転期間:日_説明変数
財務比率_売上債権回転期間:日_説明変数
120
100
80
60
40
20
0
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_売上債権回転期間:日_説明変数
財務比率_売上債権回転期間:日_説明変数
3000
2500
2000
1500
1000
500
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

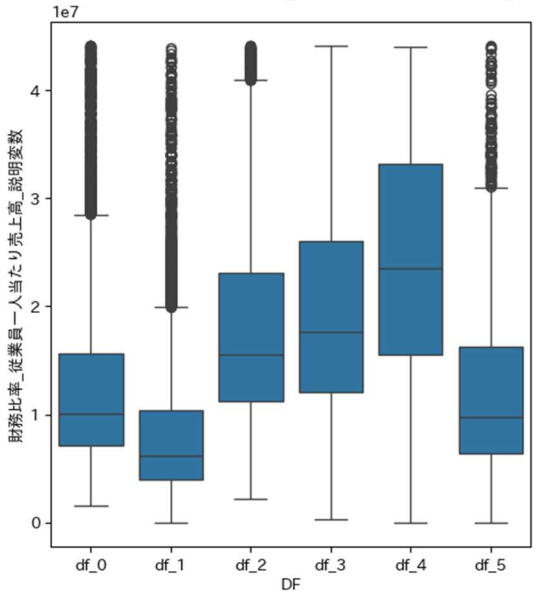

箱ひげ図（外れ値除外）：財務比率_棚卸資産回転期間:日_説明変数
財務比率_棚卸資産回転期間:日_説明変数
70
60
50
40
30
20
10
0
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_棚卸資産回転期間:日_説明変数
財務比率_棚卸資産回転期間:日_説明変数
200000
150000
100000
50000
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

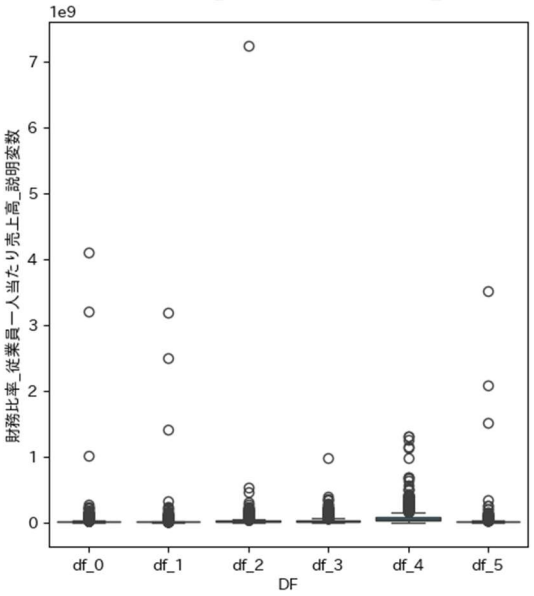

箱ひげ図（外れ値除外）：財務比率_労働生産性_説明変数
1e7
財務比率_労働生産性_説明変数
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_労働生産性_説明変数
1e8
財務比率_労働生産性_説明変数
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
df_0
df_1
df_2
df_3
df_4
df_5
DF

