ISSN 2758-0520

THE INSTITUTE OF ECONOMICS RESEARCH

Working Paper Series

No. 267

Dynamic Industry Segmentation for Anonymized Data: Insights from the Entertainment Industry

Shoka HAYAKI (Kobe University, Japan)
Makoto MORISADA (Kansai University, Japan)
Kenta YAMANOUCHI (Kagawa University, Japan)

May 2025

KAGAWA UNIVERSITY

Takamatsu, Kagawa 760-8523
JAPAN

Dynamic Industry Segmentation for Anonymized Data: Insights from the Entertainment Industry*

Shoka HAYAKI#
Kobe University, Japan

Makoto MORISADA
Kansai University, Japan

Kenta YAMANOUCHI
Kagawa University, Japan

Abstract

We propose a statistical industry classification method for unlisted firms, where stock-based approaches are infeasible. By modeling unlevered beta as a function of stable financial ratios, we identify subindustry structures through clustering. To handle anonymized, non-panel data, we combine UMAP, HDBSCAN, and MONIC for nonlinear dimension reduction, density-based clustering, and temporal tracking. Applying this framework to Japan's entertainment industry, we uncover stable clusters that align with official subindustry definitions and macro indicators, even amid COVID-19 disruptions. Our method provides a practical alternative for classifying firms without market or textual data.

Keywords: Industry classification; Machine learning; Risk exposure; Entertainment industry

JEL Classification: C38, G32, L82

* We acknowledges support of a PIA Research Institute grant. The views expressed in this research are those of the authors. No responsibility for them should be attributed to the PIA Research Institute. All remaining errors are ours.

Corresponding author: Shoka Hayaki; Address: Kobe University, 2-1, Rokkodai-cho, Nada-ku, Kobe, Hyogo, Japan; E-mail: s-hayaki@port.kobe-u.ac.jp.

1. Introduction

Industry classification plays a critical role in both academic research and practical applications. According to Weiner (2005), industry classifications—such as industry dummy variables—were used in 45.7% of articles published between 1995 and 2003 in leading finance and accounting journals, including the *Journal of Finance, Journal of Financial Economics, Journal of Accounting and Economics, and Journal of Accounting Research*. Industry-based analysis enables a more accurate assessment of the information value provided by sell-side analysts to investors (Chung et al., 2017).

In this context, various studies have proposed statistical industry classification methods based on textual data (Bernstein et al., 2003; Hoberg & Phillips, 2016; Kim et al., 2022), stock return correlations (Chan et al., 2007; Kakushadze & Yu, 2016), and financial statement variables (Elliot et al., 2000; Bagnara & Goodarzi, 2023). However, these approaches are not readily applicable to unlisted firms, as they typically rely on standardized disclosures (e.g., Form 10-K) and market price data. This limitation is especially salient in industries such as food service, retail, and entertainment, where non-listed companies account for a significant share of both total sales and employment.

We focus on two main challenges in statistically classifying unlisted firms. First, the absence of stock returns and standardized textual disclosures renders conventional approaches inapplicable. Second, government datasets—though valuable for their unified and high-frequency (e.g., quarterly) coverage—are anonymized, making it impossible to track firms across time and thus hindering the identification of temporal transitions in industry structure.

To overcome the first challenge, we develop a model linking unlevered beta (asset beta) to a set of stable financial ratios (e.g., the ratio of cost of goods sold to sales, the ratio of non-current assets to total assets). Specifically, we assume that stable financial ratios function as deep parameters: they govern firms' cash flow dynamics through their influence on capital and labor inputs, which themselves follow stochastic processes. These ratios, by shaping the structure and volatility of cash flows, ultimately determine the unlevered beta.

Our model yields three key implications. First, it clarifies how cross-sectional variation in stable financial ratios explains the heterogeneity of steady-state unlevered beta, highlighting the fundamental connection between financial structure and business risk. Second, although it is difficult to estimate capital and labor betas at the firm level, these betas—representing the sensitivities of capital and labor inputs to market risk—can reasonably be assumed to be common across firms operating in the same industry. Since stable financial ratios determine the weights applied to these common betas, clustering firms based on these ratios facilitates the identification of subindustry structures within a given industry. Third, compared to clustering based on stock prices, which inevitably reflects noise from highly volatile variables such as sales and cash flows, our approach reduces such noise and enables a more stable and fundamental classification of firms based on their intrinsic business risk.

To address the second challenge, we develop a dynamic clustering framework for anonymized data of unlisted firms. We begin by selecting stable financial ratios that can serve as proxies for the deep parameters in our model. Since the relationship between these ratios and unlevered beta is likely nonlinear, we apply Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) to reduce dimensionality while preserving the manifold structure. This improves the identifiability of density-based clustering algorithms such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al., 2013). The approach allows us to detect nonlinear, local relationships and to distinguish outliers as noise. To track changes in cluster structure over time despite data anonymization, we employ Modelling and Monitoring Cluster Transitions (MONIC; Spiliopoulou et al., 2006). MONIC treats two adjacent time periods as "data panes," performs clustering in each, and matches clusters across data panes based on membership overlap. This enables us to trace the evolution of industry groupings even when firm identities are masked.

We apply this framework to Japan's entertainment industry to construct a statistical subindustry classification. This industry presents a compelling case due to its structural heterogeneity and its exposure to the COVID-19 crisis. Measuring the true impact of the pandemic is challenging: firms formally categorized as entertainment industry may derive substantial revenue from unrelated activities such as real estate, potentially leading to underestimation of the industry's downturn. Conversely, the inclusion of amusement and recreation facilities—especially pachinko parlors—may exaggerate the impact, as their activity had already been declining prior to the pandemic. These complexities underscore the importance of identifying financially homogeneous groups of firms to disentangle the core dynamics of the entertainment industry from the influence of peripheral business activities and long-term structural trends.

Our results can be summarized as follows. By selecting stable financial ratios and performing dimensionality reduction with UMAP, we achieve clear separation of firms, overcoming time-series noise caused by sales fluctuations, especially during the COVID-19 pandemic. Our framework, using HDBSCAN with optimal parameters, reveals a stable underlying cluster structure, capturing persistent dynamics over time. The results show that excluding unstable financial ratios (e.g., sales-to-total-assets ratio) leads to more reliable cluster identification. In particular, the number of dynamic cluster groups, our statistical subindustries, consistently converges to five, which approximates the six fundamental subindustries defined in the Japanese Standard Industry Classification (JSIC), suggesting a correspondence between our statistical classification and fundamental industry classification.

Building on these findings, we further validate the relevance of the five statistical subindustries by examining their financial and temporal coherence. Structurally, the statistical subindustries correspond well to the fundamental subindustries with similar financial characteristics, as confirmed by comparisons with subindustry-level data from the Economic Census—an infrequent but reliable benchmark conducted every four to five years. Temporally, the statistical subindustries exhibit dynamic patterns that closely resemble those of the corresponding fundamental subindustries, as shown by vector autoregression (VAR) and state space analyses. These analyses show that the trajectories of sales-to-total-assets ratios within each cluster closely align with those of the Tertiary Industry Activity Index for corresponding subindustries in the entertainment industry. These results suggest that our framework not only identifies meaningful structural differences across subindustries, but also successfully replicates their dynamic evolution over time.

The remainder of the paper is structured as follows. Section 2 develops a model linking unlevered beta to financial ratios, justifying the use of stable accounting indicators in place of stock data. Section 3 reviews key features of Japanese government statistics. Section 4 presents our classification framework, combining theory and data characteristics. Section 5 applies it to the entertainment industry, identifying stable dynamic clusters. Section 6 examines their temporal structure and alignment with fundamental subindustries. Section 7 concludes.

2. Modeling Unlevered Beta with Financial Ratios

The statistical industry classification methods proposed by Chan et al. (2007) and Kakushadze and Yu (2016) rely on the correlation structure of stock returns. These approaches are based on the intuition that firms with similar business models should exhibit similar risk exposures. Accordingly, if alternative information that captures a firm's risk exposure is available, stock prices may not be essential for industry classification.

In this section, we model the relationship between a firm's unlevered beta and several stable financial ratios. Unlevered beta reflects the firm's intrinsic exposure to market risk. By formulating this relationship, we can identify financial ratios that account for the heterogeneity in risk exposures across firms.

To simplify the model, we assume a complete capital market with a zero corporate tax rate. Under the conditional Capital Asset Pricing Model (CAPM), the conditional unlevered beta (or asset beta) of firm i at time t is given by:

$$\beta_{i,t-1}^{U} = \frac{\mathrm{Cov}_{t-1}\left[\frac{\Delta C_{it}}{C_{i,t-1}}, R_t\right]}{\mathrm{Var}_{t-1}[R_t]}. \tag{1}$$

where C_{it} denotes the free cash flow of firm i, and R_t denotes the market return.[2] We assume that cash flow is generated from two input factors: capital K_{it} and labor L_{it} (defined as the number of employees including executives multiplied by working hours). These variables follow a discrete-time stochastic process:

$$\begin{pmatrix} R_t \\ \Delta K_{it}/K_{i,t-1} \\ \Delta L_{it}/L_{i,t-1} \end{pmatrix} = \begin{pmatrix} \mu_{t-1}^{R} \\ \mu_{i,t-1}^{K} \\ \mu_{i,t-1}^{L} \end{pmatrix} + \begin{pmatrix} \varepsilon_t^{R} \\ \varepsilon_{it}^{K} \\ \varepsilon_{it}^{L} \end{pmatrix}. \tag{2}$$

