informs
http://pubsonline.informs.org/journal/inte

INFORMS JOURNAL ON APPLIED ANALYTICS
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0092-2102, EISSN 1526-551X

**Submitted to *INFORMS Journal on Applied Analytics***

# Automating Procurement Practices using Artificial Intelligence

**(Authors' names blinded for peer review)**

**Abstract.** **Problem definition:** Conducting a spend analysis of procurement practices is a challenging task for manufacturers. It requires deciphering large-scale spend data in the form of unstructured texts and identifying opportunities for savings. This process relies on procurement experts' know-how and is often performed manually, a laborious task often leading to missed savings opportunities. Automating spend analysis through natural language processing and machine learning presents several challenges, such as (i) a lack of true detailed category labels for suppliers, (ii) a lack of sufficiently large sets of training data, (iii) hierarchical taxonomies that vary across manufacturers, and (iv) the reduced accuracy of hierarchical categorization algorithms beyond two levels.

**Methodology:** Our novel three-component classification model tackles these issues, facilitating the automation of spend analysis and the replication of procurement experts' decision-making processes. By processing input data composed of unstructured spend texts from Cranswick plc, a leading UK food producer, our model delivers accurate supplier categorizations that pinpoint areas ripe for substantial savings.

**Results and managerial implications:** This approach not only shows greater accuracy compared to existing benchmark models but also aids in identifying key product categories and suppliers for cost-saving initiatives. By simulating the application, we project that our method could bring annual savings of £16-22 million ($20-28 million) for Cranswick plc, illustrating the significant advantages of automating spend analysis.

**Key words:** spend analysis, data-driven procurement, natural language processing, machine learning

## 1. Introduction

Procurement in large manufacturers is a complex operation, often involving the purchase of tens of thousands of products from thousands of suppliers, with annual purchase costs in the hundreds of millions of dollars. In 2020, US manufacturers alone spent $2.8 trillion in procurement, more than half of their revenue (U.S. Census Bureau 2022). However, such large-scale procurement processes often lack transparency across the entire company and rely on a medley of heterogeneous legacy systems that operate in silos, many

of which were assimilated through previous mergers and acquisitions. The complexity of regularly updating the digital infrastructure and procedures of even a single business unit, let alone standardizing them across the entire corporation, is daunting. This situation results in many inefficiencies and missed opportunities. For example, it is typical to observe that the same types of products are procured from different suppliers at significantly different prices.

Identifying savings opportunities in large-scale procurement practices can bring big rewards. Therefore, manufacturers periodically conduct a spend analysis across their business units, often assisted by a third-party (consulting) firm. An accurate spend analysis requires collecting and making sense of a vast amount of purchase order records, as well as organizing suppliers and products into detailed hierarchical categories. However, this is often challenging due to the unstructured and inconsistently formatted nature of the data, which is typically in the form of text. For instance, the same suppliers may be referred to by different names or descriptions across various business units, and the purchased products may be recorded using acronyms or abbreviations that can be difficult to understand without proper supplier context. Moreover, practically speaking, relying on supplier self-classification can be unreliable without verification. For example, in the UK, suppliers' self-reported Standard Industrial Classification (SIC) codes are highly inaccurate. In our samples, we have found less than 20% to be accurate. Often, they only specify general level characteristics or choose "other category."

As a result, a spend analysis is typically performed manually (with the aid of spreadsheets) by procurement experts, who possess the requisite nuanced industry knowledge to understand the nature of procurement transactions and read between the lines. As such, this process is both time-consuming and expensive, and often involves multiple external procurement experts and can take several months to complete. Given the impracticality of manually filtering through thousands of suppliers, the scope of such manual analysis is often limited to a subset of suppliers with high procurement spending. This approach can be biased and prone to errors, ultimately limiting the identification of cost-saving opportunities.

A natural evolution in enhancing procurement practices, consistent with the theme of Industry 4.0, would be to automate (or semi-automate) spend analysis (Olsen and Tomlin 2020). By adopting an automated approach, businesses can easily integrate, connect, and interface their procurement practices. This can be particularly advantageous for large manufacturers frequently acquiring or restructuring new business units, or SMEs (Small and Medium-sized Enterprises) facing cost barriers for spend analysis. As such, this proposition has generated substantial attention during the pandemic in general (Dittrich et al. 2020), and significant investments from leading consulting firms in particular (McKinsey 2021, Garcia 2021).

However, while machine learning and natural language processing (NLP) are well-suited for automating these tasks, integrating these technologies into procurement to match the nuanced insights of experts presents four significant challenges. First, there exist no true hierarchical category labels for any given set of suppliers. In other words, no third-party organization accurately classifies unlisted suppliers into specific

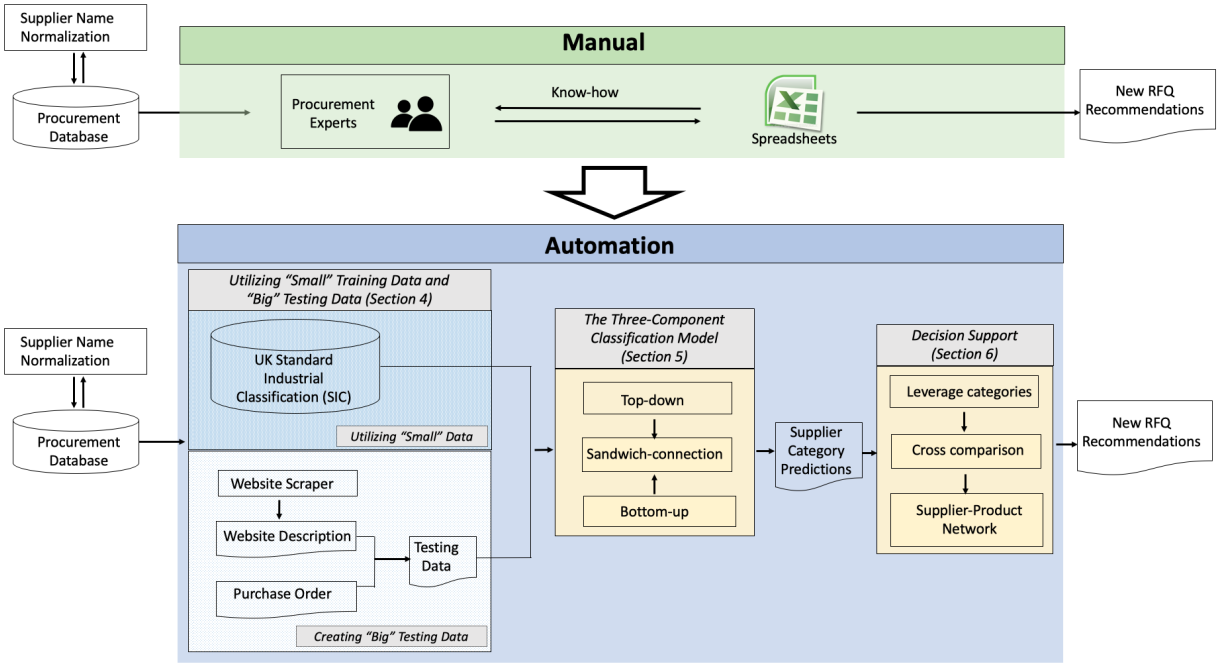
industry categories, and existing self-reported categories (e.g., in the case of the UK) are highly inaccurate and unreliable for practical use. Second, for any given supplier category, sufficiently large pre-existing training data does not exist. For example, for any group of suppliers providing specific product categories, there may be at most a few dozen suppliers, which is not enough to train a reliable classifier. Third, the automatic spend analysis needs to be flexible and applicable in different contexts, not just in a specific supplier category. Since machine learning classifiers depend heavily on the training data set, a hierarchical classifier developed for one manufacturer may not be appropriate or applicable to another. Finally, building a hierarchical classification algorithm that is accurate beyond two levels of hierarchy is fundamentally difficult because classification errors propagate between levels (Dumais and Chen 2000). To identify savings opportunities accurately in practice, however, five to six levels of hierarchical sub-categories are needed.

In this paper, we present a comprehensive methodology that employs NLP and machine learning to automate spend analysis that successfully replicates the procurement expert's know-how. Our methodology categorizes suppliers' data into a detailed and extensive hierarchical taxonomy with 6 levels and 15,574 distinct categories. This structured categorization helps to identify leverage suppliers that can result in realized cost savings of 5-10% of the current invoice values when the request-for-quote (RFQ) process is initiated. The architecture is depicted in Figure 1, and we provide a detailed expansion for each main section in the subsequent paragraphs.

Our methodology begins with addressing the lack of proper training data (in the section 'Utilizing "Small" Training Data and "Big" Testing Data'). To do so, we utilize "small data" from a detailed hierarchical taxonomy structure alongside "big data" to enrich the testing data. Specifically, we train a classifier to learn the UK's Standard Industrial Classification (SIC) guide (UK Office for National Statistics 2007), a standardized hierarchical taxonomy guide describing the company's nature of business. This guide can be easily replaced by other standardized taxonomy guides such as the United Nations Standard Product and Services Code (UNSPSC), the North American Industry Classification System (NAICS), or any company's internal taxonomy. This "taxonomy" can be an evolving document that manufacturers can customize over time, e.g., add more detailed descriptions for their own procurement needs. This "small data" from hierarchical taxonomy mirrors the procurement expert's logic and intuition that are applied when classifying suppliers. Then, we enrich the suppliers' information by incorporating web-scraped texts (containing general contextual supplier information and detailed product information) in addition to the raw purchase order data from the procurement database. This "big data" mirrors procurement experts' general contextual knowledge and familiarity with specific product descriptions.

Next, in the section "Three-Component Classification Model," we propose a three-component classification model. It utilizes (a) traditional top-down classification, effective for higher levels of the taxonomy (e.g., level 1); (b) traditional bottom-up classification based on word similarity, targeting lower taxonomy levels (e.g., level 6); and (c) an innovative "sandwich-connection" component that merges predictions from

**Figure 1** **The spend analysis system allows a transition from a manual process, which relies on expert knowledge and spreadsheets, to an automated process that uses "small data" from a hierarchical taxonomy structure and "big data" from web-scraped text. In this automated system, a three-component classification model–combining top-down, bottom-up, and a novel "sandwich-connection"–generates supplier category classifications, while a decision support tool identifies potential savings opportunities with new RFQ recommendations.**




the two traditional methods, leveraging parent-child node relationships within the taxonomy. Our three-component classification model identifies the most likely categories for each supplier through levels 1 to 6. This creates a unique "DNA" profile for each supplier, represented as a binary vector that includes all sub-categories from levels 1 to 6. To evaluate our model's accuracy, we derived a partial list of true labels for a small, randomized subset of suppliers through a series of carefully designed experiments, utilizing both crowdsourcing and expert validation. Against these best estimates of the true labels, we demonstrate that the three-component model significantly enhances prediction accuracy (as measured by the F1 score) compared to existing benchmark models in hierarchical classification.

Lastly, in the section "Decision Support Tools," our methodology is accompanied by a decision support tool that can convert the classification into potential savings opportunities. It performs a Kraljic analysis (Kraljic 1983) to identify the "leverage" categories of suppliers (those with low risk and high economic volume). Moreover, using the supplier's DNA, it enables the cross-comparison of all suppliers to identify a supplier's competitors or find a list of suppliers that can provide certain products. This approach provides a detailed understanding of the supplier-product networks within manufacturers, aiding in the strategic

selection of suppliers. It also helps in crafting targeted RFQs, which can lead to potential cost savings through volume or price discounts.

Practically, in the section "Advantage of Automated Spend Analysis: A Simulation Study," we report the implementation of our automated spend analysis in Cranswick plc, a leading food manufacturer in the UK. It provided us with detailed data on its procurement transactions over two years from January 2019 to December 2020, which amounted to a total invoice value of £1.571 billion. Our automated spend analysis was able to examine all 2,170 suppliers and accurately classify them into hierarchical categories. Together with the decision support tool, our methodology was instrumental in identifying the "leverage" supplier categories and generating a list of target suppliers to issue RFQs and the estimated cost-savings within days (instead of months).

If Cranswick plc follows through on the RFQ recommendations, significant cost savings are achievable. To estimate the cost-savings that are attributable to automation, we performed a simulation analysis (based on a model calibrated using Cranswick plc data) that incorporates (i) an improved scope of analysis, (ii) increased accuracy in classifications, and (iii) increased frequency of spend analysis performed. Specifically, over a two-year period, the automation can generate additional savings of 2-3% of total procurement costs compared to traditional manual spend analysis methods. This translates into £16-22 million in annual savings, underscoring the significant financial advantages of adopting automated spend analysis techniques.

To the best of our knowledge, this paper is the first academic study to formalize the automation of spend analysis. We introduce a comprehensive, practically implementable methodology and demonstrate its effectiveness. This methodology leverages a standardized hierarchical taxonomy, reducing the dependency on large datasets traditionally required for machine learning algorithms, thereby enhancing the effectiveness and efficiency of spend analysis in real-world applications. For example, our cost-effective and efficient methodology can make the spend analysis accessible to mid-sized companies who previously faced cost barriers, and help digitization to spread across industrial sectors. By developing the three-component classification model, our methodology provides an accurate understanding of procurement, helping firms paint a detailed picture of their supply chain networks. Additionally, our methodology can be beneficial in other areas, such as legal technology and financial services, where there might be a scarcity of extensive training data. Instead, these fields often have an abundance of documents detailing compliance matters or descriptions of services, making the methodology adaptable and valuable for broader applications.

## 2. Literature Review

Our paper contributes to the field of hierarchical classification, specifically in the context of text data. This area of research delves into three types of classifiers (Silla and Freitas 2011). The first approach, known as the top-down approach, involves constructing a separate flat classifier for each level of the hierarchy. This approach only predicts a node if its ancestor nodes have also been predicted, a strategy documented in various studies (Dumais and Chen 2000, Cesa-Bianchi et al. 2006, Esuli et al. 2008, Cerri et al. 2014). However,

a notable drawback of the top-down approach is the propagation of classification errors from upper to lower levels within the hierarchy. This flaw becomes particularly problematic in hierarchies that extend beyond two levels, where errors at higher levels inevitably affect the accuracy of lower-level classifications.

The second approach, known as the bottom-up approach, starts by predicting labels at the leaf nodes and then infers their ancestors' labels through heuristics (Ceci and Malerba 2007). However, this approach ignores the information about the parent-child relationship along the hierarchy, and thus leads to low accuracy when the number of leaf classes becomes large.

The third approach is the so-called big-bang approach where a single and comprehensive classifier is built to address the entire classification problem. Unlike top-down or bottom-up approaches that break down the classification problem into smaller tasks, the big-bang approach aims to leverage the full complexity of the data structure and class hierarchy from the outset. Most of the studies focus on optimizing traditional machine-learning algorithms (e.g., neural network, support vector machine, decision tree); these require a large set of pre-labeled training data (e.g., McCallum et al. 1998, Cai and Hofmann 2004, Peng et al. 2018, Mao et al. 2019), and are therefore infeasible in the context when such training data is not available, as is the case in procurement.

Addressing these challenges, our innovative three-component classification model combines the strengths of both the top-down and bottom-up approaches through what we call the "sandwich connection"—a methodology that effectively bridges the two approaches, to be detailed in the section "Three-Component Classification Model." This methodology achieves significantly improved accuracy across hierarchical structures that are both deep and broad. As a result, we introduce a classification methodology that is appealing to a wide range of practitioners. Our approach effectively overcomes the constraints associated with traditional classification methods, paving the way for precise and efficient classification of hierarchical data. To our knowledge, this is the first instance of creating a hierarchical classification model that operates without relying on pre-defined true labels to categorize suppliers on a large scale.

Accurate hierarchical classification is not merely an academic concern but a practical necessity in procurement, where selecting the right suppliers and products is critical for issuing RFQs effectively. The design of effective procurement mechanisms has long been a prominent area of research in the operations management field (Vickrey 1961, Laffont and Tirole 1993, Elmaghraby 2000, Hasenbein et al. 2010). Recent developments have focused on identifying the optimal sourcing strategies under different market attributes and environments (e.g., Chaturvedi et al. 2014, Li and Wan 2017, Beil et al. 2018), as well as examining the best way to issue the RFQs (e.g., Beil and Wein 2003, Wan and Beil 2009, Duenyas et al. 2013). Our paper complements these theoretical studies by examining the upper stream concerns of which supplier/product categories among its vast suppliers to issue RFQs in the first place. Although the field of data analytics is burgeoning (Mišić and Perakis 2020), its reach has been somewhat limited in the procurement space. We contribute by incorporating industrial-level data to address practical procurement problems and bringing analytics into this classic operations space.

