Chen, Lin; Houle, Stephanie

Working Paper

Turning words into numbers: Measuring news media coverage of shortages

Bank of Canada Staff Discussion Paper, No. 2023-8

Provided in Cooperation with:
Bank of Canada, Ottawa

Suggested Citation: Chen, Lin; Houle, Stephanie (2023) : Turning words into numbers: Measuring news media coverage of shortages, Bank of Canada Staff Discussion Paper, No. 2023-8, Bank of Canada, Ottawa, https://doi.org/10.34989/sdp-2023-8

This Version is available at:
https://hdl.handle.net/10419/297077

BANK OF CANADA
BANQUE DU CANADA

Staff Discussion Paper/Document d'analyse du personnel—2023-8

Last updated: March 31, 2023

Turning Words into Numbers: Measuring News Media Coverage of Shortages

by Lin Chen and Stephanie Houle

Canadian Economic Analysis Department
Bank of Canada, Ottawa, Ontario, Canada K1A 0G9
LChen@bank-banque-canada.ca, SHoule@bank-banque-canada.ca

DOI: https://doi.org/10.34989/sdp-2023-8 | **ISSN** 1914-0568

Acknowledgements

We offer a special thank you to Tatjana Dahlhaus and Alexander Ueberfeldt for comments and feedback on this paper. We would also like to thank Pierre-Yves Yanni, Jerome Lyons, Laurent Martin, Michael Francis, Alexander Chernoff and Benjamin Strauss for fruitful discussions, as well as Boyan Bejanov and André Binette for providing us with their code as a reference for the initial query and cleaning of the Cision news media database.

Abstract

We generate high-frequency and up-to-date indicators to monitor news media coverage of supply (raw, intermediate and final goods) and labour shortages in Canada. We use natural language processing to construct two news-based indicators and time-varying topic narratives to track Canadian media coverage of these shortages from 2000 to 2022. This makes our indicators an insightful alternative monitoring tool for policy. Notably, our indicators track well with monthly price indexes and measures from the Bank of Canada's Business Outlook Survey, and they are highly correlated with commonly tracked indicators of supply constraint. Moreover, the news-based indicators reflect the attention of the public on pressing issues.

Topics: Coronavirus disease (COVID-19), Econometrics and statistical methods, Monetary policy and uncertainty, Recent economic and financial developments
JEL codes: C55, C82, E37

Résumé

Nous élaborons des indicateurs à haute fréquence pour surveiller en temps quasi réel la couverture médiatique des ruptures d'approvisionnement (matières premières, produits intermédiaires et produits finis) et de la pénurie de main-d'œuvre au Canada. Nous utilisons le traitement automatique des langues pour établir deux indicateurs fondés sur les actualités ainsi que la trame narrative des sujets abordés, variable au fil du temps, pour suivre la couverture des pénuries dans les médias canadiens de 2000 à 2022. Ces indicateurs offrent un autre type de surveillance pertinente en matière de politiques. Ils cadrent bien avec les indices des prix mensuels et les mesures tirées de l'enquête sur les perspectives des entreprises de la Banque du Canada, en plus de suivre de près les indicateurs courants des contraintes d'approvisionnement. Ces indicateurs reflètent aussi l'attention publique accordée à ces enjeux pressants.

Sujets : Maladie à coronavirus (COVID-19), Méthodes économétriques et statistiques, Incertitude et politique monétaire, Évolution économique et financière récente

Codes JEL : C55, C82 et E37

Section 1: Introduction

High-frequency data obtained in real time have the potential to revolutionize economic monitoring. One example of these data is news-based indicators, which can work as an alternative source to track a broad range of topics. These indicators can help monitor, for instance, the COVID-19 pandemic that resulted in broad disruptions in supply and labour due to quickly changing containment measures aimed at controlling the spread of the virus. Similarly, they can track shortages impacting supply chains and labour markets that put upward pressure on prices and inflation. News-based indicators can also generate narratives to describe media attention on current economic issues.

We use news media to track these various shortages, such as supply and labour, with up-to-date reports on the development of these shortages. The general population widely reads the news, which thus serves as an "information intermediary" that may influence people's views and expectations of the economy (Larsen and Thorsrud 2017; Larsen, Thorsrud and Zhulanova 2021). While news reports tend to be mostly reliable accounts of current events, they can have a bias toward stories that are more likely to grab the public's attention. This makes them effective at providing early warning signals of new developments. A less-desirable implication is that the public attention and related behaviour changes based on news can be short-lived (Slovic et al. 2017). Hence, we expect a decline over time in the attention paid by news media to specific topics.

We use the Cision news media database to build a corpus (a collection of written texts) of Canadian English economic and management news articles published between January 1, 2000, and July 4, 2022. We extract sentences containing the keyword "shortage" and collect them into daily documents. Then, we use unsupervised natural language processing (NLP) to decompose the news corpus into different shortage topics. These topics reflect media attention on, among others, the supply and labour shortages that affected Canada's economy from the early 2000s to the 2020–22 period of the COVID-19 pandemic.

We show that the real-time indicators for supply and labour shortage topics correlate well with measures obtained with a lag. Our indicators correlate reasonably well with related consumer price index (CPI) categories and extremely well with Statistics Canada's Industrial Producer Price Index (IPPI) and survey measures from the Bank of Canada's Business Outlook Survey (BOS). Our indicator for supply shortages is also highly correlated with various indicators of supply constraints, such as freight cost indexes. Taking the pandemic as an example, we find that our supply shortages indicator started to increase in 2020 and remained elevated as of July 2022. Likewise, our indicator for labour shortages reached an all-time high in November 2021 and remained elevated until at least July 2022. These news-based indicators allow us to look through the lens of news media to obtain rich information on various shortage issues.

Using large-scale, high-frequency news media data and NLP methods is an emerging trend, creating real-time macroeconomic variables for policy monitoring. Some examples of this include:

- using news media to model business cycle and economic fluctuations (Bybee et al. 2021; Larsen and Thorsrud 2018, 2019);
- building an index of climate change transition risks to forecast commodity currencies (Kapfhammer, Larsen and Thorsrud 2020);
- forecasting crude oil market prices (Li et al. 2019) or gross domestic product, CPI and unemployment (Kalamara et al. 2020); and
- constructing an uncertainty index (Baker et al. 2020).

Media text analytics have also been used to gauge inflation expectations based on media mentions of prices (Angelico et al. 2022), topics (Larsen, Thorsrud and Zhulanova 2021) or tone (Lamla and Lein 2014; Shapiro, Sudhof and Wilson 2022; Pfajfar and Santoro 2013).

A large portion of the literature on shortages comes from the early 1990s in the context of how fiscal price controls or market manipulations could lead endogenously to shortages; this literature includes Qian (1994) and Weitzman (1991). However, the COVID-19 pandemic presented a unique exogenous shock where containment measures for the virus disrupted supply chains and restricted business activities. Guerrieri et al. (2022) discuss how negative supply shocks, such as those caused by lockdowns and firm closures during the pandemic, could lead to excess demand and spillover into other sectors. The impact of supply and labour shortages increases the risk of inflation and a wage-price spiral. Shapiro (2022) shows that supply shortages drove more than half of the elevated level of inflation for personal consumption expenditures during the COVID-19 pandemic.

Looking more specifically at how much media focuses on shortage, Lamont (1997) uses the frequency of mentions of the word "shortage" in news headlines to forecast short-run inflation. Pitschner (2022) highlights the advantage of isolating supply shortages from demand issues using textual analysis. The contextual information extracted from US corporate filings reflects the labour and supply chain disruptions during the pandemic (Pitschner 2022).

However, not much has been done to create timely indicators showing the media's main focus on shortages and their evolution over time. To the best of our knowledge, we are taking a step further by extracting the main topics related to shortages from large-scale news media text data to provide indicators with rich context. Specifically, we use NLP techniques to collect sentences containing the keyword "shortage," and then create alternative, news-based indicators with time-varying narratives. These narratives reveal the media attention on various goods supply and labour shortages. For this paper, we focus on shortages; however, the methodology could be widely applied to other pressing economic issues, such as inflation expectations, employment or housing.

We thus believe that the daily high frequency of media text-based indicators makes them a useful proxy to monitor pressing issues such as supply and labour shortages. These media-based indicators thus provide early warning signs and affect public expectations.