The conditional variance-covariance matrix of the error terms $\varepsilon_t^R, \varepsilon_{it}^K$, and ε_{it}^L is given by:

$$\Sigma_{i,t-1} = \begin{pmatrix} (\sigma_{t-1}^{R})^2 & \rho_{i,t-1}^{RK}\sigma_{t-1}^{R}\sigma_{i,t-1}^{K} & \rho_{i,t-1}^{RL}\sigma_{t-1}^{R}\sigma_{i,t-1}^{L} \\ \rho_{i,t-1}^{RK}\sigma_{t-1}^{R}\sigma_{i,t-1}^{K} & \left(\sigma_{i,t-1}^{K}\right)^2 & \rho_{i,t-1}^{KL}\sigma_{i,t-1}^{K}\sigma_{i,t-1}^{L} \\ \rho_{i,t-1}^{RL}\sigma_{t-1}^{R}\sigma_{i,t-1}^{L} & \rho_{i,t-1}^{KL}\sigma_{i,t-1}^{K}\sigma_{i,t-1}^{L} & \left(\sigma_{i,t-1}^{L}\right)^2 \end{pmatrix}. \tag{3}$$

Next, we express several key financial quantities as functions of input levels and deep parameters. First, sales S_{it} are assumed to follow a Cobb-Douglas production function:

$$S_{it} = A_{it} K_{it}^{\alpha_i} L_{it}^{1-\alpha_i}, \tag{4}$$

where A_{it} denotes total factor productivity and α_i the capital elasticity. Earnings before interest and taxes (EBIT) X_{it} is defined as:

$$X_{it} = S_{it} - \omega_i S_{it} - \gamma_i \zeta_i L_{it}, \tag{5}$$

[2] This constitutes a natural extension of Hamada's (1972) formulation of unlevered beta to a conditional model.

where ω_i is the cost of goods sold (COGS) to sales ratio, γ_i is the selling, general and administrative expenses (SG&A) to labor cost ratio (assuming SG&A depend entirely on labor input for simplicity), and ζ_i is the average total hourly wage, so $\zeta_i L_{it}$ represents total personnel expenses. The second term represents costs that vary proportionally with sales (i.e., variable costs), while the third term reflects costs that are largely fixed relative to sales (i.e., SG&A costs driven by labor input). Inventories V_{it} and fixed assets F_{it} are defined as:

$$V_{it} = \phi_i K_{it}, \tag{6}$$

$$F_{it} = \kappa_i K_{it}, \tag{7}$$

where ϕ_i and κ_i are the inventory-to-capital and fixed-asset-to-capital ratios, respectively. Assuming stable receivables and payables, changes in working capital equal changes in inventories. Hence, the free cash flow is given by:

$$C_{it} = X_{it} - \Delta V_{it} - \Delta F_{it} = (1-\omega_i) A_{it} K_{it}^{\alpha_i} L_{it}^{1-\alpha_i} - \gamma_i \zeta_i L_{it} - (\phi_i + \kappa_i)(K_{it} - K_{i,t-1}). \tag{8}$$

Using a log-linear approximation under the above assumptions, the conditional unlevered beta can be decomposed as:

$$\beta_{i,t-1}^{U} \approx a_{i,t-1}^{K} \beta_{i,t-1}^{K} + a_{i,t-1}^{L} \beta_{i,t-1}^{L}, \tag{9}$$

where:

$$\beta_{i,t-1}^{K} = \frac{\sigma_{i,t-1}^{K} \rho_{i,t-1}^{RK}}{\sigma_{t-1}^{R}}, \qquad \beta_{i,t-1}^{L} = \frac{\sigma_{i,t-1}^{L} \rho_{i,t-1}^{RL}}{\sigma_{t-1}^{R}}, \tag{10}$$

$$a_{i,t-1}^{K} = \left\{ (1-\omega_i) \alpha_i \frac{S_{i,t-1}}{K_{i,t-1}} - (\phi_i + \kappa_i) \right\} \frac{K_{i,t-1}}{C_{i,t-1}}, \tag{11}$$

$$a_{i,t-1}^{L} = \left\{ (1-\omega_i)(1-\alpha_i) \frac{S_{i,t-1}}{L_{i,t-1}} - \gamma_i \zeta_i \right\} \frac{L_{i,t-1}}{C_{i,t-1}}. \tag{12}$$

The beta weights $a_{i,t-1}^{K}$ and $a_{i,t-1}^{L}$ are homogeneous of degree zero, hence independent of firm size. Consequently, unlevered beta itself is also size-invariant. $\beta_{i,t-1}^{K}$ and $\beta_{i,t-1}^{L}$ represent the capital and labor betas, respectively, capturing the sensitivity of each input factor to market risk. To further simplify the model, we assume that these input betas are common across firms within the same industry. Formally, let $\mathcal{I}_j$ denote the set of firms belonging to industry j. We impose the following restriction:

$$\beta_{i,t-1}^{K} = \beta_{j,t-1}^{K}, \qquad \beta_{i,t-1}^{L} = \beta_{j,t-1}^{L}, \qquad \forall\, i \in \mathcal{I}_j \tag{13}$$

This assumption reflects the idea that firms operating in the same industry face similar technological conditions and market environments, which shape how the growth of capital and labor inputs co-moves with aggregate market risk.

Equations (9) and (13) suggest that, given capital and labor betas, a firm's risk exposure within a given industry is determined by the beta weights $a_{i,t-1}^{K}$ and $a_{i,t-1}^{L}$. Note that even firms with similar financial ratios may exhibit substantially different unlevered betas if they belong to industries with distinct capital or labor betas. This highlights the importance of applying financial ratio–based clustering within a given industry, rather than across structurally dissimilar industries.

Equations (11) and (12) also highlight a drawback of stock-price-based classification. The beta decomposition includes ratios such as capital turnover S/K, revenue per worker S/L, and the inverse of capital and labor cash flow productivity K/C, L/C. Cash flows are often highly volatile and deviate from average levels due to firm-specific shocks. Even sales can become unstable during major disruptions, such as the COVID-19 crisis. As a result, stock-price-based methods are highly sensitive to noise embedded in conditional unlevered beta. In contrast, our approach, which leverages stable financial ratios as deep parameters, aims to minimize such noise and yield a more robust and economically meaningful industry classification. Moreover, Equations (11) and (12) aligns with the perspective that a firm's business risk is shaped by its degree of operating leverage (Lev, 1974; Novy-Marx, 2011) and its intensities of human and physical capital (Elliot et al., 2000).

3. Data

For our analysis, we use the Surveys for the Financial Statements Statistics of Corporations by Industry (the FSSC).[3] The FSSC consists of annual and quarterly surveys conducted by the Ministry of Finance in Japan. These surveys collect the financial information such as sales, assets, liabilities, various costs, including personnel expenses. The FSSC is a sampling survey, with the sample comprising all large corporations with capital of 500 million yen or more, as well as a subset of corporations with capital below 500 million yen.

One major limitation of the FSSC data is that the industry classification is restricted to the major groups of JSIC, with no more detailed classifications available. This limitation may introduce bias in tracking trends within subindustries, as the sampling is based on major industry groups. Consequently, the composition of subindustry samples may vary significantly over time. Additionally, the FSSC data is anonymized data at least for us, meaning that the corporations are not labeled. As a result, we cannot identify individual firms or track sample changes over time.

In this study, we focus on firms classified within the entertainment industry. Our FSSC sample covers the period from 2016 to 2022. While the time-series analysis is conducted using the quarterly data, our industry classification method is applied to annual data to avoid time-series instability due to seasonality effects.

The goal of this study is to achieve a more detailed industry classification with a sufficient variety of financial data using the FSSC. To this end, we supplement our approach with additional data sources and compare the information obtained from the FSSC.

The first supplementary data source is the Economic Census for Business Activity (the Census), conducted approximately every five years by the Ministry of Internal Affairs and Communications. Although the Census is less frequently updated, it provides comprehensive coverage of all establishments and enterprises in Japan, except for individual proprietorships in certain industries. A key advantage of the Census is its narrowly defined industry classification, allowing for more detailed descriptive statistics. We compare the industry statistics in the Census with those of the clusters constructed from the FSSC to establish the correspondence between our detected clusters and specific subindustries.

Additionally, we examine time-series trends using the Indices of Tertiary Industry Activity (the Indices). The indices aim to capture the overall production activity within the service sector. The indices are compiled and reported monthly by the Ministry of Economy, Trade and Industry. While these Indices provide useful insights into economic activity trends in the service industry, their reliability at the subindustry level is limited

---

[3] Tsuruta (2019) is one of the major studies using the FSSC. See Tsuruta (2019) for the details of the data.

due to the scarcity of data sources for each category. We compare the trends observed in the Indices with those of our detected FSSC clusters to assess the validity of our classification method.

4. Methodology

This section outlines the framework for classifying firms within a given industry into subindustries. Two key challenges arise when conducting statistical industry classification for unlisted firms. First, the available information is typically limited to financial statements. Second, in the case of the FSSC, the data are anonymized, and a portion of the sample is rotated periodically.

To address the first issue, we adopt a set of stable financial ratios theoretically linked to unlevered beta, as established in Section 2. A critical point here is that unlevered beta is not a linear weighted average of these ratios, but rather a nonlinear function of them. This suggests that even though the data may appear to be distributed in a high-dimensional space, they may in fact lie on a lower-dimensional nonlinear manifold.[4] Based on this manifold hypothesis, we apply Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018) as a preprocessing step prior to clustering. UMAP preserves local topological structure while mapping high-dimensional data onto a lower-dimensional latent space, assuming that the data lie on a Riemannian manifold. This transformation enhances the identifiability of cluster boundaries when using density-based clustering methods such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN; Campello et al., 2013) and is expected to improve the temporal stability of cluster structures.[5]

To address the second issue, we employ Modelling and Monitoring Cluster Transitions (MONIC; Spiliopoulou et al., 2006), which enables tracking and detecting transitions in cluster structures over time.[6] In FSSC, firm names are anonymized, and individual firms cannot be directly tracked across periods, while the partial rotation of the sample further complicates analysis of firm dynamics over time. MONIC compares how clusters evolve over adjacent time periods and identifies transitions such as survival, split, merge, disappearance, and emergence. This enables us to assess the continuity and stability of clustering results over time, even under sample turnover, thereby facilitating the tracking of structural changes in subindustries.

