Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on "Are post-crisis statistical initiatives completed?"

Basel, 30-31 August 2018

Can media and text analytics provide insights into labour market conditions in China?[1]

Jeannine Bailliu, Xinfen Han, Mark Kruger,
Yu-Hsien Liu and Sri Thanabalasingam,
Bank of Canada

[1] This paper was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

Can Media and Text Analytics Provide Insights into Labour Market Conditions in China?[1]

Jeannine Bailliu, Xinfen Han[2], Mark Kruger, Yu-Hsien Liu and Sri Thanabalasingam

Abstract

The official Chinese labour market indicators have been seen as problematic, given their small cyclical movement and their only partial capture of the labour force. In our paper, we build a monthly Chinese labour market conditions index (LMCI) using text analytics applied to mainland Chinese language newspapers over the period from 2003 to 2017. We use a supervised machine learning approach by training a support vector machine classification model. The information content and the forecast ability of our LMCI are tested against official labour market activity measures in wage and credit growth estimations. Surprisingly, one of our findings is that the much maligned official labour market indicators do contain information. However, their information content is not robust and, in many cases, our LMCI can provide forecasts that are significantly superior. Moreover, regional disaggregation of the LMCI illustrates that labour conditions in the export oriented coastal region are sensitive to export growth, while those in inland regions are not. This suggests that text analytics can, indeed, be used to extract useful labour market information from Chinese newspaper articles.

Keywords: China labour market, text mining in Chinese, text classification, SVM

JEL classification: C38, C55, E24, E27

[1] The authors would like to thank Oleksiy Kryvtsov, Christopher Hajzler, Reinhard Ellwanger, Stephen Hansen and Michelle Alexopoulos for helpful comments, as well as participants at seminars at the Bank of Canada. They also thank Alexander Lam for his excellent research assistance. The views expressed in this paper are solely those of the authors.

[2] Corresponding author contact: xhan@bankofcanada.ca

Contents

1 Introduction

Assessing labour market conditions is a prerequisite for careful analysis of macroeconomic dynamics. A reliable and regularly released labour market indicator can offer insights into an economy's cyclical position, supporting macroeconomic analysis and policymaking. Moreover, such an indicator is also essential for the design of appropriate labour market policies. In this paper, we construct a labour market conditions index (LMCI) using text analytics applied to mainland Chinese-language newspapers over the period from 2003 to 2017. More specifically, we apply a supervised machine learning approach by training a support vector machine (SVM) classification model. In this context, this paper seeks to answer the following questions: Can we train a classifier to discern labour market sentiment from Chinese newspaper articles? Can a news-based index of labour market sentiment reasonably track key historical developments in China's labour market? Compared with the official data, does this index have superior information content and forecasting ability in a Phillips curve framework for wage growth and a McCallum rule framework for credit growth?

Our paper yields several interesting findings. First, the comportment of our LMCI appears to be consistent with the economic shocks that have impacted the Chinese labour market. Second, the regional disaggregation of the LMCI illustrates that labour conditions in the export-oriented coastal region are sensitive to export growth, while those in inland regions are not. Third, while each of the official labour market indicators does contain some information either for wage or for credit growth, the information in our LMCI is more consistent. Only our LMCI provides significant information in the two wage and the credit estimations. Moreover, our LMCI provides wage and credit forecasts that are better than those from any single official labour market indicator. These results suggest that the text analytics can be used to extract useful labour market information from Chinese newspapers. This paper contributes to the literature in three ways. First, we create a novel dataset and use text analytics to develop a monthly LMCI for China that can be updated in real time. Second, we build on the methodology of Tobback et al. (2018) by applying a supervised machine learning technique to Chinese-language documents and by using a two-stage approach in training our SVM classification model. Third, we find that our LMCI is more robust in explaining and forecasting both wage and credit growth than any single official labour market indicator.

Our paper is structured as follows. A brief literature review is contained in Section 2. Section 3 presents our dataset and describes our methodology. In Section 4, we compare the information content of our LMCI to that of other labour market activity indicators in explaining and predicting wage growth in a Phillips curve framework. Section 5 compares the ability of our LMCI to explain and forecast credit growth in a McCallum rule framework against that of the official labour market indicators. Section 6 offers some concluding remarks.

2 Literature review

Our paper relates to two strands of literature. First, it focuses on developing alternative measures of the Chinese unemployment rate to address problems associated with the official data. Second, it adds to the growing literature on developing text-based indicators that can be useful proxies for economic and policy conditions. Most text-based indices are based on predefined keyword searches (for example, see Alexopoulos and Cohen (2009) and Baker et al. (2016)). Our paper builds on the methodology of Tobback et al. (2018), who use a supervised machine learning technique to develop an economic policy uncertainty index for Belgium. As Tobback et al. (2018) point out; their methodology is an improvement over simple keyword searches, as it avoids the human bias inherent in the keyword selection process.

Official labour statistics do not seem to reflect the actual employment situation in China. Indeed, Cai et al. (2013) propose comprehensively reforming the statistical system to improve the current set of labour market indicators so as to have better data to inform policy-making. It is generally agreed that the official unemployment rate underestimates the level of unemployment in China (Wang and Sun (2014)). Moreover, the official rate has remained fairly stable over time and does not appear to capture key historical labour market developments. For example, it did not increase by much during the period of state-owned enterprise (SOE) reform (1996-2002) despite the massive layoffs triggered by the reform. Moreover, it did not fluctuate appreciably during the 2008-2009 global financial crisis in spite of the significant employment loss over that period. Indeed, Lam et al. (2015) note that, compared with the unemployment rate in other major countries, the official unemployment rate has displayed considerably less sensitivity to changes in output.

Although issues have been raised with respect to many of China's official statistics, those pertaining to the labour market are seen as particularly problematic. Indeed, *The Economist* (2008) noted that "the prize for the dodgiest figures goes to the labour market." There are three sets of official Chinese labour market indicators of relatively high frequency. The first is the urban registered unemployment rate, which is published on a quarterly basis by the Chinese Ministry of Human Resources and Social Services (MOHRSS). The second is the urban demand-supply ratio, which is also published on a quarterly basis by MOHRSS. The third is the employment sub-component of the manufacturing and non-manufacturing Purchasing Managers Indices (PMIs), which are published on a monthly basis by China's National Bureau of Statistics. The main problem with the official statistics is that while they capture formal employment, they do not appear to include migrant workers, who are typically engaged on an informal basis.[3] The omission of migrant workers in Chinese labour statistics is problematic because they represent a large share of the labour force. Wang and Wan (2014) estimate that there were over 100 million migrant workers in China in 2010, which represented about 25% of urban employment.

The issues with the official unemployment rate have been acknowledged in the literature, and several papers have developed alternative measures of the Chinese unemployment rate that more closely follow international guidelines. Alternative unemployment rates, such as that estimated by Feng et al. (2017), are more variant and do capture these key developments. A summary of the key studies in this literature is presented in Table 1. Although these papers report a range of estimates for any given year, they all suggest that the actual unemployment rate was higher than the official rate. While these alternative indicators are very useful in identifying the problems associated with the official unemployment rate in China, they are of limited use for policy and analysis, given that they are not updated and thus unavailable on a high-frequency and timely basis. Our paper aims to bridge this gap in the literature by developing an LMCI for China that can be updated in real time.

[3] Migrant workers in China are those workers who do not have a *hukou* – an urban residence permit – for the jurisdiction in which they work.

3 Methodology

The methodology that we use in this paper builds on that of Tobback et al. (2018), who employ a supervised machine learning technique to develop an economic policy uncertainty index for Belgium. They do so by training a classifier using an SVM algorithm to predict whether an article addresses economic policy uncertainty. Our methodology differs from theirs in two important ways. First, we use Chinese-language documents, which present some challenges. Notably, text analytics involves finding relevant words, and what constitutes a "word" in Chinese is not obvious by simply looking at a selection of text. Therefore, we need to go through an additional step of "segmenting" Chinese characters into words. Second, after training their SVM classifier, Tobback et al. (2018) use a single-stage methodology to identify articles that are relevant for economic policy uncertainty. After considering alternative specifications, we use a two-stage approach. In the first stage, we train our SVM classifier to find articles that are relevant to the state of the Chinese labour market. In the second stage, we train the classifier to distinguish between articles that represent positive labour market sentiment and those that evoke negative sentiment.

While a number of methodologies exist for classifying text, we selected the SVM methodology, as the literature suggests that it is superior for classifying Chinese-language documents (Tan and Zhang (2008)).

Our methodology consists of the following steps:

1. Preselecting the articles;
2. Constructing the training and testing subset;
3. Preprocessing the articles for machine learning;
4. Transforming the text into a numerical matrix;
5. Training the classifier; and
6. Constructing the LMCI.

3.1 Preselecting the articles

In this paper, we create a novel dataset drawing on Chinese-language newspapers from mainland China provided by Wisers, a Hong Kong-based company. Wisers is the world's largest database of Chinese newspapers, beginning in late 1999 and consisting of 428 Chinese-language newspapers from mainland China. The newspapers cover both regional and national news.[4] We focus on a subset of 90 Chinese

[4] More information on Wisers can be found at www.wisers.com.

newspapers, which were continuously published over the period from January 2003 to June 2017 (this list of 90 newspapers is provided in Appendix A). As shown in Figure 1, this set of newspapers provides broad geographical coverage of China.

Millions of articles were published in the 90 Chinese newspapers between 2003 and 2017. They contain a mixture of news: central government policies, local companies' news, major events in the country, human interest stories, etc. To make our dataset more manageable, we preselected a subset of articles based on keywords most relevant for news about the labour market. We drew on Antenucci et al. (2014) and translated the keywords into Chinese (see Table 2 for a list of the keywords selected). Our keyword search of 90 newspapers for the period January 2003 to June 2017 resulted in more than eight million articles, or over 1600 articles per day. We then randomly selected one day for each month between January 2003 and June 2017 and downloaded all the articles published on that day that contained any of our keywords. This process resulted in a set of a little over 266,000 potentially relevant newspaper articles in our dataset.[5]

3.2 Constructing the training and testing subset

In machine learning, the classification problem, which maps input data into given categories, is one of the typical problems solved by supervised machine learning algorithms. The algorithms that solve the classification problems are called "classifiers." The supervised machine learning uses a training dataset and seeks the best algorithms that predict well out of sample (Mullainathan and Spiess (2017)).

To create a subset of articles to be used for both training and testing the classifier, we randomly selected just under 800 articles from the 266,000 in our dataset. We read all of them and then classified them using a two-step process. First, the articles were divided into two groups: (i) those that clearly contained either positive or negative sentiment with respect to the state of the Chinese labour market (relevant articles) and (ii) those that did not (irrelevant articles). In the second step, the first group was further subdivided into articles that contained positive sentiment and those that contained negative sentiment (the grouping of the articles in the training/testing subset is depicted in Table 3).

To frame our reading of the articles, we agreed upon general rules to help each author to classify articles in our training/testing set. Articles were labelled as relevant (positive/negative) if they met the following principles:

- Directly reported instances of companies hiring (positive) or cutting jobs (negative); individuals finding (positive) or losing jobs (negative).
- Indirectly reported labour market conditions. For example, national or regional policies have been implemented to help people to find jobs, which indicated the underlying labour market conditions were poor (negative).