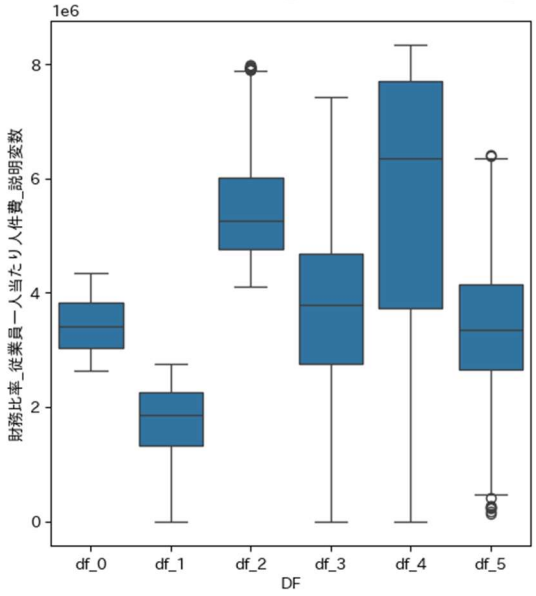
箱ひげ図（外れ値除外）：財務比率_従業員一人当たり売上高_説明変数



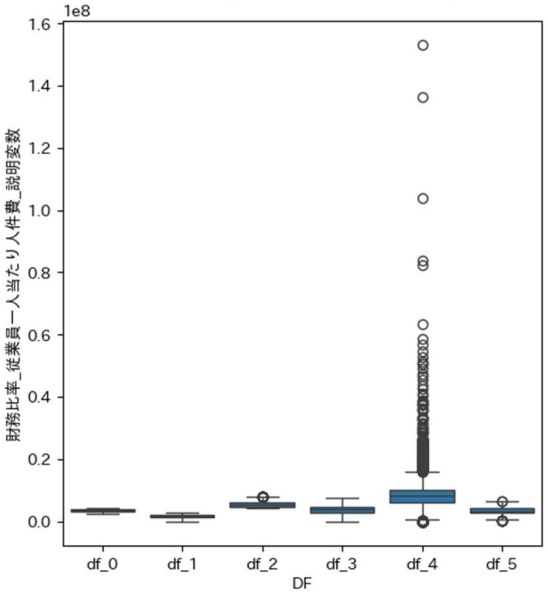
箱ひげ図:財務比率_従業員一人当たり売上高_説明変数

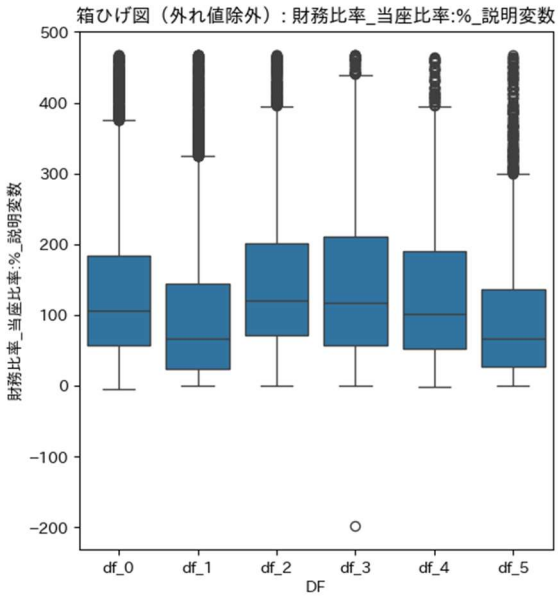


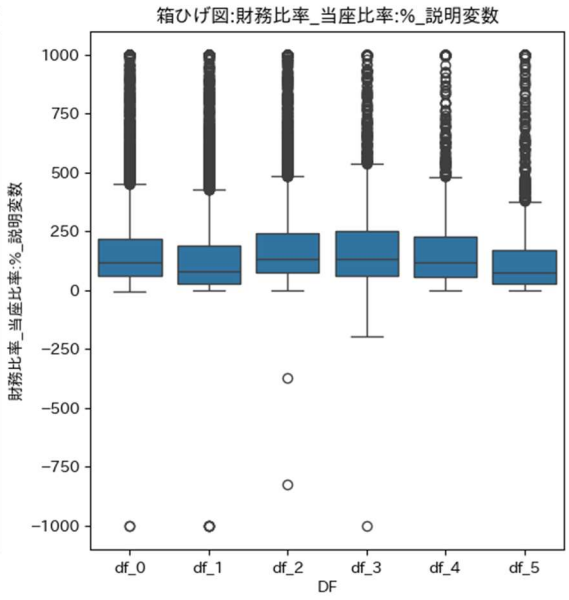
箱ひげ図（外れ値除外）：財務比率_従業員一人当たり人件費_説明変数