The paper's structure is as follows. In Section 2, we explain the workflow of our text analytics and the associated NLP methods. Specifically, we discuss three main steps: selection, processing and topic modelling. In Section 3, we describe our main findings, by showing the time series for the supply and labour shortage indicators, and by showing that those indicators correlate well with alternative, lower-frequency ones. We also showcase the usefulness of the time-varying topic narratives and illustrate how media storylines about supply and labour shortages have changed over time. We conclude in Section 4.

Section 2: Text analytics

Data

For this project, we use Canadian English news from January 1, 2000, to July 4, 2022, on the subject of economics and management from the database[1]. This subject provides around 3.8 million news articles in total, or around 200,000 news articles per year since 2004.[2] We extract essential information from the media by collecting sentences from those articles that contain the keyword "shortage" to form our shortage-related corpus. To reduce the high dimensionality of the matrix, we apply a decomposition technique known as topic modelling. The outputs are the topic time series (our news-based indicators), the word clouds that describe each topic and time-varying narratives of labour and supply shortages in the news media. The topic model for this shortage-related corpus is trained on documents from January 1, 2000, to June 30, 2021, then applied to later dates to extend the topic time series.

Methodology

We follow three main NLP steps to turn the words from the news into numbers: selection, processing and topic modelling.[3] We will discuss each in turn. A simplified example of our methodology (based on invented documents) is in Appendix 1.

1. **Selection**: We first clean articles by removing tables and symbols as well as dropping duplicated and empty articles.[4] We then further clean sentences within these articles by removing apostrophes and stop words, filtering out numbers and lemmatizing

[1] We collect economic and management news using subject labels provided by the Cision news media database.

[2] From 2000 to 2003, around 55,000 articles per year appear under the above filters.

[3] Our computations use the scikit-learn module developed by Pedregosa et al. (2011).

[4] In addition, we filter sentences within the range of 4 to 100 words to avoid sentences that are too short or too long. We also remove disclaimers, URLs, phone numbers, emails and sentences with too many numbers.

words (Bird, Klein and Loper 2009).[5, 6] Next, we select articles related to the word "shortage." We specifically keep sentences containing this keyword. This leaves us with approximately 41,490 sentences as of June 2021. These sentences are combined and collapsed together into daily documents (Larsen and Thorsrud 2019).

2. **Processing**: We identify words that are relevant to shortages using their Pointwise Mutual Information (PMI) score, based on the term frequency (TF) for each word.[7, 8] Finally, we calculate term frequency–inverse document frequency (TF-IDF) scores to convert the text into numbers.[9]

    a. We first construct a list of selected unigrams for the shortage corpus by eliminating words with low relevance to the keyword "shortage." We build a list of unigrams (single and hyphenated compound words) ranked by their co-occurrence with this keyword (Church and Hanks 1990). This score allows us to identify words that specifically co-occur with shortage, versus words that frequently occur in most documents. **Table 1** shows a sample of our list of words, with those most relevant to shortage in blue and least relevant in red. We restrict our list of unigrams to those with a daily PMI greater than 1. Thus, the number of words in our matrix decreases significantly from more than 500,000 unigrams to just over 20,000.

Table 1: Examples of term frequency scores for words in the shortage corpus

Word list (unigrams)	Relevance to shortage (PMI)
labour	3,536.73
worker	2,330.54
price	2,000.65

[5] Stop words are common words that do not provide meaningful or corpus-specific information, such as "the" and "is."

[6] Lemmatization means to return a word to its base form—its lemma—based on a morphological analysis of the word and the vocabulary. We do this using the NLTK WordNet library from Bird et al. (2019). For example, "talks" is lemmatized to "talk."

[7] We identify unigrams that co-occur with the keyword "shortage" in documents. For each document, the occurrence is calculated for each unigram (if TF > 0, then the occurrence will be 1). Then, for each day, we calculate the co-occurrence matrix based on how many times each unigram pair co-occurs in each document. Subsequently, the PMI score is calculated as the joint probability of the unigram pair (from the co-occurrence matrix), normalized by each marginal probability. Unigrams with a daily PMI score of greater than 1 with shortage are selected for further analysis.

[8] The term frequency is the count of a word in the documents.

[9] The TF-IDF is $tfidf(t,d,D) = tf(t,d) \cdot idf(t,D)$. The term frequency is the number of times term t appears in document d. The inverse document frequency $idf(t,\ D) = \log\left(\frac{N+1}{|d \in D : t \in d| + 1}\right) + 1$ is related to the total number of documents in the corpus N divided by the number of documents where the term appears; hence, the *idf* is a measure of how common the word is among all the documents.

supply	1,928.48
housing	1,547.00
skill	1,482.92
...	...
tourism-related	1.22
food-service	1.22
refuge	1.00

Note: PMI is the Pointwise Mutual Information score of each unigram that co-occurs with the keyword "shortage."

b. Based on the unigram list we get from step 2a, the sentences in each document are then split into words—known as n-grams—in particular, unigrams, bigrams and trigrams (singles as well as pairs and triplets of adjacent words). The TF-IDF scores are calculated with a Euclidean norm for each n-gram.[10] For example, skilled, labour and shortage are in the unigram list (based on PMI). Moreover, bigrams such as skilled labour and labour shortage and the trigram skilled labour shortage are identified based on the list of unigrams. This provides us with a matrix of words containing the TF-IDF scores for each n-gram for each day.

3. **Topic modelling**: For our corpus, the matrix of words is high dimensional with about 160,000 n-grams. To further reduce its dimension, we perform unsupervised topic modelling. Specifically, we use non-negative matrix factorization (NMF) on the n-grams (Lee and Seung 1999).[11] To select the optimal topic number for the model to balance the granularity and generality of the topics, we identify the upper bound using coherence measures. When the topic number is too high, topics tend to capture granular events rather than themes (e.g., Japan's tsunami-induced shortages of vehicle parts in 2011). For example, if we increase our number of topics, the recent microchip shortage would branch out from our supply shortage topic. We want to capture all goods-supply shortage issues stemming from the pandemic, which includes the microchip shortage, so we cap our number of topics at five. Moreover, we generate word clouds for each topic based on the 100 most representative unigrams as well as the top 100 bigrams and trigrams.[12] We manually label the five topics as supply, labour, health care, skill and housing shortages. We present and discuss the resulting word clouds as well as the associated daily time series in the next section.

[10] The Euclidean norm is the square root of the sum of squares.

[11] For more details, see Appendix 4.

[12] The font size of the word in topic word clouds represents the importance of that word to that topic.

Compared with standard topics derived from topic models, a unique advantage of our workflow is that, for each topic, we identify not only static top words (as in Figures 1 and 2 in the next section), but also time-varying topic narratives (similar to historical decompositions) at a weekly frequency. The top driver words for each topic per day are obtained based on their TF scores, which we aggregate to a weekly frequency by taking the average.[13] The weekly narratives around the topics allow us to assess how media stories for each topic change over time. This is important when complex and pressing shortage issues evolve. We can monitor the evolution of the drivers that dominate shortages over time.

Importantly, the topic model is trained on data from January 1, 2000, to June 30, 2021, to create a trained topic model and generate the topic-words matrix. To extend the sample and update the data, we apply this topic model to new documents. Based on the topic-word matrix and n-grams found in them, the topic model estimates the importance of our topics in each new document. We propose re-training the topic model whenever significant structural changes in vocabulary occur. In this study, any shortage-related words that have never appeared in the news media in the last 21 years would not be captured by our model.

Note that we also check the evolution over time of top driver words for each topic, which are mostly consistent, despite small changes in rank.[14] Thus, the topics are stable over time, whereas the driver words evolve. For instance, for the supply shortage topic, some top words such as chip, semiconductor and oil might peak and drive the supply shortage topic at different times. To obtain an intuitive sense of the NLP workflow, we provide a simplified example with made-up documents in Appendix 3.

Because the workflow is versatile, the keyword selection could be customized. For instance, we could adapt the workflow to build corpora centred on inflation, house prices or a combination of keywords (e.g., inflation, price and cost).

Section 3: Results

Supply and labour shortages captured by indicators

Using the methodology described in the previous section, we identify five topics for the shortage corpus. We get the following word clouds of the top unigrams, bigrams and trigrams for the supply (**Figure 1**) and labour (**Figure 2**) shortage topics. We include the word clouds and time series of the remaining three topics in appendixes 2 and 3. The font size of each word is determined by the topic importance score of words in the topic-word matrix, one of the outputs of the topic model. The first two topics there are the most relevant, given current inflationary pressures. We can see that the word clouds in Figure 1 not only contain n-grams related to shortages of various goods, commodities and raw materials but also mention some

[13] We also test computing the weekly top driver words using their TF-IDF score and get similar results.