The technical procedure of MONIC is as follows. Let the dataset be given by $\{d_t\}_{t=1}^{T}$. For each $t = 1, \dots, T-1$, define a data pane $P^{(t)} \coloneqq \{d_t, d_{t+1}\}$. We then apply clustering to each data pane, yielding a cluster set $C^{(t)} = \left\{C_1^{(t)}, C_2^{(t)}, \dots, C_{k_t}^{(t)}\right\}$. Since adjacent data panes $P^{(t)}$ and $P^{(t+1)}$ share the same d_{t+1}, the overlap between clusters across data panes can be computed as:

---

[4] In general, when trade-offs exist among cost structures or asset compositions, the feasible combinations are constrained. Such structural constraints may cause high-dimensional financial data to lie approximately on a lower-dimensional manifold. Consider the Earth in three-dimensional space: under the constraint of gravity, human activity is confined to its surface, which is a two-dimensional spherical manifold. Although unlevered beta is not itself a binding constraint, other structural limitations may lead firms with similar financial ratios to exhibit similar unlevered betas.

[5] McInnes et al. (2018) recommend using HDBSCAN after dimensionality reduction with UMAP. Compared to t-SNE (Van der Maaten and Hinton, 2008) and LargeVis (Tang et al., 2016), UMAP better preserves non-local structures while maintaining a comparable local structure.

[6] We use a R package of MONIC developed by Atif and Leisch (2022).

$$\text{overlap}\left(C_i^{(t)}, C_j^{(t+1)}\right) = \frac{\left|C_i^{(t)} \cap C_j^{(t+1)}\right|}{\left|C_i^{(t)}\right|} \text{ for } i = 1, \dots, k_t, j = 1, \dots, k_{t+1}. \tag{14}$$

A cluster $C_i^{(t)}$ is said to *survive* as $C_j^{(t+1)}$ if it overlaps with it by more than a threshold τ_{survive} and does not exceed that threshold with any other cluster. If multiple clusters from $C^{(t)}$ each overlap with $C_j^{(t+1)}$ above τ_{survive}, they are said to *merge* into $C_j^{(t+1)}$. Conversely, a cluster $C_i^{(t)}$ is said to *split* into $C_{j_1}^{(t+1)}, C_{j_2}^{(t+1)}, \dots, C_{j_M}^{(t+1)}$ if the following two conditions are met:

$$\text{overlap}\left(C_i^{(t)}, C_{j_m}^{(t+1)}\right) > \tau_{\text{split}}, \qquad m = 1,2, \dots, M, \tag{15}$$

$$\sum_{m=1}^{M} \text{overlap}\left(C_i^{(t)}, C_{j_m}^{(t+1)}\right) > \tau_{\text{survive}}. \tag{16}$$

If no such relationships are found, then $C_i^{(t)}$ is deemed to have *disappeared*, and $C_j^{(t+1)}$ is considered a *newly emerged* cluster.

Because MONIC connects clusters across adjacent data panes, excessive structural drift between data panes may impede accurate tracking of industry dynamics. Furthermore, as the theoretical model suggests, including volatile financial ratios in the clustering process introduces noise, much like stock prices. It is therefore critical to carefully select stable financial ratios to improve the robustness of the clustering structure.

Based on the discussion above, we propose the following procedure as a framework for fine-grained classification of unlisted firms using official statistics:

1. Select a set of stable financial ratios associated with free cash flow volatility, as motivated by the model.
2. Apply dimensionality reduction via UMAP to enhance clustering stability and identifiability.
3. Construct overlapping data panes from adjacent time periods and conduct clustering using HDBSCAN.
4. Track the transitions in cluster structure over time using MONIC.

A stable subindustry classification is assumed to result in a high proportion of clusters that can be continuously tracked from the initial to the final period.

5. Application

In this section, we apply our proposed method to Japan's entertainment industry. The industry was severely affected by the COVID-19 crisis in 2020. However, the magnitude of the impact may be underestimated when firms classified as entertainment businesses but earning a non-negligible portion of their revenue from other industries, such as real estate, are included in the analysis. Conversely, the inclusion of amusement and recreation facilities—mainly pachinko parlors—may overstate the impact, as their activity index, according to the Tertiary Industry Activity Index, had been declining even before the pandemic. By identifying financially homogeneous groups of firms, our framework helps isolate the core dynamics of the entertainment industry from the influence of non-core business activities and structural heterogeneity.

First, we select a set of stable financial ratios for use in our empirical analysis. Based on data availability in the FSSC, we proxy the parameters ω_i, γ_i, ζ_i, ϕ_i, and κ_i in the model introduced in Section 2 by the following financial ratios, respectively: the ratio of cost of goods sold (COGS) to sales, the ratio of selling, general and administrative expenses (SG&A) to personnel expenses, annual personnel expenses per capita, the ratio of inventory to total assets, and the ratio of non-current assets to total assets, respectively.

As a more refined specification, we also consider decomposing non-current assets into property, plant and equipment (PP&E), intangible assets, and investments and other assets (I&OA), and defining their respective shares of total assets as κ_{1i}, κ_{2i}, and κ_{3i}. Since $\kappa_i = \kappa_{1i} + \kappa_{2i} + \kappa_{3i}$, this alternative specification does not qualitatively alter the model's implications. To better capture cross-sectional heterogeneity in financial structure, we include these decomposed asset ratios in our analysis as well. In the model presented in Section 2, the effect of depreciation is offset when operating and investment cash flows are aggregated. Nevertheless, the ratio of depreciation to non-current assets reflects the share of depreciable assets and is expected to capture meaningful cross-sectional variation in firm characteristics.

Table 1 summarizes the candidate financial ratios used in the analysis. It reports the annual averages of these ratios for the entire entertainment industry from 2015 to 2022. The table also includes the results of Welch tests comparing each year's average with that of the previous year. The rightmost column presents the average of the coefficients of variation across firms for each year, capturing cross-sectional dispersion over the entire period. Observations are excluded if key financial variables (i.e., sales, total assets, or personnel expenses per capita) in the denominator of a ratio are missing. If other non-critical variables are missing in the numerator, the value is replaced with zero to retain as many observations as possible. Furthermore, all financial ratios are Winsorized at the 99th percentile to mitigate the influence of outliers.

=== Table 1 ===

Based on the results presented in Table 1, eight financial ratios are selected for clustering: (i) COGS-to-sales ratio, (ii) SG&A-to-personnel-expenses ratio, (iii) personnel expenses per capita, (iv) depreciation-to-non-current-assets ratio, (v) inventory-to-total-assets ratio, (vi) PP&E-to-total-assets ratio, (vii) intangible-assets-to-total-assets ratio, and (viii) I&OA-to-total-assets ratio. As expected from the model assumptions, these ratios appear to be stable throughout the sample period, including the COVID-19 crisis. In contrast, other ratios involving sales (except for COGS-to-sales ratio) show substantial fluctuations during the pandemic. As discussed in Section 2, using ratios that include sales or, even more so, cash flows—which tend to be highly volatile—may capture time-series noise in unlevered beta estimates. The selected ratios also exhibit sufficiently large cross-sectional coefficients of variation, suggesting that they can effectively capture firm-level heterogeneity.

To demonstrate that this variable selection is appropriate—at least within the Japanese entertainment industry—we compare the effects of different variable sets using listed firms for which labels can be tracked over time. The left panel of Figure 1 shows a two-dimensional UMAP projection of financial data that includes unstable ratios, such as the sales-to-total-assets ratio. Each circle represents the minimum enclosing circle for firms sharing the same label. Many of these circles overlap significantly, indicating that the boundaries of even the same firm become ambiguous—suggesting that meaningful subindustry clustering is infeasible under this specification. In contrast, the right panel shows the results when unstable ratios are excluded. Except for the clearly anomalous trajectory of Misonoza, there is little overlap between circles, suggesting clearer firm-level separation.

=== Figure 1 ===

Misonoza underwent a series of real estate transactions between 2013 and 2015—including the sale, reconstruction, and reacquisition of its building—resulting in large fluctuations in PP&E, which would typically be a stable variable. This indicates a substantial shift in its business model, likely due to fluctuations in rental income. Therefore, the detection of Misonoza's atypical behavior in the right panel actually highlights the strength of our variable selection in sensitively capturing such temporary shifts by removing noisy financial ratios.

Next, dimensionality reduction is performed using UMAP. The algorithm requires two key parameters: the number of neighbors (*n_neighbors*) and the minimum distance (*min_dist*). One practical objective of applying UMAP here is to enhance the separability of cluster boundaries when performing density-based clustering methods such as HDBSCAN in subsequent steps. To select appropriate parameters, we apply HDBSCAN to datasets transformed by UMAP under various parameter combinations, and calculate the average cluster membership probability across all observations. This average reflects how confidently each point is assigned to a cluster and serves as an indicator of clustering stability.

Table 2 reports the average membership probabilities under different combinations of UMAP and HDBSCAN parameters. The left half shows results excluding unstable financial ratios, while the right includes them. In both cases, the average membership probability is maximized when the UMAP parameters are set to *n_neighbors* = 10 and *min_dist* = 0.01, regardless of the chosen HDBSCAN *minPts* parameter. Based on this result, we proceed with the dataset transformed using these UMAP parameters for the subsequent analysis, as it offers the best clustering consistency over time.

=== Table 2 ===

Figure 2 shows kernel density contour plots of the UMAP-transformed data (excluding unstable financial ratios), visualizing the concentration of firms in the embedded space across time.[7] The figure is divided into seven data panes, each constructed from a pair of adjacent time periods. This design not only facilitates visual assessment of the temporal continuity and stability of the cluster structure but also prepares the data for MONIC-based dynamic cluster tracking introduced later. Across the panes, five to seven density peaks are consistently observed, with the positions of the five main peaks remaining nearly unchanged throughout the sample period. This visual pattern suggests the persistence of the underlying cluster structure over time.