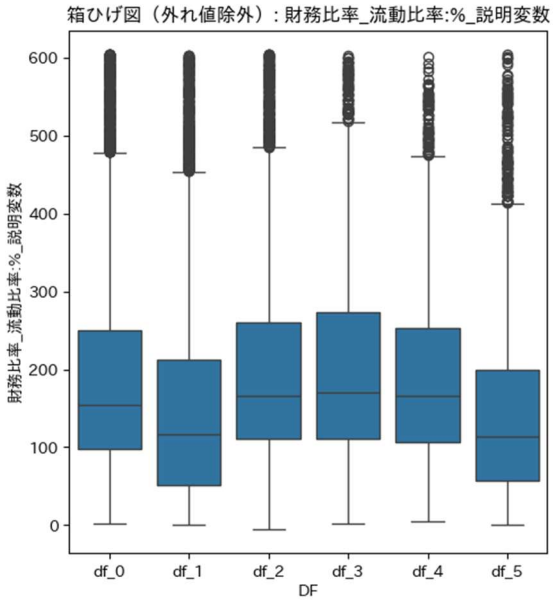


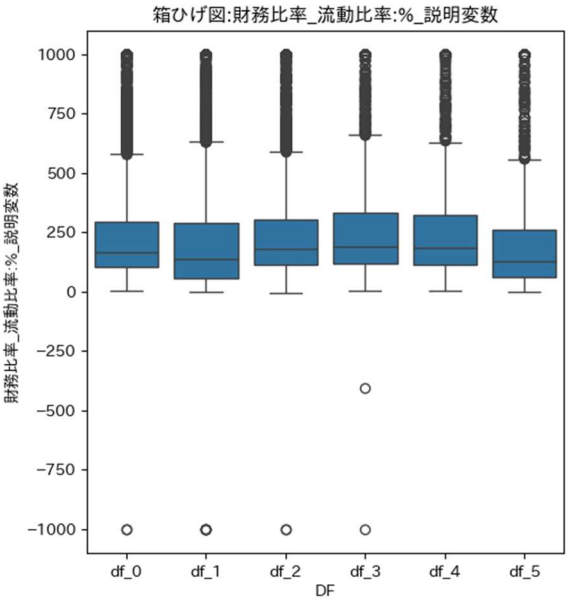
箱ひげ図:財務比率_従業員一人当たり人件費_説明変数

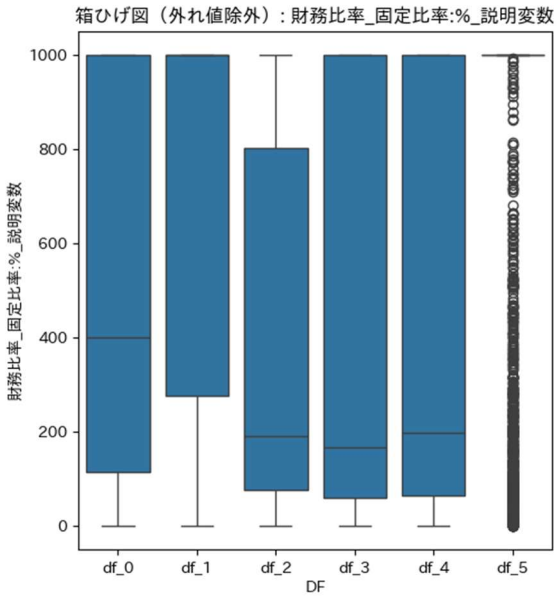


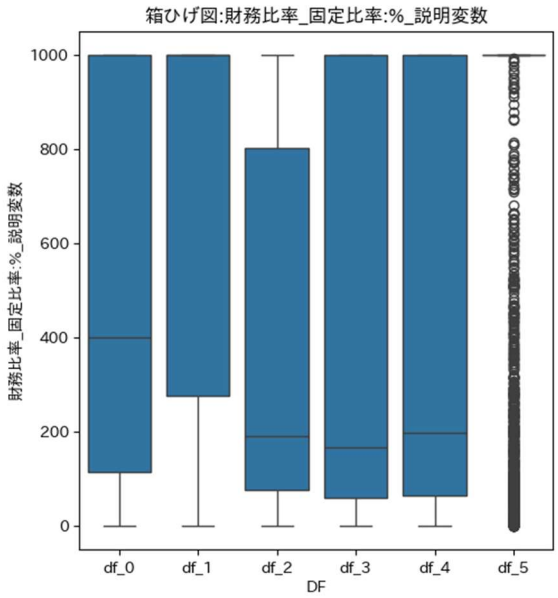

箱ひげ図（外れ値除外）：財務比率_当座比率:%_説明変数
財務比率_当座比率:%_説明変数
500
400
300
200
100
0
−100
−200
df_0
df_1
df_2
df_3
df_4
df_5
DF