[5] In order to test the robustness of our sampling method, we randomly selected a different day from each month to construct our second sample set of articles. Using this sample set, we constructed LMCI using the same methodology outlined in the paper. The resulting index is similar to the index produced from the first sample set of articles.

To ensure that the manual selection process was robust, several of the authors read and independently scored each article for both relevance and sentiment. This was followed by a discussion of those articles upon which there was a disagreement until consensus was reached as to whether an article expressed sentiment and, if so, the nature of the sentiment. It is difficult to associate the positive or negative sentiment contained in the articles with specific economic indicators: employment, labour force participation, hours worked or wage growth. We assigned articles positive or negative sentiment based on our sense of whether the article contained "positive news" or "negative news" about general labour market conditions. In this way, we see labour market sentiment as akin to consumer confidence, which could rise because of increasing employment or rising wages or the expectation of better economic times ahead. While consumer sentiment is based on survey data, our labour market sentiment is a function of whether newspaper articles report positive or negative news.

3.3 Preprocessing the articles for machine learning

The goal of preprocessing is to break the Chinese text into small, meaningful units. In English, these units are typically words, and unique words are easy to identify in a document, since they are separated from other words by spaces. In contrast, Chinese text has no spaces between characters and a character, on its own, may not form a meaningful unit. Indeed, a large proportion of Chinese words are made up of two or more characters. Since the occurrence of a Chinese word in a document is not indicated by any sort of punctuation, the meaning of a sentence is potentially ambiguous.

To illustrate how the meaning of a sentence depends on how Chinese characters are segmented into words, consider the following example. The Chinese sentence below can have different meanings depending on how the characters are segmented.

Sentence: 乒乓球拍卖完了

Segmentation 1: 乒乓球拍/ 卖/ 完/ 了 (pingpangqiupai mai wan le)

Meaning: The ping pong paddles are sold out.

Segmentation 2: 乒乓球/ 拍卖/ 完/ 了 (pingpangqiu paimai wan le)

Meaning: The ping pong ball auction is over.

In order to sort Chinese characters into words, we relied on natural language processing software called Harbin LTP (Che et al. (2010)). We tested several software packages designed for the Chinese language and found that Harbin LTP outperformed the others in terms of word segmentation accuracy and speed. We also removed "stop words" from each article at this stage. Stop words are those that are important from a grammatical perspective but do not contain independent meaning. We did this with the assistance of the *Word List with Accumulated Word Frequency Sinica Corpus 3.0*. We eliminated 63 additional words, mostly adverbs, that we felt were not independently meaningful.

3.4 Transforming the text into a numerical matrix

The next step involves transforming the text from the articles into a numerical matrix.[6] Each article can be represented as a "bag-of-words" vector $[t_1, t_2, \ldots, t_j, \ldots\ t_m]$ that contains all m unique words that are present in the training set, where t indicates how often the jth word appears in the article.

The bag-of-words vector is then used to construct the term-frequency matrix $tf\,(n, m)$, where n is the number of articles and m is the number of unique words in the training set. The term-frequency matrix essentially presents the distribution of unique words across all the articles. To diminish the weight of words that occur frequently and increase the weight of those that appear rarely, the term-frequency matrix is multiplied by the inverse document frequency (*idf)* to obtain *tfidf* matrix. The inverse document frequency measures the importance of a word in all articles in the training set and is calculated as follows:

$$idf = log\frac{Number\ of\ articles\ n\ in\ the\ training\ set}{N\,umber\ of\ articles\ in\ training\ set\ in\ which\ term\ j\ occurs} \qquad (1)$$

The re-weight of *tf* by *idf* is to diminish the importance of words that occur very frequently in the articles but that carry little meaning. It increases the importance of words that appear rarely but contain a lot of meaning. It is these words that, potentially, give the classifier more power to discriminate between different categories of articles.

Given that there are over 3000 unique words in our training set, we applied a χ^2 feature selection method to avoid model over fit. We conducted this feature selection to select the 125 most important words to train our Stage I classifier and the 200 most important words to train our Stage II classifier – the *tfidf* matrix was thus transformed into an n×125 matrix for the Stage I classifier and into an n × 200 matrix for the Stage II classifier.

3.5 Training the classifier

Having constructed the *tfidf* matrix, we can now use it as an input into the SVM algorithm. To solve any classification problem, the SVM searches for the decision boundary that maximizes the margin between the two classes. The classification problem is illustrated in Figure 2. The SVM selects two parallel hyperplanes to separate the two categories of data, so that the distance between the two hyperplanes (dashed lines) is maximized. The distance between the two hyperplanes is called the margin, and the decision boundary is the hyperplane in the middle (solid line). The circles and crosses that lie on the dashed lines are support vectors; these are the data points that are most difficult to classify. The intuition for the SVM algorithm is that if the classifier is able to separate the data points closest to the margin, it will be relatively easy to classify the data points that lie farther away.

As described in Fan et al. (2008), the linear SVM tries to solve the following optimization problem:

[6] To implement our methodology, we draw on scikit-learn, an open source machine learning library for the Python programming language (Pedregosa et al. (2011)).

$$min\frac{1}{2}\omega^T\omega + C\sum_{i=0}^{n} max(1 - y_i\omega^T x_i, 0)^2 \qquad (2)$$

Where x_i is defined as the vector of *ith* article in the training set of n articles, ω is the weight vector *[*ω_1,ω_2, ...,ω_j, ...,ω_m*]* of unique words in the training set, and y_i represents the manual classification labels (i.e., 1 or -1) for *ith* article in the training set. The term $C\sum_{i=0}^{n} max(1 - y_i\omega^T x_i, 0)^2$ is added to cover the cases that the SVM is not able to perfectly clearly classify – i.e., this term is intended to penalize classifications that fall within the margin. We performed an in-sample grid search with cross-validation to find the optimal value of *C*, the cost parameter. It controls the trade-off between limiting misclassifications and maximizing the margin. If *C* is too large, the chance of misclassification is small, but the margin will be too narrow and the chance of over-fitting is great. If *C* is too small, the margin will be too big and there will be too many misclassifications. The best value for *C* optimizes this trade-off. In our case, we trained the SVM to find an optimal linear function for each stage. The linear function (classifier) is in the following form:

$$f(x_i) = \omega_0 + \omega_1 x_{i1} + \omega_2 x_{i2} + \cdots + \omega_j x_{ij} + \omega_m x_{im} \qquad (3)$$

Where x_i represents the vector of ith article in the training set, and x_{ij} represents the *tfidf* value of jth unique term in ith article in our training set. ω_j represents the weight of jth unique term in the training set. For each article x_i, if $f(x_i) > 0$ then the article will be classified into the category labelled 1, and if $f(x_i) < 0$ then it will be grouped into the category labelled -1.

Recall that we run the SVM twice: once to find articles that are relevant (Stage I) and a second time to differentiate between positive and negative news (Stage II). In the Stage I classification, we trained the SVM to correctly identify articles that we had manually classified as either being relevant to the state of the Chinese labour market or being irrelevant. Of our training/testing sample, 80% of the articles were used as a training set and 20% as a test set (to be used to measure the out-of-sample performance of the classification model). We used an 80/20 split between the training and testing sets to ensure that we had sufficient data to conduct the ten-fold cross-validation.[7]

Following Sokolova and Lapalme (2009), we evaluated the performance of the SVM according to three metrics:

$$\textit{Accuracy:} \frac{TP + TN}{TP + FP + TN + FN}$$

$$\textit{Specificity:} \frac{TN}{FP + TN}$$

$$\textit{Sensitivity:} \frac{TP}{TP+FN}$$

where:

[7] As an example, a three-fold cross-validation would involve splitting the training set into three equal-sized subsets. The classifier would then be trained using two of the subsets, and validated using the remaining subset. The cross-validation process would then be repeated three times (the folds), with each of the three subsets used once as the validation data.

TP = True positives

TN = True negatives

FP = False positives

FN = False negatives.

The results of our Stage I testing are shown in Table 4. The SVM could achieve a high accuracy rate with 85% of the articles classified correctly. The proportion of irrelevant articles correctly identified by the classifier (i.e., specificity) was also elevated at 89%, as was the share of relevant articles properly categorized by the classifier (i.e., sensitivity) at 82%. We followed the same procedure for the Stage II classification, where we trained the SVM to separate the articles identified as relevant in Stage I into those representing positive labour market sentiment and those reflecting negative sentiment. We split the 313 articles in this sample using the same 80/20 split between the training and testing sets; this yielded 250 articles in the training set and 63 in the testing set. As shown in Table 4, reported values for the metrics suggest that the classifier performed well in the Stage II classification as well.

In an ideal case, we would like the classifier to be able to distinguish between articles as accurately as human reading does. However, there is no established standard to define the acceptable classification error rate for the Accuracy, Specificity and Sensitivity metrics. Moreover, there are trade-offs between the metrics such that improving Specificity could reduce Sensitivity. This is akin to the trade-off between Type I and Type II errors. Our reading of the literature is that an acceptable error rate for the reported metrics is project-specific and subject to the discretion of the authors.

To further assess the accuracy of our two-stage methodology, we developed two alterative classifications. We trained a one-stage classifier that sought to divide the articles into three categories: positive labour market articles, negative labour market articles, and neutral labour market ones (Method 1).[8] And we trained a two-stage classifier, in which the Stage I classifier divided all the articles into relevant and irrelevant categories and the Stage II classifier divided the relevant articles into positive sentiment, negative sentiment, or sentiment neutral categories (Method 2). The performance of our preferred methodology (Method 3) against the two alternative methods is presented in Table 5. Our preferred methodology appears to be superior. Method 1 does well at identifying labour market-neutral articles, but not at identifying labour market relevant ones. Method 2 has the same Stage I classifier as our methodology, which performs reasonably well. However, its Stage II classifier does not perform as well.

There are clear differences between the articles that express negative sentiment and those that express positive sentiment. Figures 3 and 4 show the frequency of the top 30 words that only appear in either the negative or in the positive sentiment news articles. The terms "laid-off (下岗)", "unemployment (失业)", "unemployed (失业人员)", "laid-off/unemployed (下岗失业)", "laid-off workers (下岗职工)", and "laid-off staff (下岗失业人员)" feature prominently in the negative sentiment news stories.

[8] SVM is a binary classification algorithm. A three-category classification can be done as the result of three binary classifications undertaken in one stage. See scikit-learn's documentation (Pedregosa et al. (2011)).

3.6 Constructing the LMCI

We construct a composite LMCI that is intended to capture the relative frequency of positive sentiment articles to negative sentiment articles. As a first step, we create two sub-indices: one capturing positive labour market sentiment and one indicating negative sentiment. To do so, we follow the approach used by Baker et al. (2016) to create an economic policy uncertainty index. This procedure involves first creating a monthly series of the number of positive (negative) sentiment news articles in each newspaper. The raw counts are then scaled by the total number of articles in the same newspaper and month. The resulting series for each newspaper is then standardized to unit standard deviation and summed across all the papers by month. Finally, each multi-paper index is normalized to a mean of 100.

Once the two sub-indices are created, we construct our LMCI by dividing the positive sentiment index by the negative sentiment index. We then demean the series so that the index has a mean of zero. A value above (below) zero indicates that on net, Chinese labour market sentiment is positive (negative). The standard deviation of the series is 0.1. Our LMCI is depicted in Figure 5.

4 Evaluation: Can our Chinese LMCI explain and predict wage growth?

Before we compare the usefulness of our LCMI against that of official labour market indicators, we first check to see if its evolution is consistent with changing labour market conditions.

