Corporate Social Responsibility Reports

Corporate Social Responsibility Reports:

Topic Analysis and Big Data Approach

Irina Goloshchapova, Ser-Huang Pooon, Matthew Pritchard and Phil Reed*

January 9, 2019

*Irina Goloshchapova (i.o.goloshchapova@gmail.com) is at the Lomonosov Moscow State University. Ser-Huang Poon (ser-huang.poon@mancheter.ac.uk), Matthew Pritchard (matthew.pritchard-2@manchester.ac.uk) and Phil Reed (phil.reed@manchester.ac.uk) are at the University of Manchester, UK. We would like to thank Chris Adcock, Martin Carpenter, Andreas Hoepner, and participants at the workshop on "Recent Developments in Econometrics and Financial Data Science" at ICMA Centre, Reading University in November 2017 for helpful comments. The research funding from DigitalHumanities@Manchester is gratefully acknowledged.

Corporate Social Responsibility Reports: Topic Analysis and Big Data Approach

Abstract

This paper performs topic modeling using all publicly available CSR (Corporate Social Responsibility) reports for all constituent firms of the major stock market indices of 15 industrialized countries included in MSCI Europe for the sample period from 1999 to 2016. Our text mining results and LDA analyses indicate that "employees safety", "employees training support", "carbon emission", "human right", "efficient power", and "healthcare medicines" are the common topics reported by publicly listed companies in Europe and the UK. There is a clear sector bias with industrial firms emphasizing "employee safety", Utilities concentrating on "efficient power" while consumer discretionary and consumer staples highlighting "food waste" and "food packaging." To produce these results, we used a battery of python code to organize the hundreds of reports downloaded from Bloomberg and the internet, the latest R-algorithm to estimate LDA (Latent Dirichlet Allocation) model and the $LDAvis$ interactive tool to visualize and refine the LDA model.

JEL-Classification: C11, C51, C55, C81, C88, I3, L5, Q1

Keywords: Corporate Social Responsibility, Environment Social and Governance, Latent Dirichlet Allocation, Topic Modeling

1 Introduction

The ESG regime shift dates back to 2006 when Kofi Annan, the UN Secretary at the time, led investment communities in 16 countries created the six principles of Socially Responsible Investment (SRI).[1] Since then, many pension and mutual funds, and life insurance companies have signed up for SRI either voluntarily or following the industry trend and consumers demand. Disclosure on ESG investment, or some aspects of it, has long been part of the Pension Act.[2] There are now over 30 governments, quasi-governments, stock exchanges, and professional associations who regulate ESG disclosure and assurance.[3]

In response to the EU law (Directive 2014/95/EU) on corporate ESG disclosure, which took effect from January 2017, this paper presents a comprehensive analysis of the status of CSR/ESG reporting practice in 15 industrialized countries in Europe.[4] Using LDA Topic Analysis, we will address the questions; "Which of the ESG topics that are in the headlines?", "What are the sector differences?", "What has been changed?", "Is integrated or separate report more common and by who?". To answer

[1]See https://www.unpri.org/about/the-six-principles. The first three principles center on incorporating ESG issues in investment and decision making, ownership policies and practices, and ESG disclosure, while the last three focus on promoting ESG principles in the industry, implementation and report on progress.

[2]The Pensions Act 1995 and 2005 investment regulations place a range of investment duties on trustees. For example, trustees must secure appropriate asset diversification and invest in a manner 'calculated to ensure the security, quality, liquidity, and profitability of the portfolio.' Trustees are also required to include a statement in the scheme's Statement of Investment Principles (SIP) regarding the extent they take ESG and ethical considerations into account.

[3]For ESG reporting, these organizations include most notably GRI, SDG, IRRC, and WBCSD. For special interest groups on climate and environment, there are CDP, CDSB, TCFD, the CEO Water Mandate, ICMM, IIGCC, and CERES. For country based jurisdiction, there are SASB who works closely with SEC, UK's FRC, and a whole barrage of professional organizations such as IR Society, IFC, IFAC, CCI, WICI, COSO, EFFAS, and DVFA.

[4]In this paper, we will use the two terms, CSR and ESG, interchangeably.

these questions, we first survey the current practice among big corporations in 15 industrialised countries in Europe as defined in the constituent list of MSCI Europe. We document their ESG reports and changes made since the introduction of mandatory disclosure in 2017. For each country, we include all constituent companies of the main stock market index either listed or incorporated in these 15 countries. We tracked the patterns of their ESG disclosure over the period from 1999 to 2017 in CSR (Corporate Social Responsibility) reports, ESG reports and annual reports in pdf and HTML formats, and in no less than 14 languages. While the most popular language of reporting is English, many reports are written in Spanish, French, and German. There is also a large portion of these publicly available reports that are encrypted or exist only in image form which present strong challenges for machine-based textual analysis. Due to the sheer volume of documents to be processed, the review of company ESG disclosure cannot be achieved with human effort alone. We have therefore decided to use AI (Artificial Intelligence) for greater scale, speed, and consistency. Here, we use LDA (Latent Dirichlet allocation) textual analysis to detect the number and nature of the topics included in the reports. LDA analysis has been used recently in Huang, Lehavy, Zang, and Zheng (2017, *Management Science* forthcoming) and Dyer, Lang, and Stice-Lawrence (2017, *Journal of Accounting and Economics* forthcoming).

The use of text mining and textual analysis to check accounting quality is not new. Craig Lewis from the US SEC proposed in 2012 the Accounting Quality Model together with data-driven analytics to evaluate individual registrants and consistently estimate peer-level risk metrics. Lewis's work was subsequently developed into SEC 'RoboCop', a computerized tool designed to trigger alerts concerning suspicious accounting at publicly traded companies. In this paper, we use LDA (Latent Dirichlet Allocation), a topic modeling technique, to analyze the ESG disclosure and reporting standard among UK and European listed companies.

First, all the CSR/ESG/annual reports were downloaded from Bloomberg, GRI depository and other minor websites. We use Python to handle all file management including the detection of reporting year and language, and employ google port for

language translation into English. We use *AcrobatPro*, *Xpdf*, *Tesseract*, *ImageMagick*, and *WK<html>TOpdf* to convert all reporting formats into text files for text mining. Before the LDA implementation, a series of pre-processing steps are executed which include *Tokenize*, constructing library and list of stop-words, etc. After the LDA estimation, we perform LDA visualization and topic validation. We also build a colored-word algorithm to help trace all the words for a chosen topic to their locations in the source documents. Among the 192 UK companies and 2,493 European companies, our extensive LDA analyses identified 13 CSR topics among UK companies and 16 CSR topics among the European companies.

2 Previous Studies on Textual Analysis

In this section, we first review previous literature on textual analyses and studies of company reports, in general, considering the content and technique used. In the second section, we specifically survey studies of CSR/ESG reporting practices and related issues.

2.1 Textual Analysis of Company Reports

CSR research and content analysis date back to 1999 where Milne and Adler (1999) demonstrate how inexperienced coders using the Hackston and Milne approach with little or no prior training could analyze social and environmental aggregate total disclosure in annual reports. For more detailed sub-category analysis, they recommend training the less experienced coders with at least 20 reports.

Lang and Stice-Lawrence (2015) represent one of the large scale textual analyses of accounting reports. The authors examine annual report text for over 15,000 non-US companies from 42 countries over the period 1998-2011, focusing on the length of disclosure, the presence of boilerplate, comparability with US and non-US firms, and complexity. The authors find that textual attributes are predictably associated with regulation and incentives for more transparent disclosure, and these attributes

are also correlated with economic outcomes such as liquidity, institutional ownership, and analyst following. Using mandatory IFRS adoption as an exogenous shock, annual report disclosure improved in the sense that quantity of disclosure increased, boilerplate was reduced, and comparability increased relative to both US and non-US firms. Firms with the greatest improvements in financial reporting experienced the greatest improvements in economic outcomes around IFRS adoption.

There is now greater use of techniques from computational linguistics to capture large-sample measures of disclosure quality; there has been a substantial increase in the literature devoted to textual analysis. Most of the research that has focused on the effects of readability selects from a rather limited set of existing readability measures that reflect either writing clarity (e.g., Fog Index) or disclosure quantity (e.g., the size of the filing). Bonsall et al. (2017), therefore, propose a new measure of readability, the Bog Index, which captures the plain English attributes of disclosure (e.g., active voice, fewer hidden verbs, etc.). The authors validate this new measure based on a combination of regulatory guidance, experimental validation, and archival tests surrounding a regulatory intervention related to the readability of prospectus filings. In particular, controlled experiments show that participants who receive the more readable disclosure, as measured by the Bog Index, rate the disclosure to be significantly easier to read than participants who receive a less readable disclosure. Further, the significant improvements in the Bog Index relative to other readability measures around the SEC 1998 Plain English Mandate suggests that the Bog Index best captures the writing attributes enforced by regulators after the regulation. Finally, the authors test whether readability affects capital market outcomes including future stock market volatility and equity analysts' earnings forecast properties. They find that while most readability measures are associated with future stock market volatility, the Bog Index has nearly a 25% greater association than the next closest readability measure. In contrast, the authors find that only quantity-based measures of disclosure (e.g., total words and file size) are associated with analyst forecast accuracy. These findings suggest that various types of financial statement users are affected by different read-

ability attributes. Finally, the authors caution researchers against comparing total file size over time as the file size variation could be driven by the inclusion of content irrelevant to 10-K (e.g., HTML, XML, PDFs).