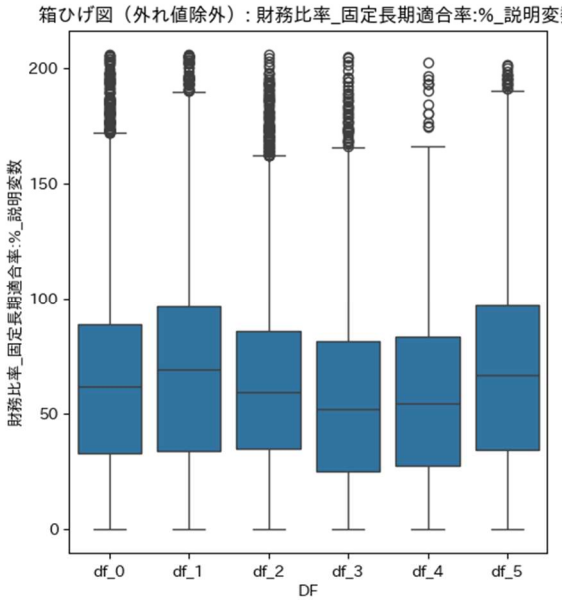

箱ひげ図:財務比率_当座比率:%_説明変数
財務比率_当座比率:%_説明変数
1000
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF

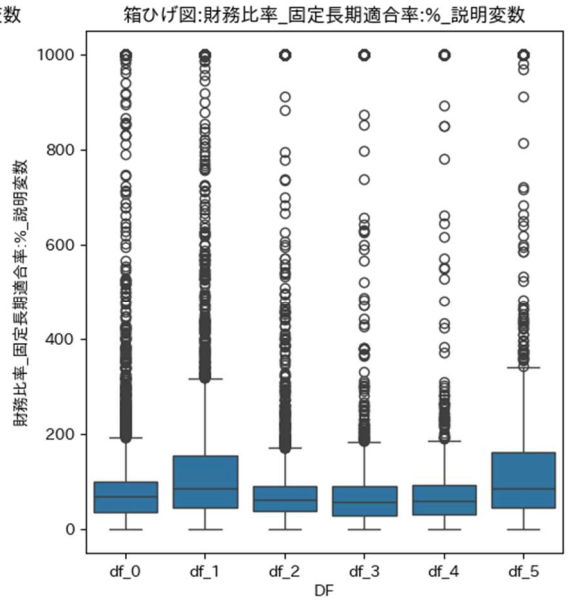

箱ひげ図（外れ値除外）：財務比率_流動比率:%_説明変数
財務比率_流動比率:%_説明変数
600
500
400
300
200
100
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

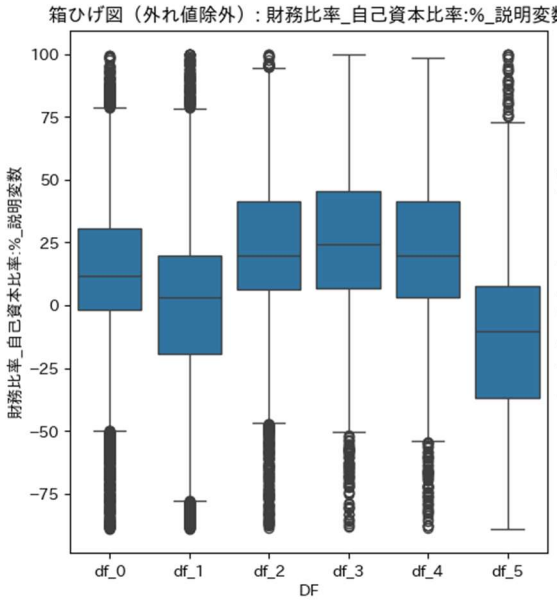

箱ひげ図:財務比率_流動比率:%_説明変数
財務比率_流動比率:%_説明変数
1000
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF

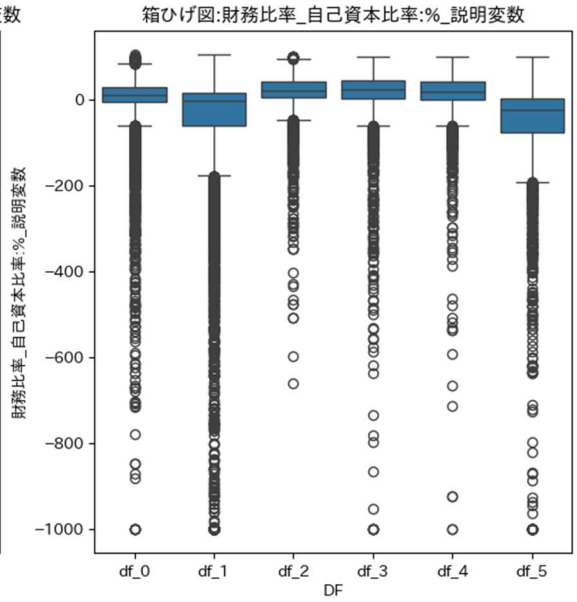

箱ひげ図（外れ値除外）：財務比率_固定比率:%_説明変数
財務比率_固定比率:%_説明変数
1000
800
600
400
200
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