[14] The top driver words are the top words (based on TF) for each topic at any given point of time.

words related to intermediate goods, such as chip shortage, part shortage and supply chain. Hence, we manually label it as our supply shortage topic. The word clouds in Figure 2 have many n-grams related to shortages of labour, such as skilled labour, worker shortage and unemployment rate. Thus, we label this our labour shortage topic.

Figure 1: Word clouds for the topic of supply shortage

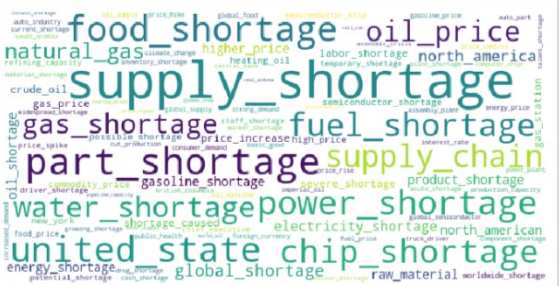
Unigrams	Bigrams and trigrams
	

Figure 2: Word clouds for the topic of labour shortage

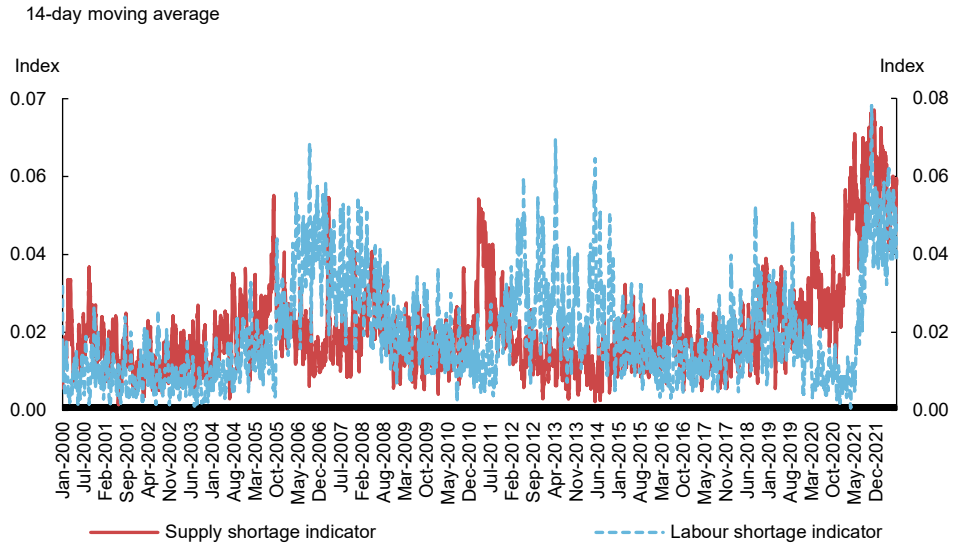
Unigrams	Bigrams and trigrams
	

The time series for the two topics, supply shortages and labour shortages, are shown in **Chart 1**. We refer to these as our indicators of shortages. The time series represent the importance (based on a 14-day moving average) of each topic over time, trained on the news media containing sentences with the keyword shortage. They capture many historical events related to each topic.

For example, for the supply shortage topic, Chart 1 captures (in red) numerous shortage issues related to raw materials and parts. An initial peak appears around 2005, when the economy was booming, oil prices were soaring and building materials for houses were in high demand. Another peak occurred after a large tsunami hit Japan in March 2011, resulting in parts shortages that impacted the supply chains for motor vehicles. By winter 2021, supply shortages

(e.g., in lumber, computer chips and semiconductors) had become prominent in shortage-related news. This topic time series indicator reached its highest level of importance in December 2021 and remained elevated until the end of the sample period in July 2022. The labour shortage topic (shown by the dotted blue line in Chart 1) reached an all-time high at the end of November 2021. As media and public attention is usually short-lived, the long-lasting elevated level of the indicators with changing narratives could not only serve as early warning signs but also flag the evolution of pressing economic issues.

Chart 1: Media attention on the supply shortage and labour shortage topics has remained elevated lately



Source: Sources: Cision media database and Bank of Canada calculations

Last observation: July 4, 2022

We now compare these indicators to price indexes, BOS indicators and alternative supply constraint indicators to determine which facets of the economy they capture. Given that our indicators capture news media's attention on the topics of supply and labour shortages, we expect them to match with surveys meant to capture these supply-side shortages from firms.

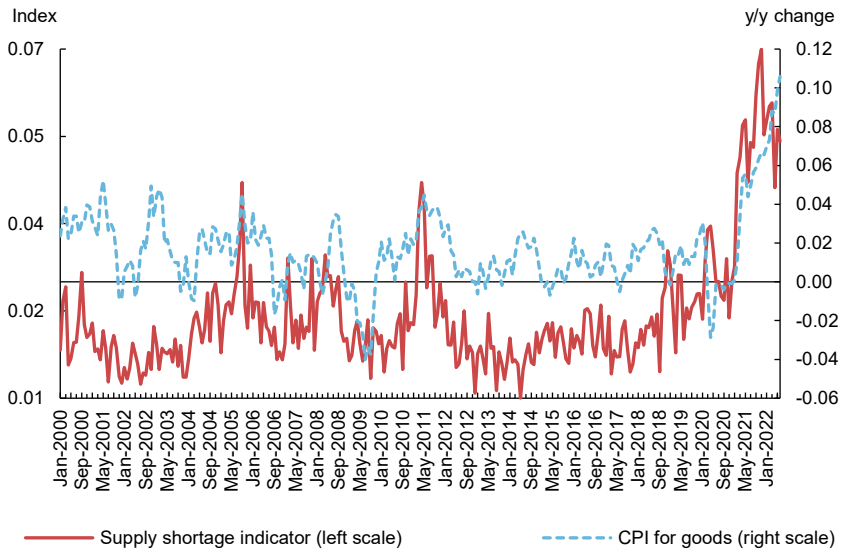
We also expect them to appear right away in the news, especially if they indicate a severe shortage. Over time, however, media attention on each shortage topic might wane as it grabs less reader attention, even if the issue has not yet been resolved.

Since we are working with high-frequency data, we expect to have extra noise, or distortions, in our data relative to lower-frequency indicators that track the long-run trend better. The detailed description and in-depth focus of the media's attention to specific topics could also affect the signal-to-noise ratio. For instance, our supply shortage indicator tracks many volatile changes that grab media attention, such as chip, semiconductor and oil shortages. This tool becomes more useful in times of rapid changes.

Indicators and related consumer price index categories

News media could capture price changes that result from supply and labour shortages impacting daily activities, such as the shortages that emerged during the COVID-19 pandemic of food, gas, lumber and semiconductors. We therefore evaluate how well our news-based indicators of supply and labour shortages track movements in related monthly CPI inflation measures. First, we average our supply shortage series to a monthly frequency and compare it with the monthly CPI measure for goods (**Chart 2**) and the IPPI (**Chart 3**), with their Pearson's correlation coefficient shown below the graph. Our supply indicator tracks well with the CPI for goods and very well with the IPPI, with high correlation coefficients of 0.57 and 0.68, respectively.