The other recently developed technique for textual analysis on big data sets is topic modeling. Huang et al. (2017), for example, use a topic modeling tool, Latent Dirichlet Allocation (LDA) developed by Blei et al. (2003), to compare the thematic content of a large sample of analyst reports issued promptly after earnings conference calls with the content of the calls themselves. LDA was used in Dyer et al. (2016) to study trends in 10-K disclosure over the period 1996-2013 concerning length, boilerplate, stickiness, and redundancy and decreases in specificity, readability, and the relative amount of hard information. They find that the new FASB and SEC requirements explain most of the increase in length and that 3 of the 150 topics, viz. fair value, internal controls, and risk factor disclosures, account for virtually all of the increase.

The LDA algorithm discovers the topic distribution for each document and the word distribution of each topic iteratively, by fitting this two-step generative model to the observed words in the documents until it finds the best set of topic and word distributions. The process is similar to cluster analysis or principal component analysis as applied to quantitative data. According to Huang et al. (2017), LDA offers three advantages over manual coding. First, it is capable of processing a large quantity of documents that would be too costly to code manually. Second, LDA provides a reliable and replicable classification of topics. Third, LDA does not require researchers to pre-specify rules or keywords for the underlying taxonomy of categories. Topics and their probabilistic relations with keywords are determined by LDA by fitting the assumed statistical model to the entire textual corpus. We will be using the LDA topic modeling technique to study the CSR reports of UK and European listed firms.

3 ESG and Corporate Social Responsibility

The World Business Council for Sustainable Development (wbcsd), among others, has initiated the Purpose Driven Disclosure (PDD) Project. The goal is to ensure that all companies will, by 2050, measure, value and report their true value, costs and profits, incorporating natural and social capital as well as financial capital in their measurement of success. In this framework the value created is shared among all stakeholders although there is an emphasis that the information is mainly for helping investors make sound business decisions. The additional disclosure will help to ensure that all capital and resources (financial, human, social, environmental) are properly measured, and their risks are considered and integrated into strategic decisions, key performance indicators (KPIs) and corporate reporting.

Lee and Moscardi (2017) (2017) reports six trends among the institutional investors placing more attention on ESG (Environment, Social, and Governance) factors.[5] The new trends coincide with the EU law (Directive 2014/95/EU), which took effect from January 2017 that requires large companies to disclose information on the way they operate and manage social and environmental challenges. The aim is to help investors, consumers, policymakers and other stakeholders to evaluate the non-financial performance of large companies and encourages these companies to develop a responsible approach to business. About 6,000 large public-interest companies in the EU with more than 500 employees are required to include non-financial statements in their annual reports from 2018 onwards. The information to be disclosed include en-

[5]The six trends are: (i) the shift towards longer-term investment policy by incorporating, e.g., climate risks or social inequality, or by tilting towards more resilient assets with a longer investment horizon; (ii) the increased awareness of physical risk, especially that relates to, e.g., climate change, and the encroaching scarcity of water; (iii) a growing trend in Asia for the level of the commitment to "Responsible Investment" to match its Western counterparts; (iv) the increased importance of ESG factors as a performance indicator; (v) the new ambition to invest with real impact and the disclosure of such an intention or information to help achieve Sustainable Development Goals; and (vi) the surge of innovation in sustainable development projects and initiatives in China, India and other emerging markets to attract foreign capital.

vironmental protection, social responsibility and treatment of employees, respect for human rights, anti-corruption and bribery, and diversity on company boards (regarding age, gender, educational and professional background). Companies were given great flexibility regarding the format of disclosure. In June 2017 the European Commission published its guidelines to help companies disclose environmental and social information. These guidelines are not mandatory, and companies may decide to use international, European or national guidelines according to their own characteristics or business environment.

The culture of ESG reporting has a long history in the UK. Since 2005 when the UK and the rest of the European Union adopted the IASB (International Accounting Standards Board) standard for accounting reports, the FRC (Financial Reporting Council) has asserted a strong influence on corporate reporting in the UK. In 2014, nine years after joining IASB, the FRC issued guidance notes on the "Strategic Report" under the UK Companies Act, requiring companies to provide shareholders with a holistic and meaningful picture of a company's business model, strategy, development, performance, position and prospects. The Guidance also encourages companies to focus on the application of materiality to disclosures and to be innovative in the structure of information to improve the clarity and conciseness of information. It is in this "Strategic Report" where disclosure on company ESG policy and ESG risk first becomes mandatory. At the same time, the Integrated Reporting Framework of 2013 represents an international attempt to connect a firm's financial and sustainability (i.e., environmental, social and governance) performance in one company report. An Integrated Report (IR) should communicate "concisely" how a firm's strategy, governance, performance, and prospects, in the context of its external environment, lead to the creation of sustainable value. Moreover, an IR needs to be "complete and balanced," i.e., broadly including all material matters, both positive and negative, in a balanced way. For publicly listed firms, the disclosure relating to ESG risk and reputation risk that may destroy shareholders' wealth rests with the "investor relation" division of these firms whose aim is to attract long-term investors to the companies.

However, a parallel development initiated by the GRI (Global Reporting Initiative) connects ESG reporting effort to the wider stakeholder community (Global Reporting Initiative (2000-2011)).

In September 2015, the UN proposed a holistic framework of Sustainable Development Goals (SDGs) to design indicators and an integrated monitoring framework in addressing all three dimensions of economic development, social inclusion, and environmental sustainability (UNSDGs report, 2015). Now the disclosure, wholly or partially, of these ESG issues have become mandatory in many countries including Australia, Austria, Brazil, Canada, China, Denmark, Finland, France, Germany, Hong Kong, India, Malaysia, Netherlands, Singapore, Sweden, South Africa, the United Kingdom, and the United States. As mentioned before, the European Commission has recently endorsed the adoption by the Council of the Directive (2014/95/EU) on the disclosure of environmental, social, and diversity information by more than 6,000 companies with effect from the 2017 financial year.

As a case study, Cho et al. (2015) examine the talk, decisions, and actions of two highly visible U.S.-based multinational oil and gas corporations during the period of significant national debate over oil exploration in the Alaskan National Wildlife Refuge. The authors argue that the prevailing economic system and conflicting stakeholder demands constrain the action choices of individual corporations. Under such circumstances, organizations engage in hypocrisy and develop facades, thereby severely limiting the quality of sustainability reports as a channel for substantive disclosures.

Kuhn et al. (2015) studied the content of the websites of the 211 sample companies from a collection of seven countries in Sub-Saharan Africa, viz. Tanzania, Uganda, Zambia, Nigeria, Kenya, Ghana, and Botswana. The authors admitted that such a research approach suffers two limitations. On the one side, companies might report a CSR-related activity that they have not undertaken just to improve their image. At the same time, some companies undertake CSR activities but lack the means or knowledge of how to report, and thus, the actual practice remains undisclosed. Despite these short-comings, the authors find that the sample African companies' CSR efforts

focus strongly on local philanthropy and therefore differ substantially from Western CSR approaches. Furthermore, there is evidence that GDP and level of governance standard positively affect CSR reporting.

Lee et al. (2017) find Chinese non-state-owned enterprises domiciled in regions with a higher level of corruption are more likely to voluntary disclose CSR information when they receive non-tax based subsidies. Their sample consists of 3,216 firm years observations for manufacturing companies listed on the Shanghai and Shenzhen Stock Exchanges from 2008-2012.

Using qualitative analysis on a sample of multinational corporations (MNCs) using face-to-face or on the telephone interviews with 67 CSR manager and 18 related personnel, Risi and Wickert (2017) show that as the institutionalization of CSR advances, the role of the CSR manager diminishes. The selected firms were listed on the German stock index (DAX, MDAX) and the Swiss Market Index (SMI) from a broad cross-industry sample of representative MNCs in each national context, including manufacturing, banking and financial services, software, automotive, construction, chemicals, and utilities. Each MNC had at least 5000 employees and operations in multiple countries. Their data indicate that although CSR managers do not become obsolete once CSR has been institutionalized, they tend to be reduced to more peripheral roles as administrators of CSR-related routines (e.g., to compile the annual sustainability report), while other professional groups take over the daily tasks of strategizing and executing CSR.

Melloni et al. (2017) study 148 reports (for the year 2013 and 2014) by 74 unique firms that were members of the official IIRC (International Integrated Reporting Council) Pilot Programme (which ended in October 2014). The IR (Integrated Reports) reports belong to firms from all continents and operating in different industry groups. There were 104 firm-year observations in the final stage due to missing data in Bloomberg and Thomson Reuters ASSET4. The authors analyzed two attributes of these reports, viz. conciseness and completeness/balance. Their textual analysis indicate that early IR adopters show that in the presence of a firm's weak financial performance, the IR

tends to be significantly longer and less readable (i.e., less concise), and more optimistic (i.e., less balanced). Moreover, firms with worse social performance provide reports that are foggier (i.e., less concise) and with less information on their sustainability performance (i.e., less complete). The finding suggests that IR early adopters employ quantity and syntactical reading ease manipulation, thematic content, and verbal tone manipulation to manage impression. The results also suggest that such strategies depend not only on the level of firms' performance but also on the type of performance (financial versus non-financial sustainability).