# 3. Problem Description

A spend analysis carried out by a manufacturer aims to create transparency in its procurement practice (i.e., the products purchased, the suppliers it purchased from, and the total quantity and cost of the purchase), identify opportunities for savings, which could then be utilized to initiate RFQ processes to reduce costs (e.g., through private negotiations with suppliers or public auctions). To do so, one must gather and organize the procurement data, group the suppliers and products into (hierarchical) categories based on their similarities, and provide nuanced insights into the procurement process. In this section, we describe the challenges associated with conducting a spend analysis using Cranswick plc as an example.

## 3.1. Cranswick plc's Transaction Data

Cranswick plc produces a range of fresh foods with a fully integrated supply chain. It sources from farmers, processes raw products, packages, and labels products, and ships them globally. It is a member of the FTSE 250, with a total reported revenue of £2 billion (approx. $2.5 billion) in 2022 (Cranswick 2023).

Cranswick plc provided us with its procurement transaction data that they assembled from multiple internal data sources from 2019 to 2020 as a spreadsheet. Each row represents a purchase order, including product-related information (e.g., item description, item code, item price), order-related information (e.g., order date, order quantity, order currency, total invoice value, delivery date), buyer-related information (e.g., a business unit within Cranswick plc), and supplier-related information (e.g., supplier name). The suppliers are not publicly listed firms and do not have an official classification that is consistently referenced.

The data covers 556,866 procurement transactions with a total invoice value of £1,571 million across two years. There are 136,190 products (identified by item description) and 2,999 suppliers (identified by supplier name). Table 1 summarizes the key information of the raw procurement data by years.

**Table 1 The Cranswick procurement transaction data used in this paper contains 556,866 transactions with a total invoice value of £1,571 million over the years 2019 and 2020. The data includes information on 136,190 products and 2,999 suppliers.**

| | 2019 | 2020 | Total |
|---|---|---|---|
| Number of suppliers | 2,161 | 2,922 | 2,999 |
| Number of products | 68,998 | 78,277 | 136,190 |
| Number of purchase orders | 265,662 | 291,204 | 556,866 |
| Number of business units (buyer) | 12 | 13 | 13 |
| Total invoice value (£ million) | 717 | 854 | 1,571 |

To interpret the raw data effectively, it is necessary to standardize supplier names due to the frequent inaccuracies or variations in recording these names across different business units (e.g., "ABC Limited," "Advanced Business Corp," "ABC"). Procurement experts achieve this by conducting a Google search for each listed supplier name and collecting the top website URL from the search results. When the search

for two or more listed suppliers leads to the same URL, they are considered to be the same entity and are consolidated under the official name listed on the website. This verification process yielded 2,170 (from 2,999) unique suppliers, of which 1,921 had an official website URL.

### 3.2. Manual Spend Analysis: Relying on the Know-How of Procurement Experts

In the process of manual spend analysis, procurement experts are tasked with understanding the scope of suppliers and their products from the raw procurement data. This undertaking is deeply dependent on the nuanced industry knowledge of the experts, challenged by several factors. Firstly, a significant portion of the purchased products are listed in a non-descriptive manner (e.g., "2 SIS ANG M&S HW BR," "JBS CBEEF"), making identification difficult. Additionally, product descriptions can sometimes be misleading. For instance, an item listed as a "1200mm wide pretzel" might initially seem related to food items. Yet, an experienced procurement expert, recognizing the unusual size (1.2 meters), would correctly deduce that it refers to a part for a conveyor belt machine.

Secondly, raw procurement data often lack vital contextual details necessary for accurate supplier classification. For example, when a purchase order lists "boilers," it is challenging to know whether the supplier is a manufacturer, wholesaler, or service provider of boilers. Distinguishing the supplier's role is vital for a comprehensive understanding of procurement practices and supplier networks. Consequently, spend analysis extends far beyond merely identifying familiar terms. Adding to the complexity, suppliers may offer a diverse range of products and services. While one supplier might focus on a specific product, another could provide a broad array of goods and services. Recognizing the extent of a supplier's product range and service offerings is also important for negotiating price and volume discounts.

Thirdly, and perhaps most importantly, a procurement transaction indicates which product was purchased from a supplier, but it does not reveal which other products could have been purchased from it. Thus, two similar suppliers could appear very different based on the procurement data alone. As a result, drawing insights into supply networks from the procurement data requires procurement experts' familiarity with the industry, and their ability to read between the lines.

After fully analyzing the transaction data, it is necessary to classify suppliers into hierarchical groups according to their similarities. The objective is to determine which categories of suppliers to approach for an RFQ process, aiming to negotiate savings via volume or price discounts. The deeper the hierarchy and the more granular categories, the better for gaining detailed insights for target identifications. Utilizing a hierarchical structure proves beneficial in practical scenarios, particularly because a manufacturer's procurement operations are often decentralized by product categories. This approach facilitates the alignment of specific procurement teams with the corresponding categories for the execution of RFQs.

However, manually comparing thousands of suppliers, even with the help of spreadsheets, is difficult. Typically adhering to the Pareto principle, a procurement expert might concentrate on a limited subset of

suppliers (e.g., 20%) that account for a significant portion of procurement expenses (approximately 80%). In the case of Cranswick plc, we find that 68 of 2,170 suppliers (3.1%) are responsible for 80% of the total invoice value. Consequently, a manual spend analysis would prioritize these 68 suppliers, constructing a detailed hierarchical classification centered around them, based on the procurement experts' knowledge. This process, driven by speculative hypotheses about the supply network, is inherently susceptible to biases and inaccuracies, overlooking the majority of suppliers and potentially missing out on savings opportunities. Moreover, the insights from one manufacturer are not transferrable to another, which requires the procurement experts to start from scratch when examining a new manufacturer. As such, this process is both time-intensive and costly, and is typically reserved only for large manufacturers with sufficient procurement spend volume to justify the costs.

### 3.3. Automation Challenges

Designing a "smart" methodology that can infer a procurement expert's nuanced understanding of the procurement setting from the vast procurement data (e.g., what specific items a particular supplier could produce) seems to be a natural evolution towards Industry 4.0. Although it may seem apparent that NLP and machine learning techniques could be employed, creating a hierarchical classification model in a procurement context faces four methodological challenges.

First, in a procurement environment, there is an absence of properly structured training data with predefined categories, meaning there are no existing true hierarchical category labels for each supplier to utilize. Second, developing a classifier that can accurately group a given supplier requires a significant volume of training data, often hundreds or thousands of samples. Yet, for each specific sub-category, the market might only offer a few dozen trustworthy suppliers, leading to a scarcity of large and pre-labeled datasets necessary for traditional machine learning approaches in procurement. Third, the effectiveness of machine learning classifiers is highly dependent on the dataset they were trained on, making a classifier designed for one sector (such as agriculture) potentially ineffective in another (such as heavy manufacturing). Fourth, even when there is enough pre-labeled data available, constructing an accurate multi-level hierarchical classification model that goes beyond two levels presents its own set of challenges. However, for the model to be of practical use, it needs to correctly classify suppliers across five to six hierarchical levels.

In the ensuing sections, we present a methodology that overcomes all of the above challenges and replicates the know-how of procurement experts.

## 4. Utilizing "Small" Training Data and "Big" Testing Data

Our aim is to find the most likely set of categories that a supplier belongs to. To address the challenge presented by the lack of extensive pre-labeled training datasets, we propose to train a classifier on a comprehensive 6-level Standard Industrial Classification (SIC) taxonomy, which we term as "small" training data. Then, we enrich the supplier's information from raw procurement data to create "big" testing data.

### 4.1. Utilizing "Small" Data

We aim to develop a machine-learning model that learns the details of the SIC guide. The SIC is a well-established taxonomy introduced by the UK government to classify business establishments by the type of their economic activity. It is a hierarchical supplier classification system, describing the activities and sub-activities in a hierarchical tree structure (a directed graph with each node having at most one parent node). The nodes in the tree contain category labels corresponding to the level of specificity. For instance, the highest level (i.e., level 1) categories represent the most general business activities, such as "manufacturing," "agriculture," "wholesale," etc. Then, each level-1 category is broken down into more specific sub-activities at level 2. For example, the level-1 category "manufacturing" is broken into level-2 categories such as "Manufacture of food products," "Manufacture of beverages," and so on. Further, every level-2 category is broken down into level 3, and then into level 4. In some cases, the level-4 category is further broken down into level 5. Therefore, the lowest-level SIC categories can be either at level 4 or level 5 of the hierarchy, describing the most specific business activities, e.g., "butter and cheese production," "Growing of citrus fruits," etc. For each lowest-level supplier category (i.e., level 4 or 5) in the guide, the SIC further details the types of specific products that the suppliers should carry. For example, level-5 category "Butter and cheese production" contains eight detailed product categories: "Butter blending," "Butter milk," "Butter oil," "Butter production," "Butterfat," "Cheese," "Curd production," and "Dairy preparation of cheese and butter." Table 2 shows two examples of SIC hierarchical categories.

**Table 2 Two detailed examples of hierarchical SIC categories illustrate the refinement from general categories (Level 1) to specific products (Level 6). Example 1 outlines levels from "manufacturing" to "butter oil production" through six levels, while Example 2 outlines levels from "agriculture" to "lemon growing" across five levels.**

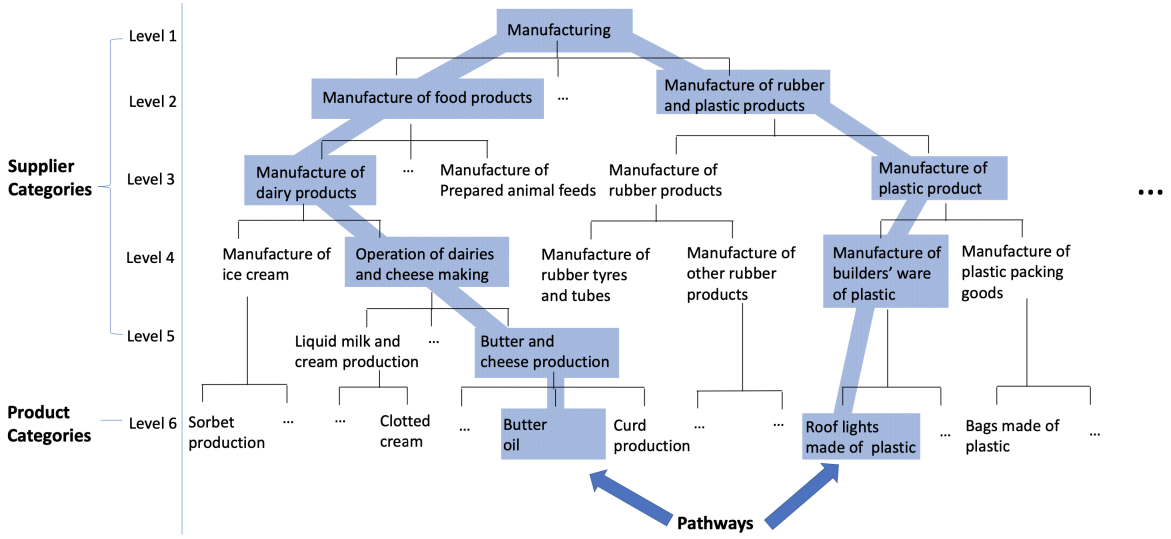
| | Example 1 | Example 2 |
|---|---|---|
| Level 1 | Manufacturing | Agriculture |
| Level 2 | Manufacture of food products | Crop and animal production, hunting and related service activities |
| Level 3 | Manufacture of dairy products | Growing of perennial crops |
| Level 4 | Operation of dairies and cheese making | Growing of citrus fruits |
| Level 5 | Butter and cheese production | — |
| Level 6 | Butter oil | Lemon growing |

We define a SIC pathway as a sequence of nodes from the level-1 category to the level-6 category (see Figure 2). In the cases of original categories that end at level 5 (e.g., Example 1 in Table 2), the pathways in expanded SIC contain nodes in six levels; in the cases of original categories that end at level 4 (e.g., Example 2 in Table 2), the pathways in expanded SIC contain five hierarchical nodes without level 5.

In our three-component classification model, we design the hierarchical taxonomy to reflect the industry structures and classification logic used by procurement experts. While our approach utilizes the SIC system

for demonstration, it is flexible and can be applied to other comprehensive taxonomies such as the United Nations Standard Products and Services Code (UNSPSC) or the North American Industry Classification System (NAICS), as well as to pre-existing internal supplier category databases. The chosen taxonomy, while serving as an initial framework, is intended to be adaptable and evolve to meet the specific needs of manufacturers. Throughout the paper, we train the model to learn the expanded 6-level SIC taxonomy so that it can classify a supplier into appropriate pathways.

**Figure 2** **The tree-structured hierarchical categories illustrate the pathway from general categories (Level 1) to specific product categories (Level 6). It highlights two example pathways: one from "manufacturing" in Level 1 to "butter oil" in Level 6, and another from "manufacturing" in Level 1 to "roof lights made of plastic" in Level 6.**




Each node of the SIC tree is a category, and associated with it is a textual label. We construct the training data for each label by joining the texts of the category labels that appear in itself and all its child nodes. Take the left-hand-side pathway in Figure 2 as an example. The training data for its level 1 category "manufacturing" is the joint text from itself and all its child nodes in levels 2-6 (i.e., all the text labels in all pathways), and the training data for its level 6 category is the label itself (e.g., *"Butter oil"*).

To facilitate text data, we preprocess all training categories and training data by splitting the text into its component words, eliminating punctuation and numbers, lemmatizing words into dictionary form, and removing single-character words and stop words. Table 3 summarizes the statistics of SIC categories/labels and training data by hierarchical levels. SIC includes 21 level-1 categories and 15,574 level-6 categories, indicating 15,574 unique pathways. We find that 2,975 (out of 15,574) pathways traverse all six hierarchical levels, whereas the remaining 12,599 pathways are without level 5. The average length of the training data is the number of words after preprocessing. We observe that higher-level categories tend to have longer

training data than lower-level categories, which is intuitive given the hierarchical structure depicted in Figure 2. Moreover, pathways that traverse all six levels tend to have longer training data than pathways that are without level 5.

**Table 3 The number of categories and the average training data length are compared across hierarchical levels from 1 to 6, between pathways that traverse all six levels and those that exclude level 5.**

| | Number of categories/labels | | | Average length of the training data | | |
|---|---|---|---|---|---|---|
| | Total | pathway traverse all six levels | pathway without level 5 | Total | pathway traverse all six levels | pathway without level 5 |
| Level 1 | 21 | 15 | 21 | 18.80 | 24.01 | 17.56 |
| Level 2 | 88 | 43 | 87 | 16.56 | 21.69 | 15.35 |
| Level 3 | 271 | 65 | 242 | 12.92 | 18.01 | 11.72 |
| Level 4 | 614 | 78 | 536 | 8.99 | 13.86 | 7.83 |
| Level 5 | 191 | 191 | – | 9.81 | 9.81 | – |
| Level 6 | 15,574 | 2,975 | 12,599 | 3.95 | 3.86 | 3.97 |

The SIC taxonomy organizes categories into a hierarchical structure, spanning from level 1 to level 6. Higher levels contain broader categories, while lower levels consist of more detailed ones. For instance, level 1 includes 21 general business categories, such as "manufacturing," whereas level 6 encompasses 15,574 specific product categories, such as "Butter oil."

For any category in the SIC taxonomy at levels 1 and 6, we create a corresponding feature representation. It it worth noting that although it is possible to create feature representations for categories at levels 2, 3, 4, and 5 as well, our model only focuses on creating these representations for categories at levels 1 and 6.

For the categories at level 1, we first extract the unique unigrams and bigrams from each of the 15,574 level-1 training data, which results in 38,189 features (i.e., 6,151 unigrams and 32,038 bigrams). Then, for each level-1 category, we compute the TF-IDF (term frequency-inverse document frequency) scores of each of 38,189 features. To enhance machine learning model performance, we select those features in the top $20^{th}$ percentile as the most informative features (Ramos et al. 2003, Domingos 2012), narrowing down to 7,729 features (i.e., 2,342 unigrams and 5,387 bigrams). Finally, for each of the 21 level-1 categories, we create a feature vector by averaging the TF-IDF scores of these features across the training data.