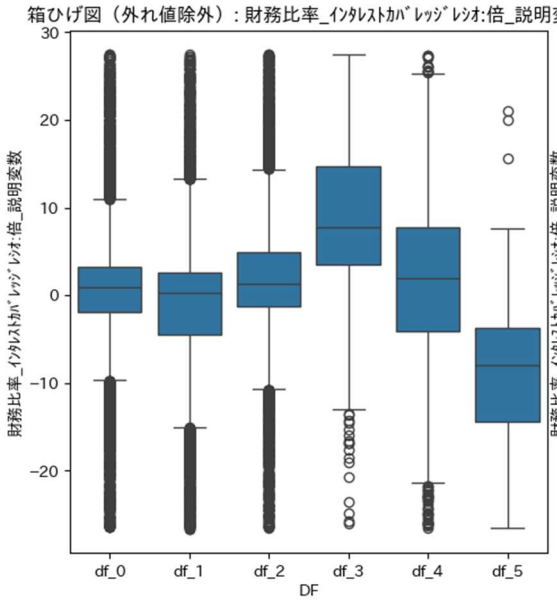

箱ひげ図:財務比率_固定比率:%_説明変数
財務比率_固定比率:%_説明変数
1000
800
600
400
200
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

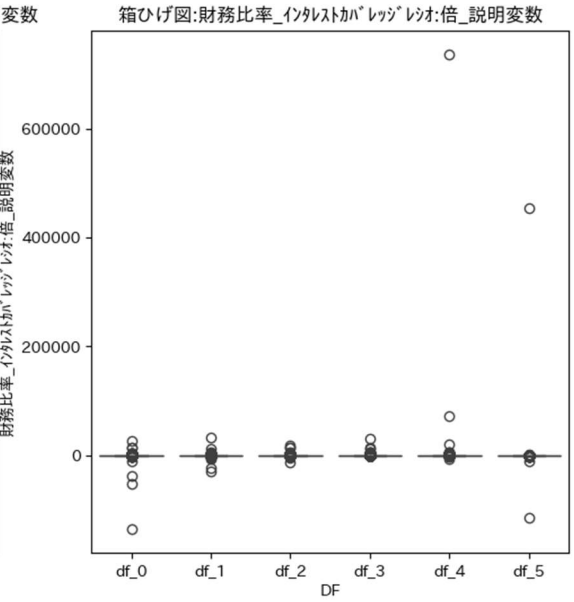

箱ひげ図（外れ値除外）：財務比率_固定長期適合率:%_説明変数
財務比率_固定長期適合率:%_説明変数
200
150
100
50
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

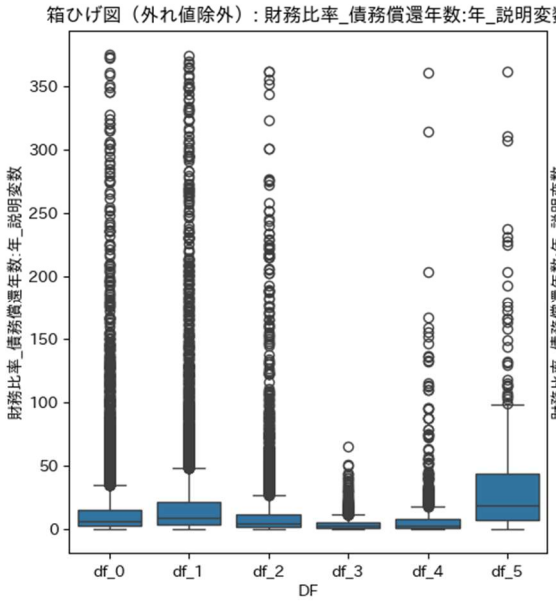

箱ひげ図:財務比率_固定長期適合率:%_説明変数
財務比率_固定長期適合率:%_説明変数
1000
800
600
400
200
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