4 Data and Pre-processing

Our sample consists of a UK sample that made up of 641 constituent stocks of the Financial Times All-Share Index (ASX) and a European sample that made up of 481 constituent stocks of 15 European Stock market indices.[6] The data sample period is from 1999 to 2016 covering 12 first-level SIC sectors.[7] The reports are downloaded from Bloomberg and GRI (Global Reporting Initiative) Sustainability Disclosure Database. These reports are mostly in pdf format with a small number in HTML format and very few in text format. The pdf and HTML files are all converted into text format before the reports can be read by the software R.

The text conversion is done using AcrobatPro, and when it fails, Xpdf is used instead. Xpdf performs very fast text extraction for text type PDF files, and it can deal with many types of encrypted PDF files. For pdf text files that are stored as images instead of text format, we use *Tesseract OCR* combined with *ImageMagick* to produce the text output. *Tesseract OCR*, sponsored by Google since 2006, is an optical charac-

[6] Austria ATX 20, Belgium BEI 20, Denmark OMXC 20, Finland OMXH 25, France CAC 40, Germany DAX 30, Iceland ISEQ 20, Italy FTSEMIBN 40, Netherland AEX 25, Norway OBX 25, Portugal PST 20, Spain IBEX 35, Sweden OMX 30, Switzerland SMI 20, and UK UKX 100.

[7] The SIC code represented are 10: Energy, 15: Materials, 20: Industrials, 25: Consumer Discretionary, 30: Consumer Staples, 35: Health Care, 40: Financials, 45: Information Technology, 50: Telecommunication Services, 55: Utilities, 60: Real Estate, and 99: Funds & Investment Co.

ter recognition system. The latest 4.00.00 alpha version is based on LSTM (Long & Short Term Memory) neural networks. For HTML files, we first use *WK<html>TOpdf* program to them into image PDF format, and then convert them into text files as before. Extracting text directly from HTML files gave wrong results because many of the HTML files appeared to have been converted into HTML format from pdf format or images.

After all the pdf and HTML files have been converted into text files, we identify reports that are non-English. To perform language detection, we use Python library *langdetect* 1.0.7, a port of Google's method, which will return a 2-character ISO code for the language detected. From here, we use Python library *googletrans* 2.1.4 to translate the non-English reports into English.

As corporate responsibility reporting is voluntary, the document submission, style, and reporting format vary from one company to another. The most critical task we face is to identify the reporting year (instead of the file creation date or the report submission date). Identifying the reporting year for all the reports downloaded proved to be quite a challenging task. First, we search the meta-data of the pdf file to identify the file creation date. Since CSR reports are often produced the year after the reporting year, we use the year from the PDF metadata creation date minus one if the report was created between January and June, or the creation date itself if the month is between July and December. One condition that must be fulfilled is that the creation year must be before the upload year. If this condition is not met, then we text mine the year from the report itself. In this regard, we have written a year-mining algorithm to detect the reporting year from the first 150 characters, or failing that first 2000 characters of the reports. Only when a year cannot be determined by all the methods above, is the upload year used. When all the above methods failed, we manually checked the reports. Where there is more than one report per company per year, we concatenate the files and merge them into one. After the year identification, we have 1,688 UK CSR reports, and 3,618 European CSR reports in total.

Various pre-processing steps on the text are necessary before the content of these

reports can be analyzed. The pre-processing includes removing non-language elements such as photo images, converting all letters into lower case, removing non-alphanumeric symbols, numbers, dots, repeat space character, one letter words, letters attached with special characters (taking care not to delete words such as o2 and co2). Also, we have constructed a list of "stopwords" which will be removed during the text analysis. Apart from the traditional stopwords list for the English language such as "a," "the," etc., we have included in the stopwords list: months, company names, words related to internet connections such as www, com, etc., and some country names such as the UK and US. Other country names, such as India and Turkey, are not removed because they could be meaningful for understanding some topics. Some popular words in company reports such as performance, sustainability, practice, etc. have also been removed. Finally, some common but non-recognisable words, such as "ve," "re," "cid," etc., appeared as top words in the result, so were removed in an iterative process.

5 Topic Modeling Algorithm

The Latent Dirichlet Allocation (LDA) algorithm is used for topic modeling. LDA is a relatively complex non-linear generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. In a topic analysis, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. Each topic has probabilities of generating various words. Words without special relevance will have a roughly even probability between classes (or can be placed into a separate category). A topic is not strongly defined, neither semantically nor epistemologically. It is identified by automatic detection of the likelihood of term co-occurrence. A lexical word may occur in several topics with a different probability, however, with a different typical set of neighboring words in each topic. The LDA model decomposes the document-term matrix into two low-rank matrices,

viz. document-topic distribution and topic-word distribution. There are several important hyper-parameters for this model: "number of latent topics," "document-topic prior" (a number less than 1), and "topic-word prior" (a number much less than 1). We set "number of latent topics" equal to 50, "document-topic prior" equal to 0.1 for a sparse topic distribution with few topics per document, and "topic-word prior" equal to 0.001 with strongly sparse word distributions and few words per topic. We change the "number of latent topics" at each iteration but keep the other two parameter values constant at 0.1 and 0.001 respectively.

We use the $text2vec$ package in R for LDA topic modeling, which is relatively new, stable and fast compared with other packages used in natural language processing. LDA in $text2vec$ is based on the state-of-the-art $WarpLDA$ sampling algorithm. One of the main advantages of $text2vec$ is that the execution time is fixed for any number of topics, which makes the processing speed very fast even in the context of Big Data.

5.1 Refining LDA Model with LDAvis

The output from *text2vec* R-algorithm contains two matrices: (i) a list of 50 keywords for each of the 50 topics selected, and (ii) the probability of topic relevance for each document, and the marginal probability for each topic. While these probabilities are useful in gauging the popularity of a topic, our manual selection of topics is not entirely based on probability. For example, if "human right" is an important ESG topic but is mentioned only once by one company, then the topic, "human right," should be selected. On the other hand, "share price performance" might have been mentioned every year and in every report, such a topic is not related to ESG and should be dropped. Furthermore, the set of keywords under each topic is reviewed as a collection for cohesiveness and with an emphasis being placed on the top words. Our experience of analyzing the LDA results for corporate social responsibility reports suggests that the top five to ten words are usually sufficient for deciding if a topic is meaningful or not as well as for identifying the title of the topic.

The keyword list and topic probability are also useful information for identifying garbage words. The garbage words appear frequently in the European list of top words whenever the translation is missed or not executed by google. In this case, a particular set of foreign language (e.g., German or Spanish) will appear together in a topic. We can then manually check the corresponding CSR report(s) or move all such top words, which are "garbage," into the stopword list.

Finally, we use the $LDAvis$ software to help fine tune the LDA model.[8] Figure 1 shows how $LDAvis$ presents the topic selection for European (50 topics) and UK (20 topics) CSR reports. When "Selected Topic" is set to 0, $LDAvis$ presents a two-dimensional display of the selected topics highlighting the prevalence of each topic (by the size of the circle) and the relationships (as reflected in the position and the amount of overlap) between topics. For Europe, it is clear that there are many overlapping topics which could be due to the large number of topics selected. All circles are approximately the same size indicating that there is no clear pattern of a dominating CSR topic in Europe and the UK.

[Figure 1 about here.]

The right panel of Figure 1 presents the *Saliency* or word frequency of the keywords. Since no specific topic is selected, the result indicates that "company" is the most common word among all 50 European topics, and has appeared over 300,000 times. In the UK, the most common word is "customer" which has appeared over 70,000 times.

Figure 2 plots the distribution of the word "emission" in the top panel and topic 10, which we identified as "carbon emission." It is clear that the word "emission" also appears frequently in topics 4, 15, 17, and to a lesser extent, topics 2, 11 and 19 as well. This result confirms that the LDA algorithm recognizes the polysemy or contextual

[8] $LDAvis$ R package is a web-based interactive visualization of topics estimated using Latent Dirichlet Allocation that is built with D3.js, a very popular javascript visualization tool (see Sievert and Shirley (2014)).

nature of words by assigning the same word to multiple topics. However, a closer inspection of these topics reveals that the word list patterns for most of the topics mentioned above are less clear-cut as compared with topic 10. Indeed, we noted later that topic 15 is related to efficient energy in general, while topic 4 strongly features supply chain sustainability, while word lists for topics 2, 11 and 19 do not present any coherent theme.

[Figure 2 about here.]

We have also investigated several validation tests for LDA topic outputs as proposed in Huang et al. (2017). The main weakness of these validation methods is that they are strongly influenced by word frequency which is not the most appropriate selection criteria for CSR keywords and topics. Due to the mechanistic adherence to word counts, these validation methods are strongly influenced by irrelevant words and are very sensitive to words trimming. That is, the results change drastically when we take out some top words that are irrelevant to CSR, and when we attempt to do the word-pair (e.g., "child labor" being replaced by "childlabor," and "child care" by "childcare," etc.) and word-trio (e.g., "chief executive officer" being replaced by "ceo") substitutions. Hence, we have decided to select the optimal number of topics by trial and error using the $LDAvis$ visualization tool and the cohesiveness of the word list. Table 1 below presents the topics identified among the UK and European CSR reports.