4.1 Does our LMCI capture the likely impact of key shocks on the Chinese labour market?

Our LMCI appears to capture the likely impact of key shocks on the Chinese labour market (Figure 5). The index dipped below zero during the 2003-2004 periods, suggesting that labour market sentiment was on net negative, likely reflecting two key events. First, the massive layoffs triggered by the SOE reforms (1996-2002) were probably still impacting the Chinese labour market in 2003-2004. It is estimated that about 45 million workers were laid off during the SOE reform period (Giles et al. (2005)) and it would have taken the labour market some time to recover from this large shock. In addition, the SARS outbreak, which started in late 2002 and went on for most of 2003, also had a negative effect on the Chinese labour market, particularly for workers in the services industry. Many migrant workers returned to their home villages during the SARS epidemic, some having been permanently laid off as their employers faced financial difficulties because of the SARS outbreak.

In the mid-2000s, the index was well above zero for several years, suggesting that labour market sentiment improved significantly and was on net positive. China's accession to the WTO (in December 2001) resulted in a rapid development in the export-oriented sector and a considerable increase in industrial employment.[9] This period of positive labour market sentiment was interrupted in 2007. This could be related to the new Labour Contract Law (LCL), which was enacted in June 2007 and became effective in January 2008. The purpose of this new law was to improve the Chinese labour contract system, clarify the rights and obligations

[9] Between 2000 and 2005, industrial employment in China increased by over 30% from 45 to 60 million workers (International Trade Organization (2011)).

of the parties, protect employees' lawful interests and strengthen labour relations (Chen and Funke (2009)). A key outcome of this law was to enforce written labour contracts. Anecdotal evidence suggests that some employers laid off informal workers in the period after the new LCL was announced but before it was implemented (i.e., in the second half of 2007) given that it became more difficult to lay off workers after the LCL became effective.

The decline in the index in 2007 was followed by a marked increase in 2008 reflecting employment gains that may be associated with the reconstruction after the Sichuan earthquake and the 2008 Olympics. In May 2008, a devastating earthquake of magnitude eight hit Sichuan province, killing more than 80,000 people and leaving more than 15 million homeless. The Chinese government responded rapidly with a reconstruction plan that involved building about 6.6 million houses, 3000 schools and 1100 medical facilities over a three-year period. Given the scope of the reconstruction effort, workers were drawn from both inside and outside the earthquake-affected area. Preparation for the 2008 Beijing Olympics is also associated with an increase in employment, particularly in the construction and transportation sectors (Wang and Zhang (2013)).

The index dipped precipitously during the global financial crisis (GFC) because of the significant employment loss over that period. It is estimated that around 23 million workers lost their jobs in China during the global financial crisis, as thousands of factories in the coastal region were closed when the number of orders filled by many export-oriented firms declined sharply (Cai and Chan (2009)). Most of these workers were migrants. In response to the GFC and to minimize its impact on the Chinese economy, the government announced a very large economic stimulus package (the headline number was US$586 billion or over 13% of GDP) in late 2008 to be invested in infrastructure and social welfare. Infrastructure-related employment helped mitigate the impact of job losses in the export-oriented sector. The recovery in the labour market is reflected in the evolution of the LMCI, which was back up to zero by mid-2010.

From 2011 onwards, the LMCI suggests that sentiment in the Chinese labour market tended to be on net positive. Several factors may account for this positive sentiment, including the solid performance of the Chinese economy over this period and the shrinking of the working-age population (starting in 2012). The index does dip below zero in late 2013 and into 2014, corresponding with the period when employment was reduced in overcapacity sectors (notably coal and steel). Although employment in the six main overcapacity sectors accounts for a small share of total employment (about 4% of total non-farming employment), companies in these sectors tend to be concentrated in certain regions/cities and hence job losses have been significant in some localities.[10] Another factor that may explain the decline in the index in 2013/2014 is the lack of job opportunities for university graduates – the Chinese media reported that 2013 was a very challenging year for university graduates seeking employment.

While the LMCI and official labour market indicators seem to follow similar trends as seen in Figures 6, 7 and 8, the LMCI captures events not observed in other indicators. For example, from around 2010 to 2013, the LMCI reflects strong growth in the Chinese economy, while the official unemployment rate displays no variation (Figure 6). Furthermore, the cuts to overcapacity sectors are not captured across the official indicators, whereas the LMCI falls sharply during this event.

[10] The Chinese government has highlighted the following six sectors as excess capacity industries: steel and iron, coal, cement, aluminium, ship building, and flat glass.

4.2 Regional LMCIs and export growth

Since we collect newspaper articles from a range of Chinese cities, we can undertake analysis to better understand how regional labour market conditions may vary. In particular, we would expect labour market conditions in the coastal region, which is more export-oriented than inland provinces, to be more sensitive to shocks emanating from abroad.

To test this, we construct two LMCI sub-indices: one for the coastal provinces (including Beijing, Tianjin, Hebei, Shanghai, Jiangsu, Zhejiang, Fujian, Shandong, Guangdong, and Hainan) and a second for the remaining inland provinces.[11] We then run a series of regressions in which the LMCI sub-indices are regressed on their first lag, a constant and export growth. The results are presented in Table 6. We find that exports are a predictor of labour market conditions in the coastal region (and for the country as a whole) but not for the inland region. This result not only sheds light on the difference between the coastal and inland labour markets, it also reinforces previous findings that the LMCI is, indeed, sensitive to actual labour market developments.

4.3 Wage movements

Ultimately, the usefulness of our LMCI will depend on the extent to which it allows us to capture direct measures of labour market outcomes. Moreover, the LMCI's value added needs to be assessed vis-a-vis the ability of the official measures of labour market activity discussed above to capture these outcomes.

A natural starting point for an indicator of labour market activity is its ability to predict wage movements. Thus, we estimate a set of Phillips curve-type equations in which wage movements are regressed on controls and various labour market indicators, including our LMCI. Our equations are of the following form:

$$w_t = \beta_0 + \beta_1 w_{t-1} + \beta_2 \pi^e + \beta_3 labour\ indicator_{t-1} + \varepsilon_t, \qquad (4)$$

where w is wage growth (in year-over-year terms), π^e is CPI inflation expectations, $labour\ indicator$ represents one of the labour market indicators that we consider and ε is an error term.

Given the lack of availability of a quarterly Chinese wage series over our sample period, we use disposable income as a proxy for wages.[12] The inflation expectations measure we use is the quarterly diffusion index created by the People's Bank of China (PBOC), which compiles survey responses on the direction of the price level in the next quarter. We convert our LMCI from monthly to quarterly frequency by taking a five-month centred average of the index and then averaging the monthly index values in each quarter. To assess the relative information content of our LMCI, we compare it against the following labour market activity measures: the urban labour demand-supply ratio, the official unemployment rate and the employment sub-components of the PMIs. Figures 6, 7 and 8 graph the official labour market indicators against our LMCI. Note that there was a trend in the urban supply-

[11] We constructed two regional LMCI sub-indices because of the limited number of articles available for each province in our dataset.

[12] We use urban disposable income as a proxy for wages. Wages and salaries data only begin in 2013Q1. In contrast, urban disposable income data begin in 2001Q1. Wage and salaries made up around 60% ofurban disposable income annually since 2013, making for a reasonable proxy.

demand ratio, which we removed by taking the four-quarter change. More details on variable construction and data sources are provided in Appendix B.

The results are presented in Table 7.[13] The results show that over the full sample period, our LMCI is the only labour market activity indicator that has explanatory information for wage growth.[14] It is the case that the employment sub-indices of the manufacturing and non-manufacturing PMIs also contain information relevant for wage growth. Note that these indicators are only available over a shorter time frame.

We next proceed to forecast wage growth and compare the forecasts using the LMCI against those using the official labour market activity measures. The forecasts were constructed as rolling four-quarter-ahead forecasts beginning in 2008Q2 and running to 2017Q1. The forecast results, which show the ratio of the root-mean-squared error (RMSE) of forecasts that use the other labour market activity variables to those using our LMCI, are presented in Table 9 (a number less than one indicates that our LMCI provides superior forecasts). The RMSEs of the forecasts are quite close. Indeed, only the four-quarter-ahead forecast using the employment sub-index of the non-manufacturing PMI is statistically superior to the others, as per the Diebold-Mariano test.

There appears to be a downward trend in wage growth beginning in 2007Q3 and continuing to the end of our sample (Figure 9). This downward trend suggests that there are longer-term structural changes affecting wage growth in addition to short-term cyclical pressures. In view of this, we estimate a second set of regressions in which the dependent variable and the lagged dependent variable are the deviations of wage growth from trend. All other variables are as defined in Equation 4. The results of these estimations are presented in Table 8. In this set up, only our LMCI and the official unemployment rate contain statistically significant information.[15] It is worth noting that the employment indices of the PMI are no longer significant in this formulation.

Once again, we undertake a forecast comparison exercise by conducting rolling four-quarter forecasts beginning in 2008Q2 and going to 2017Q1 and comparing the RMSE ratios. The results are presented in Table 10. They show that forecasts conducted with our LMCI are superior to those using other labour market activity indicators. Moreover, in about half the cases, the gain in accuracy from using our LMCI is statistically significant as per the Diebold-Mariano test.

5 Evaluation: Can our LMCI explain and predict credit growth?

In this section, we evaluate the usefulness of our labour market index by testing if it can help explain credit growth in the context of a McCallum-type monetary policy rule better than the official labour market indicators can.

[13] Unit root tests conducted using the Augmented Dickey-Fuller test suggested that all the series in equation (4) are stationary, as assumed.

[14] Our results were robust to the removal of the observations associated with the GFC (i.e., 2008Q3- 2009Q3).

[15] Here too, our results were robust to the removal of the observations associated with the GFC (i.e., 2008Q3-2009Q3).

5.1 Estimating and forecasting a McCallum-type rule

Understanding Chinese monetary policy, and attempting to represent its conduct using a monetary policy rule, is a challenging endeavour because the PBOC has many monetary policy instruments and multiple objectives. Over our sample period, the PBOC has used the following instruments to conduct monetary policy: reserve requirement ratios, benchmark interest rates, open market operations, targeted lending facilities and window guidance. Moreover, the importance accorded to individual instruments has changed over time, further complicating the task of representing the conduct of Chinese monetary policy with a rule. For instance, the PBOC relied heavily on reserve requirement ratios as a monetary policy instrument a few years ago but does less so now. More recently, it has put a bigger emphasis on the use of targeted lending facilities. While interest rates have been used as an instrument over the entire sample period, the preferred benchmark rate has changed over time: for many years, the PBOC used the one-year base lending rate but has recently been emphasizing the seven-day reverse repo rate.