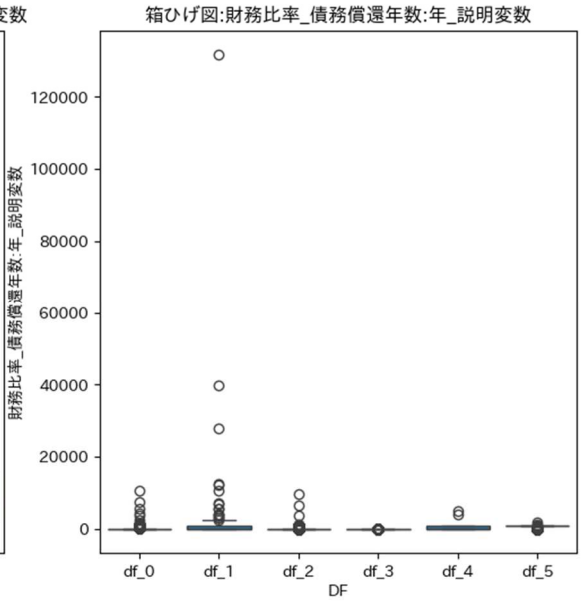

箱ひげ図（外れ値除外）：財務比率_自己資本比率:%_説明変数
財務比率_自己資本比率:%_説明変数
100
75
50
25
0
−25
−50
−75
df_0
df_1
df_2
df_3
df_4
df_5
DF

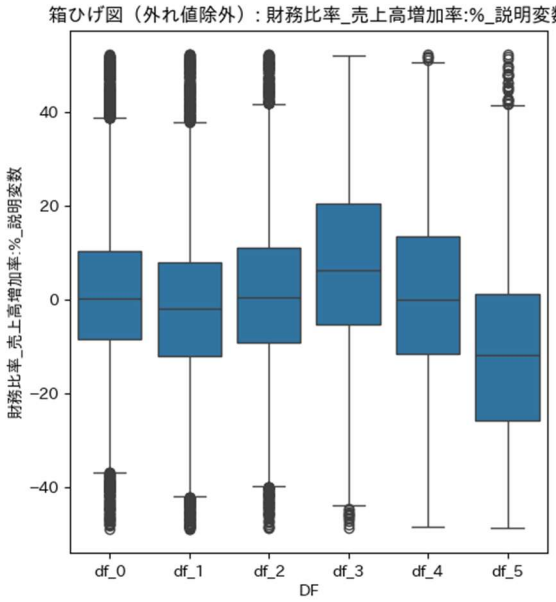

箱ひげ図:財務比率_自己資本比率:%_説明変数
財務比率_自己資本比率:%_説明変数
0
−200
−400
−600
−800
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF

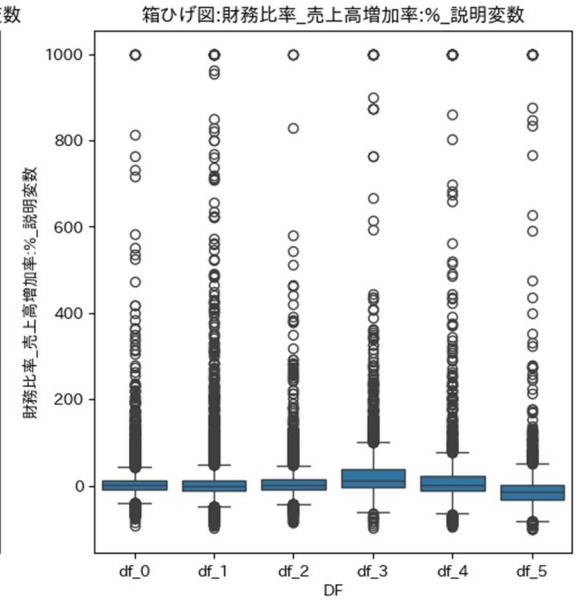

箱ひげ図（外れ値除外）：財務比率_ｲﾝﾀﾚｽﾄｶﾊﾞﾚｯｼﾞﾚｼｵ:倍_説明変数
財務比率_ｲﾝﾀﾚｽﾄｶﾊﾞﾚｯｼﾞﾚｼｵ:倍_説明変数
30
20
10
0
−10
−20
df_0
df_1
df_2
df_3
df_4
df_5
DF