[Table 1 about here.]

5.2 Topic Analyses: UK Industry and Time Trend

To conserve space, we present industry and time trend analyses for UK CSR reports only. Figure 3 presents the frequency of UK topics by industry while Figure 4 presents the frequency of UK topics by year.

[Figure 3 about here.]

It is very clear from Figure 3 that CSR topics are sector dependent. "Food waste" and "food packaging" are strongly featured in the two consumer sectors. "Healthcare product diseases" and "community world health" are prevalent in the Health Care sector, although we might wish to see "employee safety" to be strongly featured in more than just the Industrial sector. The Material sector has "paper forest" and "coal mine site safety" as key CSR topics. There is no surprise that "mobile network access" is strongly featured in Telecommunication Services while "efficient power" is strongly featured in the Utilities sector. It is intriguing that "human right" is the dominating topic in two sectors; Energy and Funds & Investment Co.

Figure 4 presents the topic frequencies for UK CSR topics year by year from 1999 to 2017. The frequency counts are much lower and sparse in 1999, 2000 and 2017, which may be due to the lack of centralized depository in the early year, and the reporting time lag for 2017. Other than that, the change in CSR topic frequencies over time is relatively stable. We will present more distinct trends for specific topics in the next subsection.

[Figure 4 about here.]

[Figure 5 about here.]

5.3 Evolution of Selected UK Topics

Figures 5 to 8 analyze the time trend for four specific topics in detail. Figure 5 shows that topic 4, "supply chain," has increasingly become more prevalent in almost all sectors and especially for Consumer Discretionary.[9]

[Figure 6 about here.]

[9]There is a lack of word count for topic 4 in Real Estate in the later year after an upsurge just before 2010.

Figure 6 shows that topic 7, "human right," occupied the headlines in Funds & Investment Co, Energy, and, to a lesser extent, Consumer Staples and Real Estate in early 2000. It seems to enjoy a mild recovering trend from around 2015 which coincides with the new Modern Slavery Act 2015.

[Figure 7 about here.]

Figure 7 shows that topic 10 (carbon emission) is prevalent in Consumer Discretionary and Utilities sectors.[10] The Utilities sector is the most critical sector as far as carbon emission, and greenhouse gas production are concerned. The fact that "carbon emission" is prevalent in Consumer Discretionary but not in Consumer Staples, Energy, and other sectors might be due to the fact the Consumer Discretionary is the most "customers facing" sector. There is a belief that CSR reporting is more important for B-to-C (business to consumer) than for B-to-B (business to business) setup.

[Figure 8 about here.]

Figure 8 shows that topic 17 (employee safety) had a very high frequency count for Industrials in the early 2000. The frequency counts have dropped throughout the years for all sectors.

[Figure 9 about here.]

6 Discussion and Conclusion

This paper is the first large-scale textual analysis of disclosure practices of European and UK CSR reports. We use a battery of Python codes to organize the thousands of reports downloaded from Bloomberg and the internet, the latest R-algorithm to estimate LDA (Latent Dirichlet Allocation) model and the $LDAvis$ interactive tool to

[10]The dip in 2017 is probably due to a reporting lag since our data collection stopped in July 2017.

visualize and refine the LDA model. The computer-based text mining technique allows us to analyze 1,663 CSR reports by 192 UK companies and 3,618 CSR reports by 2,493 European firms from 1999 to 2017. The work can be easily expanded to include all the annual reports and ESG reports of the same period.

Our text mining results and LDA analyses indicate that “employees safety,” “employees training support,” “carbon emission,” “human right,” “efficient power,” and “healthcare medicines” are the common topics reported by publicly listed companies in Europe and the UK. There is a clear sector bias with industrial firms emphasizing “employee safety,” Utilities concentrating on “efficient power” while consumer discretionary and consumer staples highlighting “food waste” and “food packaging.”

There are various weaknesses and limitations in the current version of the paper. First of all, the results are sensitive to “garbage”; garbage is treated as text in this framework and carries equal weight with normal text. To perform a competent task, we need a multi-lingual detector and subject-specific scientific knowledge to train machine models even just to recognize garbage. At the moment, the topic identified has no detailed content; we are not able to distinguish good and bad CSR performance. We need to perform further work along the line of "sentiment". There is not yet an efficient way of tracing topic keywords back to their exact locations in the source documents. At the time of writing, we have made steady progress on colored code for highlighting keywords in original documents, and future plans include a cloud network. Finally, it is clear that focusing on word frequency is “barking at the wrong tree”; many words (e.g., sustainability, social responsibility, etc.) are used too often by companies in their green washing propaganda. We plan to use a topic based dictionary to counteract this fundamental weakness of the LDA algorithm. Finally, since all these reports are produced by the companies, future work could benchmark corporate CSR disclosure in social media, news, and the analytic communities (e.g., Bloomberg, CDP, etc.) against the corporate self-reported CSR information.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003) "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022.

Bonsall, Samuel B. IV, Andrew J. Leone, Brian P. Miller, and Kristina Rennekamp (2017) "A plain English measure of financial reporting readability," *Journal of Accounting and Economics*, Vol. 63, pp. 329–357.

Cho, Charles H., Matias Laine, Robin W. Roberts, and Michelle Rodrigue (2015) "Organized hypocrisy, organizational faÃ§ades, and sustainability reporting," *Accounting, Organizations and Society*, Vol. 40, pp. 78–94.

Dyer, Travis, Mark Lang, and Lorien Stice-Lawrence (2016) "The Evolution of 10-K Textual Disclosure: Evidence from Latent Dirichlet Allocation," *Journal of Accounting and Economics*, Vol. Forthcoming.

Global Reporting Initiative (2000-2011) "Sustainability Reporting Guidelines."

Huang, Allen, Reuven Lehavy, Amy Zang, and Rong Zheng (2017) "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach," *Management Science*.

Kuhn, Anna-Lena, Markus Stiglbauer, and Matthias S. Fifka (2015) "Contents and Determinants of Corporate Social Responsibility Website Reporting in Sub-Saharan Africa: A Seven-Country Study," *Business & Society*, pp. 1–44.

Lang, Mark and Lorien Stice-Lawrence (2015) "Textual analysis and international financial reporting: Large sample evidence," *Journal of Accounting and Economics*, Vol. 60, pp. 110–135.

Lee, Edward, Martin Walker, and Colin Chen Zeng (2017) "Do Chinese State Subsidies Affect Voluntary Corporate Social Responsibility Disclosure?" *Journal of Accounting and Public Policy*, Vol. Forthcoming.

Lee, Linda-Eling and Matt Moscardi (2017) "2017 ESG Trends to watch," *MSCI ESG Research LLC*.

Melloni, Gaia, Ariela Caglio, and Paolo Perego (2017) "Saying more with less? Disclosure conciseness, completeness and balance in Integrated Reports," *Journal of Accounting and Public Policy*, Vol. 36, pp. 220–238.

Milne, Markus J. and Ralph W. Adler (1999) "Exploring the reliability of social and environmental disclosures content analysis," *Accounting, Auditing & Accountability Journal*, Vol. 12, No. 2, pp. 237–256.

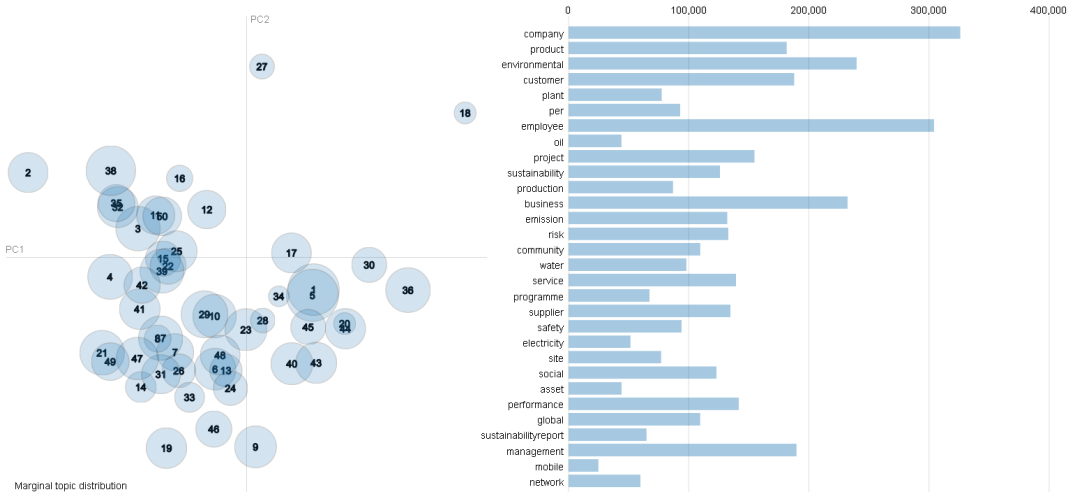
Risi, David and Christopher Wickert (2017) "Reconsidering the 'Symmetry' Between Institutionalization and Professionalization: The Case of Corporate Social Responsibility Managers," *Journal of Management Studies*, Vol. 54, No. 5, pp. 613–646.