Chart 2: **The supply shortage topic indicator correlates well with CPI goods**



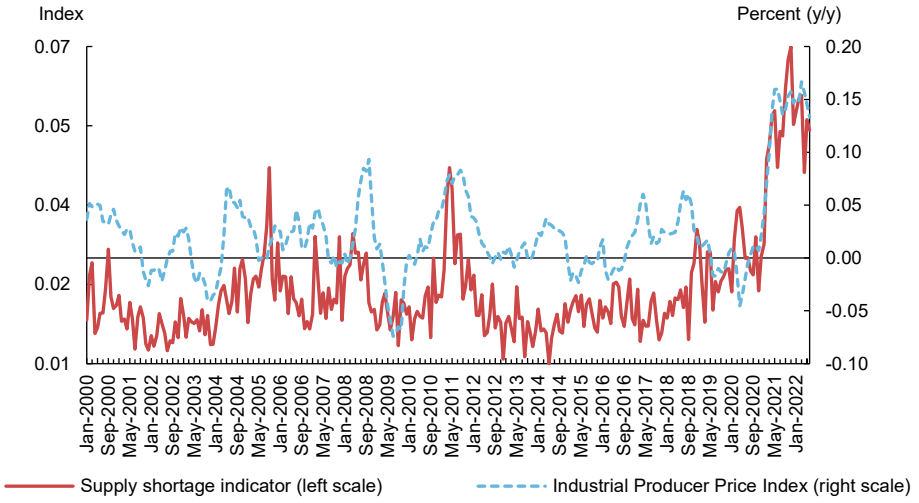
Correlation coefficient = 0.57

Sources: Cision media database (topic series), Statistics Canada series from Table 18-10-0004-01 and Bank of Canada calculations

Last observation: June 2022

We do not expect all changes in goods prices or industrial producer prices to be due to supply shortages. A large, newsworthy shortage can lead to an increase in the prices of goods and inputs, but an increase in such prices is not necessarily caused by a shortage. A few other factors determine the impact of the shortage on prices, such as the number of industries impacted by it and the level of concentration in each affected industry. Hence, this changes the price increase pass-through that firms distribute further along in the supply chain. This explains why our supply shortage indicator is more closely correlated with input prices (IPPI) than it is with goods prices (CPI). Interestingly, since June 2020 our indicator has been closely correlated to both the CPI for goods and the IPPI. This is because supply issues are an important driver for the price increases, and firms have had an easier time passing higher costs along the chain.

Chart 3: The supply shortage topic indicator correlates very well with the Industrial Producer Price Index



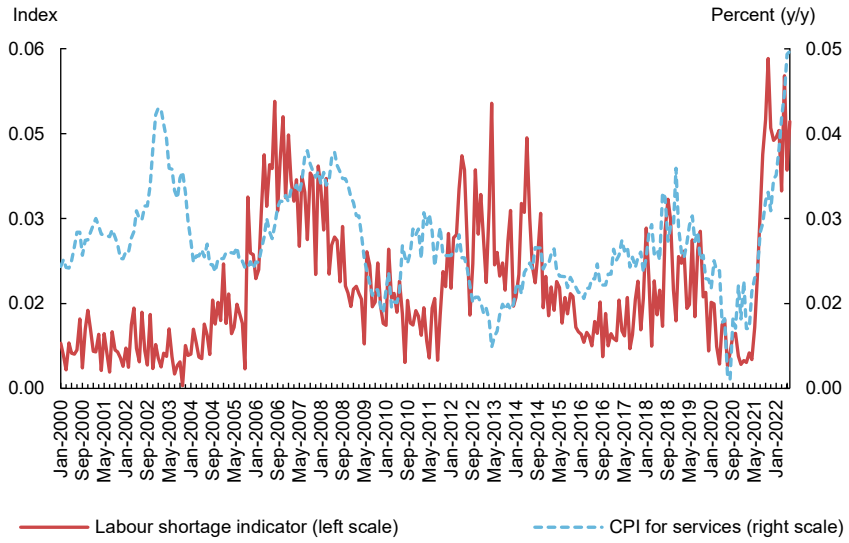
Correlation coefficient = 0.68

Sources: Cision media database (topic series), Statistics Canada series from Table 18-10-0265-01 and Bank of Canada calculations

Last observation: June 2022

Second, we compare our monthly labour shortage time series to the CPI for services (**Chart 4**). The two indicators move closely together at certain times but diverge at others. Again, we do not expect all increases in the CPI for services to be related to labour shortages. If a labour shortage occurs in an industry whose main output is not a service, then the CPI for services is not impacted. This was the case from 2012 to 2015, when our labour shortage indicator captures the shortage of workers in the Canadian oil fields. However, service industries tend to be more labour- than capital-intensive and, given that close-contact services sectors were affected more than other sectors during the pandemic, the two indicators started moving closer together in 2022.

Chart 4: **The labour shortage indicator correlates with CPI services**



Correlation coefficient = 0.26

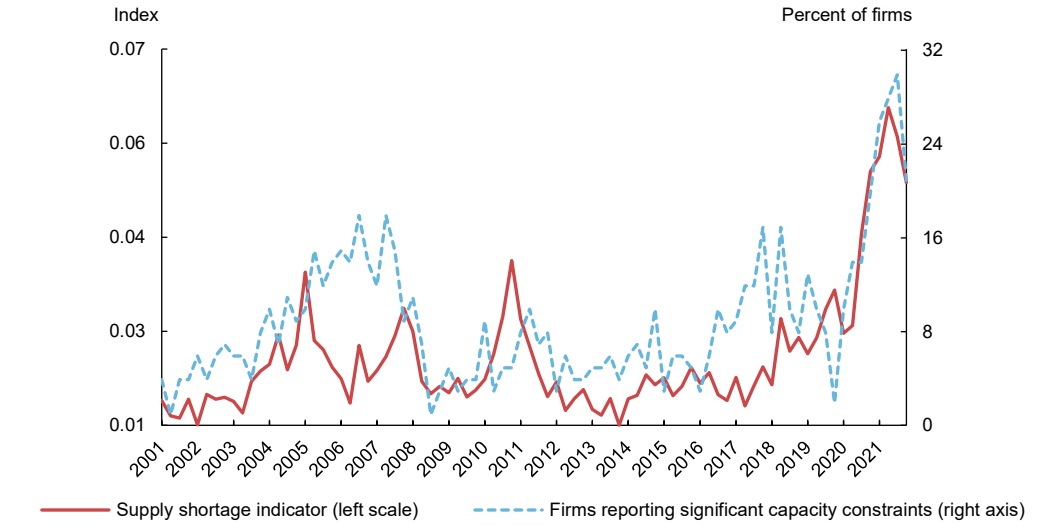
Sources: Cision media database (topic series), Statistics Canada series from Table 18-10-0004-01 and Bank of Canada calculations

Last observation: June 2022

Indicators and the Business Outlook Survey

We also evaluate the relevance of our indicators for BOS data on similar topics. We aggregate our indicators to a quarterly frequency and compare our supply shortage topic to the BOS question on the percentage of firms reporting "significant difficulty" in meeting an unexpected increase in demand (**Chart 5**). The two series have a Pearson correlation coefficient of 0.72. Hence, our indicator for supply shortages tracks well with Canadian firms' reporting of facing "significant difficulty" in meeting an unexpected increase in demand. The series matches marginally less with the more conventional measure of firms reporting "some" or "significant difficulty" in meeting demand (in **Table 2**). Hence, the news may be better at capturing critical shortages that grab more attention than milder reports of such shortages.

Chart 5: **Our supply shortage indicator correlates well with capacity constraint measures from the Business Outlook Survey**



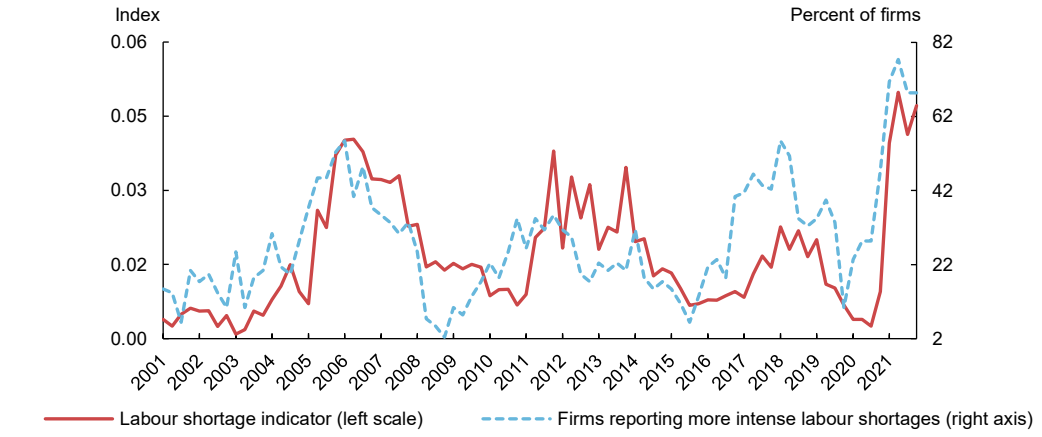
Correlation coefficient = 0.72

Sources: Cision media database (topic series), BOS series CAPACITYSIGNIF and Bank of Canada calculations

Last observation: 2022 Q2

Next, we compare the quarterly labour shortage series to the BOS question on whether firms experienced major labour shortages (**Chart 6**). Again, the series are correlated with a Pearson's coefficient of 0.63. The topic indicator appears to slightly lag the quarterly survey. It is likely that the BOS captures labour shortages, as reported by business owners, before they become mainstream in the news. As was the case with the other BOS measure, our series is slightly less correlated with the series representing the balance of opinions between firms reporting more and less intense labour shortages (in **Table 2**). Our indicator is likely better at capturing times of stress in the labour market when they are more newsworthy.