For the categories at level 6, we extract all unique single words (unigrams) from the 15,574 training examples, focusing on capturing product word similarities. This process results in 6,123 unigrams. For each level-6 category, we create a corresponding feature vector based on the TF-IDF scores of these unigrams. Since there are many categories (i.e., 15,574) compared to the number of features (i.e., 6,123) at this level, we retain all features to ensure comprehensive coverage and enhance the accuracy of our model.

## 4.2. Creating "Big" Testing Data

The testing data includes 2,170 suppliers from Cranswick plc's procurement transactions, and each supplier needs to be classified into a set of SIC hierarchical categories.

To reflect the procurement expert's contextual understanding and familiarity with specific product descriptions, we enrich the suppliers' information with "big" data available on the web. Specifically, for each of the 1,921 suppliers with identified URLs in the verification process, we scraped its general business description and detailed product information from the official website. The added website data provides us with general descriptions of the suppliers, the contexts into products, and specific product descriptions that suppliers could provide but have not been recorded in raw procurement transactions.

A supplier $m$ in the Cranswick plc testing data is represented by its testing document $d_m$. For the 1,921 suppliers with official websites, we represent supplier $m$'s testing document by $d_m \equiv (d_m^{gen}, d_m^{spe}, d_m^{PO})$, comprising text data associated with the supplier's general business description obtained from the web ($d_m^{gen}$), its specific product description from the web ($d_m^{spe}$), and the purchased item descriptions from the raw procurement purchase order data ($d_m^{PO}$). For the remaining 249 suppliers without official websites, $d_m = d_m^{PO}$.

Table 4 provides the summary statistics of the preprocessed testing data. For purchased order data in the raw procurement database, although the purchased item description is typically a long text with 474.78 words on average, it only contains 38.98 unique words. These long texts often arise from standardized instructions for restocking previous purchase orders. The limited text data from purchase orders have been significantly enriched by the scraped website data, which contains the general business description with an average of 68.15 unique words per supplier as well as the detailed product description with 90.98 unique words. Note that the testing data quality can be further improved by adding public LinkedIn page information or other private information from third-party organizations.

**Table 4 Summary statistics for the preprocessed testing data show a comparison of the average document length and the average length with unique words across various data sources, including purchased item descriptions, general business descriptions, detailed product descriptions, and suppliers' testing documents.**

| | Purchased Item Description ($d_m^{PO}$) | General Business Description ($d_m^{gen}$) | Detailed Product Description ($d_m^{spe}$) | Suppliers' Testing Document ($d_m$) |
|---|---|---|---|---|
| Number of suppliers (after verification) | 2,170 | 1,921 | 1,921 | 2,170 |
| Average length | 474.78 | 101.20 | 188.84 | 764.82 |
| Average length with unique words | **38.98** | 68.15 | 90.98 | **198.11** |

# 5. Three-Component Classification Model

We now introduce a three-component hierarchical classification model, aiming to classify testing data (e.g., a suppliers' document $d_m$) into a deep and broad hierarchical taxonomy (e.g., an expanded SIC taxonomy).

## 5.1. Methodology

The three-component classification model aims to classify a supplier into its appropriate categories across all levels, from level 1 to level 6. A supplier $m$, represented by its testing document $d_m$, is fed into both the top-down and bottom-up components of the model. Although we illustrate using $d_m$, we allow for flexibility in the testing documents. That is, any combination of $d_m^{gen}$, $d_m^{spe}$, or $d_m^{PO}$ can replace $d_m$. We will discuss the model performance under different data sources in the "Accuracy of the Three-Component Classification Model" subsection. The details of the model's three components are shown in Appendix A, where we also provide the corresponding mathematical notation in Appendix A.1. Below, we give an overview of how each component contributes to the classification process.

The top-down component constructs a flat classifier that classifies a supplier's testing document $d_m$ into the appropriate level-1 categories. In level 1, recall that we have represented each of the 21 level-1 categories by its corresponding feature vector. Thus, the testing document is first transformed into a vector of these same features. Then, the model compares the similarity between the feature vector of the testing document and those of the level-1 categories. This process helps identify which level-1 category the document most closely matches. For illustration, Appendix A.2 shows an example of the computation process for the top-down component.

The bottom-up component constructs a flat classifier that classifies a supplier's testing document $d_m$ into the appropriate level-6 categories. Recall that each level-6 category has been represented by a feature vector in dimension 6,123. Therefore, the testing document is transformed into a vector based on these 6,123 relevant features at level 6. The model then compares the similarity between the feature vector of the testing document and those of the level-6 categories to determine the best match. For illustration, Appendix A.3 shows an example of the computation process for the bottom-up component.

After identifying the set of level-1 and level-6 categories that a supplier most likely belongs to, the sandwich-connection component aims to identify the set of most likely SIC pathways (from level-1 category to level-6 category) that a testing supplier belongs to. This process combines the insights from the top-down and bottom-up components, and takes advantage of the parent-child relationships. For example, suppose that a supplier related to "boiler" is identified at level 6. If this supplier has a high level-1 similarity with "manufacturer" but a low similarity with "wholesale," then all pathways originating from the "wholesale" category would be pruned. Among all 15,574 pathways, we can isolate the most likely pathways by examining the product of the normalized similarity scores. We show the computation process of the sandwich-connection component in Appendix A.4.

As a result, our three-component model effectively yields similarity scores across all 15,574 pathways. To further refine our results, we introduce a parameter $q \in (0, 1)$, which represents a percentile threshold within these similarity scores. Setting $q$ to 0.99, for example, means that we retain only the top 1% of pathways based on their similarity scores. We explain how we select the optimal $q$ in the "Accuracy of the Three-Component Classification Model" subsection.

Ultimately, our three-component classification model classifies a supplier $m$ into a set of pathways through level-1 to level-6 categories. Note that this process automatically identifies all relevant categories at the intermediate levels (i.e., levels 2 to 5), ensuring a comprehensive hierarchical classification for each supplier. As a result, the classification output for supplier $m$ can be seen as its unique "DNA" representation.

### 5.2. Overcoming the Challenges to Accuracy Evaluation

To evaluate the accuracy of our three-component classification model, we compare the classification output of suppliers with their corresponding true labels.

As mentioned previously, one key challenge is the absence of these actual label sets for each supplier. Furthermore, accurately identifying these sets for all 2,170 suppliers out of 15,574 possible true pathways is practically infeasible. Therefore, we developed an approach that combines crowdsourcing with expert validation to gather true labels. Below is a brief overview of our approach, with the full details of the label collection process, including the mathematical representation, provided in Appendix B.

The procedure for collecting true labels includes five steps. The first two steps focus on selecting a representative sample of suppliers and determining the most accurate category pathways for this sample down to levels 4 or 5. This task was manageable due to the relatively smaller number of pathways at levels 4 and 5 (727 total) compared to the 15,574 pathways at level 6. To construct the subset of suppliers, we applied the Pareto principle, selecting 68 suppliers that account for 80% of the economic volume, along with 210 additional suppliers chosen randomly from the remaining group. This resulted in a subset of 278 suppliers for which we estimated the true hierarchical categories down to levels 4 and 5.

The next three steps focus on identifying the correct category at level 6. In our sample of 278 suppliers, 107 unique categories were identified at levels 4 and 5, which in turn included 3,258 level 6 categories. Since it would be time-consuming and costly to determine the true labels for such a large number of level 6 categories, we decided to narrow our focus to a smaller group of suppliers. To do so, in Step 3, we counted the number of suppliers in each true level 4/5 category, and selected the top four level 4/5 categories (see Table B2 of Appendix B). These top four categories included 105 unique suppliers. We then focused on this group of 105 suppliers, repeating the crowdsourcing and expert validation process to identify the most accurate level 6 categories. This allowed us to estimate the true hierarchical categories down to level 6 for these 105 suppliers.

The best estimates of the true labels for 278 suppliers are identified for levels 1 through 5, or down to level 6 for a subset of 105 suppliers. These best-estimate labels will be used to evaluate the accuracy of the

three-component classification model, as discussed in the "Accuracy of the Three-Component Classification Model" subsection.

### 5.3. Accuracy Metric

For each supplier, the three-component model classifies it into a set of categories across all levels. We aim to measure the accuracy of these predictions by comparing them to the best estimates of the true labels. It's important to note that at each level, there can be multiple predicted categories as well as multiple true categories.

To measure accuracy at each hierarchical level, we utilize the F1 score, which balances the contributions of precision and recall (Holden and Freitas 2006, Costa et al. 2007). Precision measures the proportion of correctly predicted categories out of the total number of categories predicted, while recall measures the proportion of correctly predicted categories out of all categories that should have been predicted. Perfect precision indicates that every predicted category is correct, but it does not guarantee that all correct categories have been predicted. Perfect recall, on the other hand, ensures that all correct labels are predicted, but it does not specify the number of predicted categories.

We begin by calculating the F1 score for each individual supplier at each hierarchical level. This score is derived from the precision and recall metrics, providing a balanced measure of accuracy by considering both the correctness and completeness of the predicted categories. After computing the F1 score for each supplier at each level, we then aggregate across all suppliers to obtain an overall measure of the model's performance at each hierarchical level. The detailed mathematical calculation is provided in Appendix C.

### 5.4. Accuracy of the Three-Component Classification Model

We classify all 2,170 suppliers and use the gathered true label estimates to assess classification accuracy. For the subset of 278 suppliers, we evaluate accuracy down to levels 4/5, and for the smaller subset of 105 suppliers, we assess accuracy down to level 6. These results are considered representative of the overall classification accuracy for all 2,170 suppliers. In the following sections, we first demonstrate how the three-component model outperforms benchmark models and how its structure and flexibility improve classification accuracy. We then examine how each of our model components reflects the procurement expert's know-how to improve accuracy.

First, we tested with a range of $q$ values to determine the optimal $\hat{q}$, which strikes a balance between precision and recall, thereby optimizing the accuracy of the categorization model. The F1 scores of the categorization model typically exhibit an unimodal distribution, peaking at an optimum $\hat{q}$. This is intuitive: selecting a lower $q$ value causes the model to keep more pathways per supplier, boosting recall but reducing precision, whereas a higher $q$ value keeps fewer pathways, enhancing precision at the expense of recall. Therefore, an intermediate $q$ value that strikes a balance between precision and recall achieves the most

accurate categorization model. We tested different $q \in (0, 1)$ to pinpoint the optimal $\hat{q}$ that yields the highest F1 scores (please see Appendix D for details), and found the optimal value of $q$, denoted as $\hat{q}$, to be 0.99.

Then, under the optimal value $\hat{q}$, we compare the performance of our three-component model versus two benchmark hierarchical classification models. The first benchmark is the traditional top-down model that predicts level-by-level along the SIC pathway (Dumais and Chen 2000, Cesa-Bianchi et al. 2006, Esuli et al. 2008, Cerri et al. 2014), and the second benchmark is the traditional bottom-up model that predicts level 6 and employs heuristic pruning (Ceci and Malerba 2007). See Appendix E for a detailed explanation of benchmark models.

**Table 5 F1 scores across six hierarchical levels compare the three-component model with benchmark top-down and bottom-up models, using different combinations of testing data. The three-component model shows better classification at all levels, outperforming the benchmark models.**

| | Three-Component Model | | Benchmark Models | | | |
|---|---|---|---|---|---|---|
| | | | Top-down | | Bottom-up | |
| | $d_m$ (1) | flex (2) | $d_m$ (3) | $d_m^{gen}$ (4) | $d_m$ (5) | $d_m^{spe} \cup d_m^{PO}$ (6) |
| Level 1 | 0.858 | **0.935** | 0.802 | **0.849** | 0.642 | **0.663** |
| U Level 2 | 0.569 | **0.675** | 0.561 | **0.579** | 0.403 | **0.421** |
| Level 3 | 0.449 | **0.530** | **0.438** | 0.387 | 0.321 | **0.342** |
| Level 4 | 0.382 | **0.457** | **0.356** | 0.294 | 0.283 | **0.306** |
| Level 5 | 0.377 | **0.449** | **0.342** | 0.277 | 0.281 | **0.305** |
| Level 6 | 0.367 | **0.424** | **0.204** | 0.158 | 0.230 | **0.301** |

**Notes:** F1 at levels 1-5 are aggregated with 278 suppliers, and level 6 is aggregated with 105 suppliers. In all models, $q = \hat{q} = 0.99$.

Table 5 shows the F1 scores for each level from 1 to 6, comparing the performance of the three-component model with two benchmark models using different combinations of supplier data (e.g., general business description $d_m^{gen}$, detailed product description $d_m^{spe}$, or purchased item description $d_m^{PO}$). Full results for all text combinations are provided in Appendix F. The top-down benchmark model's accuracy decreases significantly at deeper levels due to error propagation—if the prediction is wrong at level 1, it will also be wrong at subsequent levels. The bottom-up benchmark performs better at deeper levels (such as levels 5 and 6) but struggles with higher levels because it lacks contextual information. The three-component model combines the strengths of both benchmark models. When comparing models with the same input data (columns 1, 3, and 5), the three-component model consistently performs better at every level, demonstrating that its structure improves classification accuracy.

Additionally, our analysis in Table 5 explored the performance of various data combinations for each model. We found that the top-down model generally performs better when employing general business

description (i.e., column 4 with $d_m^{gen}$) instead of all combined texts (i.e., column 3 with $d_m$). On the other hand, the bottom-up model tends to perform better when using specialized and purchase order data together (i.e., column 6 with $d_m^{spe} \cup d_m^{PO}$) instead of all combined text (i.e., column 5 with $d_m$). This suggests that adding extra data can introduce noise and reduce classification accuracy.

Moreover, in Table 5, the highest accuracy for the three-component model (column 2) was achieved by using flexible data sources: applying the general business description ($d^{gen}$) to the top-down component, while using specialized and purchase order data ($d_m^{spe} \cup d_m^{PO}$) for the bottom-up component. This flexibility in data usage significantly boosts accuracy compared to the benchmark models. Notably, the accuracy at level 6 for the three-component model is on par with the level-2 or level-3 accuracy of the benchmarks.

In sum, the three-component model improves prediction accuracy by (1) taking advantage of the prediction capabilities of top-down and bottom-up components, and (2) providing the flexibility to incorporate different sources of suppliers' information (i.e., supplier contextual description, product descriptions, and supplier actual procurement) to mitigate the effects of noise in the input data.

### 5.5. Impact of Model Structure on Classification Performance

In the three-component model, the top-down component is designed to classify level-1 categories, using general descriptions that incorporate the procurement expert's contextual knowledge. In contrast, the bottom-up component focuses on matching specific words to classify level-6 categories, reflecting the expert's familiarity with specific product terms and associations. The sandwich-connection component leverages the hierarchical relationships between categories, mimicking the logical approach experts use when classifying items. Next, we examine how this structure, rooted in expert knowledge, improves classification accuracy compared to benchmark methods.

Table 6 shows the F1 scores at level 1. As previously discussed, the benchmark top-down model performs better than the bottom-up model in predicting level-1 categories. We observe that the three-component model brings visible improvement in level-1 accuracy compared to the benchmark bottom-up model. This demonstrates the importance of incorporating a top-down component in predicting higher-level categories. Specifically, by doing so, the classification accuracy improved from the bottom-up model's accuracy of $0.663$ to $0.848$ when classifying with specific product descriptions ($d_m^{spe} \cup d_m^{PO}$) in column (4). Moreover, by being able to "connect the dots," we improve from the top-down model's accuracy of $0.849$ to $0.898$ when classifying with general product descriptions ($d_m^{gen}$) in column (1). Finally, by incorporating data flexibility to reduce noise, the three-component model improved the accuracy score to $0.935$.

Table 7 presents the F1 scores for level 6, showing that the benchmark bottom-up model exceeds the performance of the benchmark top-down model at this level, across various input texts from columns (1) to (5). Also, our three-component model significantly enhances the accuracy at level 6, demonstrating the effectiveness of including a bottom-up component for predicting more specific categories. Specifically, the

**Table 6 F1 scores at Level 1 compare the three-component model with benchmark top-down and bottom-up models, using different data sources for the 278 testing suppliers. The three-component model shows improvements in classification accuracy, particularly when using flexible data sources, achieving the highest score of 0.935 at level 1.**

| | General description | Specific descriptions | | | All texts |
|---|---|---|---|---|---|
| F1 at level 1 | $d_m^{gen}$ (1) | $d_m^{spe}$ (2) | $d_m^{PO}$ (3) | $d_m^{spe} \cup d_m^{PO}$ (4) | $d_m$ (5) |
| Benchmark Top-down model | **0.849** | 0.746 | 0.610 | 0.742 | 0.802 |
| Benchmark Bottom-up model | 0.563 | 0.638 | 0.619 | **0.663** | 0.642 |
| Three-Component model (same data source for both entries) | **0.898** | 0.849 | 0.723 | 0.848 | 0.858 |
| Three-Component model (best data source for each entry) | | | **0.935** | | |

**Note:** In all models, $q = \hat{q} = 0.99$.

best performance of the benchmark top-down model is achieved with the use of comprehensive texts in column (5) (i.e., $d_m$), where the three-component model improves classification accuracy from $0.204$ to $0.367$ with the same text. The best performance of the benchmark bottom-up model is achieved with the specific product descriptions in column (4) (i.e., $d_m^{spe} \cup d_m^{PO}$), where the three-component model improves classification accuracy from $0.301$ to $0.359$. Finally, by incorporating data flexibility, the three-component model improved the level-6 accuracy score to $0.424$.