Sievert, Carson and Kenneth E. Shirley (2014) "LDAvis: A method for visualizing and interpreting topics," *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70.

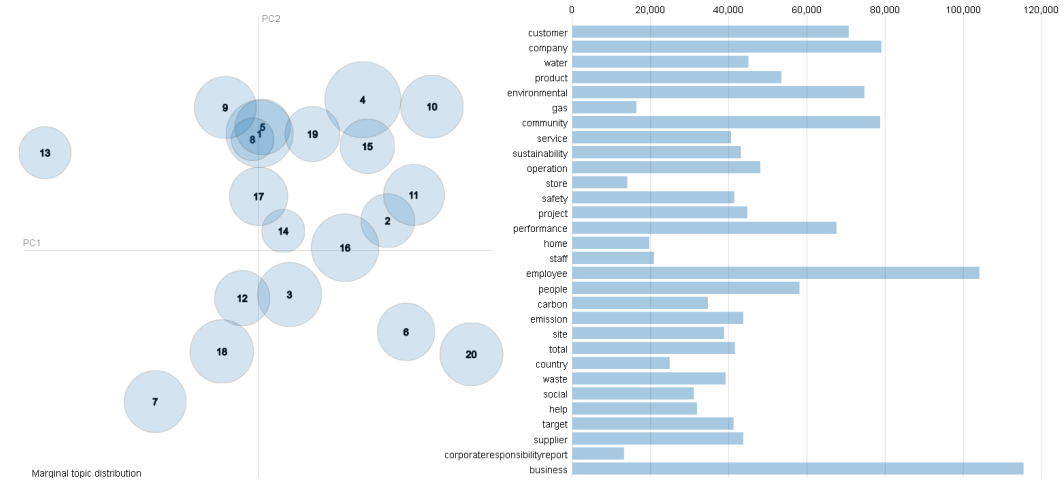
List of Figures

Figure 1: LDA Topic Analysis Using $LDAvis$ visualization

(a) EU CSR Reports with 50 Topics (left) and Top 30 Most Sailent Terms (right)



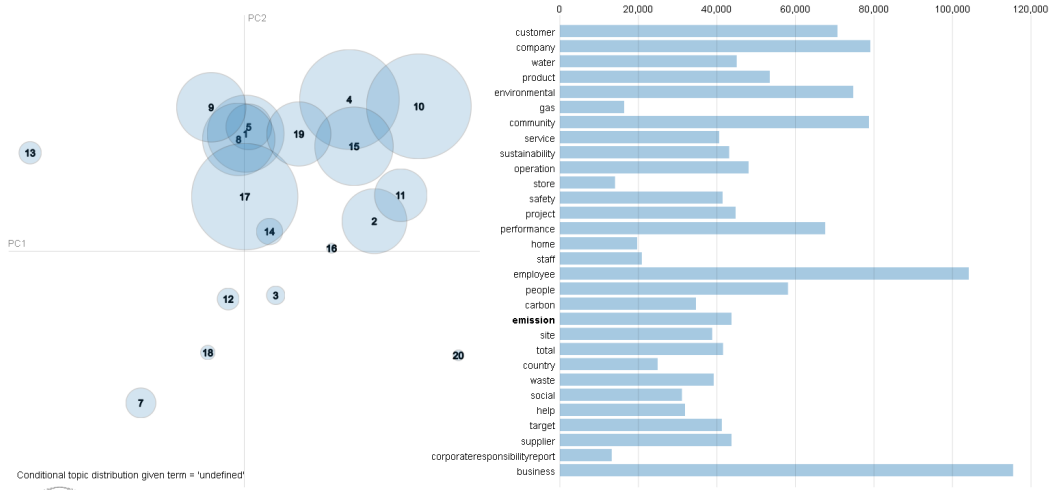
(b) UK CSR Reports with 20 Topics (left) and Top 30 Most Sailent Terms (right)



The graphs on the LHS are Intertopic Distance Map of topics via multidimensional scaling. The more overlapped of the topics means they share a larger proportion of words. Figure (1a) has more overlapping areas than Figure (1b) possibly becasue the number of topic set is higher (i.e. 50 vs. 20). Topics 2 and 18 in Figure (1a), and similarly topics 7 and 13 in Figure (1b), appear to contain very different sets of words as they have no overlapping areas. On the RHS are the actual word counts for the top 30 most important words appeared in the topics. In figure 1b, the word "business" appeared most frequently for nearly 120,000 times but it is less important then the word "customer" for the topics selected for the UK CSR reports.

Figure 2: UK $LDAvis$ visualization: Keyword "Emission" vs. Topic 10 "Carbon Emission"

(a) Distribution of the word "emission" among the 20 CSR topics



(b) UK CSR Reports Topic 10 "Carbon Emission" (left) and Top 30 Most Sailent Terms (right)

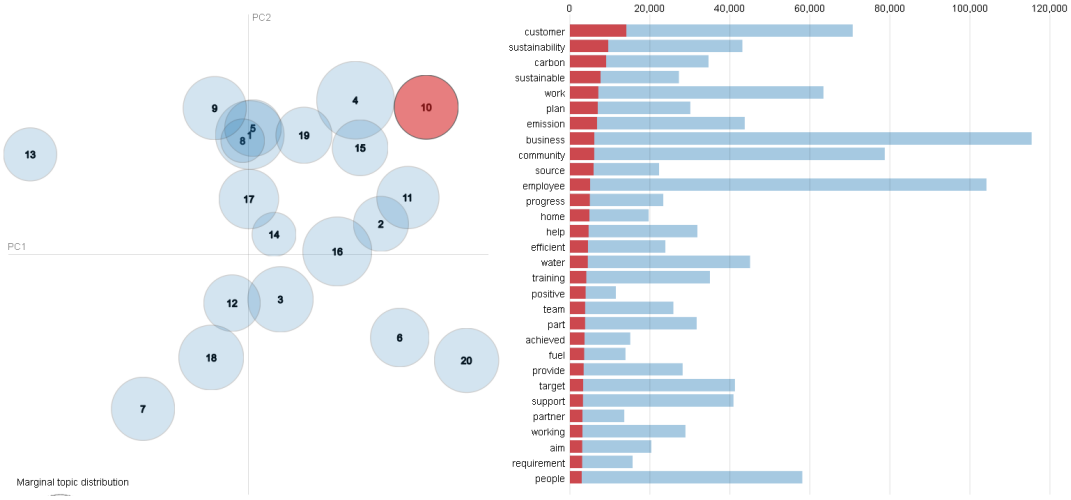
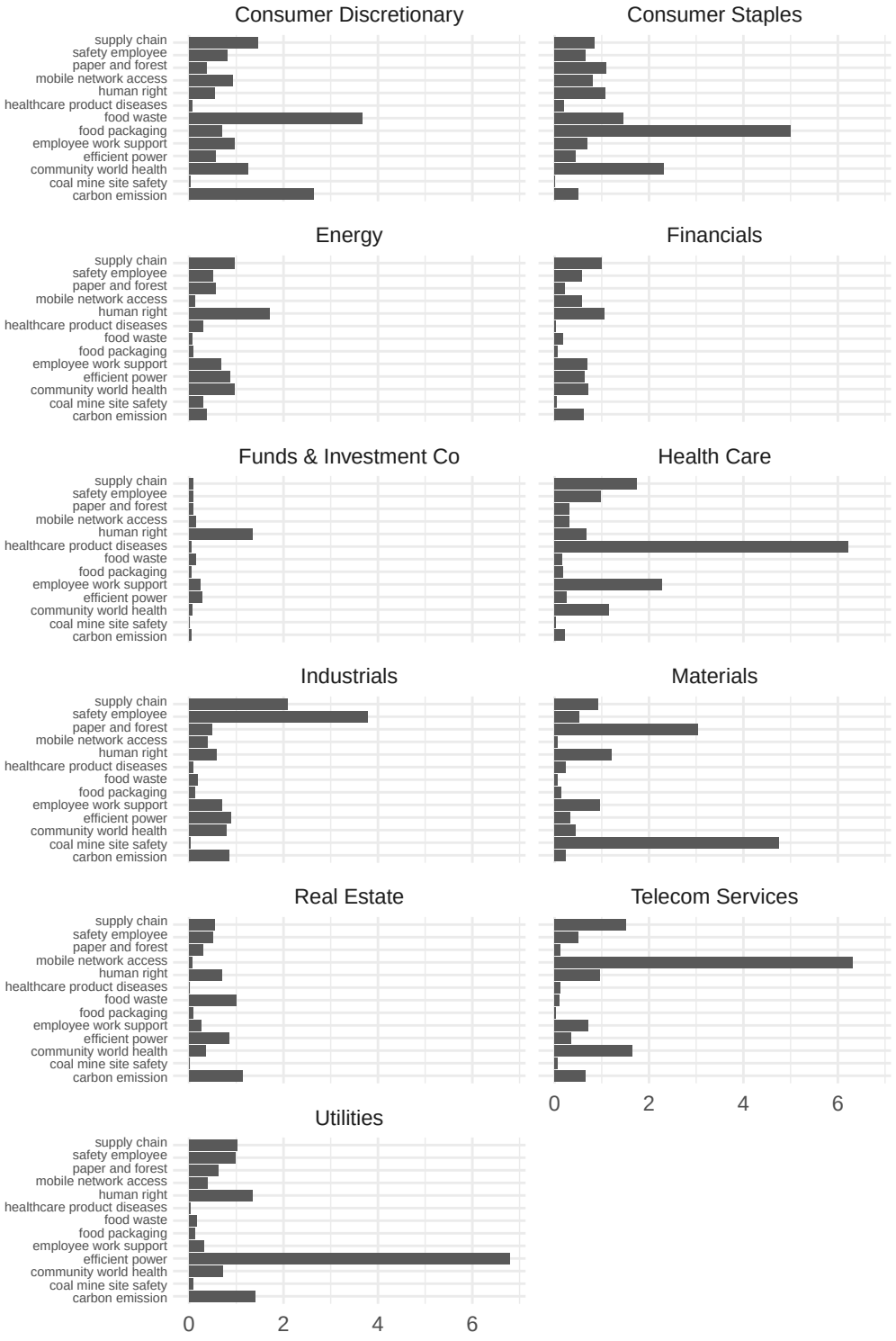