Chart 6: **Our labour shortage indicator correlates well with labour shortage measures from the Business Outlook Survey**



Correlation coefficient = 0.63

Sources: Cision media database (topic series), BOS series LSIMORE and Bank of Canada calculations

Last observation: 2022 Q2

An advantage of our indicators, relative to those in the BOS, are that they are available at a much higher frequency and with basically no publication lag. We can capture daily news on shortages, whereas the Bank interviews firms for the BOS every three months, then publishes the results at the end of the quarter. This difference in timing makes our indicators a valuable complement to the BOS. Additionally, the news indicators can pick up supply issues that originate internationally but impact Canada later on, such as the 2011 tsunami in Japan that triggered supply shortages for motor vehicle parts.

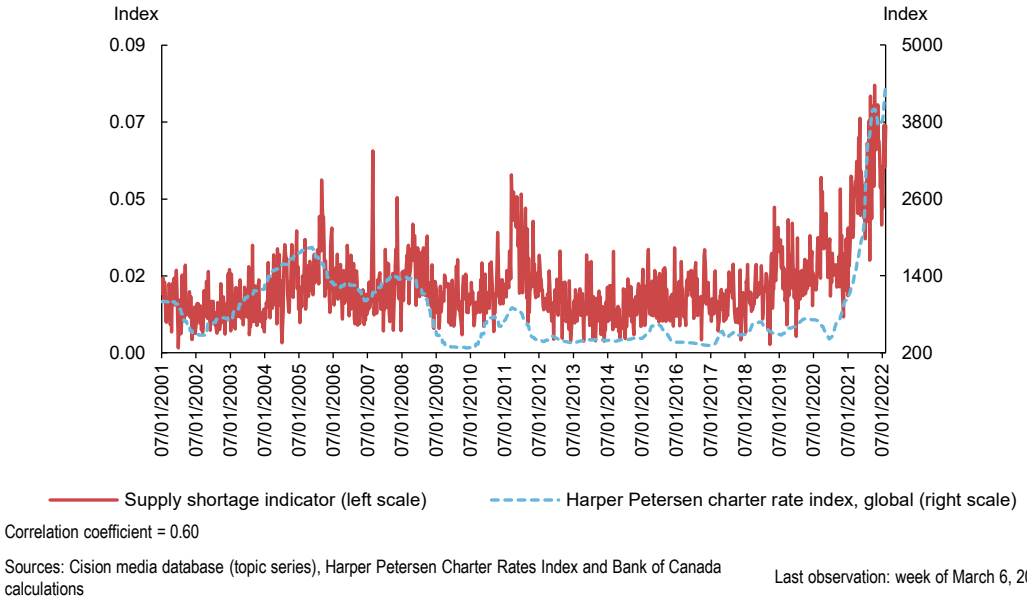
Our supply shortage indicator and other conventional indicators of supply constraints

To further assess our supply shortage indicator, we compare it with other alternative indicators of supply constraints. For example, many supply chains rely heavily on international container shipping to operate efficiently. This was strained during the global pandemic, leading to increases in freight costs coinciding with delays in shipments.

Chart 7 shows that our supply shortage topic indicator is highly correlated with the Harper Petersen Charter Rate Index, which reflects the worldwide price development on the charter market for container ships. Our indicator also has a high degree of correlation with other shipping indicators, such as the Freightos Baltic Index of shipping rates from China to North America's West Coast (correlation coefficient = 0.79) and the Canadian Federation of Independent Business' reporting of firms facing a shortage of inputs (correlation coefficient = 0.83). Below, **Table 2** summarizes these results along with those from the previous sections. Our indicator could also capture ground supply constraints within Canada, across a wide range

of supply issues, such as the car parts shortage around 2011 from the Japanese tsunami and the supply shortages starting in 2020 related to the pandemic.

Chart 7: The supply shortage topic also tracks with charter rates



Correlation coefficient = 0.60

Sources: Cision media database (topic series), Harper Petersen Charter Rates Index and Bank of Canada calculations

Last observation: week of March 6, 2022

Table 2: Summary of correlations with our supply and labour shortage indicators

Indicator	Shortage topic	Pearson's correlation coefficient (p-value)
Inflation categories		
CPI, goods (monthly)	Supply	0.57 (0.000)
CPI, services (monthly)	Labour	0.26 (0.000)
IPPI (monthly)	Supply	0.68 (0.000)
BOS categories		
Firms reporting more intense labour shortages (LSIMORE—quarterly)	Labour	0.63 (0.000)
Intensity of labour shortages—balance of opinion (LSI—quarterly)	Labour	0.54 (0.000)
Firms reporting significant capacity constraints (CAPACITYSIGNIF—quarterly)	Supply	0.72 (0.000)
Firms reporting some or significant capacity constraints (CAPACITYSIGNIF + CAPACITYSOME—quarterly)	Supply	0.67 (0.000)

Alternative tracking		
Harper Petersen Charter Rates Index, global (weekly)	Supply	0.60 (0.000)
Canadian Federation of Independent Businesses: Shortage of Inputs (monthly)	Supply	0.83 (0.000)
Freightos Baltic Index, China to North American West Coast (weekly)	Supply	0.79 (0.000)
Freightos Baltic Index, China to Europe (weekly)	Supply	0.81 (0.000)

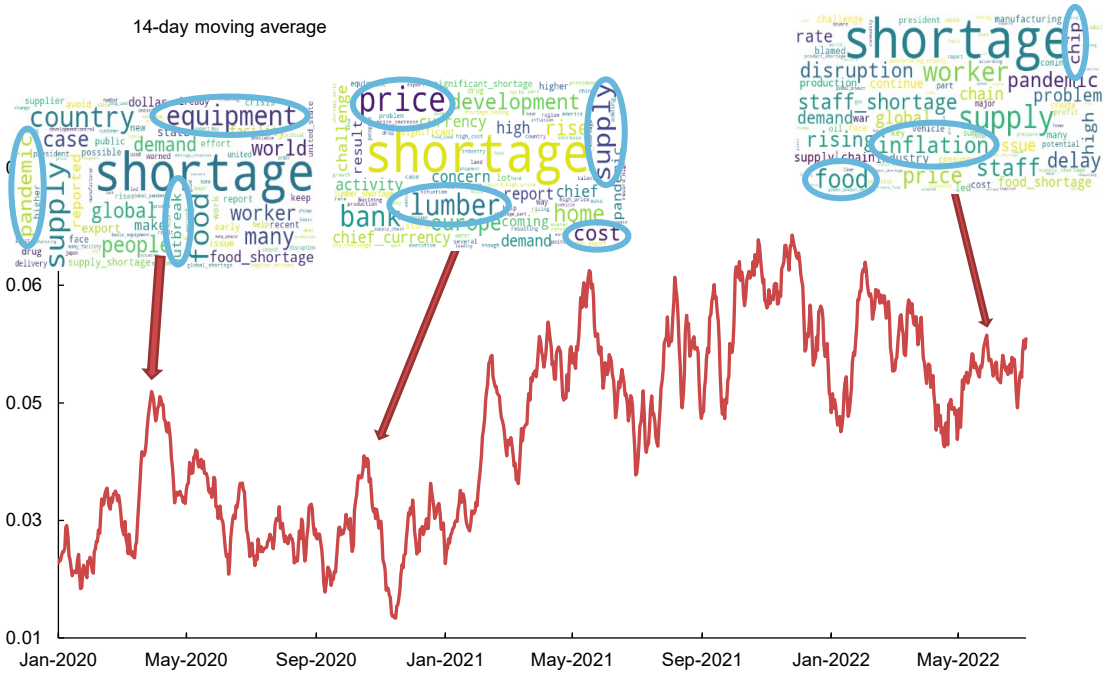
Note: CPI is the consumer price index and BOS is the Business Outlook Survey. The LISMORE, LSI, CAPACITYSIGNIF and CAPACITYSOME are series names in the BOS data.