Chinese monetary policy has also been guided by multiple objectives over our sample period: employment, GDP growth target, inflation target, monetary aggregate target, external balance, stable currency and financial stability.[16] Although it is likely that the importance of its different monetary objectives has also changed over time, its employment objective has been and continues to be very important. It is very difficult to assess the output gap in a dynamic economy like China's. The PBOC may have better information about the relationship between unemployment and it natural rate and target balance in the labour market so as to maintain aggregate demand in line with aggregate supply. Given the importance of employment as a monetary policy objective, we would expect monetary policy to adjust in a counter-cyclical fashion to labour market conditions: tighten when market conditions are buoyant and loosen when conditions deteriorate. With this hypothesis in mind, we test if our LMCI can help explain the conduct of Chinese monetary policy by modifying the following McCallum monetary policy rule to make it more applicable to the Chinese context:

$$\Delta m_t = \beta_0 + \beta_1(\Delta \mathrm{y}^* - \Delta \mathrm{y}_{t-1}) + \beta_2 \emptyset_t + \varepsilon_t, \qquad (5)$$

where Δm is monetary aggregate growth, $\Delta \mathrm{y}^*$ is the target of nominal GDP growth, $\Delta \mathrm{y}$ is nominal GDP growth, $\emptyset$ are additional relevant variables and ε is an error term. In this type of rule, a central bank is assumed to conduct monetary policy by responding to deviations in GDP growth from its target and to other additional relevant variables (for example, change in the velocity of money or deviations from an inflation target).[17]

We modified this rule so it better reflected monetary policy with "Chinese characteristics." We used deviations in credit growth from its trend instead of money growth as the dependent variable. In doing so, we are assuming that the impact of all the PBOC's monetary instruments was used to control credit growth. We took the PBOC's Total Social Financing as our broad credit measure.[18] We use deviations of

[16] "The single objective of maintaining price stability is an enviable arrangement, as it is simple and easy to measure and communicate. However, it is not yet realistic for China at this stage. For a long time, the annual objectives of the PBOC mandated by the Chinese government have been maintaining pricestability, boosting economic growth, promoting employment, and broadly maintaining balance of payments" (Zhou (2016)).

[17] The McCallum rule is described in McCallum (1988). For the application of McCallum rule to the analysis of Chinese monetary policy see Burdekin and Siklos (2008) and Klingelhöfer and Sun (2018)

[18] Total Social Financing is seen as one of the PBOC's key monetary targets. "The PBOC will implement the prudent and neutral monetary policy, control total money supply, and use multiple monetary policy instruments to maintain reasonable growth of money, lending and social

real GDP growth from its target instead of nominal GDP growth because the Chinese central government has had a real rather than a nominal GDP growth target. Finally, we used deviations of inflation from target, where the inflation target comes from data compiled by Klingelhöfer and Sun (2018) for the period 2000-2015 and from press reports for subsequent years.

The estimation results for different specifications of the McCallum-type rule applied to China using OLS are presented in Table 11.[19] In the first column, we report the results of a basic McCallum rule. In subsequent columns, we add our labour market activity indicators. Our LMCI and the urban supply-demand ratio come in significantly at the 1% level, while the employment sub-index of the non-manufacturing PMI is significant at the 10% level but the sign on this variable is wrong.

Next, we undertake a rolling four-quarter out-of-sample forecasting exercise, like those conducted in Section 4 above. The results are reported in Table 12. For each of the forecast horizons, our LMCI provides forecasts of credit deviations that are significantly better than those from equations using the official unemployment rate and the manufacturing PMI employment sub-index, as per the Diebold-Mariano test. The forecast from the urban supply-demand ratio and the non-manufacturing PMI employment sub-index beat those of our LMCI, but the improvement in forecast accuracy is not significant.

6 Concluding Remarks

Building on the methodology of Tobback et al. (2018), we constructed a Chinese LMCI using text analytics applied to Chinese-language newspapers from the mainland over the period from 2003 to 2017. Visual inspection suggests that our news-based LMCI appears to track the key historical developments in China's labour market. Regional disaggregation illustrates that labour conditions in the export-oriented coastal regions are sensitive to export growth while those in inland regions are not.

We then formally test the information content and the forecast ability of our LMCI against four official labour market activity measures: the unemployment rate, the urban supply-demand ratio and the employment sub-indices of the non-manufacturing and manufacturing PMIs. Surprisingly, one of our findings is that the much-maligned official labour market indicators do contain information. However, their information content and their forecasting ability are not robust across the two wage growth and the credit estimations. Indeed, each of the official labour market activity variables is only significant (and properly signed) in one of the three estimations. In contrast, our LMCI does well in all three estimations. Moreover, in many instances, our LMCI is able to provide forecasts that are significantly superior to those of official labour market indicators.

Economists trying to understand the Chinese economy have to overcome a number of data challenges. Chinese data on many economic variables of interest either do not exist, or the time series are short. This leads researchers to create proxies. Indeed, this is what we had to do in this paper to get at wage

financing to keep adequate and stable liquidity, improve the efficiency of financial operations and its capacity to serve the real economy, and constrain the overall leverage ratio at an acceptable level" (PBOC (2018)).

[19] Unit root tests conducted using the Augmented Dickey-Fuller test suggested that all the series used in these estimations are stationary.

growth. Since the Chinese economy is deeply dynamic, many popular proxies cease to be helpful. Consider the "Li Keqiang Index" which is made up of railway cargo volume, electricity consumption and loans disbursed by banks. It has been considered as an alternative to GDP for measuring economic activity. While this measure may have been useful in the past, China's transition to a more service-based economy and the advent of the shadow banking system makes it less relevant now. Our research suggests that text analytics can be used to extract useful labour market information from Chinese newspaper articles. More generally, the development of text analytics offers researchers the ability to turn newspapers into an alternative source of information about the Chinese economy.

References

1. Alexopoulos, M., and Cohen, J. (2009). Uncertain times, uncertain measures. University of Toronto, Department of Economics, Working Paper 352.

2. Antenucci, D., M.Cafarella, Levenstein, M., Re, C., and Shapiro, M. (2014). Using social media to measure labour market flows. NBER Working Paper 20010.

3. Baker, S., Bloom, N., and Davis, S. (2016). Measuring economic policy uncertainty. Quarterly Journal of Economics, 131 , 1593–1636.

4. Burdekin, R., and Siklos, P. (2008). What has driven Chinese monetary policy since 1990? investigating the people's bank's policy rule. Journal of International Money and Finance, 27 , 847–859.

5. Cai, F., and Chan, W. (2009). The global economic crisis and unemployment in China. Eurasian Geography and Economics, 50 , 513–531.

6. Cai, F., Du, Y., and Wang, M. (2013). Demystify the labor statistics in China. China Economic Journal , 6 , 123–133.

7. Che, W., Li, Z., and Liu, T. (2010). LTP: A Chinese language technology platform. Coling 2010: Demonstration Volume.

8. Chen, Y., and Funke, M. (2009). China's new labour contract law. China Economic Review , 20 , 558–572.

9. Economist (2008). An aberrant abacus. May 1st, 2008.

10. Fan, R., Chang, K., Hsieg, C., Wang, X., and Lin, C. (2008). LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 9 , 1871–1874.

11. Feng, S., Yingyao, H., and Moffitt, R. (2017). Long run trends in unemployment and labor force participation in urban China. Journal of Comparative Economics, 45 , 304–324.

12. Giles, J., Park, A., and Zhang, J. (2005). What is China's true unemployment rate? China Economic Review , 16 , 149–170.

13. International Trade Organization (2011). Trade and employment: From myths to facts. October 4th, 2011.

14. Klingelh¨ofer, J., and Sun, R. (2018). China's regime-switching monetary policy. Economic Modelling , 68, 32–40.

15. Knight, J., and Xue, J. (2006). How high is urban unemployment in China? Journal of Chinese Economics and Business Studies, 4 , 91–107.

16. Lam, R. W., Liu, X., and Schipke, A. (2015). China's labor market in the 'new normal'. IMF Working Paper WP/155/151.

17. Liu, Q. (2012). How high is urban unemployment in China? China Economic Review , 23 , 18–33.

18. McCallum, B. (1988). Robustness properties of a rule for monetary policy. Carnegie-Rochester Conference Series on Public Policy , 29 , 173–203.

19. Mullainathan, S., and Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31 , 87–106.

20. PBOC (2018). PBC monetary policy committee held q4 2017 meeting. URL: http://www.pbc.gov.cn/english/130721/3456056/index.html.

21. Pedregosa et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 20 , 2825–2830.

22. Sokolova, M., and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45 , 427–437.

23. Tan, S., and Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. Expert Systems with Applications, 34 , 2622–2629.

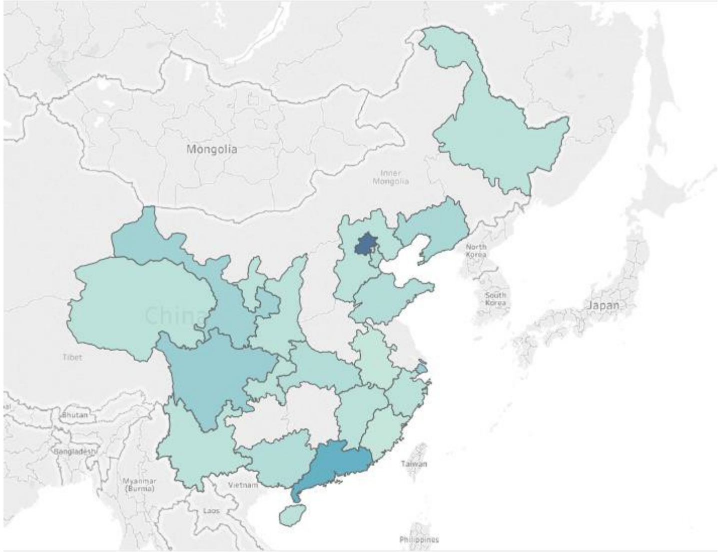
24. Tobback, E., de Fortuny, E. J., Naudts, H., and Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. International Journal of Forecasting, forthcoming.

25. Wang, J., and Zhang, J. (2013). The analysis of the economic value of the Beijing Olympic Games. Journal of Nanjing Sport Institute, 27 , 1–9.

26. Wang, X., and Sun, W. (2014). Discrepancy between registered and actual unemployment rates in China: An investigation in provincial capital cities. China and World Economy, 22 , 40–59.

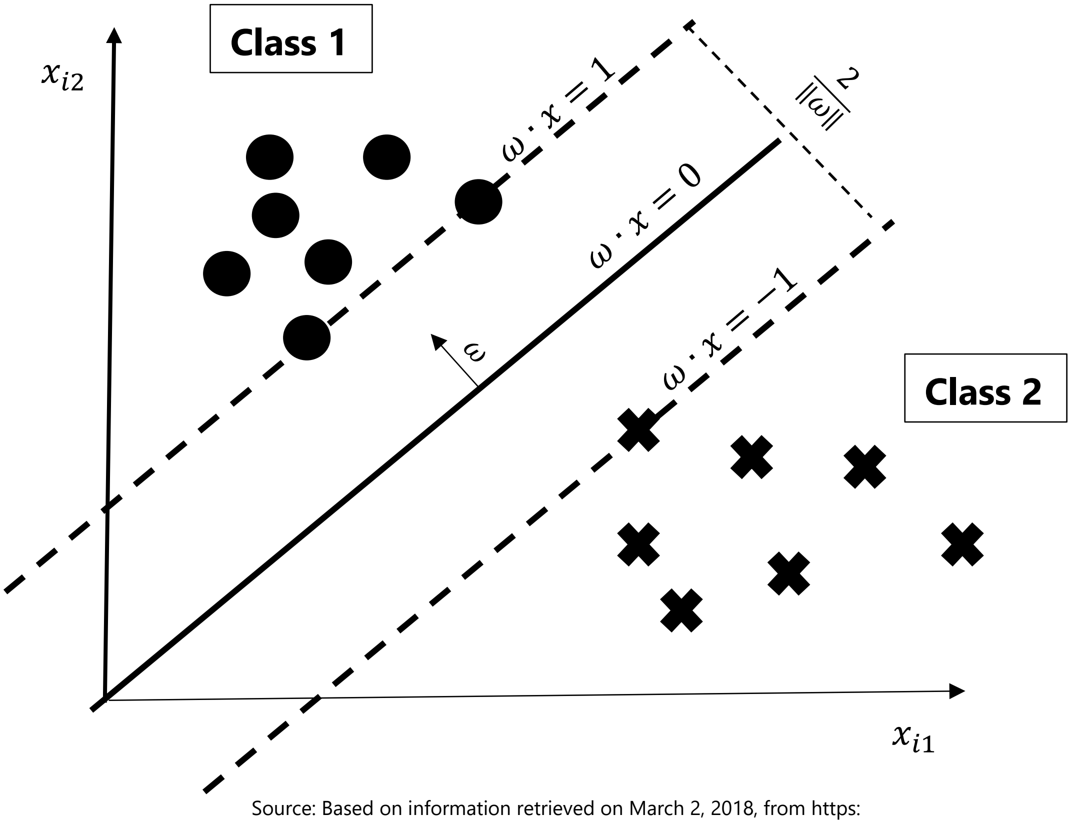
27. Wang, X., and Wan, G. (2014). China's urban employment and urbanization rate: A reestimation. China and World Economy, 22 , 30–44.