Figure (2a) shows the distribution of the word "emission" among the 20 selected UK CSR topics, and its dominance in topics 10 and 17. Figure (2b) shows the frequency counts for the top 30 most important words for topic 10. The lighter shade bars are the actual frequency counts of the words from the entire corpus whereas the darker shade bars reflect the frequency counts in topic 10.

Figure 3: Topic Frequency for selected UK CSR Topics By Industry



This figure shows the distribution of 13 selected topics across 11 industries. It shows that many topics are strongly sector biased, but there are topics, e.g. "human right", that span across all industries.

Figure 4: Topic Frequency for UK CSR Topics By Year(a) 1999-2008

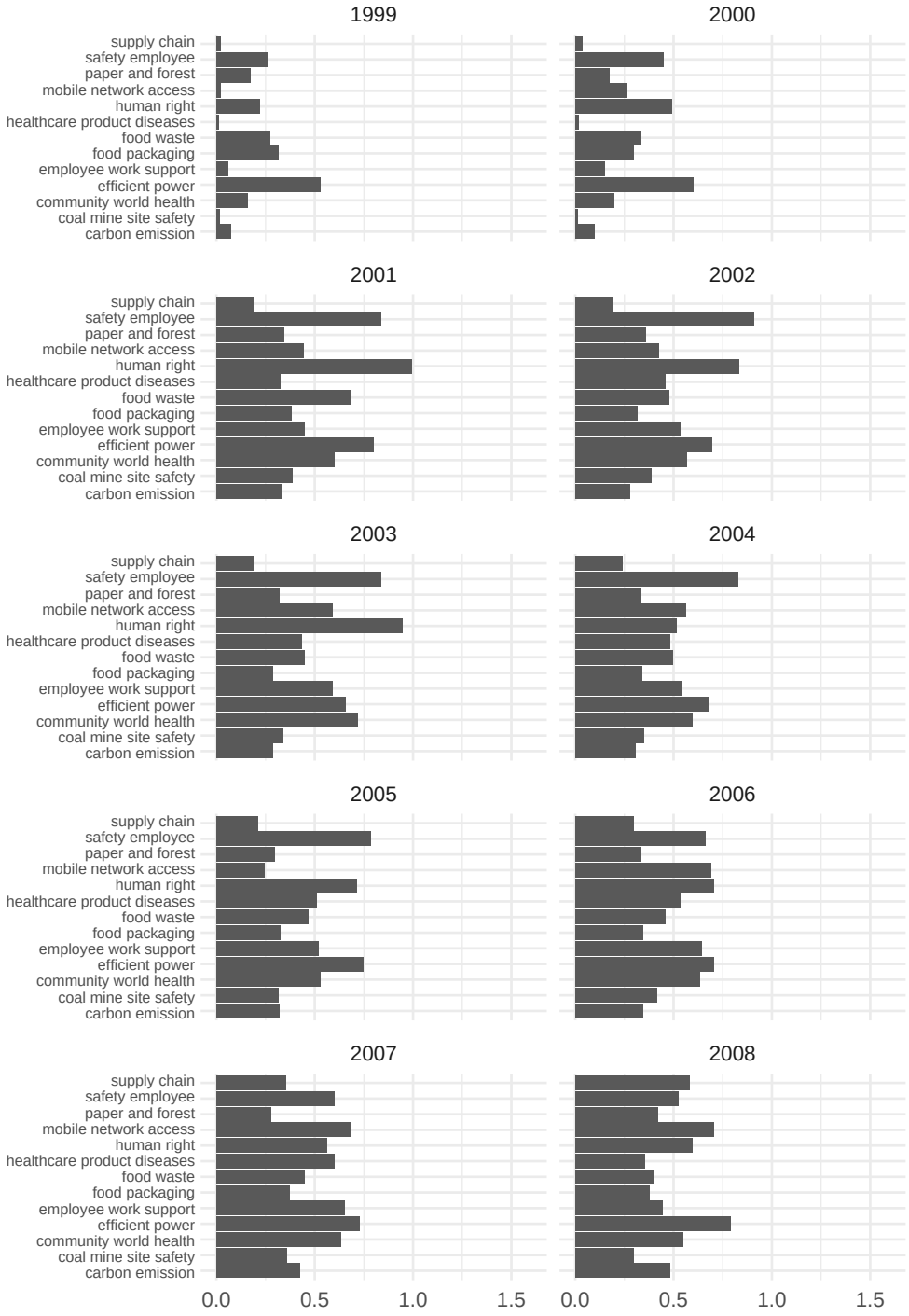
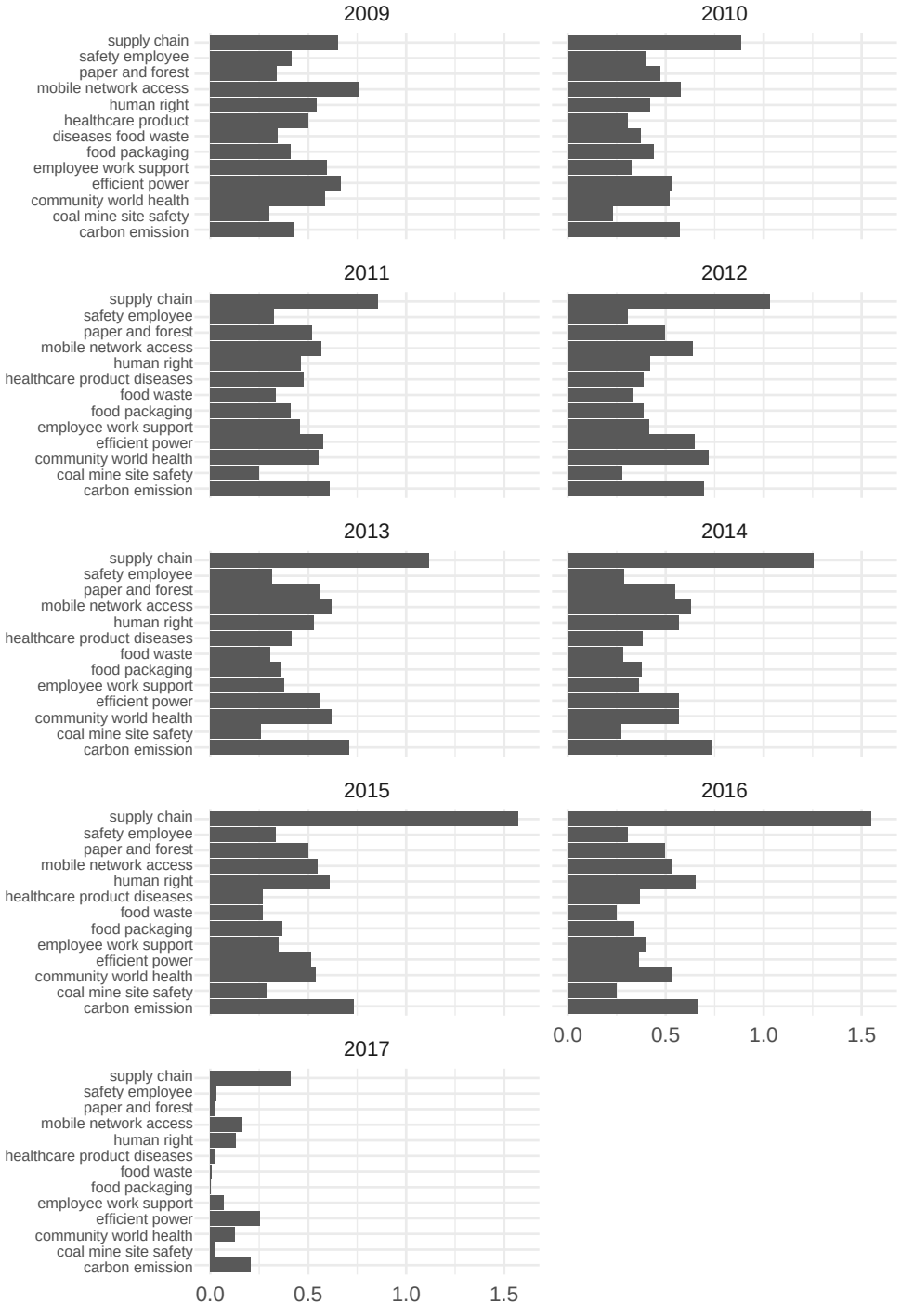
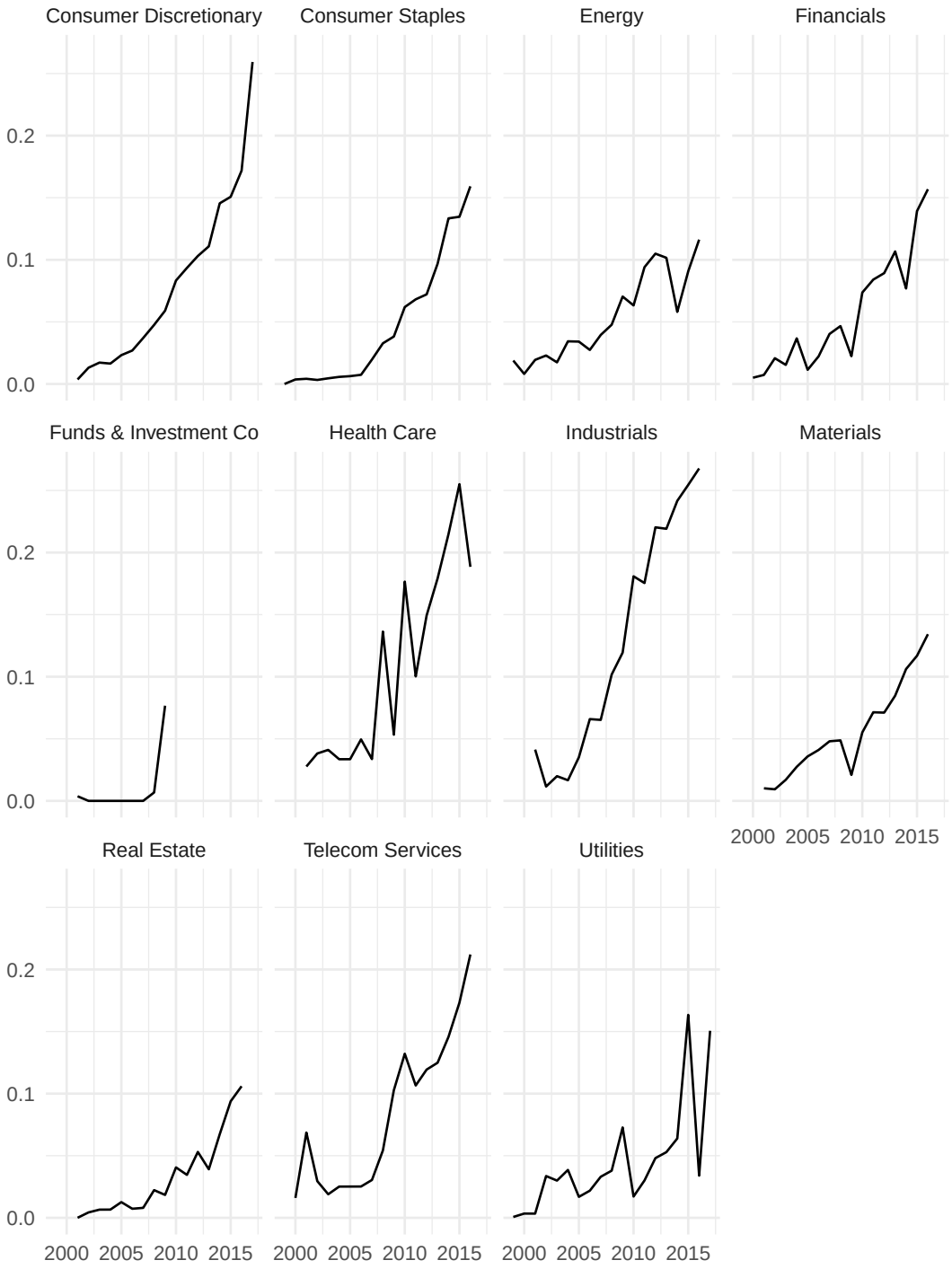