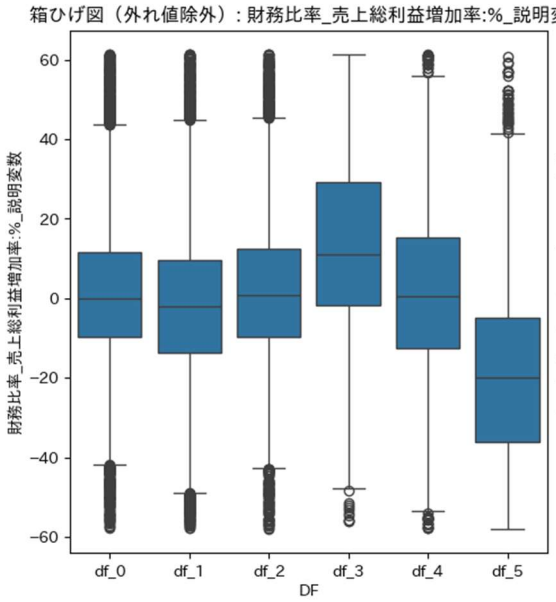

箱ひげ図:財務比率_ｲﾝﾀﾚｽﾄｶﾊﾞﾚｯｼﾞﾚｼｵ:倍_説明変数
財務比率_ｲﾝﾀﾚｽﾄｶﾊﾞﾚｯｼﾞﾚｼｵ:倍_説明変数
600000
400000
200000
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

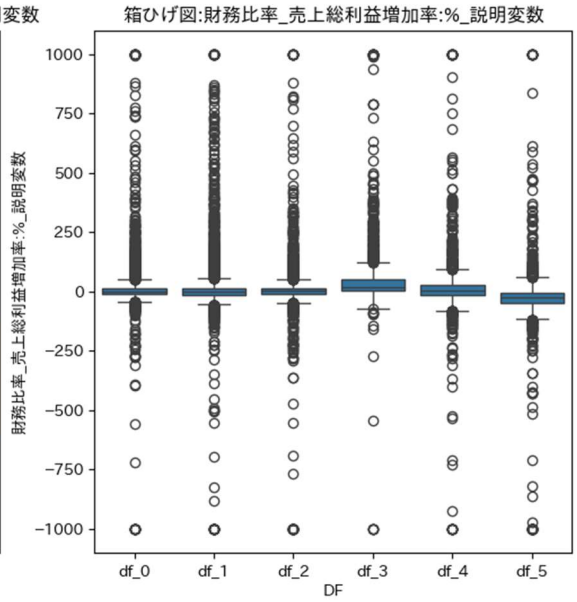

箱ひげ図（外れ値除外）：財務比率_債務償還年数:年_説明変数
財務比率_債務償還年数:年_説明変数
350
300
250
200
150
100
50
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

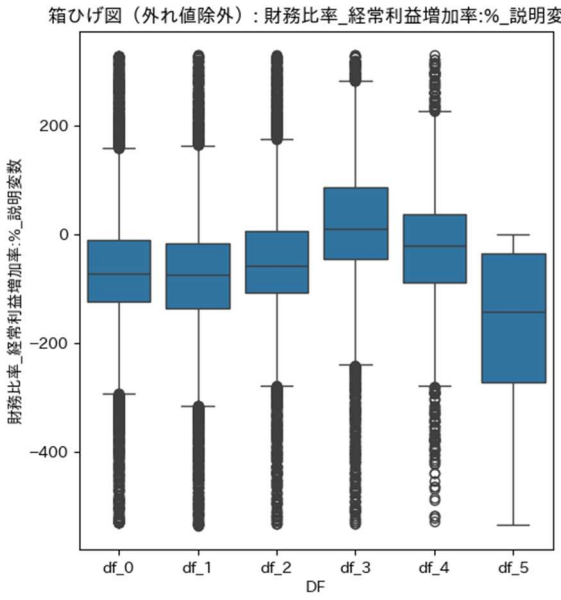

箱ひげ図:財務比率_債務償還年数:年_説明変数
財務比率_債務償還年数:年_説明変数
120000
100000
80000
60000
40000
20000
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