**Table 7 F1 scores at Level 6 compare the three-component model with benchmark top-down and bottom-up models, using different data sources for the 105 testing suppliers. The three-component model shows improvements in classification accuracy, particularly when using flexible data sources, achieving the highest score of 0.424 at level 6.**

| | General description | Specific descriptions | | | All texts |
|---|---|---|---|---|---|
| F1 at level 6 | $d_m^{gen}$ (1) | $d_m^{spe}$ (2) | $d_m^{PO}$ (3) | $d_m^{spe} \cup d_m^{PO}$ (4) | $d_m$ (5) |
| Benchmark Top-down model | 0.158 | 0.190 | 0.138 | 0.189 | **0.204** |
| Benchmark Bottom-up model | 0.242 | 0.246 | 0.274 | **0.301** | 0.230 |
| Three-Component model (same data source for both entries) | 0.349 | **0.368** | 0.315 | 0.359 | 0.367 |
| Three-Component model (best data source for each entry) | | | **0.424** | | |

**Note:** In all models, $q = \hat{q} = 0.99$.

## 6. Decision Support Tools

The methodology thus far has organized vast text data and purchase order data from 2,170 suppliers into a hierarchical taxonomy that is both deep (6 levels) and broad (15,574 leaf nodes). The classification helped

make sense of the vast amount of purchase order records in the form of unstructured text data. In this section, we present the complementary decision support tool that can help its users in converting the classification results into opportunities for savings.

The aim of conducting a spend analysis is to identify opportunities for savings and recommend target suppliers to initiate the RFQ process to negotiate lower costs. Implementing the RFQ ranges from holding private negotiations with the suppliers to designing and holding public auctions. The outcome of a successful RFQ usually involves switching suppliers and managing new relationships which entails significant commitment of the manufacturer's internal resources (and may sometimes require re-structuring parts of its procurement processes). Thus, an RFQ recommendation must present convincing evidence of the potential cost savings. We next describe how our decision support tool offers insights into the supplier and product categories the buyers should target to seek price/volume discounts (subsection "Identify Leverage Categories"), and offers tools for easy cross-comparisons of many suppliers to understand the nuances in the procurement practice (subsection "Comparison of Suppliers").

### 6.1. Identify Leverage Categories

Recall that the three-component classification model generates a unique classification profile, or "DNA," for each supplier. When a supplier offers a unique product or service that few others can (e.g., Foxconn) and it is crucial for a buyer (e.g., Apple), the buyer faces a supply risk that could markedly affect its profitability. Managing these "high-risk / high-profit" impact suppliers requires careful consideration, and they should not be approached with RFQs. In contrast, our decision support tool, using this classification alongside the invoice value, helps identify the leverage supplier categories. These are suppliers that pose a low supply risk to the manufacturer but whose products have a high impact on the manufacturer's costs. These suppliers are where the cost-savings or RFQ opportunities generally arise because the buyers hold the dominant position in the buyer-supplier relationship.

Our decision support tool applies a widely used strategic sourcing method known as Kraljic analysis (Kraljic 1983, Webb 2017), which classifies suppliers into four categories based on their impact on the buyer's risk and profit. The number of suppliers in a given category serves as a useful indicator of supply risk—more suppliers mean lower risk. Similarly, the total spending on a product category indicates its impact on profit—higher invoice values suggest a greater influence on profitability. The classification model we have developed makes it easy to identify suppliers in these leverage categories at any level of the hierarchy.

To identify leverage categories, the hierarchical classification of suppliers allows for seamless navigation through various supplier and product groups, helping pinpoint key opportunities. For example, Figure 3 shows the categories in hierarchy level 3 (left panel), level 4/5 (middle panel), and level 6 (right panel) for Cranswick plc's supplier/product categories. Presenting each level is very useful for strategic and organizational reasons. For example, in large-scale procurement settings, there are dedicated departments or teams

for managing different supplier relationships. A department that handles animal feed suppliers would differ from one managing meat product suppliers, and within a department, different teams could be dedicated to poultry and sausage meat. This level of detail, as illustrated in Figure 3, supports targeted and effective management within the organization. The x-axis denotes the number of suppliers per category, the y-axis shows the total invoice values for all suppliers within a category, and the bubble size indicates the product diversity within the category or the count of sub-categories.

The left panel of Figure 3 illustrates the categories in level 3. We can identify three leverage categories positioned on the top-right: "Manufacture of other food products," "Processing and preserving of meat and production of meat products," and "Manufacture of prepared animal feeds." Focusing on "Processing and preserving of meat and production of meat products," the middle panel zooms in to show its level-5 sub-categories. We observe that it consists of three detailed supplier categories: "Production of meat and poultry meat products," "Processing and preserving of poultry meat," and "Processing and preserving of meat." The right panel further narrows down to reveal product categories within "Production of meat and poultry meat products," showing that they can be distinguished by specific product types, such as "ham curing" and "sausage meat." Note that level-6 categories represent the most detailed classification with no sub-categories, and thus all bubbles appear uniform in size.

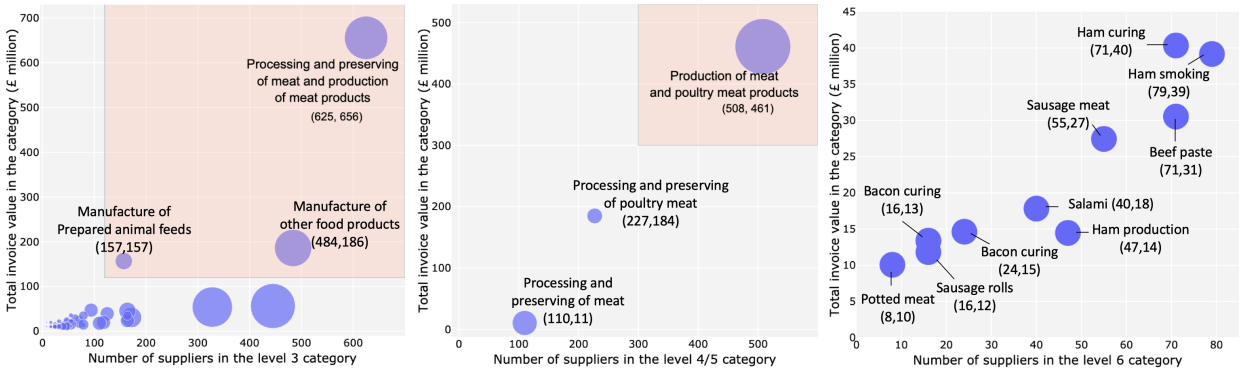
### 6.2. Comparison of Suppliers

For the identified leverage categories, the decision support tool enables cross-comparisons of their suppliers and product categories to help in developing a fine-grained picture of the supplier-product relationships. By comparing the classifications (or "DNA") of two suppliers, the tool can measure their similarity using techniques such as cosine similarity or weighted differences in their classification profiles.

For instance, in the left panel of Figure 3, there were 625 suppliers categorized under level-3 as "Processing and preserving of meat and production of meat products." Within this category, we can list those who compete with "A.B.P. GROUP LIMITED" based on their similarities. This is illustrated in the left panel of Figure 4, which identifies "2 SISTER FOOD GROUP LIMITED" emerges as the most closely matched competitor. Furthermore, our decision support tool is capable of listing suppliers not just by their past supply records but also by potential supply capabilities. The right panel of Figure 4 demonstrates this by ranking suppliers that are capable of providing "sausage meat," arranged according to their total invoice values. This figure offers valuable insights that are not easily obtained through manual methods.

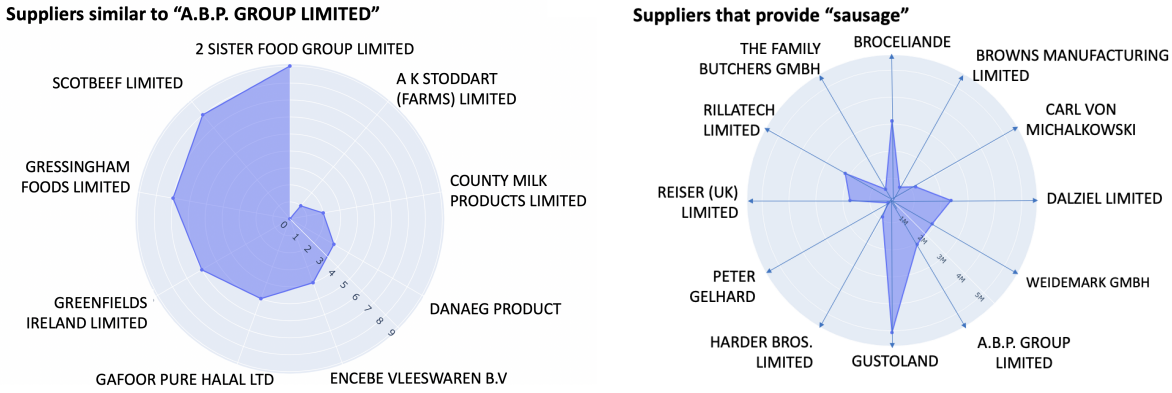
Furthermore, the decision support tool not only enables detailed cross-comparisons between suppliers but also uncovers opportunities for cost savings through strategic sourcing. For instance, in the right panel of Figure 3, there were 71 suppliers within the "Ham curing" category. We found that "Supplier 383" dominates the category, with a significantly higher invoice value than all of its competitors combined. Meanwhile, "Supplier 1390" offers a similar product range but contributes only a fraction of the invoice

**Figure 3** **The Kraljic Analysis illustrates Cranswick plc's supplier/product categories across hierarchy levels 3 (left panel), 4/5 (middle panel), and 6 (right panel). Each panel illustrates the relationship between the number of suppliers and total invoice values, with bubble size indicating product diversity within each category. Leverage categories are positioned in the top-right in each panel. For example, "Processing and preserving of meat and meat products" is one of the leverage categories in Level 3 (left panel), and it further divided into detailed subcategories at Level 4/5 in the middle panel. The right panel further refines the analysis, breaking down the leverage category "Production of meat and poultry meat products" from middle panel into specific product types, such as "ham curing" and "sausage meat."**




**Notes:** Since a supplier can be categorized into multiple nodes in a hierarchy, the sum of the suppliers can exceed 2,170. $q = \hat{q} = 0.99$.

**Figure 4** **Radar charts show suppliers similar to "A.B.P. Group Limited" (left panel) and suppliers that could provide "sausage" (right panel). The left panel indicates "2 Sister Food Group Limited" as the closest competitor. The right panel ranks suppliers capable of providing "sausage" based on their total invoice values, offering insights into potential supply capabilities.**




value compared to "Supplier 383." Increasing the purchase variety and volume from "Supplier 1390" would create competitive pressures that can help lower purchase costs for the buyer in this product category. Such

information can be utilized to inform the development of RFQ recommendations and realize its savings potential.

## 7. Advantage of Automated Spend Analysis: A Simulation Study

The automation discussed thus far enhances classification by providing an initial, high-fidelity categorization. It is designed to complement, rather than replace, manual input. There are three key reasons why automation is beneficial: it expands scope, improves accuracy, and boosts adaptability by enabling more frequent analysis. First, automation enables a manufacturer to analyze all its suppliers rather than just a subset, thereby expanding the scope of spend analysis. Second, automation improves the accuracy of classifications. Instead of creating classifications from scratch, the manufacturer can focus its efforts on reviewing and correcting any errors in the initial automated classification. Third, automation significantly increases the speed of spend analysis. With automated processes, what once took months can now be accomplished in a matter of days. This efficiency allows firms to conduct spend analysis more regularly, enabling them to respond more swiftly to market conditions. The improved adaptability that comes with more frequent analysis helps firms stay competitive and make timely adjustments to their strategies based on the latest data.

In this section, we estimate the savings achieved through the automation of a company's spend analysis. To do this, we conducted a simulation study, introducing a model of Cranswick's supply chain and calibrating it with the provided data. We differentiate between the savings generated by automation and those typically realized through manual analysis alone.

### 7.1. Simulation Experiment Design

Cranswick interacts with a large number of suppliers to order a wide range of level-6 product categories. Specifically, Cranswick's supply chain includes 2,171 suppliers and covers 3,258 distinct product categories. The relationship between suppliers and the product categories they provide, along with the economic value of these supplies, can be represented by two matrices that track this distribution. While the detailed mathematical notation is provided in Appendix G, where we also provide the corresponding mathematical notation, we offer an overall description below.

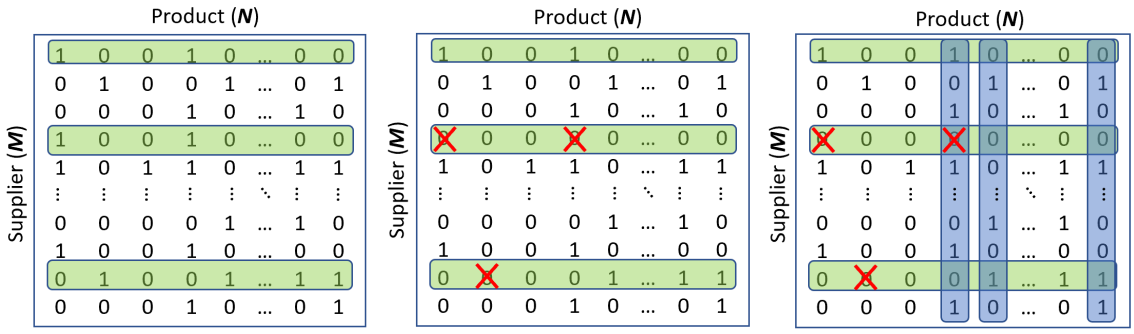
The first matrix, referred to as the supplier-product matrix, captures the relationship between suppliers and the product categories they provide. If a supplier can supply a specific product category, the corresponding entry is marked as 1; otherwise, it is marked as 0. By examining the classification data, we find that, on average, each supplier provides 5.03 different level-6 product categories, with a range from 1 to 13 categories. To build this matrix, we assign each supplier a random product range based on a binomial distribution, and then randomly select the appropriate number of products from the total set of available categories.

The second matrix, the purchase order matrix, records the value of orders placed with suppliers for specific product categories. In our analysis, we reviewed 556,866 purchase orders over a two-year period, totaling £1.57 billion, sourced from 2,171 suppliers. On average, each supplier had an invoice value of £723,776 over two years, or approximately £361,888 annually. Since suppliers typically provide around 5.03 product categories, the average invoice value for a single product category was £72,377. We model the invoice value for each product category using an exponential distribution, based on this average value.

### 7.2. Model of Spend Analysis: Scope, Accuracy, and Adaptability.

Next, we explain how we assess the scope, categorization accuracy, and adaptability of spend analysis for both manual and automated spend analysis. To illustrate this, we use Figure 5, which highlights the supplier-product matrix, our key tool for analyzing these aspects.

**Figure 5** **An illustration of the differences between manual and automated spend analysis through a supplier-product matrix highlights key distinctions. The left panel shows the scope differences between manual and automated spend analysis, with manual analysis covering only a subset of suppliers (shaded rows), while automated analysis considers all suppliers. The middle panel shows accuracy differences, where misclassified supplier-product links are marked by "×", indicating potential errors in manual classification, whereas automated analysis eliminates these errors. The right panel shows the final comparison, highlighting the refined matrix in automated analysis, which demonstrates a broader scope and higher accuracy compared to the manual approach.**




***Scope.*** Manual spend analysis and automated spend analysis differ primarily in their supplier scope. In manual spend analysis, procurement experts typically focus on a small subset of suppliers that represent a large proportion of the total spend. For example, in the case of Cranswick plc, 68 out of 2,171 suppliers are responsible for 80% of the total invoice value. For our normalization, we assume that automated spend analysis can examine all of the suppliers corresponding to 100% of the total spend, while manual spend analysis covers only a subset of suppliers that represent 70-90% of the total spend. For illustration, the left panel of Figure 5 shows the scope of analysis. The highlighted rows represent the scope for manual analysis, whereas all rows represent the scope for automated spend analysis.