[7] Due to data anonymization policies by the FSSC, scatter plots cannot be published; therefore, kernel density estimates are used as a substitute.

=== Figure 2 ===

Moving forward, HDBSCAN is applied to each data pane constructed from two adjacent time periods. HDBSCAN requires the specification of a minimum number of points (*minPts*) needed to form a cluster.[8] A distinctive feature of HDBSCAN is its ability to designate points that do not meet the density criteria of any cluster as noise. This characteristic aligns well with the structure of standard industry classifications, such as the "miscellaneous amusement and recreation services" category in JSIC, which groups firms that do not clearly belong to any specific subindustry. While the presence of noise points can enhance the stability of the resulting clusters by excluding ambiguous cases, an excessive number of such points may indicate that the clustering fails to capture the underlying structure of the data.

The number of clusters generated in each data pane is also a key indicator. Significant fluctuations in cluster count across time may suggest instability in the cluster structure, potentially due to frequent events of cluster absorption, splitting, disappearance, or emergence. Moreover, the number of clusters should remain within a reasonable range. In the case of Japan's entertainment industry, six subindustries are defined in the JSIC.[9]

Panel A of Table 3 presents the number of clusters and the proportion of noise points for each data pane when UMAP and HDBSCAN are applied to a dataset excluding unstable financial ratios. When the parameter *minPts* is set in the range of 10 to 20, the minimum number of clusters across the data panes is about 5 to 7, which is roughly consistent with the number of subindustries in the JSIC. Approximately 10 to 20% of noise points are observed overall, with the tendency for the number of noise points to be minimized when *minPts* is set to 15.

=== Table 3 ===

Panel B of Table 3 shows the results when UMAP and HDBSCAN are applied to a dataset including unstable financial ratios. When *minPts* is set to 15 or 20, the noise rate is nearly the same as in the case without unstable financial ratios. However, in all cases, the number of clusters decreases to 2 in the 2021-22 data pane. This suggests that time-series noise introduced by unstable financial ratios may prevent adequate cluster identification in the final data pane. Additionally, the time-series stability of the cluster structure appears questionable, making the results less reliable.

Panel C of Table 3 presents the results when Principal Component Analysis (PCA) is applied for diagonalization, followed by clustering using all principal components with HDBSCAN on a dataset excluding unstable financial ratios. This approach does not involve dimensionality reduction, and by using all principal components obtained from PCA, information loss is avoided. However, considering the structural constraints

---

[8] For consistency in clustering, it is desirable to keep the *minPts* parameter consistent across all data panes. The use of a sliding window method for the MONIC data panes helps suppress variations in sample size between data panes and maintain the consistency of *minPts*.

[9] While there are technically seven subindustries, one corresponds to "miscellaneous amusement and recreation services," which conceptually matches the noise category in HDBSCAN.

that reflect the non-linear relationships between financial ratios and unlevered beta, applying a linear transformation such as PCA may fail to capture the non-linear manifold structure effectively. In fact, when *minPts* is set to 10 or higher, the number of clusters in one of the data panes becomes 0, with all points classified as noise. Even when *minPts* is set to 5, the number of clusters decreases to 2 to 3, and the noise rate in the final data pane reaches 45%, indicating extreme instability in the cluster structure.

Based on the above analysis, we focus hereafter on the results obtained by applying UMAP to the dataset excluding unstable financial ratios and performing HDBSCAN clustering with *minPts* set to 10, 15, or 20. Panels A, B, and C of Figure 3 respectively present the MONIC-based tracking results for the clusters generated under each of these *minPts* settings. For MONIC, we set the survival threshold τ_{survive} to 0.25 and the split threshold τ_{split} to 0.05.[10]

=== Figure 3 ===

From Panel A of Figure 3, we observe that the cluster structure under the setting of *minPts* = 10 undergoes highly complex changes. In particular, the clusters that eventually connect to C71 and C72 repeatedly split and merge. A major issue with this parameter setting is the presence of cluster groups—such as those connecting to C76—that consist of many clusters but fail to link with any group in the initial period. Given the nature of the entertainment industry as a relatively mature industry, it is unlikely that entirely new business models would emerge rapidly during the observation period. Thus, such discontinuities in cluster structure are more plausibly interpreted as artifacts of inappropriate clustering. Based on the intuition from Figure 2, we expect cluster transitions to be more stable over time, suggesting that the *minPts* = 10 setting may fail to produce meaningful clusters.

Panel B of Figure 3 shows that the cluster structure with *minPts* = 15 is temporally more stable. Among the eight clusters present in the initial period, two disappear midway, while four continue without merging or splitting. The remaining two merge in the second period and remain stable until the final period. In total, five dynamic cluster groups can be tracked from the initial to the final period, which is consistent with the six subindustries defined in the JSIC classification.

Panel C of Figure 3 displays the case with *minPts* = 20, which also yields a stable clustering structure similar to the *minPts* = 15 case. However, one issue is the disappearance of the cluster group descended from C16, which had otherwise remained stable. In this case, four cluster groups are traceable from the initial to the final period.

Overall, the *minPts* = 15 setting achieves the best balance between temporal stability of cluster structures and alignment with intuitive expectations. Indeed, when treating all clusters that cannot be tracked from the initial to the final period as noise, the rate of noise points across all time periods are 0.42, 0.10, and 0.20 for

[10] Given our data pane configuration, even a perfect match between clusters in adjacent data panes would yield an overlap ratio of approximately 0.50. Considering potential changes in business models among firms near cluster boundaries, as well as sample replacement in the Corporate Enterprise Statistics Survey, we regard an overlap ratio of around 0.25 as sufficient to treat clusters across adjacent data panes as continuous. Furthermore, when clusters undergo division, many data points near their boundaries tend to be labeled as noise. Therefore, even if a cluster splits into 2 or 3 subgroups, using a substantially lower split threshold than 0.25—such as 0.05—enables finer detection of such transitions.

minPts = 10, 15, and 20, respectively.

In the case of the *minPts*=15 parameter setting, we define dynamic cluster groups based on their continuity from the initial to the final period. Group A starts with cluster C11 in the initial period and ends with cluster C71 in the final period. Similarly, group B starts from C12; group C from C13, C14, and C45; group D from C16; and group E from C18.[11] Table 4 reports the average values of key financial ratios for each of these groups.

=== Table 4 ===

We summarize the results in Table 4 along two dimensions: capital/labor intensity and operating leverage. The former is related to the structural parameters ϕ_i, κ_i, γ_i, and ζ_i, which exert fixed effects on the weight terms of capital and labor betas ($a^K_{i,t-1}$ and $a^L_{i,t-1}$), while the latter corresponds to ω_i, which influences these weights in proportion to sales S_t constructed from combinations of capital and labor input. Although these are only relative indicators, groups A, D, and E tend to be more capital-intensive, whereas groups B and C tend to be more labor-intensive. Similarly, groups A, B, and E are characterized by relatively low operating leverage, while groups C and D exhibit higher operating leverage.

Table 5 summarizes the financial characteristics of JSIC-defined subindustries based on the 2015 Economic Census. Evaluated along the two axes—capital/labor intensity and operating leverage—the fundamental subindustries defined by JSIC and the statistical subindustries derived from our framework correspond approximately as follows: Group A and Group E align with Amusement and Race; Group B with Park; Group C with Cinema and Sports; and Group D with Performance.[12] It should be noted that these correspondences are based on relative groupings in terms of financial characteristics, rather than on definitional alignment with JSIC categories.

=== Table 5 ===

Taken as a whole, this section demonstrates that our framework effectively captures stable structural differences in risk exposures based on financial ratios across firms through clustering. In the following section, we analyze the correspondences between the fundamental and statistical subindustries using time-series information.

[11] A small number of firms (only two across the entire period) belong to multiple groups. For analytical convenience, we assign them to the group in alphabetical order. Assigning them differently, or weighting them equally across groups, does not qualitatively affect the results.

[12] “Cinema” and “Performance” denote cinemas, and performances (except otherwise classified) and theatrical companies in JSIC, respectively. Similarly, “Race” denotes bicycle, horse, motorcar and motorboat racetrack operations and companies and “Sports” denotes sports facilities. “Park” and “Amusement” denote public gardens and amusement parks, and amusement and recreation facilities in JSIC, respectively.

6. Discussion

Due to the anonymity of the FSSC, evaluating the results obtained in the previous section is quite challenging. Here, we explore an evaluation method that takes advantage of the quarterly financial data, a key strength of the FSSC. Currently, we have access to data covering 32 periods, from the second quarter of 2015 to the first quarter of 2023, based on Japan's fiscal year calendar. While this data is not suitable for analyzing complex time-series models with many parameters, it is sufficient for verifying relationships within the data using a very simple time-series model.

Next, we perform an analysis using impulse response functions. The sales-to-total-assets ratio from the dynamic cluster group obtained in the previous section is denoted as x_{it} for $i = A, B, \dots, E$, and the Tertiary Industry Activity Index for the six subindustry groups within the entertainment industry is denoted as y_{jt} for $j = 1, \dots, 6$. To isolate the effects specific to each subindustry, we first calculate the residuals u_{it}^{x} by regressing x_{it} on the average sales-to-toral assets $\bar{x}_t$ for each firm, and u_{it}^{y} by regressing y_{it} on the overall industry activity index $\bar{y}_t$. For various combinations of i and j, we estimate a VAR model with the residuals $\left(u_{it}^{x}, u_{jt}^{y}\right)^{\top}$ and compute the impulse response functions. The results are presented in Figure 4.

=== Figure 4 ===

Figure 4 suggests a time-series relationship between group A and Amusement, and between group D and Performance. Although no significant relationships are observed at the same point in time, a significant positive hump-shaped relationship is observed in the subsequent period. Notably, a positive hump-shaped relationship, although not significant, is observed between group B and Park, and between group C and Cinema in the subsequent period. These relationships are similar to the results of the matching based on financial ratios evaluated along the two axes of capital/labor intensity and operating leverage, as discussed in the previous section.