28. Zhou, X. (2016). Managing multi-objective monetary policy: From the perspective of transitioning Chinese economy. Michel Camdessus Central Banking Lecture.



Note: There are between 1 and 23 newspapers in each region (the darker the colour, the larger the number of newspapers in that region)

Figure 1: Geographical distribution of newspapers



Source: Based on information retrieved on March 2, 2018, from https: //en.wikipedia.org/wiki/Support vector machine.

Figure 2: Graphical representation of a support vector machine classifier

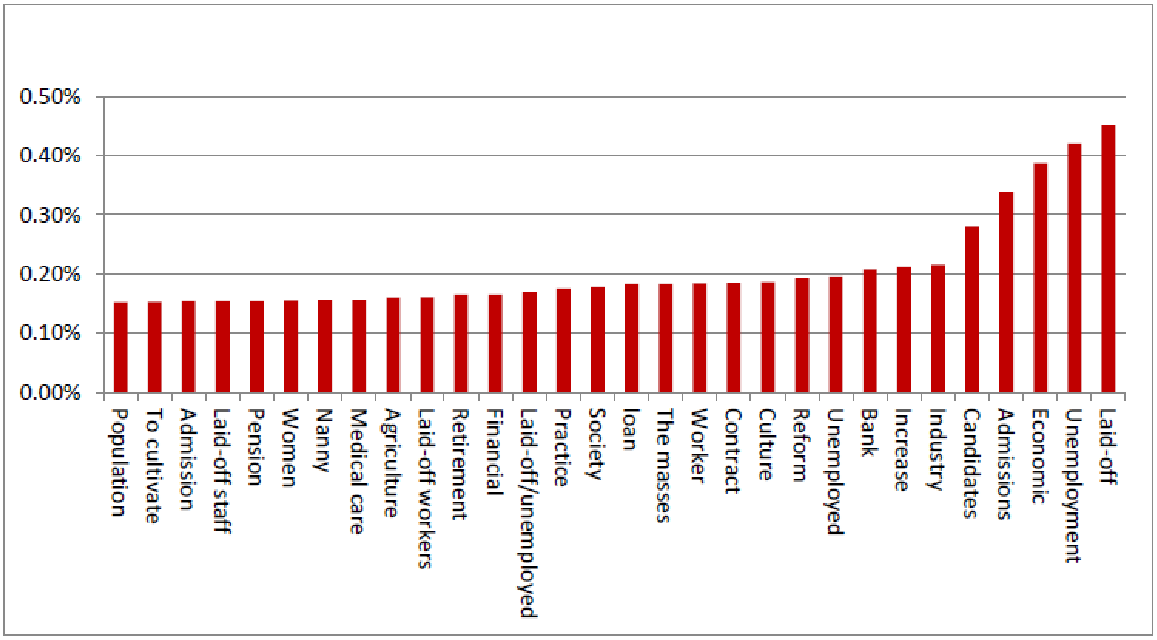


Figure 3: Negative sentiment news articles: word weights of top 30 words

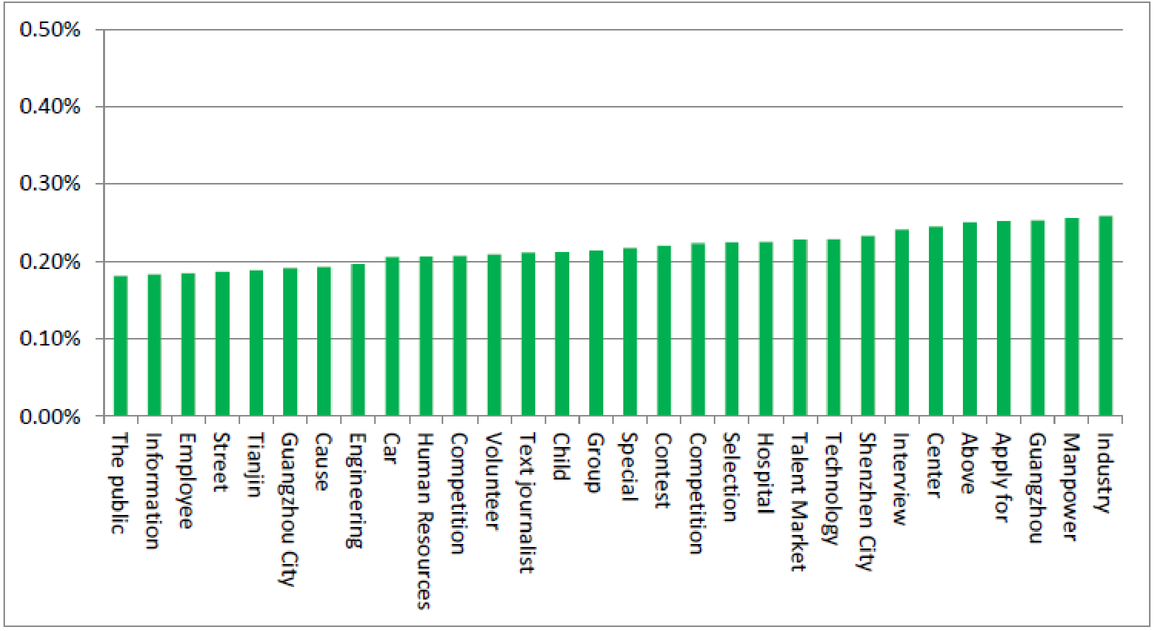


Figure 4: Positive sentiment news articles: word weights of top 30 words

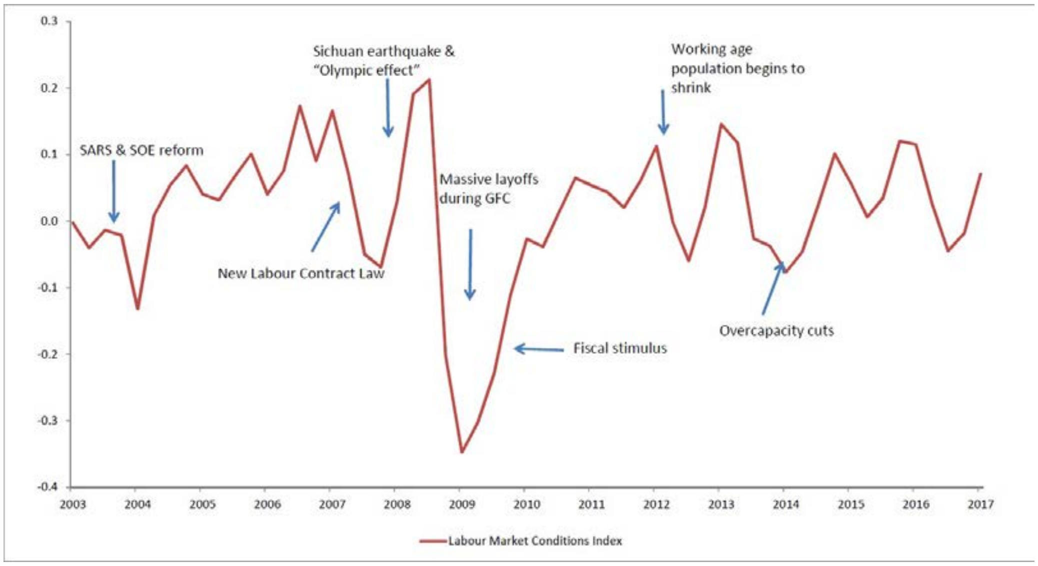


Figure 5: Key historical shocks and the labour market conditions index in China

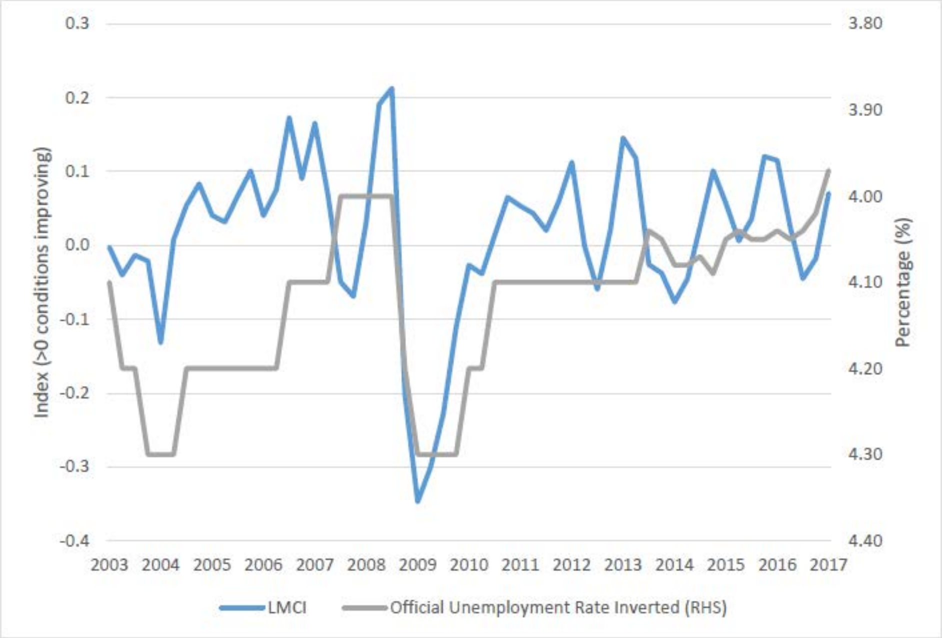


Figure 6: Labour market indicator comparisons (1)