***Accuracy.*** Within the given scope of analysis, manual and automated spend analysis differ in the accuracy of suppliers' classification. In automation-aided spend analysis, procurement experts are provided with an initial categorization of suppliers, which they manually check and amend as necessary. In contrast, manual analysis requires the firm to conduct the initial categorization from scratch, which takes more time and effort and has a greater risk of misclassification. Thus, with automation, classification tends to be more accurate than when it is performed via manual methods alone. In our simulation, to represent the level of accuracy, we will randomly misclassify the suppliers by altering the 1's and 0's as illustrated in the mid panel of Figure 5. The elements marked with "$\times$" correspond to misclassified supplier-product links. Manual spend analysis results in an accuracy between 70%-90%, meaning 10%-30% of columns will be marked with $\times$" in the highlighted rows. In contrast, for automation-aided spend analysis, we normalize the accuracy to be 100% (i.e., none of the columns will have "$\times$").

***Adaptability.*** One of the key benefits of automating spend analysis is the ability to perform it quickly and cost-effectively. This allows for frequent analysis, enabling better adaptability to changing market conditions, including fluctuations in suppliers, products, volumes, and prices. To model this effect, we consider the years 2019 and 2020. We assume that manual spend analysis can only be conducted once during this period, while automated spend analysis can be performed once in 2019 and once in 2020.

### 7.3. Model of Saving Realization.

Once the suppliers are categorized, the procurement team must (i) identify savings opportunities by pinpointing leverage suppliers and (ii) conduct a request-for-quote (RFQ) process. Next, we describe how these steps are modeled and calibrated in our simulation.

***Identification of Saving Opportunities.*** Recall that to identify savings opportunities, a Kraljic analysis is conducted to identify product categories with high spending volume and a large number of suppliers. In our simulation, these categories are identified by ranking the total spending and the number of suppliers for each product category. The highlighted columns in the right panel of Figure 5 represent these high-priority categories, where savings opportunities can be found.

***Implementing RFQ and the Estimated Cost Savings.*** Once the target level-6 product categories are identified, Cranswick plc will initiate an RFQ for all the suppliers in the product categories. The RFQ can involve either a private negotiation with an individual supplier or a public auction. Initiating RFQs can be a costly process, so they would typically do so only if they expect a savings of between 5-10% of the current invoice values. For each supplier identified in the leverage categories, we apply a random discount, ranging between 5% and 10%, to estimate the new invoice value. This 5-10% savings range is based on industry partner validation, reflecting their real-world experience.

***Overall Savings from Manual Spend Analysis.*** Implementing RFQs on all selected leverage product categories recommended by manual spend analysis typically translates into 2-3% in overall procurement

cost savings in a successful industry practice. To reflect this percentage, we calibrate the Kraljic analysis to identify the product categories that are in the top 33% for economic volume and the top 33% for the number of suppliers. In examining the effect of adaptability in the second period, we produce another randomized invoice matrix to reflect the changes in the market. However, to reflect the lasting benefit of spend analysis from the first period, we will employ lower average invoice prices (e.g., 2-3% lower than £72,377 of the first stage).

### 7.4. Simulation Results

In what follows, we will observe that automation of spend analysis creates additional savings over the current manual spend analysis by enabling the analysis of a greater scope of suppliers with increased categorization accuracy. Moreover, by enabling spend analysis to be performed more frequently, automation allows these benefits to compound over time.

In each simulation iteration, the model randomly selects the supplier-product matrix and the purchase order matrix according to the calibrated distributions and performs nine manual spend analyses. These analyses combine different scopes of supplier coverage (70%, 80%, and 90% of invoice value) with varying classification accuracies (70%, 80%, and 90%). Additionally, automated spend analysis is conducted with both scope and accuracy normalized to 100%. For each combination of these parameters, we conducted 1,000 iterations.

Table 8 presents the results for the 1-year analysis, while Table 9 presents the results for the 2-year analysis. From both tables, we observe that the total invoice cost from all suppliers across all product categories ranges between £780 million and £796 million annually (See Table 8), and over a two-year period, the total cost ranges from £1.565 billion to £1.600 billion (See Table 9). These figures align with Cranswick plc's annual procurement expenditure, as indicated in Table 1.

In Table 8, manual spend analysis achieves annual savings ranging from 2.25% to 3.00%, consistent with the RFQ implementation benchmarks observed in the industry. As expected, we observe that increased scope and accuracy in manual spend analysis contribute to higher savings. Examining the value of automated spend analysis, we find that it leads to overall annual savings of 3.27% to 3.43%. Compared to manual spend analysis, this results in an additional 0.43% to 1.28% savings (equivalent to £3.42 million to £10.20 million) in the first year.

In Table 9, the automated spend analysis generates savings between 4.91% and 5.07% over a two-year period, providing an additional 2.02% to 2.82% compared to manual spend analysis over the same duration. This corresponds to an additional £31.98 million to £44.53 million in savings over two years, or an extra £16 million to £22 million annually compared to manual spend analysis.

**Table 8 A summary of 1-year simulation results compares cost savings between manual and automated spend analysis. Annual invoice costs range between £780 million and £796 million, with manual analysis savings varying based on scope and accuracy. Automated analysis, assuming 100% scope and 100% accuracy, consistently yields higher savings, with the final column showing the percentage and monetary difference between the manual and automated analysis.**

| | Manual Analysis (1-year) | | | Automated Analysis (1-year) | |
|---|---|---|---|---|---|
| Cost (£ million) | Manual Scope | Manual Accuracy | Manual Saving (%) | Auto Saving (%) (100% Scope & 100% Accuracy) | %-Δ(Auto-Manual) (£ million) |
| 785 | 70% | 70% | 2.25% | 3.43% | +1.18% (+9.26) |
| 796 | 70% | 80% | 2.42% | 3.37% | +1.28% (+10.20) |
| 786 | 70% | 90% | 2.50% | 3.38% | +0.88% (+6.92) |
| 780 | 80% | 70% | 2.33% | 3.35% | +1.02% (+7.88) |
| 783 | 80% | 80% | 2.54% | 3.37% | +0.83% (+6.50) |
| 792 | 80% | 90% | 2.76% | 3.33% | +0.57% (+4.51) |
| 783 | 90% | 70% | 2.44% | 3.27% | +0.83% (+6.50) |
| 795 | 90% | 80% | 2.76% | 3.38% | +0.62% (+4.93) |
| 795 | 90% | 90% | 3.00% | 3.43% | +0.43% (+3.42) |

**Table 9 A summary of 2-year simulation results compares cost savings between manual and automated spend analysis. Annual invoice costs range between £1,565 million and £1,600 million, with manual analysis savings varying based on scope and accuracy. Automated analysis, assuming 100% scope and 100% accuracy, consistently yields higher savings, with the final column showing the percentage and monetary difference between the manual and automated analysis.**

| | Manual Analysis (2-year) | | | Automated Analysis (2-year) | |
|---|---|---|---|---|---|
| Cost (£ million) | Manual Scope | Manual Accuracy | Manual Saving (%) | Auto Saving (%) (100% Scope & 100% Accuracy) | %-Δ(Auto-Manual) (£ million) |
| 1579 | 70% | 70% | 2.25% | 5.07% | +2.82% (+44.53) |
| 1600 | 70% | 80% | 2.42% | 5.00% | +2.58% (+41.28) |
| 1577 | 70% | 90% | 2.50% | 4.99% | +2.49% (+39.27) |
| 1570 | 80% | 70% | 2.33% | 4.99% | +2.66% (+41.76) |
| 1565 | 80% | 80% | 2.54% | 5.00% | +2.46% (+38.50) |
| 1587 | 80% | 90% | 2.76% | 4.93% | +2.17% (+34.43) |
| 1578 | 90% | 70% | 2.44% | 4.91% | +2.47% (+38.97) |
| 1575 | 90% | 80% | 2.76% | 5.03% | +2.27% (+35.75) |
| 1583 | 90% | 90% | 3.00% | 5.02% | +2.02% (+31.98) |

## 8. Discussion

To the best of our knowledge, our methodology is the first academic work to formalize the automation of spend analysis using NLP and machine learning. We highlight its potential contributions to a path towards the evolution of Industry 4.0 (Olsen and Tomlin 2020).

### 8.1. Impact on Procurement Practice

Our methodology has the potential to democratize access to spend analysis to many small and medium-sized enterprises (SMEs). The supply chain complexity of many SMEs can be comparable to that of large firms. However, due to their volume of purchases being smaller, the estimated value of potential savings from conducting a spend analysis often does not justify the cost of hiring multiple procurement consultants over an uncertain prolonged duration. Automation of spend analysis removes these costs and makes accurate spend analysis accessible.

For example, our methodology has also been applied to conduct a spend analysis for a mid-cap company in the industrial sector. A private equity firm who had recently acquired it wanted to estimate the potential value improvement (e.g., by restructuring the supply chain) and provided us with the company's raw procurement data for 2020. The complexity of data was comparable to that from Cranswick plc, and consisted of 86,629 purchased orders, 25,025 products, and 1,829 suppliers. However, the total annual procurement spend was significantly lower at approximately £80 million. Our methodology identified four "leverage categories," whose combined value sums to roughly £22 million per year. Utilizing the decision tools, we were able to generate a list of supplier targets to recommend RFQs, which we estimated would translate into approximately 1.5-2.5% of the overall cost (approx. £2 million) in annual savings if implemented.

Since the manual spend analysis was previously not accessible for such firms, the value of the automated spend analysis would directly correspond to the savings that could be achieved. Thus, automated spend analysis could permeate throughout the industrial sectors. For example, a digital platform that provides automated spend analysis once buyers upload their transaction records can be offered, which would enable them to monitor their spending in real time. Such developments would further accelerate the automation of spend analysis across different industries and transform the way in which we monitor how physical "things" are produced and distributed.

### 8.2. Generalizeability of Methodology

Our paper introduces a methodology that relaxes the reliance on extensive datasets commonly needed for machine learning algorithms, and instead trains small but informative data efficiently. This approach enhances the flexibility of our methodology in manufacturer settings. For example, our methodology has been applied in a merger and acquisition setting where a German industrial manufacturer acquired a Swedish company. The German acquirer had a detailed internal supplier classification database and taxonomy and wanted to classify the Swedish firm's extensive supplier list according to their existing system. The German company shared with us its own hierarchical taxonomy and its supplier database. Instead of using SIC taxonomy as the training data (as shown in "Utilizing 'Small' Data" in Figure 1), we utilized the German company's hierarchical database to train our three-component model. Our three-component model was then able to provide the classification of the Swedish firm's suppliers according to the German company's taxonomy, facilitating its timely integration.

Also, our methodology takes large sets of unstructured data and converts them into a structured format that can provide strategic insights. This feature allows for applicability across various industries beyond manufacturing, particularly where structured data is limited but abundant descriptive documentation is available. For example, in the financial or legal services sectors, there is extensive documentation of regulations, codes of practice, and compliance reports. Our three-component model may map these processes into a hierarchical structure. Once the hierarchical structure of the regulations, sections, clauses, and sub-clauses are understood, the three-component model could be trained on the extensive regulatory documentation. It could then classify new documents (e.g., live cases) into the relevant categories within the regulatory framework. After categorization, the model pinpoints which sub-clauses or sections are most frequently associated with live cases. Such classification supports decision-making by highlighting areas that require attention, allowing for proactive measures to address compliance issues or service needs.

While we recognize the growing importance and capabilities of large language models (LLMs) in various applications, LLMs often struggle with categorization tasks requiring specific industry knowledge. LLMs must be supplemented with large, industry-specific datasets. Thus, an LLM is not a substitute, but a complement of our methodology that can help improve it further. For instance, rather than relying on static SIC documentation as we currently do, LLMs could be employed to incorporate the ability to learn industry-specific settings and update detailed databases of suppliers and product information. How to incorporate the capabilities of generative AI to complement our method may lead to fruitful directions for research and development.

### 8.3. Conclusion

This study has introduced a solution for automating the spend analysis process by leveraging large-scale industrial procurement data. Our findings illustrate that the developed hierarchical classification model, consisting of three components, successfully categorizes documents into any hierarchical taxonomy with high accuracy. The model outperforms current benchmarks, showing particular strength in handling taxonomies that are both deep and broad, as often required in real-world scenarios. Furthermore, the model's design to accommodate various types of textual data (e.g., general vs. specific) enhances its precision in classification. We have effectively demonstrated that spend analysis can be automated, incorporating the expertise of procurement professionals even in the absence of large datasets and precise supplier labels. This proof of concept paves the way for further research and development initiatives that could revolutionize the practice of spend analysis.

Due to the vast scale of the digital infrastructures in large manufacturing firms and their links to people and processes, a drastic change is prohibitively costly and risky, and often infeasible within a reasonable time frame. Similar to re-purposing and re-connecting existing physical infrastructures of a large housing complex (e.g., pipes and wires), we presented a methodology that utilizes existing digital infrastructures by

gathering them and generating insights. Such a solution represents the spirit of gradual process improvement (Fine and Porteus 1989) that is necessary to evolve towards Industry 4.0.

While data analytics involving NLP and machine learning, and artificial intelligence more generally, is a burgeoning field, its reach has been comparatively limited in many of the industrial sectors, such as large-scale procurement. We believe that many other B2B operations management contexts are currently untapped by data analytics. For example, an important social agenda is for manufacturers to reduce their carbon emissions. While firms are making important strides in reducing their direct (Scope-1) and indirect (Scope-2) carbon emissions, significant challenges remain in addressing emissions in their supply chain (Scope-3), which represent the vast majority of emissions. Our research method enhances transparency in the supply chains of the firms, and could pave ways for them to accurately track and manage their Scope-3 carbon emissions. We hope that operations management scholars can help lead the effort to modernize the industrial process.

## References

Beil DR, Chen Q, Duenyas I, See BD (2018) When to deploy test auctions in sourcing. Manufacturing & Service Operations Management 20(2):232–248.

Beil DR, Wein LM (2003) An inverse-optimization-based auction mechanism to support a multiattribute rfq process. Management Science 49(11):1529–1545.

Bragg J, Weld DS, et al. (2013) Crowdsourcing multi-label classification for taxonomy creation. First AAAI conference on human computation and crowdsourcing.

Budak C, Goel S, Rao JM (2016) Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis. Public Opinion Quarterly 80(S1):250–271, ISSN 0033-362X, URL http://dx.doi.org/10.1093/poq/nfw007.

Cai L, Hofmann T (2004) Hierarchical document categorization with support vector machines. Proceedings of the thirteenth ACM international conference on Information and knowledge management, 78–87.

Ceci M, Malerba D (2007) Classifying web documents in a hierarchy of categories: a comprehensive study. Journal of Intelligent Information Systems 28(1):37–78.

Cerri R, Barros RC, De Carvalho AC (2014) Hierarchical multi-label classification using local neural networks. Journal of Computer and System Sciences 80(1):39–56.

Cesa-Bianchi N, Gentile C, Zaniboni L (2006) Hierarchical classification: combining bayes with svm. Proceedings of the 23rd international conference on Machine learning, 177–184.

Chaturvedi A, Beil DR, Martínez-de Albéniz V (2014) Split-award auctions for supplier retention. Management Science 60(7):1719–1737.

Costa E, Lorena A, Carvalho A, Freitas A (2007) A review of performance evaluation measures for hierarchical classifiers. Evaluation methods for machine learning II: Papers from the AAAI-2007 workshop, 1–6.

Cranswick (2023) Cranswick plc annual report accounts. URL https://s3.eu-west-1.amazonaws.com/cranswick-2021/Interim-statement-FY23.pdf.

Dittrich J, Julka R, Mercker BU, Riedstra P (2020) The role of spend analytics in the next normal. McKinsey Insights. URL https://www.mckinsey.com/business-functions/operations/our-insights/the-role-of-spend-analytics-in-the-next-normal.

Domingos P (2012) A few useful things to know about machine learning. Communications of the ACM 55(10):78–87.

Duenyas I, Hu B, Beil DR (2013) Simple auctions for supply contracts. Management Science 59(10):2332–2342.

Dumais S, Chen H (2000) Hierarchical classification of web content. Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 256–263.