Subsequently, we analyze the impact of the COVID-19 shock and its adjusted trends. Specifically, we compare the sales-to-total-assets ratio in the dynamic cluster groups with the Indices of Tertiary Industry Activity at the subindustry level. The decomposition of the COVID-19 shock impact and the shock-adjusted trends is performed using a local linear trend model.

The model is expressed as follows:

$$z_t = \mu_{1,t} + \theta_t + \psi_t D_t + \epsilon_t, \tag{17}$$

$$\mu_{1,t} = \mu_{1,t-1} + \mu_{2,t-1} + \eta_{1,t}, \tag{18}$$

$$\mu_{2,t} = \mu_{2,t-1} + \eta_{2,t}, \tag{19}$$

$$\theta_t = -\theta_{t-1} - \theta_{t-2} - \theta_{t-3} + \xi_t, \tag{20}$$

$$\theta_{2,t} = \theta_{1,t-1} \tag{21}$$

$$\theta_{3,t} = \theta_{2,t-1} \tag{22}$$

$$\psi_t = \psi_{1,t-1} + \nu_t, \tag{23}$$

where z_t represents the sales-to-total-assets ratio from the dynamic cluster groups and the Indices of Tertiary Industry Activity at the subindustry level. Equation (17) linearly decomposes z_t into a level trend ($\mu_{1,t}$), seasonality (θ_t), time-variant COVID-19 shock (ψ_t), and noise (ϵ_t). D_t is a dummy variable equal to 1 from the second quarter of 2020 onward and 0 otherwise. Equation (18) further decomposes the level trend into its previous level, slope of the trend ($\mu_{2,t-1}$), and noise ($\eta_{1,t}$). Equations (20), (21), and (22) captures the quarterly seasonality, while Equation (23) models the time-variant COVID-19 shock. All noise terms are assumed to be normally distributed such as $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, $\eta_{1,t} \sim N(0, \sigma_{\eta,1}^2)$, $\eta_{2,t} \sim N(0, \sigma_{\eta,2}^2)$, $\xi_t \sim N(0, \sigma_\xi^2)$, and $\nu_t \sim N(0, \sigma_\nu^2)$.

We estimate these equations for both fundamental and statistical subindustries. Figure 5 presents the estimated impacts of the COVID-19 shock, $\hat{\psi}_t D_t$. The shock appears to have had a more prolonged effect on groups B, C, and D, as well as on the subindustries Cinema, Park, and Performance. In terms of the relative magnitude of the shock, group B corresponds to Park, group C to Cinema, and group D to Performance. By contrast, the impacts were relatively short-lived for groups A and E, and for the Amusement and Sports subindustries. In terms of the trajectory of shock impacts, group A shows a pattern similar to those of Amusement and Sports, suggesting that it may include firms from both subindustries. Group E and the Race subindustry exhibit notably different dynamics compared to the others, making them more difficult to interpret without further analysis.[13] Overall, these correspondences are consistent with earlier findings based on financial ratio-based matching and dynamic clustering results.

=== Figure 5 ===

Figure 6 presents the shock-adjusted trends $\hat{\mu}_{1,t}$ for groups A through E and their corresponding subindustries in the Indices. A clear upward trend is observed for groups C and D, as well as for the Cinema, Performance, and Race subindustries. Among these, group D and Performance exhibit the strongest growth, whereas the trends for group C, Cinema, and Race are somewhat more moderate. In contrast, a noticeable downward trend can be observed for group A and the Amusement and Sports subindustries. The Park subindustry requires careful interpretation, as it may still be affected by residual COVID-19 shock components. Nevertheless, a downward trend can be observed even prior to the pandemic, followed by a subsequent upward trend in the post-pandemic period—a pattern that is also observed in group B. Taken together, these observations reinforce the plausibility of the proposed correspondence between statistical and fundamental subindustries, further validating the stability and interpretability of our clustering results.

=== Figure 6 ===

Finally, we demonstrate the validity of our framework by analyzing its relationship with ticket sales data from Pia Corporation. Pia mainly handles ticket sales related to live entertainment events, and the total settlement amount for Pia tickets is well-suited to capture core trends in the entertainment industry. By

[13] Both group E and Race contain relatively few firms, which likely contributes to greater statistical noise and hinders straightforward interpretation.

examining the time-series relationship between the sales-to-total-assets ratios of the five dynamic cluster groups derived from our framework and the ticket sales data, we assess whether our classification effectively captures movements in the “live” entertainment industry.

Figure 7 presents the impulse response functions from total ticket sales p_t to x_{it}.[14] We find that among the groups identified by our framework, groups B, C, and D exhibit significantly positive responses to p_t, with groups C and D in particular showing especially strong sensitivities. This indicates that these groups are highly responsive to fluctuations in ticket sales, suggesting that they possess (sub)industrial characteristics closely tied to live entertainment demand. Consistent with our earlier discussion, group C and group D can be interpreted as corresponding to Cinema and Performance, respectively, which aligns well with the primary focus of the tickets sold by Pia.

=== Figure 7 ===

7. Conclusion

We propose a simple theoretical framework that links financial ratios to unlevered beta and use its implications to classify broader industry groups into subindustry groups based on a set of stable financial indicators. Since our approach is applicable to anonymized datasets, such as official government statistics, it is particularly well-suited for fine-grained classification in industries with a high prevalence of unlisted firms.

To illustrate the utility of our method, we apply it to Japan’s entertainment industry. The analysis shows that the five dynamic clusters identified through statistical classification closely correspond to groups with similar financial characteristics and temporal dynamics. The former is supported by comparisons with subindustry-level data from the Economic Census, which, although conducted only once every four to five years, provides a valuable benchmark for financial traits across subindustries. The latter is demonstrated by aligning the trajectories of sales-to-total-assets ratios within our dynamic clusters with those of the Tertiary Industry Activity Index for corresponding subindustries in the entertainment industry, as well as with the Pia ticket settlement index, which reflects broader ticketing trends in the industry. These results suggest that our method not only captures structural differences across subindustries but also replicates their distinct behavioral patterns over time.

While our framework offers a practical and flexible approach to industry classification, further validation is needed. Future research should compare the performance of our unsupervised method with supervised learning techniques to evaluate its robustness and applicability across different industries.

[14] The data used in this study are based on the total settlement amount of Pia tickets provided by Pia Research Institute, indexed by dividing by the value in a designated base period. Ticket settlements are typically made several months prior to the actual events, and the figures may be retrospectively adjusted in the case of cancellations and refunds. Therefore, to proxy the demand trends for live entertainment, it is necessary to apply a one-quarter lag. In addition, to eliminate seasonal effects, all variables are smoothed using annual averages.

Reference

Atif, M., & Leisch, F. (2022). clusTransition: An R package for Monitoring Transition in Cluster Solutions of Temporal Datasets. *Plos one*, 17(12), e0278146.

Bagnara, M., & Goodarzi, M. (2023). Clustering-Based Sector Investing. SAFE Working Paper No. 397.

Bernstein, A., Clearwater, S., & Provost, F. (2003). The Relational Vector-Space Model and Industry Classification. In *Proceedings of the Learning Statistical Models from Relational Data Workshop at the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 8-18.

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 160-172). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chan, L. K., Lakonishok, J., and Swaminathan, B. (2007). Industry Classifications and Return Comovement. *Financial Analysts Journal*, 63(6), 56-70.

Chung, D. Y., Hrazdil, K., & Li, X. (2017). The Effect of Industry Classification on Analyst Following and the Properties of Their Earnings Forecasts. *Applied Economics Letters*, 24(6), 417-421.

Elliott, R., Greenaway, D. & Hine, R. (2000). Tests for Factor Homogeneity and Industry Classification. *Review of World Economics*, 136, 355–371.

Hamada, R. S. (1972). The Effect of the Firm's Capital Structure on the Systematic Risk of Common Stocks. *Journal of Finance*, 27(2), 435-452.

Hoberg, G., & Phillips, G. (2016). Text-Based Network Industries and Endogenous Product Differentiation. *Journal of Political Economy*, 124(5), 1423-1465.

Kakushadze, Z., & Yu, W. (2016). Statistical Industry Classification. *Journal of Risk & Control*, 3*,* 17–65.

Kim, D., Kang, H. G., Bae, K., & Jeon, S. (2022). An Artificial Intelligence-Enabled Industry Classification and Its Interpretation. *Internet Research*, 32(2), 406-424.

Lev, B. (1974). On the Association Between Operating Leverage and Risk. *Journal of Financial & Quantitative Analysis*, 9(4), 627-641.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv*:*1802.03426*.

Novy-Marx, R. (2011). Operating Leverage. *Review of Finance*, 15(1), 103-134.

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., & Schult, R. (2006). MONIC: Modeling and Monitoring Cluster Transitions. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 706-711).

Tang, J., Liu, J., Zhang, M., & Mei, Q. (2016). Visualizing Large-Scale and High-Dimensional Data. *In Proceedings of the 25th International Conference on World Wide Web*, 287-297.