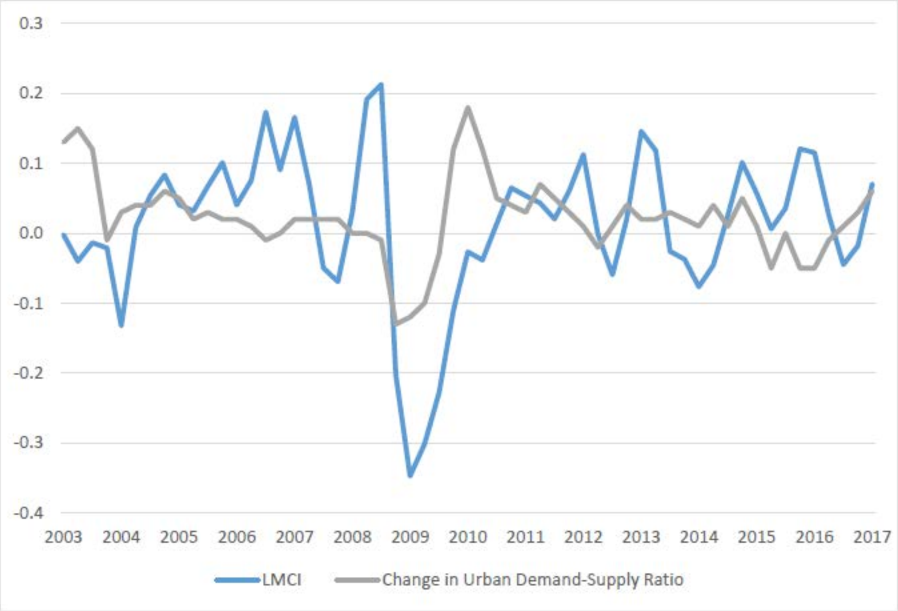


Figure 7: Labour market indicator comparisons (2)

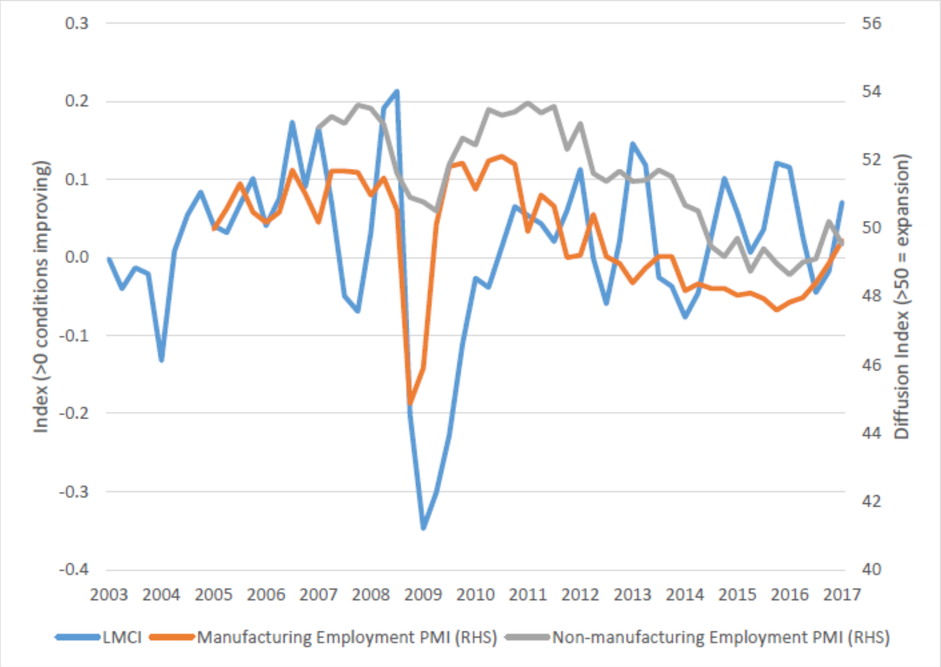


Figure 8: Labour market indicator comparisons (3)

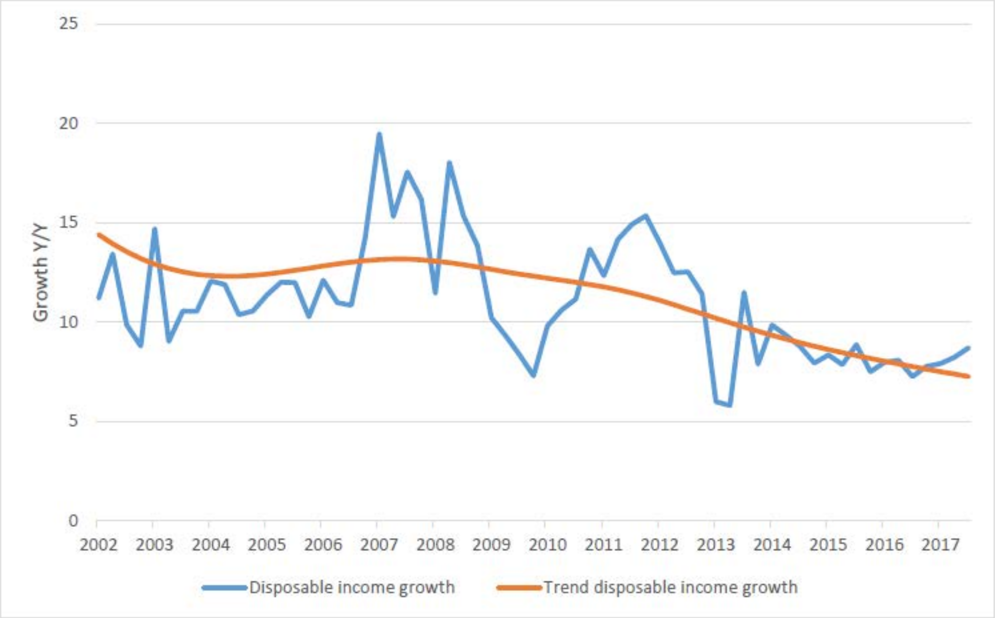


Figure 9: Disposable income growth vs. trend

Table 1: Estimates of China's Unemployment Rate

Reference	Data	1996	2000	2002	2007	2009
NBS	Based on urban residents registered as unemployed in unemployment centres	3%	3.1%	4%	4%	4.3%
Giles et al. (2005)	Based on China Urban Labour Survey and population census data	6.8%	10%	14%		
Knight and Xue (2006)	Based on adjusted administrative statistics (i.e., official rate), Urban Household Survey (1999) and population census data (1982, 1990, 1995, 2000)	8.5%	11.5%			
Liu (2012)	Based on Chinese Household Income Project Surveys (1988, 1995, 2002)			9.5%		
Wang and Sun (2014)	Based on household survey undertaken by Unirule Institute of Economics and the Horizon Research Inc. in 30 provincial capital cities (2007)				13.4%	
Feng et al. (2017)	Based on Urban Household Survey (1988-2009)	4.1%	7.8%	10.4%	8.1%	8.9%

Table 2: Keywords Used to Identify Articles Relevant to Labour Market

Category	Sub-category	Chinese	English translation
Negative labour market sentiment	Job loss	下岗	Layoff
	Unemployment	资遣	layoff
		裁员	layoff
		请辞	resign
		辞职	resignation
		失业	unemployment
		失业保险	unemployment insurance
	Wage reduction Dismissal	减薪	pay cut
		开除	dismissed
		解雇	dismissal
		革职	dismissed
		辞退	dismiss
		解聘	dismissed
		解聘	retired/resign
		解除劳动合同	dismiss the labour contract
	Getting fired	炒鱿鱼	fried squid
		被 炒	getting fired
		卷铺盖	rolling up bed sheets
		丢饭碗	lost rice bowl
	Bankruptcy	破产	bankruptcy
		公司破产	company bankruptcy
		企业破产	enterprise bankruptcy
Positive labour market sentiment	Job search	应征	application
	Employment Recruitment	求职	job search
		找工作	find a job

Continued on next page

Table 2 – continued from previous page

Category	Sub-category	Chinese	English translation
		雇用	employ
		招聘	recruitment
		招工	recruitment
		职缺	job vacancies
		岗位	job position
		招聘会	job fair
		再就业	re-employment
Labour market	Migrant workers	农民工	migrant workers
		民工	migrant workers
	Staff	职工	staff
		员工	employee
		职员	staff
		上班族	office worker
	Job/positions	工作	jobs
		职位	position
		职业	career
		饭碗	rice bowl
		金饭碗	golden rice bowl

Table 3: Manual Classification of Articles in Training/Testing Subset

Number of articles representing positive labour market sentiment	187
Number of articles representing negative labour market sentiment	140
Number of articles relevant to the labour market	327
Number of articles irrelevant to the labour market	450
Total number of articles	777

Table 4: Performance of Support Vector Machine Classifier

Classifier	Total # of labelled articles	# of articles in training set	# of articles in testing set	Accuracy rate	Specificity rate	Sensitivity rate
Stage I	777	620	157	85%	89%	82%
Stage II	313	250	63	83%	94%	72%

Note: The discrepancy between the number of articles relevant to the labour market reported in Table 3 and the number of labelled articles reported for Stage II above reflects the removal of 14 articles that were related to Hong Kong, Macao or Taiwan (and not mainland China).

Table 5: Performance of Our Classification Methodology against Alternative Methods

Method	Classifier	Total # of labelled articles	# of articles in training set	# of articles in testing set	Overall Accuracy rate	(# of classified negative)/ (# of actual negative)	(# of classified positive)/ (# of actual positive)	(# of classified neutral)/ (# of actual neutral)
Method 1*	3-class classifier	777	621	156	73%	55%	57%	88%
Method 2*	Stage I binary classifier	777	621	156	85%	89%	82%	—
	Stage II 3-class classifier	293	234	59	52%	62%	48%	25%
Method 3	Stage I	777	620	157	85%	89%	82%	—
	Stage II	313	250	63	83%	94%	72%	—

*Note: The three-class classifier was designed to balance out our training set by using different weights for each class.

Table 6: Regional LMCIs' Relationship with Export Growth

Variables	Coastal LMCI	Inland LMCI	Overall LMCI
Export growth	0.00125** (0.00056)	0.00067 (0.00050)	0.00195** (0.00085)
Coastal LMCI (-1)	0.70677*** (0.09087)		
Inland LMCI (-1)		0.7259*** (0.09233)	
Overall LMCI (-1)			0.58715*** (0.10532)
Observations	54	54	54
Estimation Period	03Q2-16Q3	03Q2-16Q3	03Q2-16Q3
R-squared	0.58	0.56	0.47
Adjust R-squared	0.56	0.54	0.45

Notes: (1) The dependent variables are coastal LMCI, inland LMCI, and overall LMCI, respectively. (2) Standard errors are in parentheses. (3)The constant term is not shown. (4)***, ** and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Table 7: Estimation Results for Various Wage Phillips Curve Specifications (1)

Variables	1	2	3	4	5	6
Disposable income growth (-1)	0.63***	0.6***	0.63***	0.63**	0.61***	0.46***
	(0.095)	(0.094)	(0.098)	(0.097)	(0.103)	(0.119)
Inflation expectations	0.12***	0.12**	0.13**	0.12**	0.03	0.04
	(0.046)	(0.045)	(0.049)	(0.047)	(0.068)	(0.062)
Urban demand-supply ratio (-1)			-0.7			
			(5.527)			
Official unemployment rate (-1)				-0.04		
				(3.184)		
Manufacturing employment PMI (-1)					0.54**	
					(0.267)	
Non-Manufacturing employment PMI (-1)						0.75**
						(0.318)
LMCI (-1)		5.05*				
		(2.578)				
Observations	56	56	56	56	48	40
Estimation Period	03Q2-17Q1	03Q2-17Q1	03Q2-17Q1	03Q2-17Q1	05Q2-17Q1	07Q2-17Q1
R-squared	0.55	0.58	0.55	0.55	0.62	0.68
Adjust R-squared	0.53	0.56	0.53	0.52	0.59	0.66