Narratives surrounding supply and labour shortages

One of the key advantages of using daily news data to track labour and supply shortages is that we can assess how the media narratives about them start and evolve over time. By tracking the top words for each topic in a given week, we can pinpoint which sectors, jobs, industries, services or goods are impacted by shortages. The narratives give a more thorough understanding of and complement the high-frequency time series.

To illustrate the usefulness of these time-varying topic narratives, we focus on the period of elevated shortages that began in mid-2020. **Chart 8** shows our supply shortage indicator, with three selected word clouds highlighting the topic driver words for the weeks of April 6, 2020; September 28, 2020; and July 4, 2022. In these word clouds, the size of the word represents the term frequency—in other words, how frequently that word was mentioned in the news about a particular topic for that week.

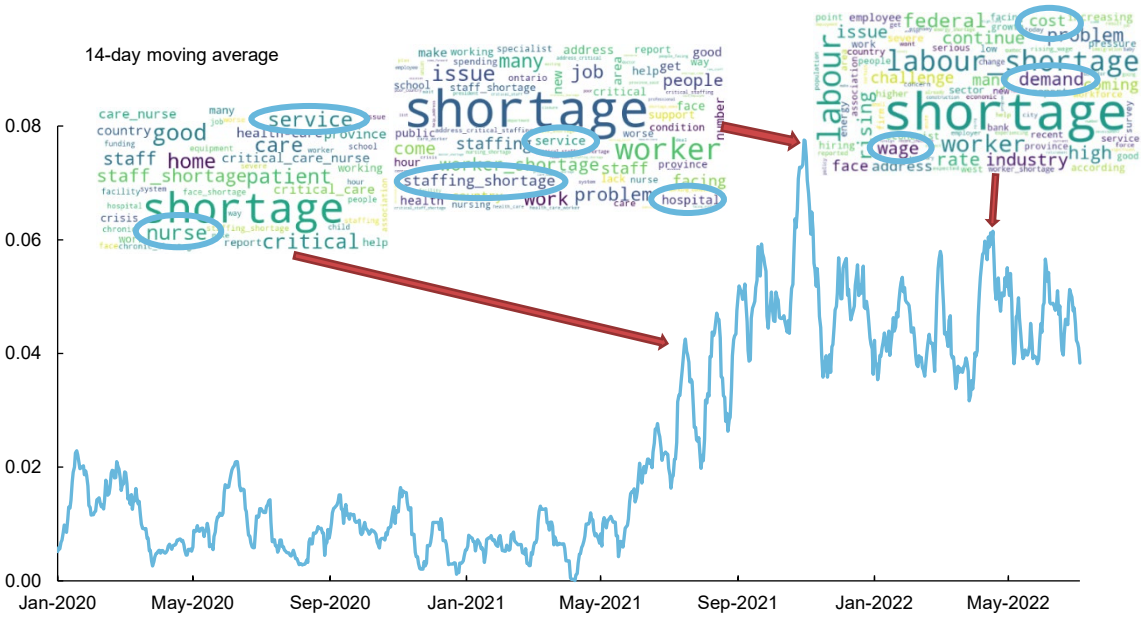
Chart 8: **Time-varying topic narratives describe the evolution of supply shortages**



At the start of the pandemic, words such as pandemic, outbreak and equipment (shortages) emerge. However, by the autumn of 2020, the narratives had shifted. The word cloud highlights lumber, pandemic, price and cost, while some of the supply chain issues from the pandemic were spreading to other sectors. By July 2022, because the supply shortage topic was still persistently appearing in news stories at elevated levels, high-frequency narratives further describe the evolution of the pressing issues, such as inflation, chip shortages and food shortages. We also see mentions of staffing shortages feeding into supply shortages at a time when job vacancies were elevated. Therefore, the topic driver words allow us to understand the focus of the media around shortages in near real time.

Our labour shortage indicator also provides useful narratives throughout the pandemic. **Chart 9** shows word clouds for the weeks of June 28, 2021; November 1, 2021; and July 4, 2022. In June 2021, some COVID-19 restrictions started to ease, and the labour shortage indicator started to rise. Therefore, we see that the word cloud for that week points to the services sector facing labour shortages, as businesses were strained to rehire workers quickly. There were also mentions of nurse shortages in a health care system already struggling from early in the pandemic. When the indicator was at its peak in autumn 2021 and health care services were facing heavy staffing shortages from the Delta wave of the virus, we see mentions of hospital, service and staffing shortage. In the first week of July 2022, the narratives focus on demand and cost, as well as wage pressures. The time-varying narratives could highlight information on how fast shortages are changing and whether they could be long-lasting or transient.

Chart 9: **Time-varying topic narratives describe the evolution of labour shortages**



Sources: Cision media database and Bank of Canada calculations

Last observation: July 4, 2022

The examples above of time-varying narratives show the versatility of using news media to track the evolutions of different aspects of the pressing economic issues. One limitation of these top driver words is that news articles do not use a standardized vocabulary. This makes splitting the top words for industries and goods categories challenging. Nonetheless, we extract valuable information from our high-frequency topic indicators.

Section 4: Conclusion

We develop our supply and labour shortage topic indicators in response to news stories about shortages, in particular, the supply shortages from the 2020–21 pandemic-related wave putting upward pressure on inflation. Our high-frequency indicators are correlated with monthly CPI and IPPI measures as well as quarterly BOS and alternative supply constraint indicators. Given their high frequency, these near-real-time news-based indicators provide early warning signs for supply and labour shortages. A unique advantage is that, besides showing static top words for each topic, we can also extract narratives by determining which words are included in the news for each topic indicator at any given time. Our workflow is adaptable and can be used to monitor other key economic issues from news media, such as inflation expectations and house prices. This shows that text data from news media contain relevant and timely information on economic drivers and development, making them a powerful policy monitoring tool for a broad range of economic activities and issues.

References

Angelico, C., J. Marcucci, M. Miccoli and F. Quarta. 2022. "Can We Measure Inflation Expectations Using Twitter?" *Journal of Econometrics* 228 (2): 259–277.

Baker, S. R., N. Bloom, S. J. Davis and S. J. Terry. 2020. "COVID-Induced Economic Uncertainty." NBER Working Paper No. 26983.

Bird, S., E. Klein and E. Loper. 2009. *Natural Language Processing with Python*. Sebastopol, California: O'Reilly Media Inc.

Bybee, L., B. T. Kelly, A. Manela and D. Xiu. 2021. "Business News and Business Cycles." NBER Working Paper No. 29344.

Church, K. W. and P. Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16 (1): 22–29.

Cichocki, A. and A. H. Phan. 2009. "Fast Local Algorithms for Large Scale Nonnegative Matrix and Tensor Factorizations." *IEICE transactions on fundamentals of electronics, communications and computer sciences* 92 (3): 708–721.

Dhillon, I. S. and S. Sra. 2005. "Generalized Nonnegative Matrix Approximations with Bregman Divergences." In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05)*. Cambridge, Massachusetts: MIT Press.

Févotte, C. and J. Idier. 2011. "Algorithms for Nonnegative Matrix Factorization with the Beta-divergence." *Neural Computation* 23 (9).

Guerrieri, V., G. Lorenzoni, L. Straub and I. Werning. 2022. "Macroeconomic Implications of COVID-19: Can Negative Supply Shocks Cause Demand Shortages?" *American Economic Review* 112 (5): 1437–1474.

Kalamara, E., A. Turrell, C. Redl, G. Kapetanios and S. Kapadia. 2020. "Making Text Count: Economic Forecasting Using Newspaper Text." Bank of England Staff Working Paper No. 865.

Kapfhammer, F., V. H. Larsen and L. A. Thorsrud. 2020. "Climate Risk and Commodity Currencies." CESifo Working Paper No. 8788.

Kuang, D., P. J. Brantingham and A. L. Bertozzi. 2017. "Crime Topic Modeling." *Crime Science* 6 (12).

Lamla, M. and S. Lein. 2014. "The Role of Media for Consumers' Inflation Expectation Formation." *Journal of Economic Behavior & Organization* 106: 62–77

Lamont, O. 1997. "Do "Shortages" Cause Inflation?" In *Reducing Inflation: Motivation and Strategy*, 281–306. Chicago: University of Chicago Press.

Larsen, V. H. and L. A. Thorsrud. 2017. "Asset Returns, News Topics, and Media Effects." Norges Bank Working Paper No. 17/17.