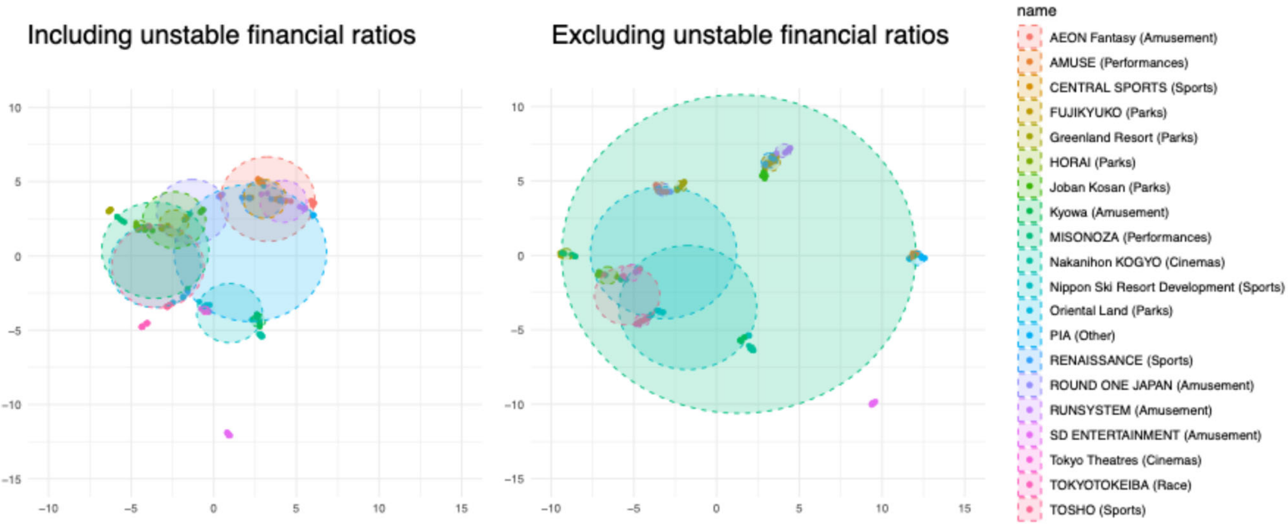
Tsuruta, D. (2019). Working capital management during the global financial crisis: Evidence from Japan. *Japan and the World Economy*, 49, 206-219.

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(11), 2579-2605.

Weiner, C. (2005). The Impact of Industry Classification Schemes on Financial Research. *Available at SSRN 871173*.

Figures and Tables

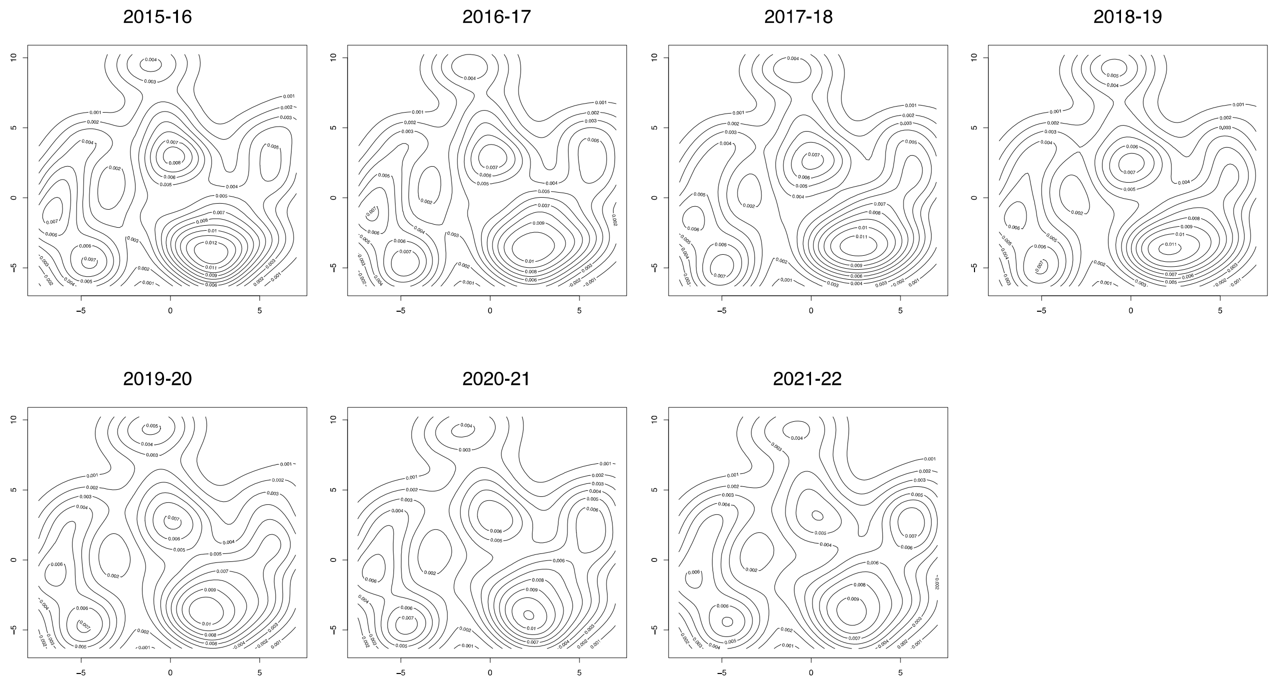
Figure 1 Comparison of the impact of unstable variables



Note: Since personnel expense-related data cannot be obtained from the securities reports of Japanese companies, the analysis uses only the variables from Table 1 that exclude those related to personnel expenses.

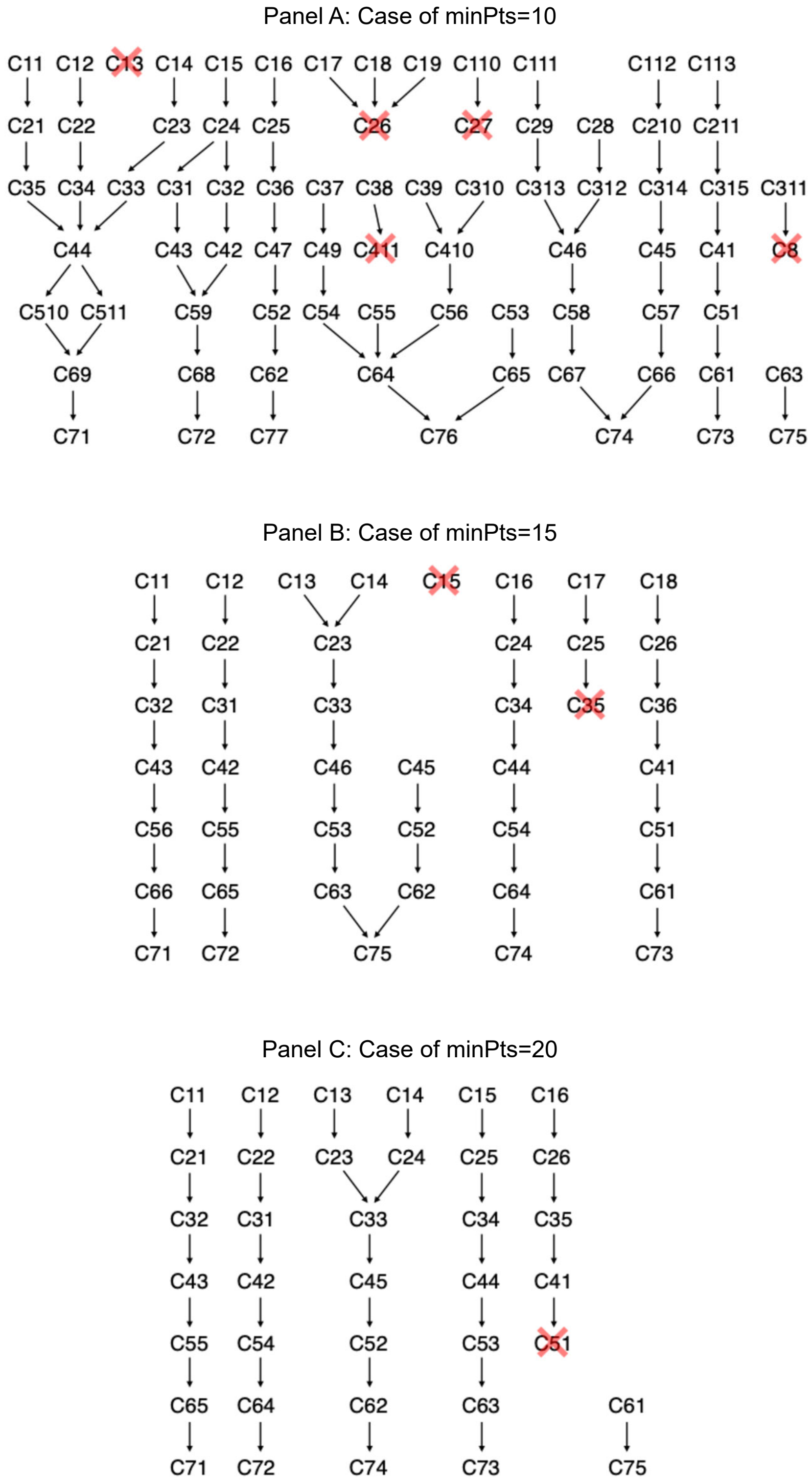
Source: authors' calculation.