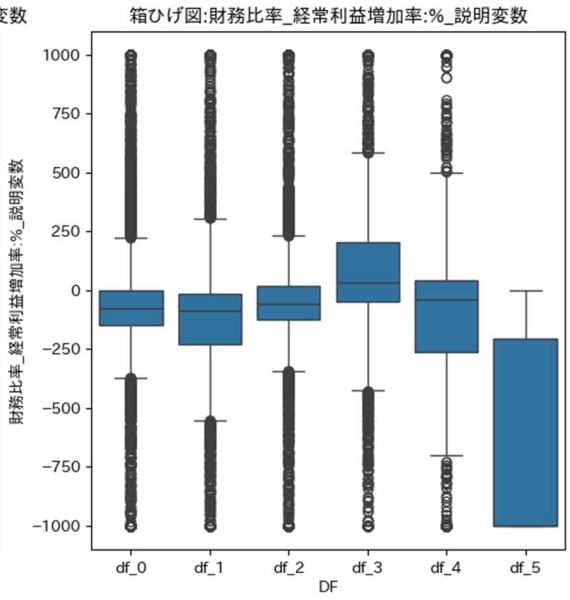

箱ひげ図（外れ値除外）：財務比率_売上高増加率:%_説明変数
財務比率_売上高増加率:%_説明変数
40
20
0
−20
−40
df_0
df_1
df_2
df_3
df_4
df_5
DF

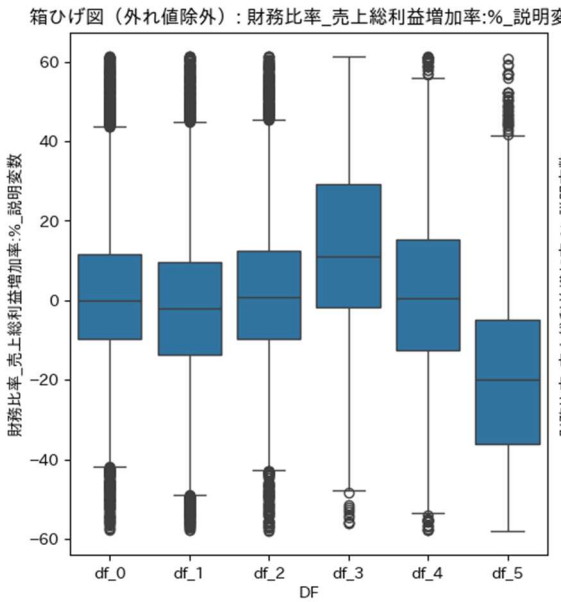

箱ひげ図:財務比率_売上高増加率:%_説明変数
財務比率_売上高増加率:%_説明変数
1000
800
600
400
200
0
df_0
df_1
df_2
df_3
df_4
df_5
DF

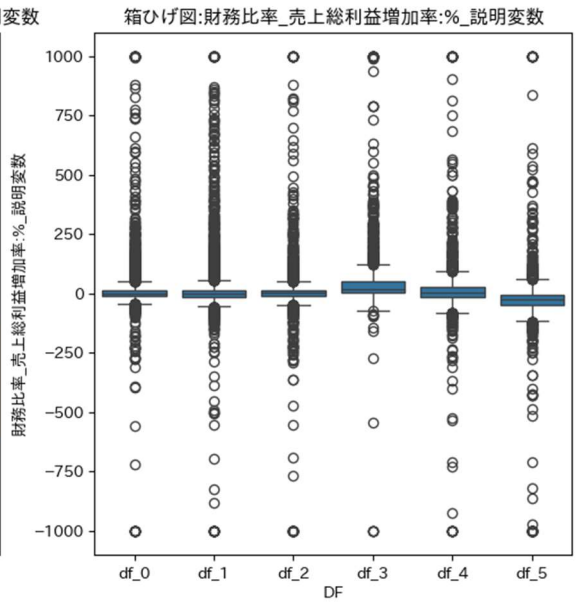

箱ひげ図（外れ値除外）：財務比率_売上総利益増加率:%_説明変数
財務比率_売上総利益増加率:%_説明変数
60
40
20
0
−20
−40
−60
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_売上総利益増加率:%_説明変数
財務比率_売上総利益増加率:%_説明変数
1000
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図（外れ値除外）：財務比率_経常利益増加率:%_説明変数
財務比率_経常利益増加率:%_説明変数
200
0
−200
−400
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_経常利益増加率:%_説明変数
財務比率_経常利益増加率:%_説明変数
1000
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図（外れ値除外）：財務比率_売上総利益増加率:%_説明変数
財務比率_売上総利益増加率:%_説明変数
60
40
20
0
−20
−40
−60
df_0
df_1
df_2
df_3
df_4
df_5
DF


箱ひげ図:財務比率_売上総利益増加率:%_説明変数
財務比率_売上総利益増加率:%_説明変数
1000
750
500
250
0
−250
−500
−750
−1000
df_0
df_1
df_2
df_3
df_4
df_5
DF