Figure 4: (b) 2009-2017



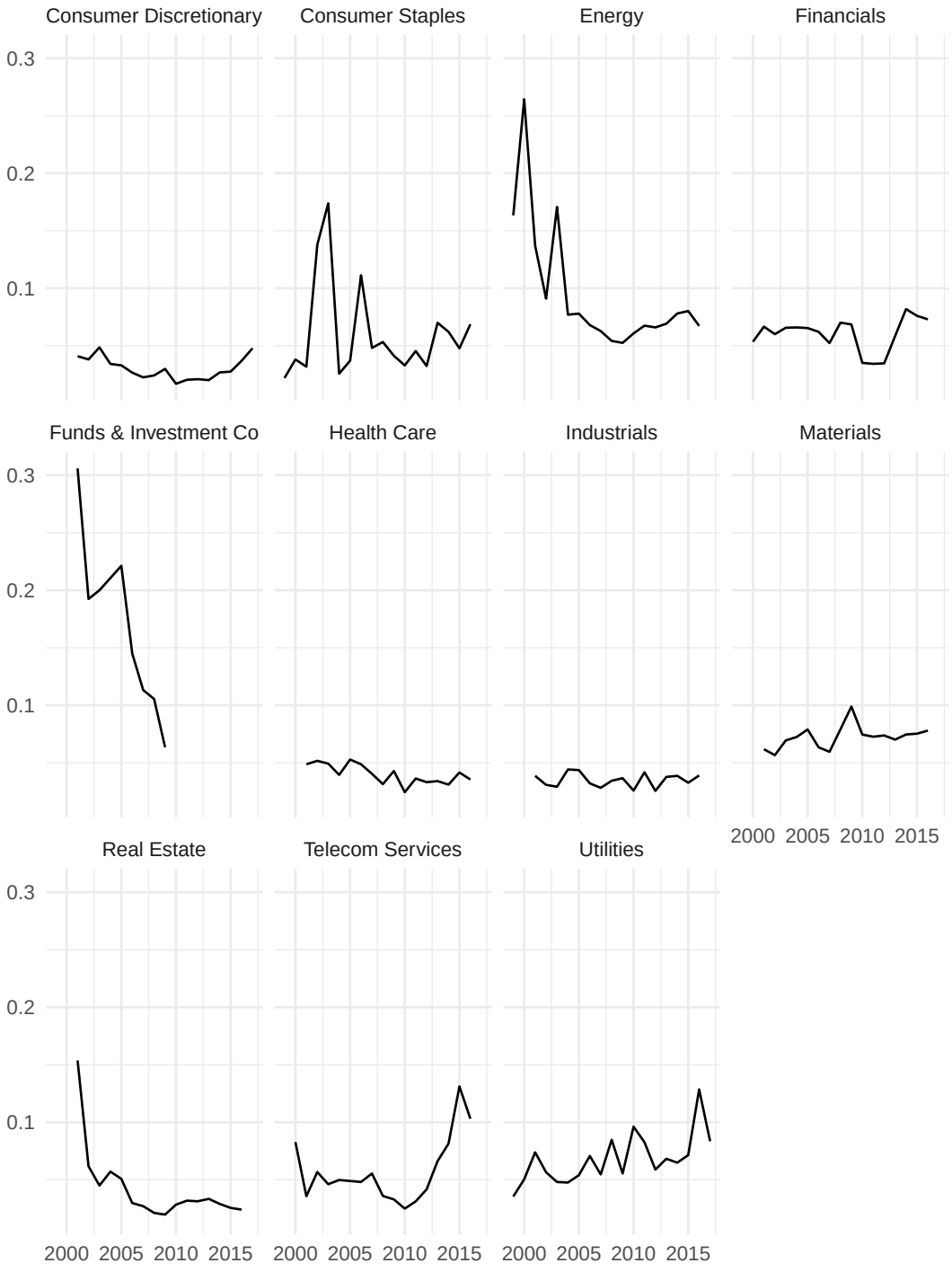
This figure shows the distribution of 13 selected topics by year. The distribution is thinner in earlier years (i.e. 1999 and 2000) as there were fewer reports, and it is thinner in 2017 as many reports were not released at the time of data collection. It is clear that e.g. "supply chain" has become a prominent topic since 2011 probably due to its connection with modern slavery and the UK Modern Slavery Act that finally became law in 2015.