Elmaghraby WJ (2000) Supply contract competition and sourcing policies. Manufacturing & Service Operations Management 2(4):350–371.

Esuli A, Fagni T, Sebastiani F (2008) Boosting multi-label hierarchical text categorization. Information Retrieval 11(4):287–313.

Fine CH, Porteus EL (1989) Dynamic process improvement. Operations Research 37(4):580–591.

Garcia LV (2021) BCG: Procurement of the future. Boston Consulting Group, Procurement URL https://procurementmag.com/digital-procurement/bcg-procurement-future.

Hasenbein JJ, Gray P, Greenberg HJ (2010) Risk and Optimization in an Uncertain World (INFORMS).

Holden N, Freitas AA (2006) Hierarchical classification of g-protein-coupled receptors with a pso/aco algorithm. Proceedings of the IEEE Swarm Intelligence Symposium (SIS'06), 77–84 (IEEE Press).

Kraljic P (1983) Purchasing must become supply management. Harvard business review 61(5):109–117.

Laffont JJ, Tirole J (1993) A theory of incentives in procurement and regulation (MIT press).

Li C, Wan Z (2017) Supplier competition and cost improvement. Management Science 63(8):2460–2477.

Maas A, Daly RE, Pham PT, Huang D, Ng AY, Potts C (2011) Learning word vectors for sentiment analysis. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies, 142–150.

Mao Y, Tian J, Han J, Ren X (2019) Hierarchical text classification with reinforced label assignment. arXiv preprint arXiv:1908.10419 .

McCallum A, Rosenfeld R, Mitchell TM, Ng AY (1998) Improving text classification by shrinkage in a hierarchy of classes. ICML, volume 98, 359–367.

McKinsey (2021) Actionable spend insights with orpheus. URL https://www.mckinsey.com/business-functions/operations/how-we-help-clients/product-development-procurement/actionable-spend-insights-orpheus.

Mišić VV, Perakis G (2020) Data analytics in operations management: A review. Manufacturing & Service Operations Management 22(1):158–169.

Olsen TL, Tomlin B (2020) Industry 4.0: Opportunities and challenges for operations management. Manufacturing & Service Operations Management 22(1):113–122.

Peng H, Li J, He Y, Liu Y, Bao M, Wang L, Song Y, Yang Q (2018) Large-scale hierarchical text classification with recursively regularized deep graph-cnn. Proceedings of the 2018 world wide web conference, 1063–1072.

Ramos J, et al. (2003) Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning, volume 242, 29–48 (Citeseer).

Silla CN, Freitas AA (2011) A survey of hierarchical classification across different application domains. Data Mining and Knowledge Discovery 22(1):31–72.

UK Office for National Statistics (2007) Standard Industrial Classification. URL https://www.ons.gov.uk/methodology/classificationsandstandards/ukstandardindustrialclassificationofeconomicactivities/uksic2007.

US Census Bureau (2022) Statistics for industry groups and industries. annual survey of manufactures: 2020. URL https://www.census.gov/library/publications/2020/econ/e20-asm.html.

Vickrey W (1961) Counterspeculation, auctions, and competitive sealed tenders. The Journal of finance 16(1):8–37.

Wan Z, Beil DR (2009) Rfq auctions with supplier qualification screening. Operations Research 57(4):934–949.

Webb J (2017) What is the kraljic matrix? Forbes. URL https://www.forbes.com/sites/jwebb/2017/02/28/what-is-the-kraljic-matrix/?sh=24f54588675f.

# Appendix

## A. Detailed Explanation for Three-Component Model

### A.1. Mathematical Notation

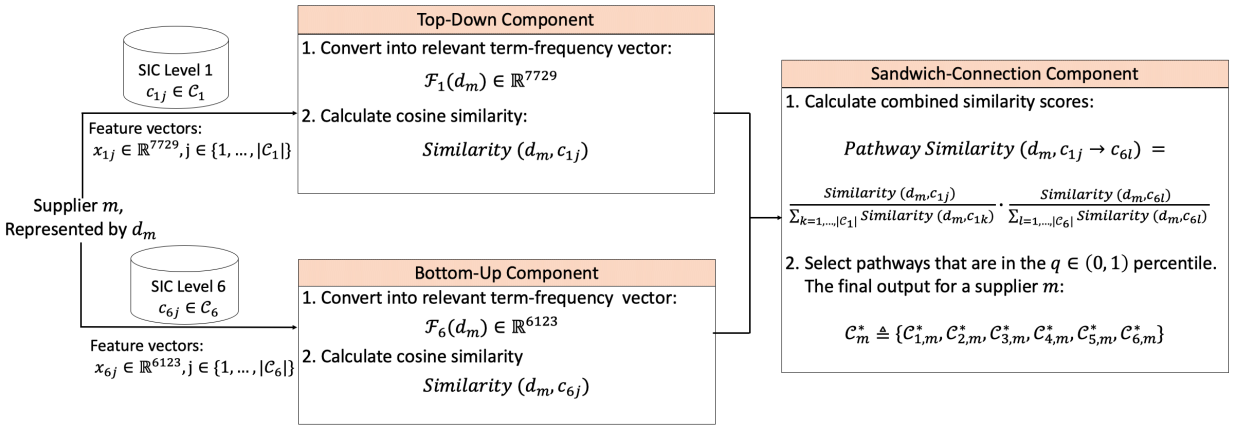
In the SIC taxonomy, let $\mathcal{C}_i$ denote the set of categories at level $i \in \{1, 2, 3, 4, 5, 6\}$, with $|\mathcal{C}_i|$ representing the number of categories at each level. For example, $\mathcal{C}_1$ consists of $|\mathcal{C}_1| = 21$ broad categories (e.g., "manufacturing"), and $\mathcal{C}_6$ contains $|\mathcal{C}_6| = 15,574$ specific categories (e.g., "Butter oil").

We let $c_{ij} \in \mathcal{C}_i$ denote the $j^{th}$ category at level $i$ in the hierarchy. For example, an element $c_{1j} \in \mathcal{C}_1$ may represent a general business category (e.g., "manufacturing") at level 1; and an element $c_{6j} \in \mathcal{C}_6$ may represent a specific product category (e.g., "Butter oil") at level 6.

We let $x_{ij}$ denote the feature vector for each category $c_{ij} \in \mathcal{C}_i$ at levels $i = \{1, 6\}$. For instance, at level 1, there are 7,729 features, meaning each of the 21 level-1 categories $c_{1j} \in \mathcal{C}_1$ was represented by a feature vector $x_{1j} \in \mathbb{R}^{7729}$. Similarly, at level 6, with 6,123 features, each category $c_{6j} \in \mathcal{C}_6$ was represented by a feature vector $x_{6j} \in \mathbb{R}^{6123}$.

The model's three components are illustrated in Figure A1. In the following subsections, we provide detailed explanations and examples for each component.

**Figure A1** **An illustration of the three-component classification model shows the top-down and bottom-up components used to calculate cosine similarity between supplier documents and SIC categories, and the sandwich-connection component that combines these similarities to select the most relevant pathways.**




### A.2. Top-Down Component

First, we convert the testing supplier $m$'s text document $d_m$ into the relevant term-frequency vector, denoted by $\mathcal{F}_1(d_m) \in \mathbb{R}^{7729}$. To do so, we extract the unigrams and bigrams from $d_m$, restrict the extracted terms to the 7,729 selective informative features, and count the frequency of the overlapped terms. For example, suppose $d_m =$ *"We manufacture chicken wing chicken thigh."* We extract ten terms (five unigrams and five bigrams). Say, only five terms (e.g., "manufacture," "chicken," "wing," "thigh," and "manufacture chicken") are overlapped with level-1 features.

| Unigrams | | | | | Bigrams | | | | |
|---|---|---|---|---|---|---|---|---|---|
| we | manufacture | chicken | wing | thigh | we manufacture | manufacture chicken | chicken wing | wing chicken | chicken thigh |
| 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | ✓ | ✓ | ✓ | ✓ | | ✓ | | | |

Thus, the term-frequency vector $\mathcal{F}_1(d_m) \in \mathbb{R}^{7729}$ is represented as [1, 2, 1, 1, 1, 0,....,0], where only five (out of 7,729) elements have non-zero frequencies.

Second, we determine which level-1 categories $c_{1j} \in \mathcal{C}_1$ the document $d_m$ of the testing supplier belongs to most closely. To reduce the bias against the size of the documents when creating term-frequency vectors and to handle sparse vectors (Maas et al. 2011), we employ cosine similarity between the feature vector $\mathcal{F}_1(d_m) \in \mathbb{R}^{7729}$ and the corresponding feature vector $x_{1j} \in \mathbb{R}^{7729}$ of each $c_{1j}$:

$$\text{Similarity}(d_m, c_{1j}) = \frac{\mathcal{F}_1(d_m) \cdot x_{1j}}{||\mathcal{F}_1(d_m)|| \cdot ||x_{1j}||}.$$

As an example output from the top-down component, the normalized cosine similarity is 0.53 for "Manufacturing," 0.28 for "Wholesale and retail trade," 0.19 for "Agriculture, forestry and fishing," and zero for the other level-1 categories.

### A.3. Bottom-Up Component

First, we extract the unigrams from $d_m$, and convert it into the relevant term-frequency vector $\mathcal{F}_6(d_m) \in \mathbb{R}^{6123}$. For example, suppose $d_m =$ *"charalambides christis edam cheese charalambides butter"*, which contains five unique unigrams. If only two of these unigrams, i.e., "cheese" and "butter," overlapped with level-6 features, then $\mathcal{F}_6(d_m)$ = [1, 1, 0, 0, 0, ..., 0], wherein only two elements have non-zero frequencies out of a total of 6,123 dimensions.

| charalambides | christis | edam | cheese | butter |
|---|---|---|---|---|
| 2 | 1 | 1 | 1 | 1 |
| | | | ✓ | ✓ |

Second, we compare the cosine similarity between the testing supplier $\mathcal{F}_6(d_m) \in \mathbb{R}^{6123}$ with each level-6 category $c_{6j}$'s corresponding feature vector $x_{6j} \in \mathbb{R}^{6123}$:

$$\text{Similarity}(d_m, c_{6j}) = \frac{\mathcal{F}_6(d_m) \cdot x_{6j}}{||\mathcal{F}_6(d_m)|| \cdot ||x_{6j}||}.$$

As an example output from the bottom-up component, "Butter oil" has a normalized cosine similarity of 0.0021, "Butter milk" has a normalized cosine similarity of 0.0011, "Lime growing" has a cosine similarity of 0, and so on.

### A.4. Sandwich-Connection Component

The sandwich-connection component aims to identify the set of most likely SIC pathways (from level-1 category to level-6 category) that a testing supplier belongs to. Among all 15,574 pathways $c_{1j} \rightarrow c_{6\ell}$, we can isolate the most likely pathways by examining the product of the normalized cosine similarity scores:

$$\text{Pathway Similarity}(d_m, c_{1j} \to c_{6\ell}) \triangleq \frac{\text{Similarity}(d_m, c_{1j})}{\sum_{k=1,\ldots,|\mathcal{C}_1|} \text{Similarity}(d_m, c_{1k})} \times \frac{\text{Similarity}(d_m, c_{6\ell})}{\sum_{\ell=1,\ldots,|\mathcal{C}_6|} \text{Similarity}(d_m, c_{6\ell})}.$$

An example output was shown in Table A1. As the last step, we select the pathways that are in the $q \in (0,1)$ percentile based on their pathway similarity. Ultimately, our model classifies a supplier $m$ into multiple pathways through level-1 to level-6 categories, i.e., $\mathcal{C}_m^* \triangleq \{\mathcal{C}_{1,m}^*, \mathcal{C}_{2,m}^*, \mathcal{C}_{3,m}^*, \mathcal{C}_{4,m}^*, \mathcal{C}_{5,m}^*, \mathcal{C}_{6,m}^*\}$

**Table A1 An illustration of the Sandwich-Connection component for hierarchical classifications displays supplier category classifications with normalized cosine similarity scores for level 1 and level 6. The table details pathway similarity calculations and intermediate connections (levels 2–5).**

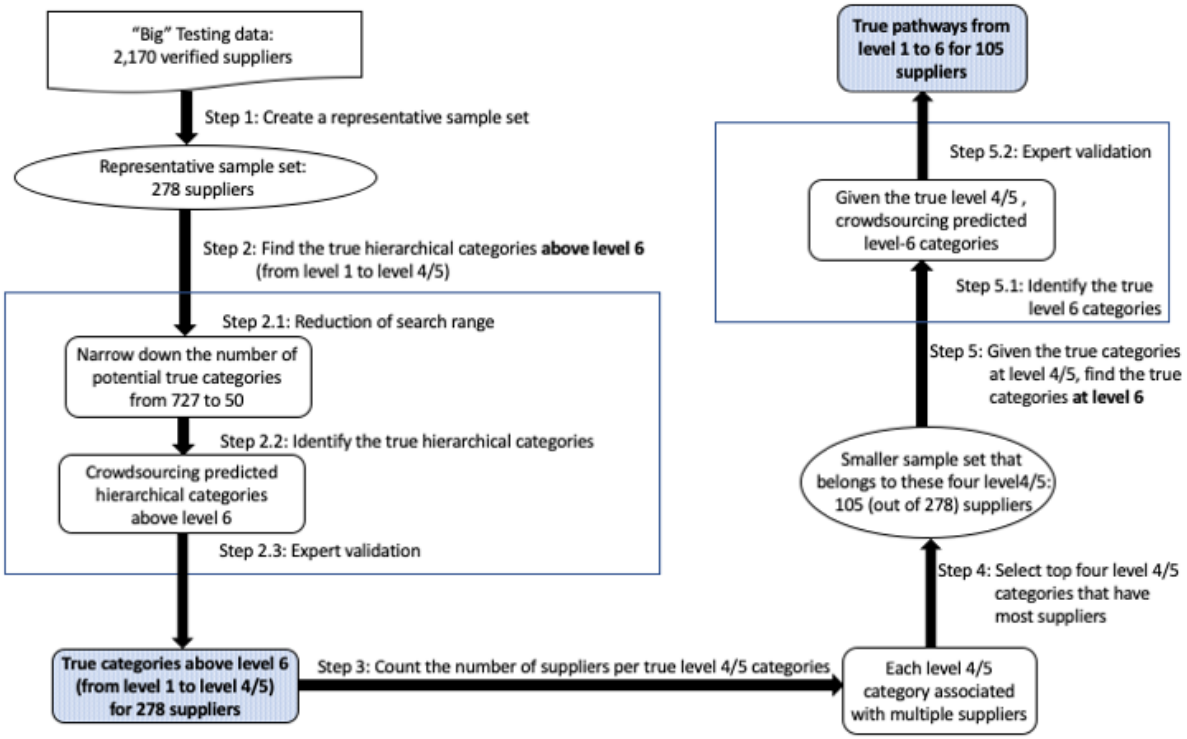
| | Level 1 predicted | Level 1 normalized cosine similarity | Level 2-5 connected | Level 6 predicted | Level 6 normalized cosine similarity | Pathway Similarity |
|---|---|---|---|---|---|---|
| 1 | Manufacturing | 0.53 | ... | Butter oil | 0.0021 | 0.53×0.0021 |
| 2 | Manufacturing | 0.53 | ... | Butter milk | 0.0011 | 0.53×0.0011 |
| 3 | Manufacturing | 0.53 | ... | Butter production | 0.0010 | 0.53×0.0010 |
| ... | Manufacturing | 0.53 | ... | ... | ... | ... |
| 9,189 | Manufacturing | 0.53 | ... | Roof lights made of plastic | 0 | 0 |
| 9,190 | Wholesale and retail trade | 0.28 | ... | Butter | 0.0020 | 0.28×0.0020 |
| 9,191 | Wholesale and retail trade | 0.28 | ... | Cheese | 0.0019 | 0.28×0.0019 |
| 9,192 | Wholesale and retail trade | 0.28 | ... | Milking machines | 0.0001 | 0.28×0.0001 |
| ... | Wholesale and retail trade | 0.28 | ... | ... | ... | ... |
| 10,943 | Wholesale and retail trade | 0.28 | ... | Cinema kiosk | 0 | 0 |
| 10,944 | Agriculture, forestry and fishing | 0.19 | ... | Butter | 0.0020 | 0.19×0.0020 |
| 10,945 | Agriculture, forestry and fishing | 0.19 | ... | Lemon growing | 0 | 0 |
| ... | Agriculture, forestry and fishing | 0.19 | ... | ... | ... | ... |
| 11,456 | Agriculture, forestry and fishing | 0.19 | ... | Bean growing | 0 | 0 |
| 11,457 | ... | 0 | ... | ... | ... | 0 |
| ... | ... | 0 | ... | ... | ... | 0 |
| 15,574 | ... | 0 | ... | ... | ... | 0 |

## B. Finding true labels

Prior research (e.g., Bragg et al. 2013, Budak et al. 2016) has emphasized the effectiveness of crowdsourcing in obtaining high-quality labels. In this study, we leveraged crowdsourcing via Amazon Mechanical Turk (MTurk) to construct true labels (i.e., true pathways) for a small sample of suppliers. To enhance the accuracy and reliability of the labels, we further incorporated complementary validation checks by two procurement experts.