Figure 2 Kernel density of the data



Source: authors' calculation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

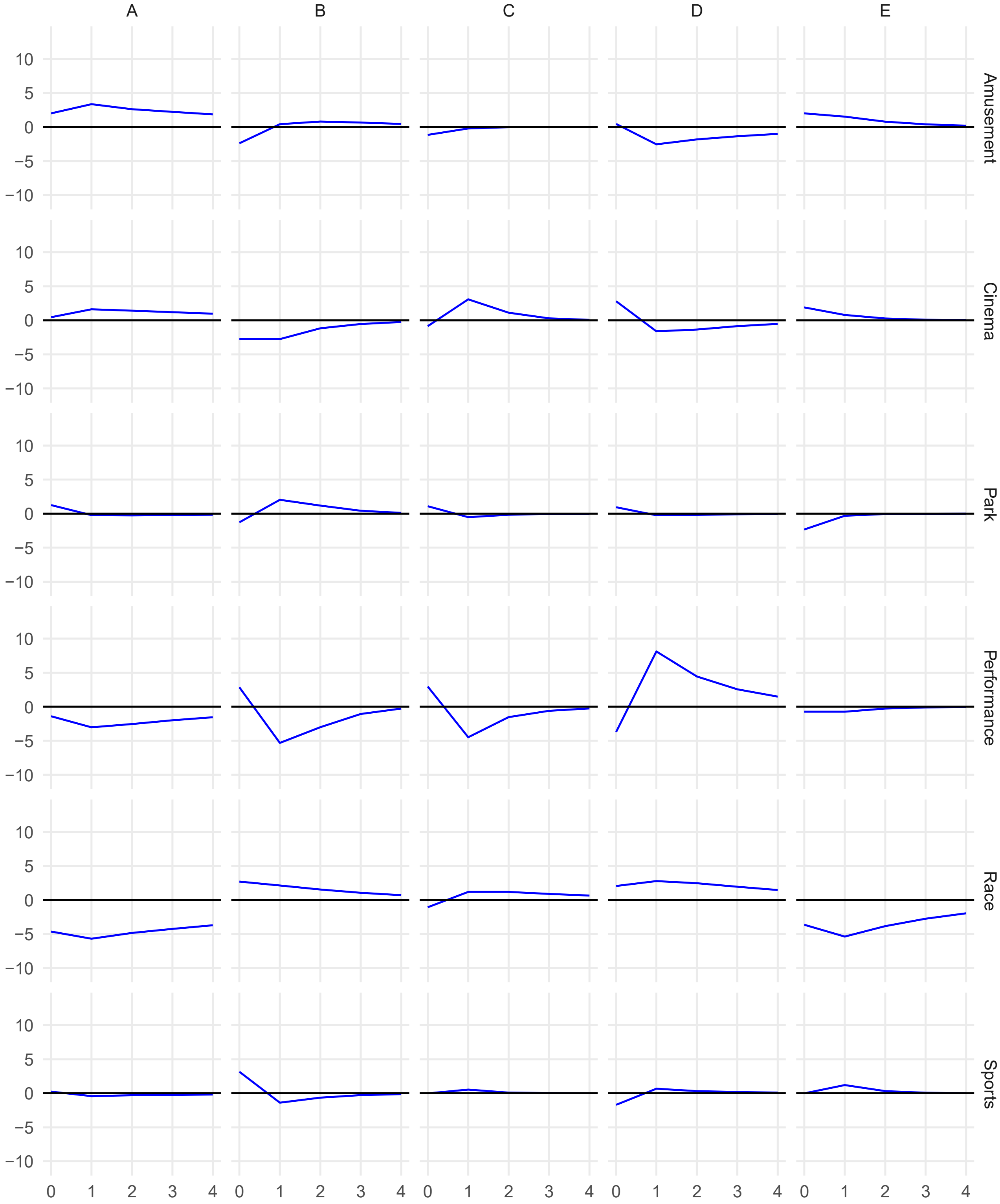
Figure 3 Monitoring dynamic clustering with MONIC



Note: Cij means the j-th cluster in the i-th data pane.

Source: authors' calculation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

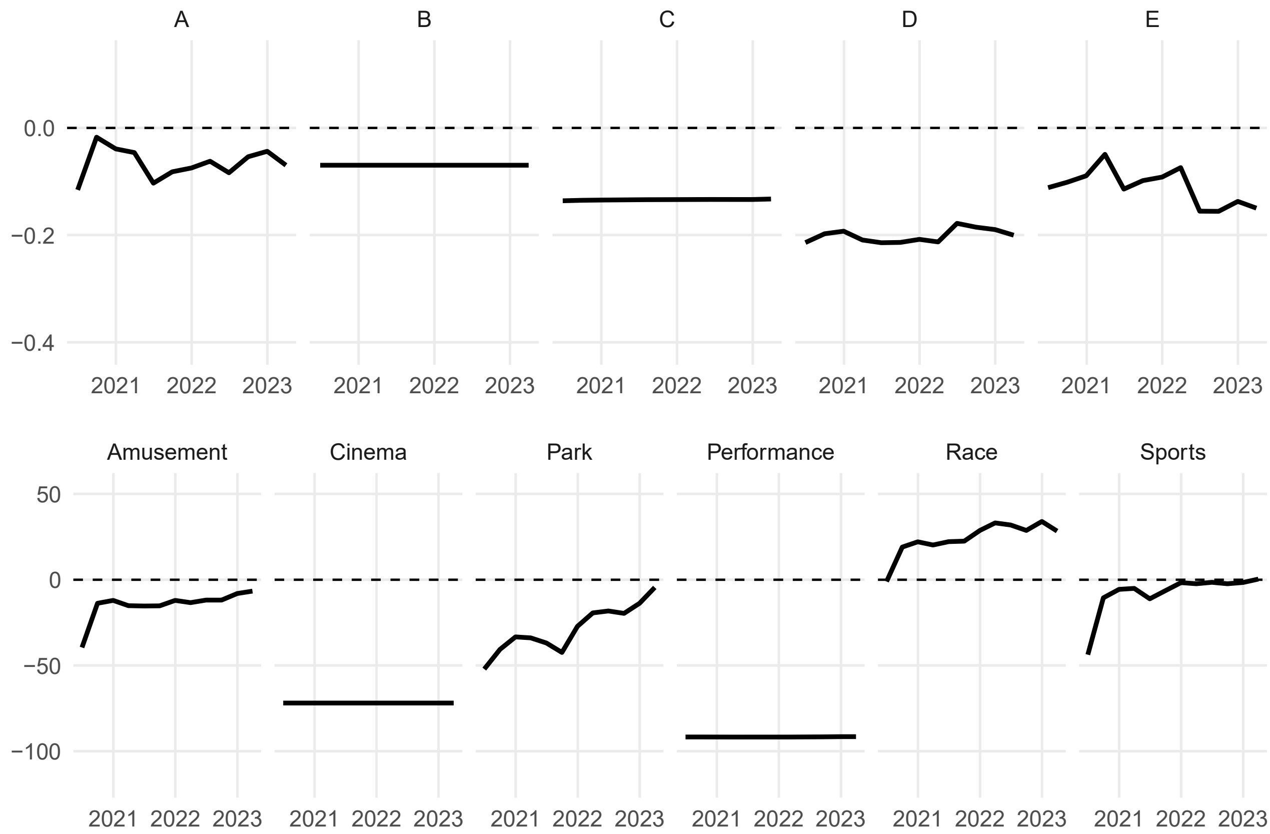
Figure 4 Impulse response



Note: The horizontal axis indicates the impulse variable, and the vertical axis indicates the response variable. The shaded gray areas represent 95% confidence intervals based on 300 bootstrap replications.

Source: authors' calculation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

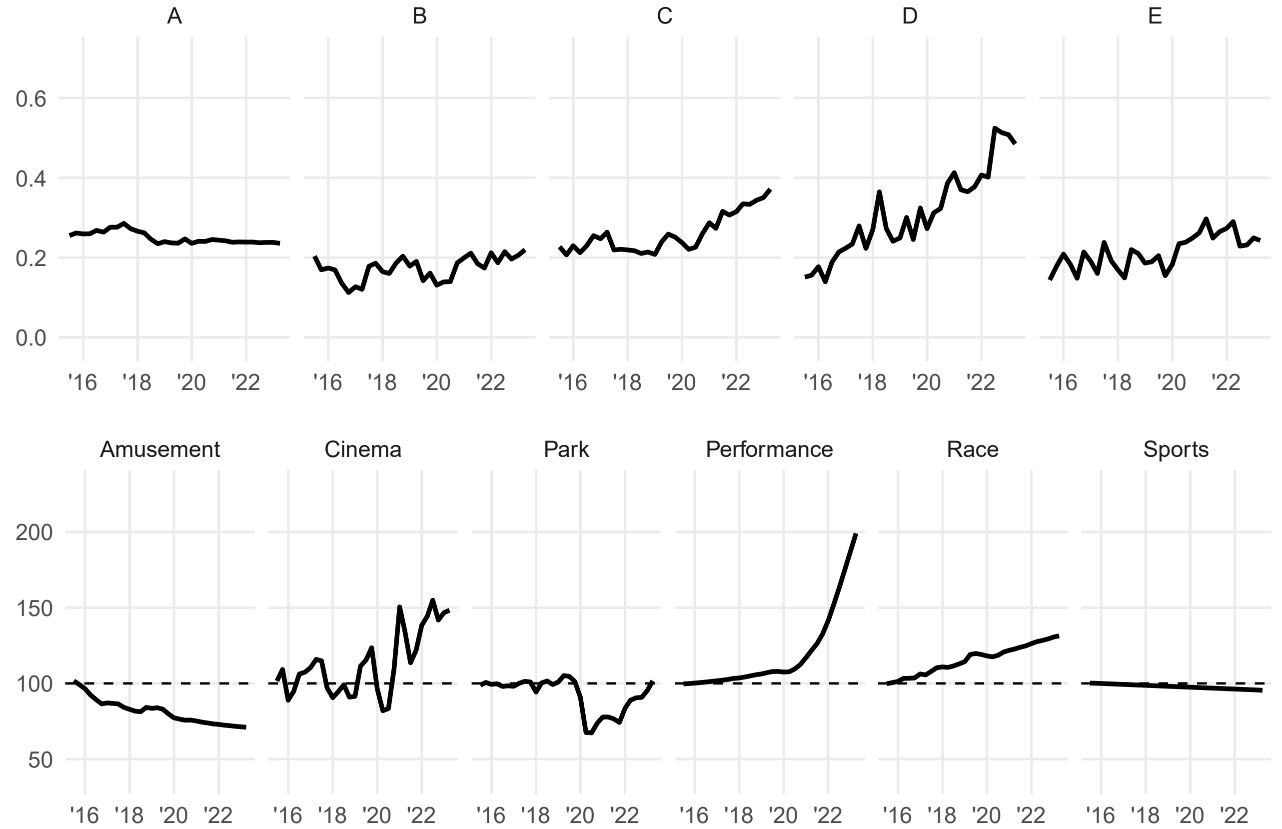
Figure 5 Dynamics of COVID-19 shock coefficient



Note: The shaded gray areas represent 95% confidence intervals.