Figure 5: UK Topic 4 (Supply Chain) frequency evolution by industry



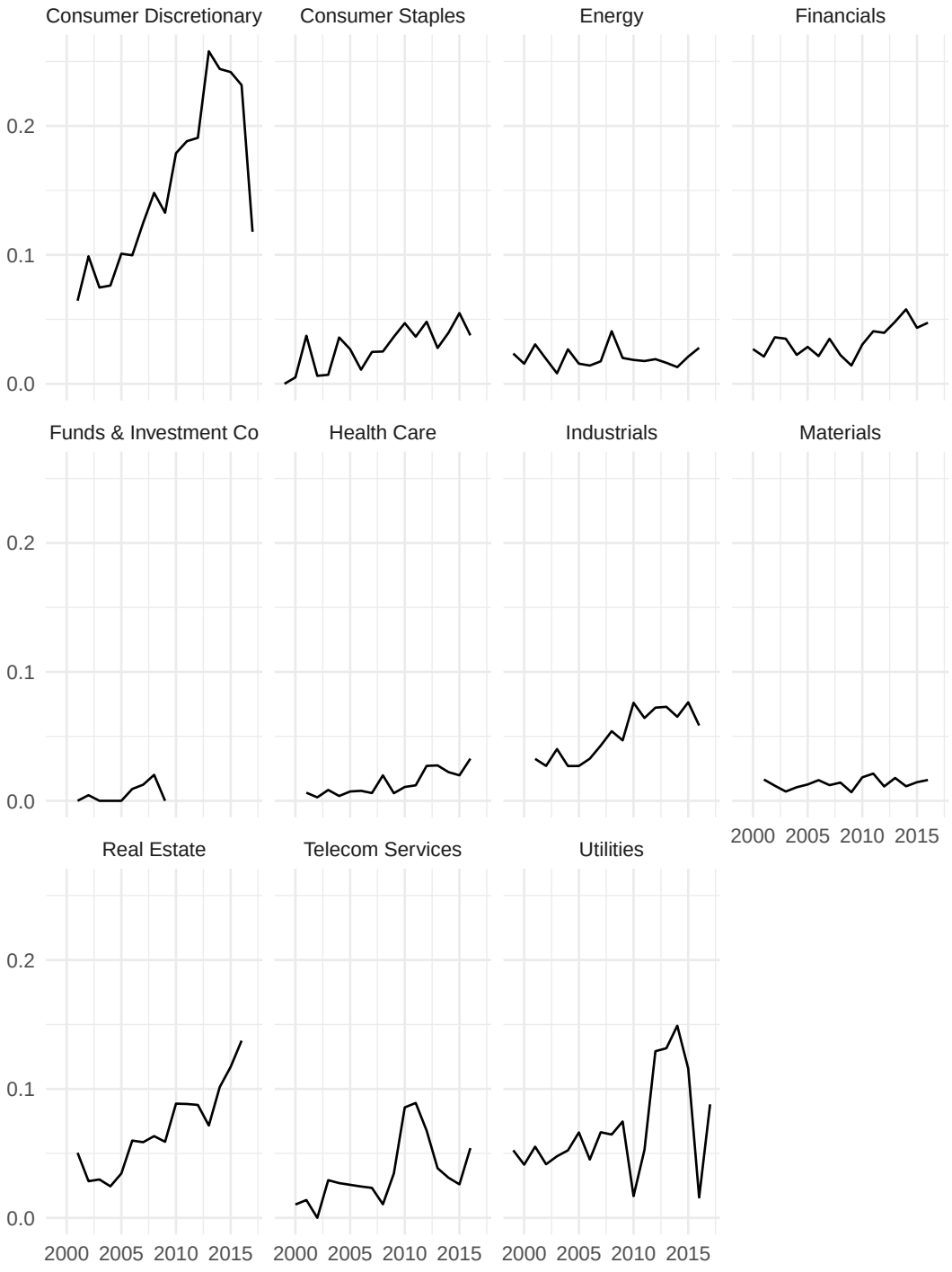
This figure shows the frequency distribution of topic 4 "supply chain" across the 11 industry from 1999 to 2017. It clearly shows the increased importance of the topic over the years.

Figure 6: UK Topic 7 (Human Right) frequency evolution by industry



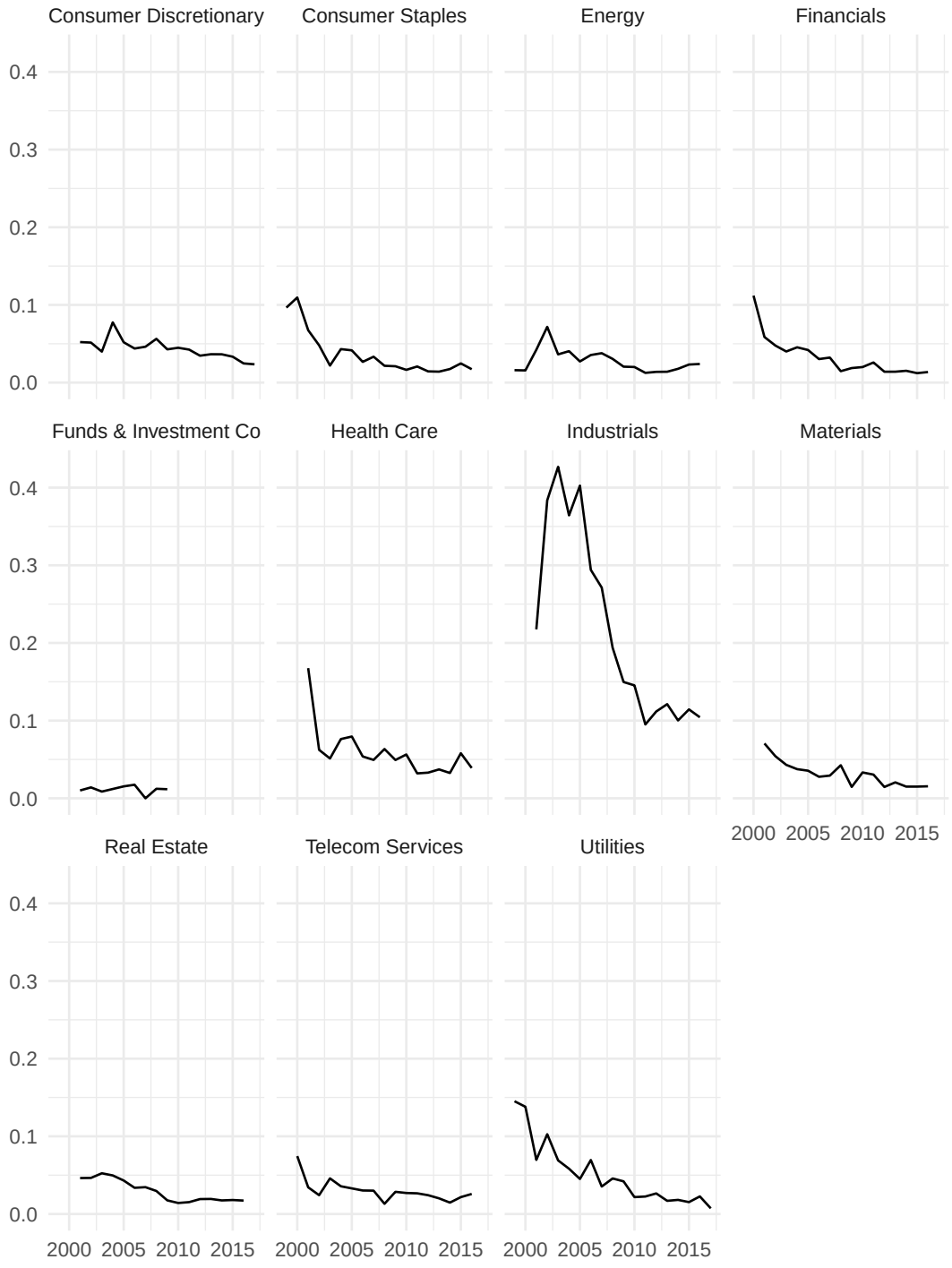
This figure shows the frequency distribution of topic 7 "human right" across the 11 industry from 1999 to 2017. The topic peaked in early 2000. It seems to enjoy a mild recovering trend from around 2015 which coincides with the new Modern Slavery Act 2015.

Figure 7: UK Topic 10 (Carbon Emission) frequency evolution by industry



This figure shows the frequency distribution of topic 10 "carbon emission" across the 11 industry from 1999 to 2017. Despite the universal importance, the topic appears strongly among utility and consumer discretionary, the strongest B-t-C type industries that are most sensitive to carbon emission issue.

Figure 8: UK Topic 17 (Employee Safety) frequency evolution by industry



This figure shows the frequency distribution of topic 17 "employee safety" across the 11 industry from 1999 to 2017. This topic had a very high frequency count for Industrials in the early 2000. The frequency counts have dropped throughout the years for all sectors.

List of Tables

Table 1: UK and European CSR Topics

UK		Europe	
topic 1:	paper forest	topic 4:	carbon emission efficient reduce
topic 2:	mobile network access	topic 5:	employee social training initiative
topic 4:	supply chain	topic 9:	electricity efficient fuel nuclear transmission
topic 5:	food packaging	topic 12:	employee support education children school germany
topic 6:	employee work support	topic 13:	oil fuel naturalgas emission
topic 7:	human right	topic 19:	employee product safety
topic 9:	coal mine site safety	topic 21:	water nutrition food farmer
topic 10:	carbon emission	topic 22:	human right mining coal health
topic 12:	food waste	topic 23:	climate strategy initiative
topic 14:	healthcare product diseases	topic 31:	airport aircraft transport safety
topic 15:	efficient power	topic 33:	healthcare medicines patients safety
topic 17:	safety employee	topic 38:	human right principle social
topic 20:	community world health	topic 41:	recycling waste system material management
		topic 45:	biodiversiy spanish commitment
		topic 46:	paper mill water forest
		topic 49:	carbon footprint target sustainability