Larsen, V. H. and L. A. Thorsrud. 2018. "Business Cycle Narratives." CESifo Technical Report No. 7468.

Larsen, V. H. and L. A. Thorsrud. 2019. "The Value of News for Economic Developments." *Journal of Econometrics* 210 (1): 203–218.

Larsen, V. H., L. A. Thorsrud and J. Zhulanova. 2021. "News-driven Inflation Expectations and Information Rigidities." *Journal of Monetary Economics* 117 (C): 507–520.

Lee, D. D. and H. Seung. 1999. "Learning the Parts of Objects by Non-Negative Matrix Factorization." *Nature* 401 (6755): 788–791.

Li, X., and S. Wang, 2019. "Text-based crude oil price forecasting: A deep learning approach." *International Journal of Forecasting*, 35(4): 1548–1560.

Mikolov, T., K. Chen, G. Corrado and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In *International Conference on Learning Representations*, 1–12. Proceedings of workshop, Scottsdale, Arizona, May 2–4.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Pfajfar, D. and E. Santoro. 2013. "News on Inflation and the Epidemiology of Inflation Expectations." *Journal of Money, Credit and Banking* 45 (6): 1045–1067.

Pitschner, S. 2022. "Supply Chain Disruptions and Labor Shortages: COVID in Perspective." *Economics Letters* 221: 110895.

Qian, Y. 1994. "A Theory of Shortage in Socialist Economies Based on the 'Soft Budget Constraint'." *The American Economic Review* 84 (1): 145–156.

Řehůřek, R. and P. Sojka. 2010. "Software Framework for Topic Modelling with Large Corpora." International Conference on Language Resources and Evaluation, Workshop on New Challenges for NLP Frameworks, Valetta, Malta, May 22.

Shapiro, A. H. 2022. "How Much Do Supply and Demand Drive Inflation?" Federal Reserve Bank of San Francisco FRBSF Economic Letter 2022 (15): 1–6.

Shapiro, A. H., M. Sudhof and D. J. Wilson. 2022. "Measuring News Sentiment." *Journal of Econometrics* 228 (2): 221–243.

Slovic, P., D. Västfjäll, A. Erlandsson and R. Gregory. 2017. "Iconic Photographs and the Ebb and Flow of Empathic Response to Humanitarian Disasters." *PNAS* (Proceedings of the National Academy of Sciences of the United States of America) 114 (4): 640–644.

Weitzman, M. L. 1991. "Price Distortion and Shortage Deformation, or What Happened to the Soap?" *American Economic Review* 81 (3): 401–414.

Appendix 1: Simplified example from the natural language processing workflow

The following workflow illustrates the main NLP steps for this study, using a simplified example that consists of four documents, each on different days.

1) In each document, we select sentences containing the keyword "shortage" (or "shortages"). We collapse these sentences together into daily documents. Here, we present a simplified example containing one sentence in each daily document.

Dates	Document Text
2020-01-01	The global semiconductor chip shortage continues to constrain vehicle production in the auto industry.
2020-01-02	The chip shortage and inflation impacts the auto industry.
2020-01-03	Multiple sectors and industries were impacted by labour shortages.
2020-01-04	Some daycares are reducing services to deal with labour shortages in the industry.

2) a) We narrow down the list of unigrams using PMI scores. The PMI is calculated as the joint probability of each unigram co-occurring with the word "shortage" in each document, normalized by the marginal probability of each word occurring per day. Normally however, from large-scale text data with multiple documents per day, we select the n-grams with a daily PMI score greater than 1 for the unigram list. These daily PMI scores are then summed up across the days. Indeed, calculating PMI scores for hundreds of documents per day helps us greatly to reduce the dimensions of the words we analyze (Table 1). However, for our simplified example, we have only one document per day, and shortage occurs in every daily document; thus, none of the unigrams co-occur multiple times per day with shortage. This makes every unigram's PMI score equal to one each day, and the sum of these is the total count of the times that word co-occurs with shortage over the four days.

Word list (unigrams)	PMI score with shortage
industry	4
labour	2
…	…
service	1
vehicle	1

b) We break down the sentences into a matrix of documents and words consisting of unigrams, bigrams and trigrams, based on whether they occur in step 2a). Here, the elements in the matrix capture the daily frequency of each word normalized by its

occurrence across all documents (the normalized TF-IDF score mentioned in Section 3). This allows for words that are specific to some documents (like chip shortage) to have a relatively higher score than more common words (like industry).

Docs/words	chip_shortage	labour_shortage	vehicle	industry	...
2020-01-01	0.16	0	0.2	0.11	
2020-01-02	0.26	0	0	0.17	
2020-01-03	0	0.22	0	0.15	
2020-01-04	0	0.2	0	0.13	

3) We reduce the dimension by decomposing this daily matrix of documents and words into a much lower number of fewer topics. In our simplified example, we choose to narrow down to two topics. This allows us to produce two outputs, the word clouds (**Figure A-1**) and the time-series graphs (**Figure A-2**). The word clouds on the left show that the narratives for Topic 1 and Topic 2 are around chip shortages and labour shortages, respectively. The time-series graphs (right-hand side) show the topic importance over time. Because Topic 1 (supply shortage, and specifically chip shortage) was mentioned only in the first two documents, the Topic 1 time series is elevated in the first two days. Similarly, Topic 2 is elevated only in the latter two days.

Figure A-1: **Word clouds for the simplified example**

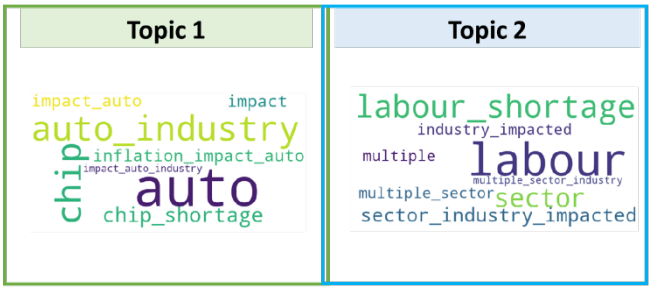
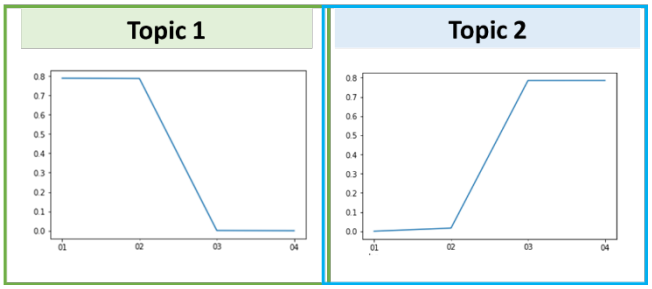


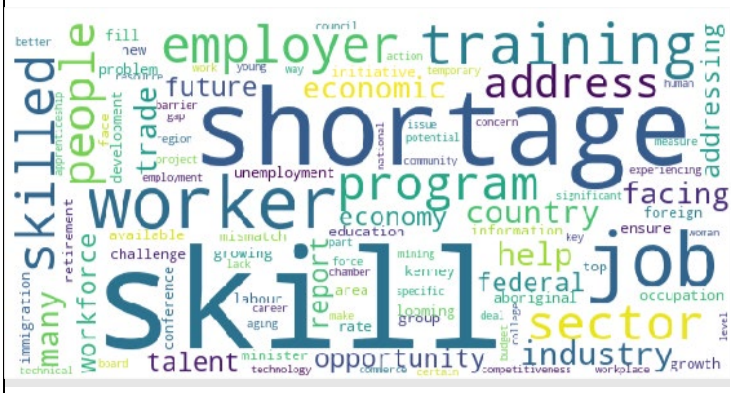
Figure A-2: **Time-series graphs for the simplified example**



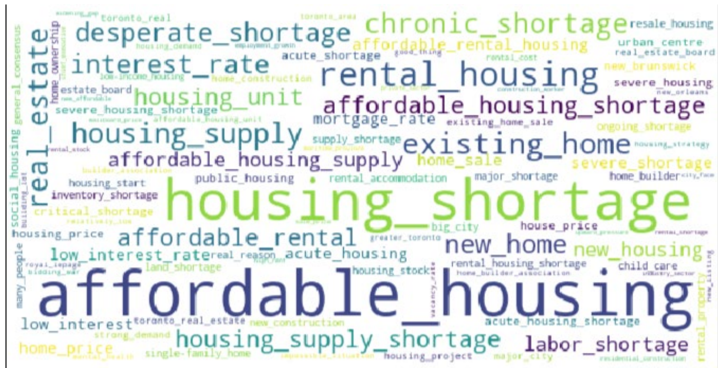
Appendix 2: Topic word clouds for other shortage topics

Figure 1 and Figure 2 track n-grams for the first two topics of the shortage corpus, namely supply and labour, respectively. The word clouds in **Figure A-3** contain the unigrams, bigrams and trigrams for the three remaining shortage topics (health care, skill and housing). The font size of each word represents its score of importance in the topic-words matrix. The skill topic might be related to the labour topic (Figure 2), in the sense that the media may cover specific skills and talents needed in labour markets. However, as the topic modelling is unsupervised, the skill topic—with its own specific sets of words—may focus on different aspects of the shortage topic than the labour topic does.