Source: authors' compilation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

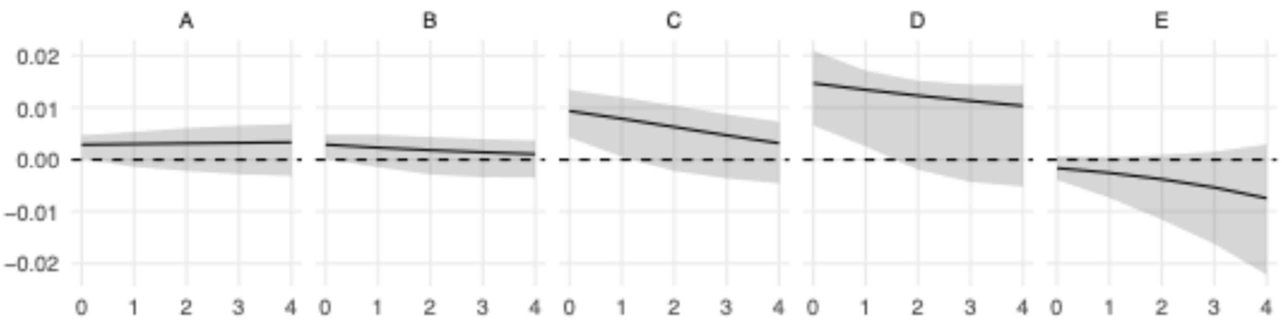
Figure 6 Dynamics of shock-adjusted trends



Note: The shaded gray areas represent 95% confidence intervals.

Source: authors' compilation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

Figure 7 Impact from Live Entertainment Industry



Note: The shaded gray areas represent 95% confidence intervals.

Source: authors' compilation, using data for total ticket sales index from PIA Research Institute, and Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

Table 1 Selection of time-series variables

	Annual Cross-Sectional Average								Avg. CS CV
	2015	2016	2017	2018	2019	2020	2021	2022	
Sales / Total Assets	0.950	0.955	0.919	0.888	0.965	0.609**	0.689	0.803	1.115
COGS / Sales	0.442	0.458	0.456	0.462	0.508	0.475	0.462	0.437	0.823
SG&A / Sales	0.578	0.610	0.616	0.615	0.562	0.817**	0.637**	0.664	0.815
Personnel Expenses / Sales	0.302	0.321	0.313	0.312	0.296	0.418**	0.328**	0.309	0.724
SG&A / Personnel Expenses	2.484	2.439	2.434	2.415	2.314	2.509	2.517	2.681	0.807
Personnel Expenses Per Capita	4.039	4.108	4.070	3.981	4.240	3.971	4.147	4.278	0.510
Depreciation / Sales	0.077	0.078	0.090	0.080	0.081	0.118**	0.088*	0.097	1.220
Depreciation / Non-Current Assets	0.061	0.063	0.070	0.071	0.066	0.058	0.063	0.073	1.337
Inventory / Sales	0.020	0.025	0.025	0.025	0.029	0.043*	0.034	0.029	2.056
Inventory / Total Assets	0.014	0.015	0.016	0.016	0.021	0.017	0.016	0.016	1.902
PP&E / Total Asset	0.586	0.555	0.549	0.543	0.529	0.545	0.522	0.488	0.566
Intangible Assets / Total Assets	0.015	0.017	0.015	0.013	0.017	0.017	0.017	0.022	3.729
I&OA / Total Assets	0.101	0.111	0.103	0.111	0.106	0.097	0.099	0.089	1.399

Notes: ** and * denote significance at the 1% and 5% levels, respectively, based on the Welch test for no difference from the previous year. COGS, SG&A, PP&E, and I&OA represent "cost of goods sold," "selling, general and administrative expenses," "property, plant and equipment," and "investments and other assets," respectively. Avg. CS CV refers to the average cross-sectional coefficient of variation.

Source: authors' calculation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

Table 2 Membership probability under UMAP parameters

	excluding unstable variables				including unstable variables			
minPts	UMAP1	UMAP2	UMAP3	UMAP4	UMAP1	UMAP2	UMAP3	UMAP4
10	0.249	0.159	0.296	0.208	0.295	0.225	0.330	0.295
20	0.260	0.162	0.470	0.237	0.233	0.171	0.377	0.284
30	0.389	0.131	0.473	0.191	0.227	0.202	0.519	0.294
40	0.378	0.093	0.445	0.153	0.198	0.349	0.558	0.242
50	0.340	0.160	0.433	0.220	0.156	0.291	0.518	0.410
60	0.270	0.179	0.380	0.231	0.360	0.299	0.268	0.378
70	0.281	0.130	0.319	0.225	0.315	0.276	0.347	0.376
80	0.245	0.191	0.291	0.200	0.318	0.247	0.299	0.346
90	0.202	0.112	0.264	0.178	0.295	0.244	0.265	0.316
100	0.193	0.102	0.236	0.155	0.310	0.251	0.259	0.291

Note: The minimum distance (*min_dist*) is set to 0.10 for UMAP1 and UMAP2, and to 0.01 for UMAP3 and UMAP4. The number of neighbors (*n_neighbors*) is 10 for UMAP1 and UMAP3, and 50 for UMAP2 and UMAP4. The *minPts* parameter refers to the minimum number of points in HDBSCAN.

Source: authors' calculation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

Table 3 Dynamic clustering and noise

Panel A	minPts=5		minPts=10		minPts=15		minPts=20		minPts=25	
	Clusters	Noise	Clusters	Noise	Clusters	Noise	Clusters	Noise	Clusters	Noise
2015-16	44	0.222	13	0.151	8	0.103	6	0.127	6	0.256
2016-17	34	0.212	11	0.101	6	0.079	6	0.119	6	0.228
2017-18	40	0.193	15	0.283	6	0.073	5	0.092	4	0.135
2018-19	43	0.212	11	0.171	6	0.184	5	0.136	4	0.122
2019-20	31	0.169	11	0.258	6	0.159	5	0.141	4	0.157
2020-21	33	0.189	9	0.082	6	0.092	5	0.259	4	0.251
2021-22	35	0.205	7	0.048	5	0.105	5	0.198	3	0.135

Panel B	minPts=5		minPts=10		minPts=15		minPts=20		minPts=25	
	Clusters	Noise	Clusters	Noise	Clusters	Noise	Clusters	Noise	Clusters	Noise
2015-16	50	0.146	13	0.155	8	0.136	6	0.198	4	0.205
2016-17	39	0.099	15	0.147	8	0.212	5	0.095	4	0.145
2017-18	45	0.148	11	0.128	8	0.163	4	0.169	4	0.165
2018-19	36	0.196	10	0.166	7	0.124	5	0.260	4	0.152
2019-20	28	0.134	11	0.295	6	0.250	4	0.106	2	0.035
2020-21	34	0.159	10	0.172	6	0.236	4	0.134	3	0.127
2021-22	38	0.168	10	0.188	2	0.010	2	0.010	2	0.010

Panel C	minPts=5		minPts=10		minPts=15		minPts=20		minPts=25	
	Clusters	Noise	Clusters	Noise	Clusters	Noise	Clusters	Noise	Clusters	Noise
2015-16	2	0.026	0	1.000	0	1.000	0	1.000	0	1.000
2016-17	3	0.038	2	0.800	2	0.818	0	1.000	0	1.000
2017-18	2	0.034	0	1.000	0	1.000	0	1.000	0	1.000
2018-19	2	0.053	2	0.781	0	1.000	0	1.000	0	1.000
2019-20	2	0.061	0	1.000	0	1.000	0	1.000	0	1.000
2020-21	3	0.206	0	1.000	0	1.000	0	1.000	0	1.000
2021-22	2	0.450	2	0.815	0	1.000	0	1.000	0	1.000

Note: Clusters and Noise represent number of clusters and the rate of noise points, respectively.

Source: authors' calculation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

Table 4 Financial characteristics of dynamic cluster groups

	A	B	C	D	E
COGS / Sales	0.859	0.775	0.240	0.360	0.616
SG&A / Personnel Expenses	0.551	0.739	3.393	2.607	1.936
Personnel Expenses Per Capita	4.827	2.665	2.975	4.793	9.449
Depreciation / Non-Current Assets	0.074	0.056	0.064	0.074	0.045
Inventory / Total Assets	0.020	0.014	0.015	0.016	0.018
PP&E / Total Assets	0.524	0.598	0.574	0.517	0.380
Intangible Assets / Total Assets	0.009	0.032	0.009	0.021	0.040
I&OA / Total Assets	0.126	0.095	0.091	0.077	0.182
Sales Per Capita (¥M)	29.8	10.3	20.2	32.1	76.0
Total Assets Per Capita (¥M)	55.4	34.4	51.9	57.2	218.2
Core Revenue Ratio	0.720	0.850	0.839	0.806	0.622
N	293	196	797	208	120

Note: Core Revenue Ratio refers to the proportion of entertainment-related revenue to total revenue. N denotes the number of firms belonging to the dynamic cluster group.

Source: authors' calculation, using Surveys for the Financial Statements Statistics of Corporations by Industry (Ministry of Finance).

Table 5 Financial characteristics of fundamental subindustries

	Cinema	Performance	Race	Sports	Park	Amusement
COGS / Sales	0.435	0.533	0.011	0.324	0.522	0.738
SG&A / Personnel Expenses	4.730	1.875	0.498	1.791	0.844	4.179
Personnel Expenses Per Capita	1.521	5.013	3.243	2.089	2.816	3.666
Sales Per Capita (¥M)	15.4	36.8	168.0	7.4	13.5	79.7
N	153	2,225	541	6,684	549	7,395

Notes: “Cinema” and “Performance” denote cinemas, and performances (except otherwise classified) and theatrical companies in JSIC, respectively. Similarly, “Race” denotes bicycle, horse, motorcar and motorboat racetrack operations and companies and “Sports” denotes sports facilities. “Park” and “Amusement” denote public gardens and amusement parks, and amusement and recreation facilities in JSIC, respectively.

Source: authors' calculation, using Economic Census for Business Activity (Ministry of Internal Affairs and Communications).