We begin with an overview of the steps of our true label gathering that combines crowdsourcing and expert validation in Figure B1. Step 1 draws a sample set of testing suppliers of Cranswick plc, denoted by the subset $\mathcal{M}$. In Step 2, we estimate the true hierarchical categories for the supplier sample for levels 4/5. This step involves (a) the design of simple experiments, (b) the aggregation of responses taking advantage of the wisdom of crowds, and (c) expert validation. In Step 3, the number of suppliers per category is counted. In Step 4, the top four level 4/5 categories that include the maximum number of suppliers are selected to create a smaller sample set (denoted as $\mathcal{M}' \subset \mathcal{M}$) for identifying the true level 6. Finally, in Step 5, we repeat Step 2 to estimate the true level-6 categories for the smaller sample of suppliers. Below we describe each of these steps in detail.

**Figure B1** **The workflow for identifying true hierarchical categories for suppliers uses a representative sample set. The left panel shows the beginning of the process for predicting and validating categories above level 6 for 278 suppliers, while the right panel narrows down the process for identifying true pathways down to level 6 for 105 suppliers.**



***Step 1: Create a representative sample set.*** A sample of 278 suppliers was drawn from a total of 2,170 suppliers using a two-stage sampling approach. The first stage involved the selection of 68 suppliers who contributed to 80% of the total invoice value, while the second stage involved the random sampling of 10% (i.e., 210) of the remaining 2,102 suppliers. The resulting sample size of 278 suppliers represents 12.8% of the total supplier population. So we estimate the true hierarchical categories for $|\mathcal{M}| = 278$ suppliers down to levels 4/5.

***Step 2: Find the true hierarchical categories for levels 4/5.*** Table 3 shows that before the expansion of SIC to level 6, there were 727 unique level 4/5 categories, with 191 at level 5 and 536 at level 4. For each of the 278 sampled suppliers, we asked MTurkers to find the most likely SIC categories from level 1 to level 4/5 the supplier belongs to, and asked two procurement experts to validate the provisional true labels. To ensure reliable predictions from level 1 to level 4/5, we segmented the tasks into three sub-steps.

*Step 2.1: Reduction of Search Range.* First, we narrowed down the number of potential level 4/5 categories for each supplier from 727 to a more manageable number. For each sampled supplier, we grouped the 727 level 4/5 categories into 73 groups based on the combined similarity scores from the three-component model. These groups were constructed such that each group contained 10 categories, with group 1 comprising the most similar ten level 4/5 categories (1st to 10th) and group 73 containing the least similar level 4/5 categories (721st to 727th). Figure B2 shows an example of a MTurk task with the first two groups for a focal supplier.

**Figure B2 An MTurk task example for narrowing down level 4/5 categories for a supplier in Step 2.1 shows the first 2 groups out of 73. Each group contains 10 categories, ranked by similarity scores from the three-component model, with Group 1 containing the most similar categories.**

| Level1 | Level 4/5 | Group_number |
|---|---|---|
| AGRICULTURE, FORESTRY AND FISHING | Growing of vegetables and melons, roots and tubers | 1 |
| AGRICULTURE, FORESTRY AND FISHING | Mixed farming | 1 |
| MANUFACTURING | Other processing and preserving of fruit and vegetables | 1 |
| AGRICULTURE, FORESTRY AND FISHING | Gathering of wild growing non-wood products | 1 |
| AGRICULTURE, FORESTRY AND FISHING | Growing of other tree and bush fruits and nuts | 1 |
| AGRICULTURE, FORESTRY AND FISHING | Growing of spices, aromatic, drug and pharmaceutical crops | 1 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Wholesale of fruit and vegetables | 1 |
| MANUFACTURING | Processing and preserving of potatoes | 1 |
| MANUFACTURING | Manufacture of sugar confectionery | 1 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of fruit and vegetables in specialised stores | 1 |
| MANUFACTURING | Manufacture of agricultural and forestry machinery (other than agricultural tractors) | 2 |
| AGRICULTURE, FORESTRY AND FISHING | Plant propagation | 2 |
| AGRICULTURE, FORESTRY AND FISHING | Growing of other non-perennial crops | 2 |
| AGRICULTURE, FORESTRY AND FISHING | Post-harvest crop activities | 2 |
| MANUFACTURING | Manufacture of soft drinks; production of mineral waters and other bottled waters | 2 |
| MANUFACTURING | Manufacture of cider and other fruit wines | 2 |
| MANUFACTURING | Manufacture of prepared feeds for farm animals | 2 |
| MANUFACTURING | Tea processing | 2 |
| MANUFACTURING | Manufacture of condiments and seasonings | 2 |
| MANUFACTURING | Manufacture of wine from grape | 2 |

To ascertain the likelihood of a supplier belonging to at least one of the categories within each group, we allocated five different MTurkers to check per supplier and each MTurker was asked 73 randomly ordered TRUE/FALSE questions. We launched 1,390 MTurk tasks (5 MTurkers per supplier $\times$ 278 suppliers, \$1.5 per task) and collected

101,470 TRUE/FALSE answers (73 answers per task × 1,390 tasks). To ensure no potentially relevant pathway was unintentionally removed, we kept the groups for which at least two MTurkers agreed TRUE. Table B1 shows that the agreed (at least 2 out of 5) "TRUE" answers fall within group 1 through group 5 among 278 sampled suppliers. The preliminary task suggested that MTurkers can be asked to search for true level 4/5 categories among 50 (not 727) without compromising accuracy.

**Table B1 The reduction of search range in Step 2.1 shows MTurk results for narrowing down potential level 4/5 categories per supplier from 727 to 50. Suppliers' categories were grouped into 73 sets, and each set was reviewed by 5 MTurkers. The table shows the number of suppliers for which at least 2 out of 5 MTurkers agreed on a 'TRUE' classification within the first five groups, suggesting that focusing on the top 50 most similar level 4/5 categories reduces the search range without compromising accuracy.**

| | Number of suppliers (Each tagged by 5 MTurkers) | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0/5 TRUE | 1/5 TRUE | 2/5 TRUE | 3/5 TRUE | 4/5 TRUE | 5/5 TRUE | Total |
| Group 1 | 9 | 11 | 44 | 68 | 60 | 86 | 278 |
| Group 2 | 23 | 48 | 68 | 45 | 35 | 59 | 278 |
| Group 3 | 51 | 81 | 43 | 53 | 19 | 31 | 278 |
| Group 4 | 105 | 121 | 27 | 10 | 9 | 6 | 278 |
| Group 5 | 125 | 114 | 32 | 3 | 4 | 0 | 278 |
| Group 6 | 183 | 95 | 0 | 0 | 0 | 0 | 278 |
| ... | ... | ... | ... | ... | ... | ... | 278 |
| Group 73 | 278 | 0 | 0 | 0 | 0 | 0 | 278 |

*Step 2.2: Identifying the true hierarchical categories.* Based on 50 level 4/5 categories per supplier, we requested MTurkers to determine whether a focal supplier belonged to each category by responding to 50 TRUE/-FALSE questions. Figure B3 shows an example of such MTurk task. To provide MTurkers with relevant information for their assessments, we provided the supplier's official website URL (269 out of 278 have official website URLs and website text data) and a snippet of purchase order records.

We ensured a sufficient number of responses per supplier by allocating five different MTurkers to label each supplier, resulting in the completion of 1,390 MTurk tasks (5 MTurkers per supplier × 278 suppliers), with each task being compensated at a rate of $2. The resulting dataset consisted of 69,500 TRUE/FALSE responses (50 responses per task × 1,390 tasks). Based on the majority of TRUE answers received from the MTurkers for each supplier (i.e., at least 3 out of 5), we identified the level 4/5 categories that served as the provisional true labels for each supplier.

**Figure B3** **An MTurk task example for determining level 4/5 category assignments for suppliers (Step 2.2) involves asking MTurkers to classify each supplier into one of 50 potential categories by responding to 50 TRUE/FALSE questions. Each supplier's website and purchase order information were provided for reference to ensure accurate classification.**

| Level1 | Level 4/5 | Option index |
|---|---|---|
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of antiques including antique books, in stores | 1 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of fish, crustaceans and molluscs in specialised stores | 2 |
| MANUFACTURING | Manufacture of wire products, chain and springs | 3 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of meat and meat products in specialised stores | 4 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of electrical household appliances in specialised stores | 5 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of fruit and vegetables in specialised stores | 6 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of music and video recordings in specialised stores | 7 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of flowers, plants, seeds, fertilisers, pet animals and pet food in specialised stores | 8 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale of beverages in specialised stores | 9 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Other retail sale of food in specialised stores | 10 |
| ... | ... | ... |
| ... | ... | ... |
| ... | ... | ... |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Non-specialised wholesale of food, beverages and tobacco | 46 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Retail sale via stalls and markets of other goods | 47 |
| MINING AND QUARRYING | Support activities for petroleum and natural gas extraction | 48 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Agents involved in the sale of furniture, household goods, hardware and ironmongery | 49 |
| WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES | Wholesale of petroleum and petroleum products | 50 |

*Step 2.3: Expert validation.* To improve the reliability of the initial true labels gathered via crowdsourcing for the 278 suppliers, we engaged two procurement experts for verification. These experts undertook the same labeling task and reviewed the answers provided by the MTurkers. Through this verification process, we identified a total of 655 true level 4/5 categories for the 278 suppliers, including 107 unique categories. Among these 278 suppliers, 84 (30%), 99 (36%), 58 (21%), and 37 (13%) were respectively associated with one, two, three, and more than three true level 4/5 categories. Verifying the true labels for these suppliers was an exhaustive effort, requiring three months to complete. However, this rigorous process was critical in ensuring the accuracy of the true labels for our sample of suppliers.

***Step 3: Count the number of suppliers for each true level 4/5 categories.*** The aforementioned Steps 1 and 2 obtained accurate hierarchical categories at level $i \in \{1, 2, 3, 4, 5\}$. The next three steps aim to identify the correct category at level 6. In our sample set of 278 suppliers, 107 (out of 727) unique categories were identified at level 4/5, which in turn encompassed 3,258 level 6 categories. Due to the substantial time and cost involved in obtaining true labels for a vast number of potential level 6 categories, we further narrowed down our focus to a smaller set of suppliers. To do so, in Step 3, we counted the number of suppliers in each true level 4/5 category. The top 10 level 4/5 categories, along with their corresponding supplier count, are presented in Table B2.

**Table B2 A summary of the top 10 level 4/5 true categories for 278 suppliers shows the number of suppliers in each category. These categories were identified through hierarchical classification steps, and the selected categories (top 4) were prioritized for further analysis at level 6.**

| | Level 4/5 true category for 278 suppliers | # of suppliers (total 278) | Selected |
|---|---|---|---|
| 1 | Production of meat and poultry meat products | 47 | ✓ |
| 2 | Processing and preserving of meat | 30 | ✓ |
| 3 | Other processing and preserving of fruit and vegetables | 23 | ✓ |
| 4 | Butter and cheese production | 21 | ✓ |
| 5 | Repair of machinery | 19 | × |
| 6 | Processing and preserving of poultry meat | 18 | × |
| 7 | Manufacture of plastic packing goods | 16 | × |
| 8 | Manufacture of prepared feeds for farm animals | 17 | × |
| 9 | Manufacture of paper and paperboard containers other than sacks and bags | 14 | × |
| 10 | Manufacture of other articles of paper and paperboard | 14 | × |

***Step 4: Select top four 4/5 categories and create a smaller sample set.*** In Step 4, our objective was to restrict the sample size to a more manageable level to find true level 6, while still maintaining an adequate degree of statistical rigor. Based on these considerations, we selected a subset of four level 4/5 categories that comprised more than 20 suppliers. These categories are as follows: "Production of meat and poultry meat products," "Processing and preserving of meat," "Other processing and preserving of fruit and vegetables" and "Butter and cheese production." The selected categories comprised a total of 121 suppliers, of which 105 were unique. Therefore, we reduced the number of suppliers in our sample set from 278 to 105. We thus set the size of $\mathcal{M}'$ as 105 (i.e., $|\mathcal{M}'| = 105$). It is worth noting that each level 4/5 category comprises different numbers of level 6 categories, with a range from a minimum of 9 to a maximum of 179. Table B3 provides a summary of this information.

**Table B3 A summary of selected level 4/5 categories for identifying true level 6 categories (Step 4) highlights the subset of four level 4/5 categories, comprising 105 unique suppliers, chosen to balance sample size and statistical rigor. The number of level 6 categories ranges from 9 to 179.**

| Level 4/5 category | Number of suppliers | Number of level 6 | | |
|---|---|---|---|---|
| | | Min | Max | Mean |
| Production of meat and poultry meat products | 47 | 46 | 179 | 76 |
| Processing and preserving of meat | 30 | 62 | 179 | 92 |
| Other processing and preserving of fruit and vegetables | 23 | 52 | 151 | 69 |
| Butter and cheese production | 21 | 9 | 55 | 29 |
| Total | 105 (unique) | 9 | 179 | 64 |

***Step 5: Find the true level 6 categories conditioned on true level 4/5.*** In Step 5, we conducted an additional phase of validation using MTurk and expert assessments for 105 selected suppliers, focusing on pinpointing their accurate categories at level 6, conditional on their confirmed categories at levels 4/5. To facilitate this, we created a MTurk task that presented participants with 40 to 75 TRUE/FALSE questions, each aimed at verifying whether a specific supplier was associated with a given level 6 category. Due to suppliers being associated with multiple

categories at levels 4/5, the array of potential level 6 categories they could belong to varied. For instance, as illustrated in Figure B4, one particular supplier was linked to two level-4/5 categories ("Butter and cheese production" and "Liquid milk and cream production"), which together encompass 25 distinct level 6 categories.

**Figure B4** **An MTurk task example for validating level 6 categories is based on confirmed level 4/5 categories. It shows MTurkers answering TRUE/FALSE questions to verify the association between suppliers and their corresponding Level 6 categories. This example shows two level 4/5 categories, "Butter and cheese production" and "Liquid milk and cream production," linked to 25 distinct level 6 categories.**

| Level 4/5 | Level 6 | Option index |
|---|---|---|
| Butter and cheese production | Butter blending | 1 |
| Butter and cheese production | Butter milk | 2 |
| Butter and cheese production | Butter oil | 3 |
| Butter and cheese production | Butter production | 4 |
| Butter and cheese production | Butterfat | 5 |
| Butter and cheese production | Cheese | 6 |
| Butter and cheese production | Curd production | 7 |
| Butter and cheese production | Dairy preparation of cheese and butter | 8 |
| Butter and cheese production | Processed cheese | 9 |
| Liquid milk and cream production | Clotted cream | 10 |
| Liquid milk and cream production | Cream (sterilised) | 11 |
| Liquid milk and cream production | Cream from fresh homogenized liquid milk | 12 |
| Liquid milk and cream production | Cream production | 13 |
| Liquid milk and cream production | Double cream | 14 |
| Liquid milk and cream production | Heat treatment of milk | 15 |
| Liquid milk and cream production | Homogenised milk production | 16 |
| Liquid milk and cream production | Milk sterilising | 17 |
| ... | ... | ... |
| ... | ... | ... |
| Liquid milk and cream production | Sterilised cream | 25 |

To ensure that each MTurk task contained 40 to 75 questions, we grouped two or more suppliers into one task if the supplier had less than 40 potential level 6 categories, or split a supplier into two or three tasks if the supplier had more than 75 potential level 6 categories. The MTurk task allocation process is detailed in Table B4, showing that 105 suppliers comprised 122 MTurk tasks.

A total of 610 MTurk tasks were completed, with five different MTurkers assigned to label each task, and each task being compensated at a rate of $2. Based on the majority of TRUE answers received from the MTurkers for each supplier (i.e., at least 3 out of 5), we further asked two procurement experts to validate the responses. As a result, a total of 416 true level 6 categories were identified for 105 suppliers.