Figure A-3: **Word clouds for the other shortage topics**

Topics	Unigrams	Bigrams and trigrams
Health care shortages		
Skill shortages		
Housing shortages		

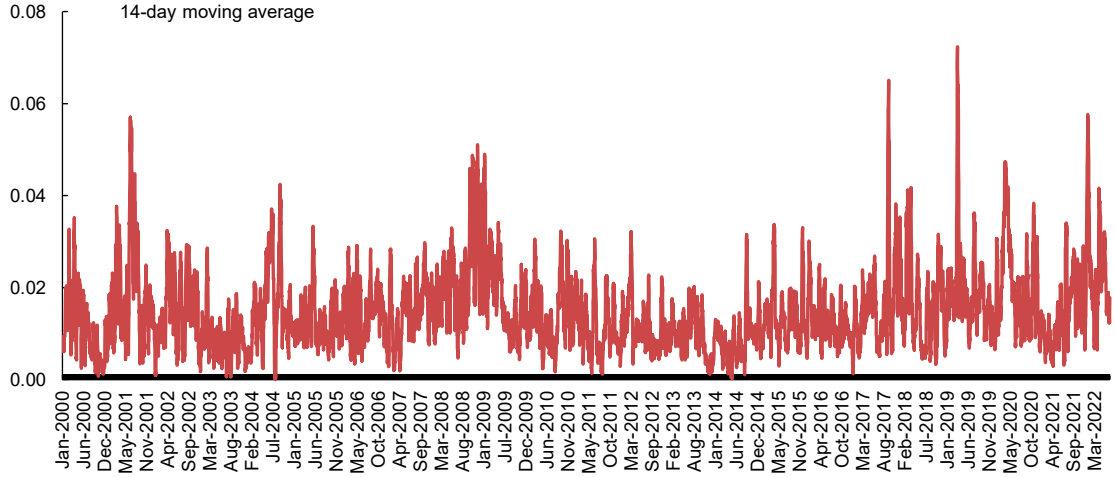

housing
affordable
shortage
rental
home
toronto
city
new
sale
land
house
existing
listing
construction
property
community
builder
buyer
income
family
development
building
unit
rate
address
people
price
federal
estate
lot
mortgage
cost
real
many
rent
help
severe
apartment
developer
strong
poverty
growth
shelter
interest
resale
urban
chronic
number
desperate
ottawa
labor
space


affordable_housing
housing_shortage
desperate_shortage
chronic_shortage
rental_housing
interest_rate
affordable_housing_shortage
housing_supply
existing_home
affordable_housing_supply
housing_unit
real_estate
affordable_rental
new_home
housing_supply_shortage
labor_shortage
low_interest_rate
acute_housing
new_housing
mortgage_rate
low_interest
home_price
affordable_rental_housing
severe_shortage
home_sale

Appendix 3: Topic time series

The following charts show the topic time series generated for the remaining three themes from the shortage news corpus: health care, skill and housing. The series for the first two topics—supply and labour shortages—are shown in **Chart 1**.

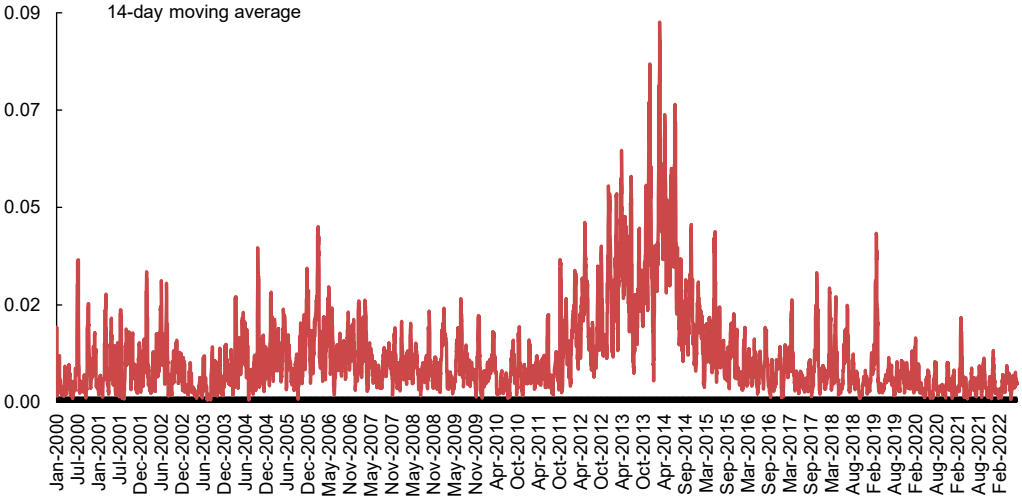
Chart A-1: **Health care shortage topic series**



Sources: Cision media database and Bank of Canada calculations

Last observation: July 4, 2022

Chart A-2: **Skill shortage topic series**

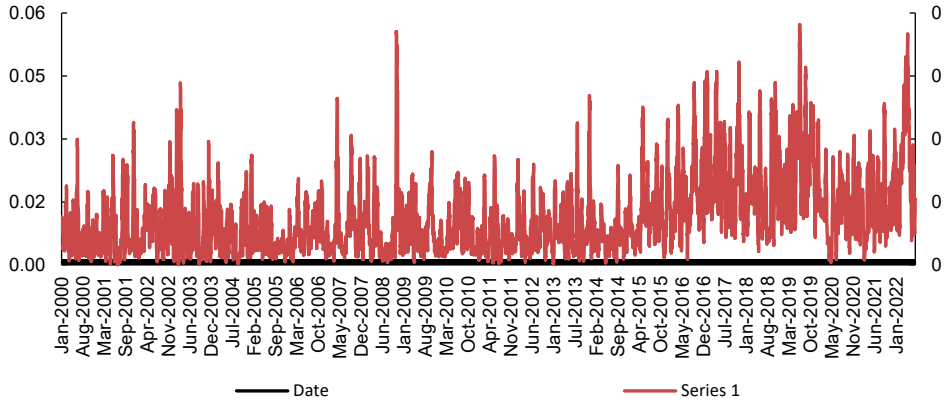


Sources: Cision media database and Bank of Canada calculations

Last observation: July 4, 2022

Chart A-3: **Housing shortage topic series**

14-day moving average



Sources: Cision media database and Bank of Canada calculations

Last observation: July 4, 2022

Appendix 4: Topic modelling illustrations

Non-negative matrix factorization (Dhillon and Sra 2006; Cichocki and Phan 2009; Févotte and Idier 2011) is a decomposition technique to reduce dimensions. It is similar to principal component analysis, except all values in NMF in the matrix are non-negative (Lee and Seung 1999). Here, we use the scikit-learn NMF module (Pedregosa et al. 2011). Below is the schematic from Kuang, Brantingham and Bertozzi (2017) to illustrate the decomposition process. Specifically, the documents-words matrix containing the n-gram TF-IDF scores (shown in matrix A, below) is decomposed into two separate matrices: topics-words, where words represent the n-grams (matrix W), and documents-topics (matrix H). Here, the high-dimensional document-word matrix is reduced to just a few topics. To identify the optimal number of topics, we use coherence-based topic number selection. We first use the Word2Vec module from Řehůřek and Sojka (2010) to build similarity scores for n-grams pairwise, based on word-embedding (Mikolov et al. 2013). Then, for each topic number, we obtain the average coherence score across all topics based on the mean similarity score of n-gram pairs. The coherence-based measure identifies the optimal topic number as 16 topics for the shortage corpus. We use this as an upper bound for the number of topics.