Notes: (1) The dependent variable is disposable income growth. (2) Standard errors are in parentheses. (3) The constant term is not shown. (4)***, **, and * indicate statistical significance at the 1%, 5% and 10% levels

Table 8: Estimation Results for Various Wage Phillips Curve Specifications (2)

Variables	1	2	3	4	5	6
De-trended disposable income growth (-1)	0.48***	0.4***	0.46***	0.31**	0.51***	0.47***
	(0.112)	(0.11)	(0.112)	(0.126)	(0.122)	(0.13)
Inflation expectations	0.1**	0.1**	0.11**	0.1**	0.06	0.07
	(0.043)	(0.04)	(0.045)	(0.041)	(0.067)	(0.063)
Urban demand-supply ratio (-1)			-3.36			
			(5.351)			
Official unemployment rate (-1)				-8.63**		
				(3.347)		
Manufacturing employment PMI (-1)					0.2	
					(0.252)	
Non-Manufacturing employment PMI (-1)						0.16
						(0.268)
LMCI (-1)		6.81***				
		(2.431)				
Observations	56	56	56	56	48	40
Estimation Period	03Q2-17Q1	03Q2-17Q1	03Q2-17Q1	03Q2-17Q1	05Q2-17Q1	07Q2-17Q1
R-squared	0.34	0.43	0.35	0.42	0.35	0.37
Adjust R-squared	0.32	0.39	0.31	0.38	0.31	0.32

Notes: (1) The dependent variable is de-trended disposable income growth. (2) Standard errors are in parentheses. (3) The constant term is not shown. (4)***, **, and * indicate statistical significance at the 1%, 5% and 10% levels,

Table 9: Wage Growth Forecasting Performance

RMSE Ratios: dynamic out-of-sample evaluation (2008Q2-2017Q1)

Model	T+1	T+2	T+3	T+4
Official unemployment rate	0.96	0.95	0.96	1.00
Urban demand-supply ratio	0.98	0.97	0.99	1.03
Manufacturing employment PMI	0.98	1.00	1.10	1.12
Non-manufacturing employment PMI	1.04	1.10	1.17	1.26**

Table 10: De-trended Wage Growth Forecasting Performance

RMSE Ratios: dynamic out-of-sample evaluation (2008Q2- 2017Q1)

Model	T+1	T+2	T+3	T+4
Official unemployment rate	0.98	0.96	0.95	0.97
Urban demand-supply ratio	0.86*	0.74**	0.72*	0.76
Manufacturing employment PMI	0.89	0.80*	0.79*	0.79**
Non-manufacturing employment PMI	0.87*	0.77**	0.75**	0.76**

Notes: (1) RMSE ratios are constructed by dividing the RMSE of the specification with the LMCI by the specification with the labour market indicator listed in the table. (2) ***, **, and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Table 11: Estimation Results for Various McCallum-type Monetary Policy Rules

Variables	1	2	3	4	5	6
De-trended TSF growth (-1)	0.82***	0.75***	0.92***	0.83***	0.84***	0.57***
	(0.065)	(0.066)	(0.059)	(0.065)	(0.075)	(0.085)
Deviation of real GDP from target (-1)	0.123	0.12	-0.05	0.06	-0.04	-0.06
	(0.125)	(0.118)	(0.114)	(0.140)	(0.203)	(0.158)
Deviations in inflation from target (-1)	0.46***	0.31**	0.32**	0.56***	0.39***	0.85***
	(0.134)	(0.138)	(0.119)	(0.167)	(0.142)	(0.161)
LMCI (-1)		-6.7***				
		(2.49)				
Urban demand-supply ratio (-1)			-21.31***			
			(4.853)			
Official unemployment rate (-1)				-3.59		
				(3.429)		
Manufacturing employment PMI (-1)					-0.26	
					(0.268)	
Non-Manufacturing employment PMI (-1)						0.38*
						(0.211)
Observations	53	53	53	53	49	41
Estimation Period	04Q2-17Q2	04Q2-17Q2	04Q2-17Q2	04Q2-17Q2	05Q2-17Q2	07Q2-17Q2
Adjust R-squared	0.79	0.82	0.85	0.79	0.78	0.79
Sum of squared residuals	153.3	133.16	109.35	149.88	140.18	75.02

Notes: (1) The dependent variable is de-trended total social financing growth. (2) Standard errors are in parentheses. (3) The constant term is not shown. (4)***, ** and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Table 12: De-trended TSF Growth Forecasting Performance

RMSE Ratios: dynamic out-of-sample evaluation (2008Q2- 2017Q2)

Model	T+1	T+2	T+3	T+4
Official unemployment rate	0.82*	0.78*	0.77*	0.78*
Urban demand-supply ratio	1.12	1.14	1.14	1.11
Manufacturing employment PMI	0.85**	0.79**	0.79**	0.85**
Non-manufacturing employment PMI	0.97	1.01	1.06	1.18

Notes: (1) RMSE ratios are constructed by dividing the RMSE of the specification with the LMCI by the specification with the labour market indicator listed in the table. (2) ***, **, and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Appendix A: List of Newspapers Covered in Database

Newspaper name	Province	Newspaper Name	Province
21st Century Business Herald	Guangdong	Sichuan Daily	Sichuan
Shanghai Securities News	Shanghai	Sichuan Economic Daily	Sichuan
China Business Times	Beijing	Dazhong Daily	Shandong
China Enterprises News	Beijing	Da Lian Daily	Liaoning
China PetroChemical news	Beijing	Tianjin Daily	Tianjin
China Taxation News	Beijing	Ningxia Daily	Ningxia
China Economic Times	Beijing	Guo Xi Daily	Guangxi
China Business	Beijing	Modern Life Daily	Guangxi
China Green Times	Beijing	Chengdu Business Daily	Sichuan
China Securities Journal	Beijing	Chengdu Daily	Sichuan
China Trade Journal	Beijing	Cheng Du Wan Bao	Sichuan
China High-Tech Industry Herald	Beijing	Baokan Wenzhai	Shanghai
Yunnan Daily	Yunnan	Wen Hui Bao	Shanghai
Ren Min Zheng Xie Bao	Beijing	New Express	Guandong
People's Daily Overseas Edition	Beijing	Modern Evening Times	Heilonggjiang
Evening Today	Tianjin	Xin Min Evening News	Shanghai
Jian Kang Shi Bao	Beijing	Shanghai Morning Post	Shanghai
Guangming Daily	Beijing	Chunchen Evening News	Yunnan
Lanzhou Daily	Gansu	Chutian Metropolis Daily	Hubei
Lanzhou Evening News	Gansu	Daily Update	Tianjin
Lanzhou Morning News	Gansu	Shantou Daily	Guangdong
Beijing Daily	Beijing	Shantou Tequ Wanbao	Guangdong
Beijing Morning Post	Beijing	Shantou DushiBao	Guangdong
Beijing Youth Daily	Beijing	Jiang Nan City News	Jiangxi
Ban Dao Morning News	Liaoning	Jiangxi Daily	Jiangxi
Hua Xi Du Shi Bao	Sichuan	The Mirror	Beijing
Nan Guo Zao Bao	Guangxi	Zheijiang Daily	Zheijiang
Nan Fang Daily	Guangdong	Hainan Daily	Hainan
Southern Metropolis Daily	Guangdong	Haikou Evening News	Hainan
Hefei Evening News	Anhui	China Light Industries Post	Beijing
Harbin Daily	Heilongjiang	Shenzhen Economic Daily	Guangdong

Newspaper name	Province
Shenzhen Evening News	Guangdong
Shenzhen Special Zone Daily	Guangdong
Hubei Daily	Hubei
Yanzhao Evening News	Hebei
Global Times	Beijing
Gan Su Nong Min Bao	Gansu
Gansu Daily	Gansu
Shen Huo Ri Bao	Shandong
Shijiazhuang Daily	Hebei
Fujian Daily	Fujian
Economic Information Daily	Beijing
Economic Daily	Beijing
Yangcheng Evening News	Guangdong
Xi An Daily	Shaanxi
Xi'an Evening News	Shaanxi
XiHai DuShi Bao	Qinghai
Jiefang Daily	Shanghai
Securities Times	Guangdong
Liaoning Daily	Liaoning
Liao Shen Evening News	Liaoning
Chongqing Economic Times	Chongqing
Chongqing Evening News	Chongqing
Qian Jiang Wan Bao	Zhejiang
Yinchuan Evening News	Ningxia
Changjiang Daily	Hubei
Qing Hai Daily	Qinghai
Qilu Evening News	Shandong

Appendix B: Variable and Data Description

Variable	Description/Source
Composite LMCI	The composite LMCI is constructed following the methodology described in Section 3. The index is normalized to a mean of 0. Data source: Wisers (see Section 2 for more details).
Headline CPI inflation	Constructed as the year-over-year growth rate in the headline CPI. Data source: National Bureau of Statistics of China / Haver Analytics.
Inflation expectations	Quarterly diffusion index based on survey responses on the direction of the CPI in the next quarter. An index of 50 or higher indicates that the price level in the next quarter is expected to increase (and the higher the index the higher the expectation of a rising price level in the next quarter). Data source: PBOC / Haver Analytics.
Urban labour demand-supply ratio	Ratio of demand for labour to supply of labour. Data source: Ministry of Human Resources and Social Security of China / Haver Analytics.
Official unemployment rate	Urban registered unemployment rate. Data source: Ministry of Human Resources and Social Security of China / Haver Analytics.

Employment sub-indices of the manufacturing and non-manufacturing PMIs	Monthly diffusion index based on survey responses on the direction of employment conditions in the current month. An index of 50 or higher indicates employment conditions are improving. Data source: National Bureau of Statistics of China / Haver Analytics.
Growth rate of per capita disposable income	Constructed as the year-over-year growth rate in urban per capita disposable income. Data source: National Bureau of Statistics of China / Haver Analytics.

Irving Fisher Committee on
Central Bank Statistics

BANK FOR INTERNATIONAL SETTLEMENTS

Ninth IFC Conference on “Are post-crisis statistical initiatives completed?”

Basel, 30-31 August 2018

Can media and text analytics provide insights into labour market conditions in China?[1]

Jeannine Bailliu, Xinfen Han, Mark Kruger,
Yu-Hsien Liu and Sri Thanabalasingam,
Bank of Canada

[1] This presentation was prepared for the meeting. The views expressed are those of the authors and do not necessarily reflect the views of the BIS, the IFC or the central banks and other institutions represented at the meeting.

BANK OF CANADA
BANQUE DU CANADA

Can Media and Text Analytics Provide Insights into Labour Market Conditions in China?

Jeannine Bailliu, Xinfen Han, Mark Kruger, Yu-Hsien Liu, Sri Thanabalasingam
International Department

Disclaimer

- Bank of Canada staff working papers provide a forum for staff to publish work-in-progress research independently from the Bank's Governing Council.
- This research may support or challenge prevailing policy orthodoxy.
- Therefore, the **views expressed in this paper are solely those of the authors** and may differ from official Bank of Canada views.
- No responsibility for them should be attributed to the Bank.