**Table B4 A summary of MTurk task allocation for validating level 6 categories is presented. Suppliers with fewer than 40 potential level 6 categories were bundled into one task, while suppliers with more than 75 categories were split into multiple tasks. A total of 105 suppliers were assigned across 122 MTurk tasks for validation, with tasks varying in size based on the number of potential level 6 categories.**

| | Bundle of four suppliers (each with $< 20$ level 6) | Bundle of two suppliers (each with 20-39 level 6) | 1 task (each with 40-75 level 6) | Split into 2 tasks (each with 76-120 level 6) | Split into 3 tasks (each with $> 120$ level 6) | Total |
|---|---|---|---|---|---|---|
| # of suppliers | 8 | 4 | 68 | 19 | 6 | 105 |
| # of MTurk tasks | 2 | 2 | 68 | 38 | 12 | 122 |

## C. Details on Accuracy Metric

Our three-component classification model classifies a supplier $m$ into a set of pathways through level-1 to level-6 categories. This hierarchical classification is represented by $\mathcal{C}^*_m$, defined as $\mathcal{C}^*_m \triangleq \{\mathcal{C}^*_{1,m}, \mathcal{C}^*_{2,m}, \mathcal{C}^*_{3,m}, \mathcal{C}^*_{4,m}, \mathcal{C}^*_{5,m}, \mathcal{C}^*_{6,m}\}$, where $\mathcal{C}^*_{i,m}$ denotes the set of categories $c_{ij} \in \mathcal{C}_i$ that the supplier $m$ is most likely to belong to at each level $i$ (where $i = 1, 2, 3, 4, 5, 6$).

Also, we denote the best-estimate true labels for supplier $m$ as $\mathcal{T}^*_m = \{\mathcal{T}^*_{1,m}, \mathcal{T}^*_{2,m}, \mathcal{T}^*_{3,m}, \mathcal{T}^*_{4,m}, \mathcal{T}^*_{5,m}\}$ if $m \in \mathcal{M}$, or down to level 6 as $\mathcal{T}^*_m = \{\mathcal{T}^*_{1,m}, \mathcal{T}^*_{2,m}, \mathcal{T}^*_{3,m}, \mathcal{T}^*_{4,m}, \mathcal{T}^*_{5,m}, \mathcal{T}^*_{6,m}\}$ if $m \in \mathcal{M}'$. For suppliers in $\mathcal{M}$, accuracy is assessed against the true labels down to level 4/5, while for suppliers in $\mathcal{M}'$, accuracy is compared against the true labels down to level 6.

Let $TP_{i,m}$ (True Positive) denote the number of categories that are correctly predicted, and $FN_{i,m}$ (False Negative) denote the number of categories that should have been predicted but were not. Formally, these can be expressed as $TP_{i,m} \equiv |\mathcal{C}^*_{i,m} \cap \mathcal{T}^*_{i,m}|$ and $FN_{i,m} \equiv |\overline{\mathcal{C}^*_{i,m}} \cap \mathcal{T}^*_{i,m}|$. To assess the accuracy of the classification model, we utilize the precision and recall metrics:

$$\text{Precision}_{i,m} \equiv \frac{TP_{i,m}}{|\mathcal{C}^*_{i,m}|}, \quad \text{Recall}_{i,m} \equiv \frac{TP_{i,m}}{TP_{i,m} + FN_{i,m}}.$$

To balance the precision and recall metrics, we employ the widely-used F1 score. At hierarchical level $i$, the F1 score for an individual supplier $m$ and for the average across all $M$ suppliers are respectively:

$$F1_{i,m} \equiv \frac{2 \cdot \text{Precision}_{i,m} \cdot \text{Recall}_{i,m}}{\text{Precision}_{i,m} + \text{Recall}_{i,m}}, \; F1_i = \frac{1}{M} \sum_{m=1}^{M} F1_{i,m}.$$

## D. Optimal Value for Parameter $q$

In Table D1, we evaluated three distinct $q$ values across three models. For the three-component model, it was observed that $q = 0.99$ resulted in the optimal F1 scores for levels 1, 4, 5, and 6. Conversely, $q = 0.999$ achieved the highest F1 scores for levels 2 and 3, though the difference from the scores at $q = 0.99$ was minimal. For the benchmark Top-down model, it is evident that $q = 0.99$ leads to the highest F1 scores across all levels. As for the benchmark Bottom-up model, $q = 0.99$ achieves the best F1 scores for levels 5 and 6, whereas $q = 0.999$ secures the highest F1 scores for the top four levels. Overall, we determined $q = 0.99$ to be the optimal $q$ value, so $\hat{q} = 0.99$.

**Table D1 An evaluation of F1 scores across different $q$ values for the three-component model and benchmark models (top-down and bottom-up) is presented. The table shows optimal F1 scores for each model and hierarchical level, highlighting that $q = 0.99$ consistently performs best across most levels.**

| | Three-Component Model | | | Benchmark Models | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Top-down | | | Bottom-up | | |
| | q=0.90 | q=0.99 | q=0.999 | q=0.90 | q=0.99 | q=0.999 | q=0.90 | q=0.99 | q=0.999 |
| Level 1 | 0.826 | **0.858** | 0.856 | 0.478 | **0.802** | 0.773 | 0.230 | 0.642 | **0.698** |
| Level 2 | 0.460 | 0.569 | **0.610** | 0.366 | **0.561** | 0.558 | 0.121 | 0.403 | **0.492** |
| Level 3 | 0.360 | 0.449 | **0.451** | 0.335 | **0.438** | 0.388 | 0.092 | 0.321 | **0.385** |
| Level 4 | 0.318 | **0.382** | 0.346 | 0.290 | **0.356** | 0.283 | 0.083 | 0.283 | **0.296** |
| Level 5 | 0.315 | **0.377** | 0.320 | 0.281 | **0.342** | 0.276 | 0.084 | **0.281** | 0.274 |
| Level 6 | 0.263 | **0.367** | 0.152 | 0.132 | **0.204** | 0.189 | 0.063 | **0.230** | 0.136 |

**Note:** F1 at levels 1-5 are aggregated with 278 suppliers, and level 6 is aggregated with 105 suppliers.

## E. Detailed Explanation for Benchmark Models

### E.1. Top-down Model

The benchmark top-down model classifies supplier $m$ level-by-level along the SIC pathway until reaching level 6. At any level $i \in \{1, 2, 3, 4, 5, 6\}$, we use the selectively parameter $q \in (0, 1)$ to choose predicted categories $c_{ij}$ whose cosine similarities are in the $q^{th}$ percentile.

It starts with level 1 prediction by calculating the cosine similarities between $d_m$ and each $c_{1j} \in \mathcal{C}_1$ ($|\mathcal{C}_1| = 21$). Any $c_{1j}$ will be kept if its normalized cosine similarity is in the $q \in (0, 1)$ percentile. Table E1 shows an example. Say $q = 0.99$ gives us the cut-off value as 0.311, then only one level-1 category "Manufacturing" is kept. Consequently, the top-down model narrows the choices at level 2, only considering the subset of $\mathcal{C}_2$ that belongs to survived $c_{1j}$.

**Table E1 An example of level 1 predictions for an individual supplier using the benchmark top-down model shows normalized cosine similarities for each predicted level-1 category. Categories with cosine similarities above the $q = 0.99$ threshold are retained, with 'Manufacturing' selected for further narrowing at subsequent levels.**

| | Level 1 predicted | Level 1 normalized cosine similarity | Keep |
|---|---|---|---|
| 1 | Manufacturing | 0.312 | ✓ |
| 2 | Agriculture, forestry and fishing | 0.307 | × |
| 3 | Professional, scientific and technical activities | 0.143 | × |
| 4 | Transportation and storage | 0.085 | × |
| ... | ... | ... | × |
| ... | ... | ... | × |
| 20 | Construction | 0.000 | × |
| 21 | Wholesale and retail trade | 0.000 | × |

In the level 2 prediction, we calculate the cosine similarities between $d_m$ and each $c_{2j}$ in the survived subset, and use the same parameter $q \in (0, 1)$ to choose predicted categories $c_{2j}$ whose cosine similarities are in the $q^{th}$ percentile. Continuing with the example above, we show the similarities between $d_m$ and 24 $c_{2j}$ within "Manufacturing" in Table E2. Applying $q = 0.99$ leads to two categories "Manufacture of food products" and "Manufacture of basic metals" survived. Consequently, the top-down model narrows the choices at level 3 within these two survived level-2 categories. We repeat the prediction down to level 6.

**Table E2 An example of level 2 predictions for an individual supplier using the benchmark top-down model shows normalized cosine similarities for categories within the "Manufacturing" sector. Categories with cosine similarities above the $q = 0.99$ threshold, such as "Manufacture of food products" and "Manufacture of basic metals," are retained for further narrowing at subsequent levels.**

| | Survival Level 1 | Level 2 predicted | Level 2 normalized cosine similarity | Keep |
|---|---|---|---|---|
| 1 | Manufacturing | Manufacture of food products | 0.367 | ✓ |
| 2 | Manufacturing | Manufacture of basic metals | 0.327 | ✓ |
| 3 | Manufacturing | Manufacture of textiles | 0.102 | × |
| ... | ... | ... | ... | × |
| 23 | Manufacturing | Manufacture of leather and related products | 0.000 | × |
| 24 | Manufacturing | Manufacture of tobacco products | 0.000 | × |

### E.2. Bottom-Up Model

The benchmark bottom-up model is similar to the bottom-up component. It starts by calculating the cosine similarities between $d_m$ and each $c_{6j} \in \mathcal{C}_6$ ($|\mathcal{C}_6| = 15574$). For each $c_{6j}$, we can trace the corresponding pathway up to level 1, and directly use the cosine similarity at level 6 to represent pathway similarity (See Table E3). Then, we apply the parameter $q \in (0, 1)$ to choose the pathways that have pathway similarities in the $q^{th}$ percentile.

**Table E3 An example of level 6 predictions for an individual supplier using the benchmark bottom-up model shows normalized cosine similarities for each level-6 category. Categories such as "Butter oil" and "Cocoa butter" are selected based on their level 6 cosine similarities, representing the pathway similarity traced back to level 1.**

| | Level 6 predicted | Level 6 normalized cosine similarity | Level 1-5 traced | Pathway Similarity |
|---|---|---|---|---|
| 1 | Butter oil | 0.0023 | ... | 0.0023 |
| 2 | Cocoa butter | 0.0021 | ... | 0.0021 |
| 3 | Shea butter | 0.0019 | ... | 0.0019 |
| 4 | Peanut butter | 0.0018 | ... | 0.0018 |
| ... | ... | ... | ... | |
| ... | ... | ... | ... | |
| 15573 | Pork pie | 0.0000 | ... | 0.0000 |
| 15574 | Lime Growing | 0.0000 | ... | 0.0000 |

## F. F1 Scores for All Combinations of Text

The following tables detail the average F1 scores for different text combinations utilized within the three-component model, as well as the benchmark top-down and bottom-up models. Columns that are part of the combinations previously presented in Table 5 are shaded in yellow for emphasis.

**Table F1 Average F1 scores for different text combinations in the three-component model are presented. Columns highlighted in yellow correspond to text combinations previously shown in Table 5, with "flex" representing the best data source for each entry, yielding the highest F1 scores across levels.**

| | General description | Specific descriptions | | | All texts | Best data source for each entry |
|---|---|---|---|---|---|---|
| | $d_m^{gen}$ (1) | $d_m^{spe}$ (2) | $d_m^{PO}$ (3) | $d_m^{spe} \cup d_m^{PO}$ (4) | $d_m$ (5) | flex (6) |
| Level 1 | 0.898 | 0.849 | 0.723 | 0.848 | 0.858 | 0.935 |
| Level 2 | 0.575 | 0.600 | 0.467 | 0.563 | 0.569 | 0.675 |
| Level 3 | 0.437 | 0.464 | 0.366 | 0.445 | 0.449 | 0.530 |
| Level 4 | 0.385 | 0.412 | 0.297 | 0.376 | 0.382 | 0.457 |
| Level 5 | 0.379 | 0.407 | 0.282 | 0.370 | 0.377 | 0.449 |
| Level 6 | 0.349 | 0.368 | 0.179 | 0.359 | 0.367 | 0.424 |

**Notes:** F1 at levels 1-5 are aggregated with 278 suppliers, and level 6 is aggregated with 105 suppliers. In all models, $q = \hat{q} = 0.99$.

**Table F2 Average F1 scores for different text combinations in the benchmark top-down model are presented. Columns highlighted in yellow correspond to text combinations previously shown in Table 5.**

| | General description | Specific descriptions | | | All texts |
|---|---|---|---|---|---|
| | $d_m^{gen}$ (1) | $d_m^{spe}$ (2) | $d_m^{PO}$ (3) | $d_m^{spe} \cup d_m^{PO}$ (4) | $d_m$ (5) |
| Level 1 | 0.849 | 0.746 | 0.610 | 0.742 | 0.802 |
| Level 2 | 0.579 | 0.544 | 0.385 | 0.510 | 0.561 |
| Level 3 | 0.387 | 0.407 | 0.288 | 0.414 | 0.438 |
| Level 4 | 0.294 | 0.327 | 0.231 | 0.333 | 0.356 |
| Level 5 | 0.277 | 0.312 | 0.221 | 0.319 | 0.342 |
| Level 6 | 0.158 | 0.190 | 0.138 | 0.189 | 0.204 |

**Notes:** F1 at levels 1-5 are aggregated with 278 suppliers, and level 6 is aggregated with 105 suppliers. In all models, $q = \hat{q} = 0.99$.

**Table F3 Average F1 scores for different text combinations in the benchmark bottom-up model are presented. Columns highlighted in yellow correspond to text combinations previously shown in Table 5.**

| | General description | Specific descriptions | | | All texts |
|---|---|---|---|---|---|
| | $d_m^{gen}$ (1) | $d_m^{spe}$ (2) | $d_m^{PO}$ (3) | $d_m^{spe} \cup d_m^{PO}$ (4) | $d_m$ (5) |
| Level 1 | 0.563 | 0.638 | 0.619 | 0.663 | 0.642 |
| Level 2 | 0.372 | 0.421 | 0.320 | 0.421 | 0.403 |
| Level 3 | 0.289 | 0.342 | 0.239 | 0.342 | 0.321 |
| Level 4 | 0.258 | 0.306 | 0.200 | 0.306 | 0.283 |
| Level 5 | 0.254 | 0.305 | 0.200 | 0.305 | 0.281 |
| Level 6 | 0.242 | 0.246 | 0.274 | 0.301 | 0.230 |

**Notes:** F1 at levels 1-5 are aggregated with 278 suppliers, and level 6 is aggregated with 105 suppliers. In all models, $q = \hat{q} = 0.99$.

## G. Details on Simulation Study

Cranswick interacts with $M$ suppliers to order $N$ different level-6 product categories. Specifically, Cranswick plc's supply chain includes $M = 2,171$ suppliers and $N = 3,258$ product categories. We define two random $M$-by-$N$ matrices $S$ and $P$ below.

The matrix $S$ is a binary matrix that represents the supplier-product relationship of Cranswick plc. If supplier $m$ can supply product category $n$, then its element $s_{mn} = 1$; otherwise the element $s_{mn} = 0$. Examining the classification $\mathcal{C}^*_m$, for all $m$, we find that each supplier supplied $5.03$ level-6 product categories on average, with a minimum of $1$ and a maximum of $13$. Thus, to construct this supplier matrix $S$, for each supplier $m$, we assign a random product scope (i.e., the number of products from a supplier) characterized by a binomial distribution with parameters $n = 13$ and $p = 5.03/13$. We then randomly select the corresponding number of product(s) from the set of $N$ products.

The matrix $P$ represents the purchase order $p_{mn}$ that a supplier $m$ charges when supplying a certain quantity of product category $n$. Recall from Table 1, we had examined 556,866 purchase orders over a two-year period totaling £1.57bn. These purchase orders were from 2,171 unique suppliers, and thus each supplier had an average invoice value of £$723,776$ over two-year period, which translates into an average annual invoice of £$361,888$ per supplier. Given that suppliers supplied on average $5.03$ level-6 product categories, for a single product category the average invoice value per product category $p_{mn}$ is £$72,377$. Thus, for each product-level category, we model the invoice value as an exponential distribution, with CDF, $F(p) = 1 - e^{-p/72,377}$.