Labour statistics among China’s worst

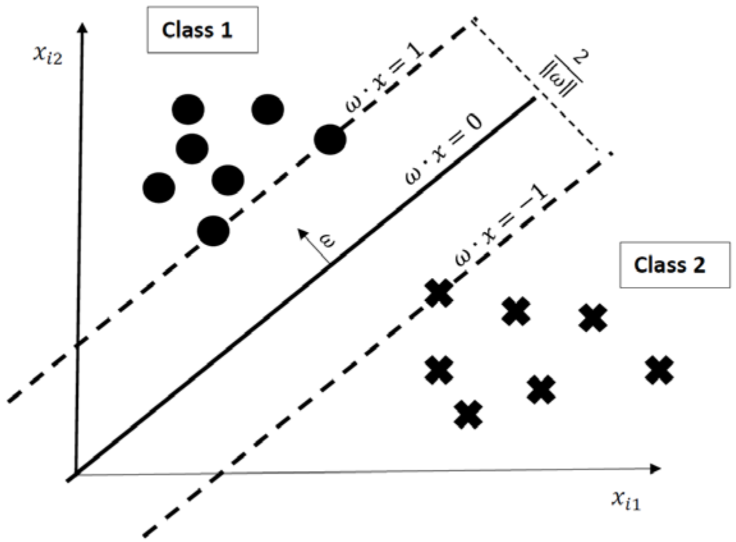
- The prize for the **dodgiest figures** goes to the labour market
  - Urban unemployment rate is “meaningless” Economist (2008)
  - Wage figures are also “lousy”
- Surveys suggest official rate **underestimates unemployment**
  - Knight and Xue (2007)
  - Wang and Sun (2014)
- Compared to other major countries, China’s official unemployment rate shows **little sensitivity** to changes in output
  - Lam et al. (2015)
- Three relatively high frequency indicators capture formal employment, but not migrant workers
  - Migrant workers could make up 25% of urban employment
    - Wang and Wan (2014)

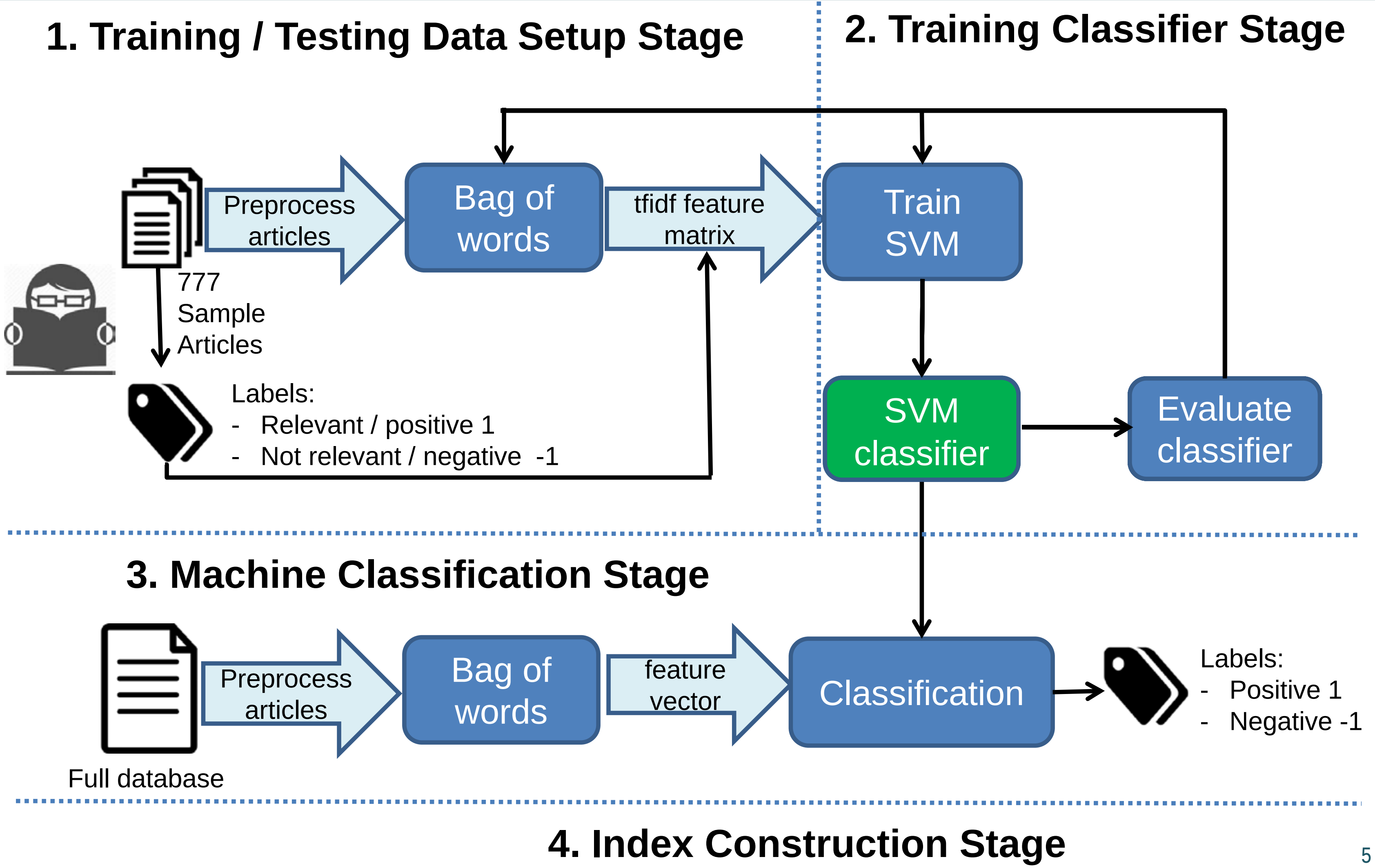
Our database

- **Chinese language** newspaper database
  - Wisers, a Hong Kong-based company
- We focus on **subset of 90** Chinese newspapers
  - Continuously published over **January 2003 to June 2017**
  - Broad geographic coverage
    - **26 out of 34 regions**
    - **77% population**
- Building the relevant article pool
  - 8 millions articles from predefined keywords search
  - downloaded all articles from randomly selected one day per month
    - **266,414 potentially relevant articles**

Text mining methodology

- Our approach is inspired by Tobback et al. (2016)
    - Use text mining to produce economic policy uncertainty index





1. Training / Testing Data Setup Stage

Preprocess articles

Bag of words

tfidf feature matrix

777 Sample Articles

Labels:
- Relevant / positive 1
- Not relevant / negative -1

2. Training Classifier Stage

Train SVM

SVM classifier

Evaluate classifier

3. Machine Classification Stage

Full database

Preprocess articles

Bag of words

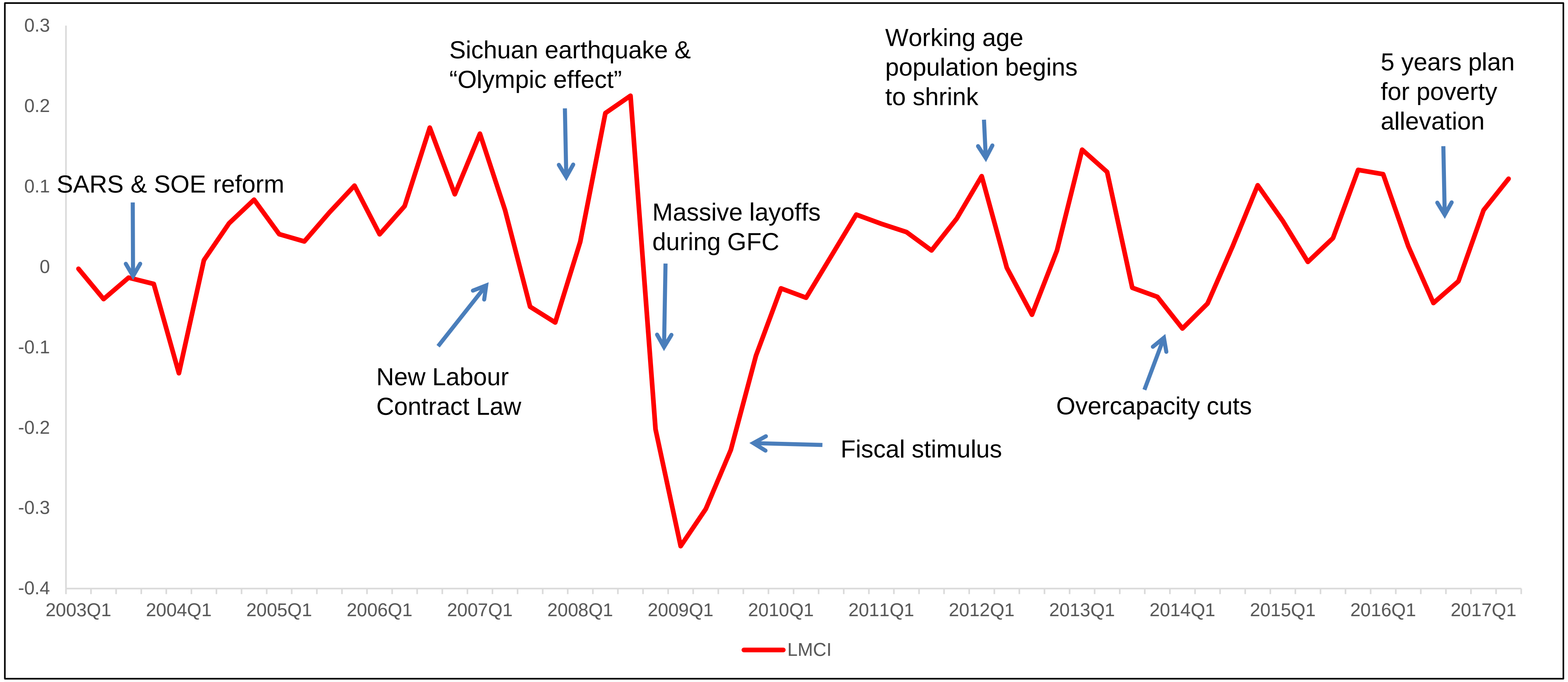
feature vector

Classification

Labels:
- Positive 1
- Negative -1

4. Index Construction Stage

Our Labour Market Conditions Index (LMCI)



SARS & SOE reform

Sichuan earthquake & “Olympic effect”

New Labour Contract Law

Massive layoffs during GFC

Fiscal stimulus

Working age population begins to shrink

Overcapacity cuts

5 years plan for poverty allevation

Text mining methodology

- **Why we use machine learning approach?**
  - Manual classification **costly**
    - 3 or 4 authors read and classify articles independently
    - Discuss disagreements until consensus reaches
  - Machine learning classification more **consistent**
- **Challenges parsing Chinese text**
  - In English, unique words are easy to identify since they are separated by spaces
  - Chinese text has no spaces between characters and a character, on its own, may not form a meaningful unit
  - Harbin LTP **natural language processing software**
- **Our methodology is generic** and can be applied to other classification problems

LMCI Validation

- Construct formal models to evaluate LMCI
  - **Wage** Phillips Curve
    - The co-movements between our LMCI and wage growth
  - **McCallum Rule** (1998) with "Chinese characteristics"
    - The PBOC responds in a counter-cyclical fashion to labour market conditions
- Construct two **regional sub-indices**
  - Our results show labour conditions in **coastal regions sensitive** to **export growth**, while in inland regions are not.

- Our study suggests that text analytics can be used to **extract useful labour market information** from Chinese media.

Questions?

Scan for more information: