HONG KONG INSTITUTE FOR MONETARY AND FINANCIAL RESEARCH

A ROBUST TEXTUAL ANALYSIS OF THE DYNAMICS OF HONG KONG PROPERTY MARKET

Ken Wong, Max Kwong, Paul Luk and Michael Cheng

HKIMR Working Paper No.08/2021

April 2021

HONG KONG
ACADEMY OF
FINANCE
金融學院

Hong Kong Institute for Monetary and Financial Research
香港貨幣及金融研究中心
(a company incorporated with limited liability)

A Robust Textual Analysis of the Dynamics of Hong Kong Property Market

Ken Wong
Hong Kong Monetary Authority

Max Kwong
Hong Kong Monetary Authority

Paul Luk
Hong Kong Institute for Monetary and Financial Research

Michael Cheng
Hong Kong Monetary Authority

April 2021

Abstract

Market sentiments influence the dynamics of Hong Kong's macro-critical property market, but the unobservable nature of market sentiments makes it difficult to assess systemically this sentiment channel. Using text-mining techniques, this paper sets up a news-based property market sentiment index and a Google Trends-based buyer incentive index for Hong Kong, and studies the sentiment channel of transmission in the Hong Kong property market. The news-based property market sentiment index can reflect the change in sentiments in past key events, with the sentiments in the primary market tending to lead that of the secondary market during the low housing supply period. For the Google Buyer Incentive Index, we find that it has value-added in forecasting (or nowcasting) the official property price index. In mapping out the sentiment channel using a structural vector-autoregressive (SVAR) model, we find that an improvement in market sentiments could stimulate buyers' incentives, which then together would affect property prices and transaction volumes.

Keywords: property market, nowcasting, textual analysis, sentiment.
JEL classification: R31.

---

* Email: Wong: kcswong@hkma.gov.hk, Kwong: mwmkwong@hkma.gov.hk, Luk: pskluk@hkma.gov.hk and Cheng: mkscheng@hkma.gov.hk

* The authors would like to thank Lillian Cheung, Giorgio Valente and Jason Wu for their helpful comments. This paper represents the views of the authors, which are not necessarily the views of the Hong Kong Monetary Authority, the Hong Kong Academy of Finance, the Hong Kong Institute for Monetary and Financial Research or its Board of Directors. All errors are the authors' own.

I. INTRODUCTION

The housing market is an integral part of the macroeconomy. While the housing literature has stressed the importance of key fundamental channels such as the collateral channel and the interest rate channel (Iacoviello, 2005; Liu, Wang and Zha, 2013) through which macroeconomic disturbances transmit to the housing market, the role played by market sentiments remain less frequently explored. In fact, the notion of market sentiments is not novel. Kindleberger (1978), Galbraith (1990), Tetlock (2007) and Shiller (2009) have argued that market sentiments are one of the key determinants of asset prices, including housing prices. However, the unobservable and elusive nature of market sentiments makes systematic studies of how they influence the housing market difficult.

The recent integration of language processing techniques and big data in economic and financial research provides an opportunity to study the role of market sentiments systematically. Using text-mining techniques on newspaper articles, researchers have found ways to identify economic policy uncertainty (Baker, Bloom and Davis, 2016; Huang et al., 2020), improve inflation expectation forecasts (Larsen, Thorsrud and Zhulanova, 2020) and gauge news sentiment (Shapiro, Sudhof and Wilson, 2020). Besides, in a series of papers, Choi and Varian (2012) and Scott and Varian (2014, 2015) have shown that internet search results can significantly improve nowcasting of a wide range of variables, including macroeconomic variables. Compared with traditional, survey-based indicators, these new lexicon-based indices are more flexible, less costly to obtain, available with very little time lag, and usually based on large samples (which means more resistant to small sample biases).

The objective of this paper is to study the extent to which macroeconomic shocks in Hong Kong transmits to the housing market through the sentiments channel. The real estate sector in Hong Kong is directly related to about 10% of domestic economic activities, and more than half of the household debt is residential mortgage loans. In addition, almost one-third of the Hong Kong Government revenue depends on the property market performance (e.g. stamp duties and land premium). Although many studies have examined Hong Kong's real estate market because of its salient features and importance to the economy (e.g. Kwan et al. (2015), Huang et al. (2018), Leung and Tse (2017), Leung et al. (2020b) and Tang (2019)), most of

these studies did not discuss the sentiments channel in detail.[2] Given the importance of the real estate sector in Hong Kong, understanding the sentiments channel of the housing market is a key matter for macro-financial stability.

We proceed in three steps. First, we use the lexicon approach to set up a news-based property market sentiment index for Hong Kong based on local news articles. In the second step, we measure the buyer's incentives to buy a house in Hong Kong. With a rationale that the internet search intensity of housing market-related information can reflect buyers' incentives, we use Google Trends to develop a Google Buyer Incentive Index. It is expected that market sentiments would influence buyers' incentives to buy a house. Last, we map out the transmission channel of market sentiments to the housing market, by estimating a structural vector-autoregressive (SVAR) model and conduct an impulse response analysis.

There are several interesting findings in this paper. The news-based property market sentiment index can reflect the change in sentiments during major economic and market specific events. On further distinguishing the market sentiments in the primary market and the secondary market, we find that the primary market sentiments tend to lead the secondary market sentiments during the low housing supply period. For the Google Buyer Incentive Index, we find that it has value-added in forecasting (or nowcasting) the official Rating and Valuation Department (R&VD) housing price index. Regarding the transmission mechanism, we discover that, in line with theories, both market sentiments and buyers' incentives could be driven by the macro environment. In particular, an improvement in market sentiments could stimulate buyers' incentives, which then together would affect housing prices and transaction volumes through the sentiments channel.

This paper is related to Wu et al. (2017) who use a Bayesian vector-autoregression model to estimate the short-run housing market dynamics in Hong Kong. Their analysis suggests that sentiments could account for about 8% of housing price variations. However, their analysis simply uses the stock market index (the Hang Seng Index) to proxy the overall financial market sentiment, which is unable to track the time-varying correlation between the

[2] Kwan et al. (2015) find that consumption-based asset pricing models, especially a recursive utility variation which separates the intertemporal substitution and uncertainty resolution, could "fit" the stock and housing markets. Huang et al. (2018) propose a search-theoretic model to explain the Hong Kong housing market's price-rent ratio and turnover rate. Leung and Tse (2017), Leung et al. (2020b), Tang (2019) stress the speculation in the real estate markets.

stock prices and housing prices. The key contribution of this paper relative to theirs is the use of novel and robust techniques in text-mining to directly measure housing market sentiments for the empirical analysis.

The remainder of this paper is structured as follows. The next section reviews Hong Kong's housing market development. Section III describes the methodology that we use to compile the news-based property market sentiments index for Hong Kong, followed by some observations and analysis. Section IV describes the methodology of compiling the Google Buyer Incentive Index with some findings. In Section V, we set out our empirical approach to identify the impacts and transmission channels of market sentiments and buyers' incentives in the Hong Kong housing market. The final section concludes.

II. FEATURES OF THE HONG KONG PROPERTY MARKET

Similar to many other economies, Hong Kong has experienced property boom-bust cycles. After a downturn between the Asian financial crisis in 1998 and the SARS outbreak in 2003, Hong Kong's property market has been on a long rally for more than ten years despite occasional declines amid regional financial market turbulences in late-2015, increasing US-China tensions since 2018, and more recently the COVID-19 pandemic.

Among the various factors driving property price dynamics, the supply-side factor is widely considered as one key contributing factor to the uptrend in Hong Kong's housing prices (Wu et al. (2017)). With the Hong Kong Government once suspending regular land sales in 2003, and with a low share of residential land use in total land area (7% in 2018, according to the Hong Kong Planning Department), new completion in Hong Kong has stayed at relatively low levels, despite some pick-ups in recent years following the resumption of regular land sales in 2010. Accordingly, primary market (i.e. first-hand property) transaction volume was generally less than secondary market (i.e. second-hand property) transaction volume, with the primary market accounting for around 30% of the total property transaction volume over the past five years.

Another feature of the supply side is there is no time requirement for property developers to sell new completions, and therefore property developers can launch new units in

accordance with property market conditions.[3] Property developers also often pursue active marketing strategies to promote new launches, including through advertisements in newspapers and various media, with a consequence of influencing property market sentiments.

III. MEASURING PROPERTY MARKET SENTIMENTS IN HONG KONG

3.1 Methodology

We begin our analysis by first constructing sentiment indices for the property market in Hong Kong using newspaper information.[4] Using textual information to compile sentiment index is a rapidly growing area (see Soo, 2018; Huang, Simpson, Ulybina and Roitman, 2019; Shapiro, Sudhof and Wilson, 2020, for example). For our purpose, we find it satisfactory to use a simple Boolean approach similar to Soo (2018) and Gao and Zhao (2018).

Put simply, in a given month t, we count the number of articles that display positive or negative sentiments for the property market $m \in \{p,s\}$, where p and s stand for the primary and secondary market respectively. Precisely, for each article a in newspaper n at time t, we define a binary variable $pos_{a,n,m,t} \in \{0,1\}$, which takes on a score of 1 when the article displays positive sentiment. Similarly, we define an indicator variable $neg_{a,n,m,t} \in \{0,1\}$, which takes on a score of 1 when the article displays negative sentiment.

We use a compound text-search approach to assign sentiment scores to a given article. For example, an article is assigned $pos_{a,n,m,t}=1$ ($neg_{a,n,m,t}=1$) when there exists a paragraph in which (a) at least one keyword about market m is mentioned; and (b) at least one keyword indicating positive (negative) sentiment is mentioned.[5] The sentiment keywords, reported in Appendix A.1, are commonly found in newspaper articles in describing the property market condition. Most of them are adjectives, while some are pheromones which would be perceived as an optimistic/pessimistic market outlook (e.g. forfeiting deposits).[6] To avoid capturing

[3] To encourage property developers to expedite the supply of first-hand private residential units, the Hong Kong Government has announced it will introduce a 'Special Rates' on vacant first-hand private residential units by amending the Rating Ordinance. However, the amendment is still in the legislation process.

[4] Newspapers are the key sources of information to the Hong Kong public. See the Yearly Survey of People's Main Source of News conducted by HKU POP https://www.hkupop.hku.hk/chinese/popexpress/press/main/year/datatables.html.

[5] Details of the keyword lists and index construction method are reported in Appendix A.1.

[6] Recent studies (e.g. Loughran and McDonald (2011), Walker (2014)) on textual analysis have argued that

housing prices dynamics information, words which might be describing price changes (e.g. new high, gain) are excluded in our dictionary. Robustness checks with alternative keyword choices are conducted in a later section.

The property market sentiment $S_{m,t}$ for market m and period t adds up all the sentiment scores across newspaper articles:[7]

$$\text{For } m = p,s, \quad S_{m,t} = \sum_{a,n}(pos_{a,n,m,t} - neg_{a,n,m,t}). \tag{1}$$

It is straightforward to define the overall market sentiments of a given period S_t as the sum of the sentiment scores of the primary and second markets:

$$S_t = S_{p,t} + S_{s,t}. \tag{2}$$

The sentiment scores are then scaled by the number of news articles in the local real estate section in the same period, with the aim of mitigating the influence from the changing number of real-estate related articles over time to the resulting sentiment index. Finally, the scaled series are then standardised to have a unit standard deviation.

Our sample of Hong Kong newspaper articles comes from Wisers Information Portal, a digital archive containing Chinese news media. The sample period starts from April 1998, the earliest available date from the archive. The scope of newspaper articles is limited to those in the real estate section, which better reflects the development in the property market. Reporting on non-domestic real estate markets are excluded from our analysis.[8]

3.2 Analysis of the news-based property market sentiment index

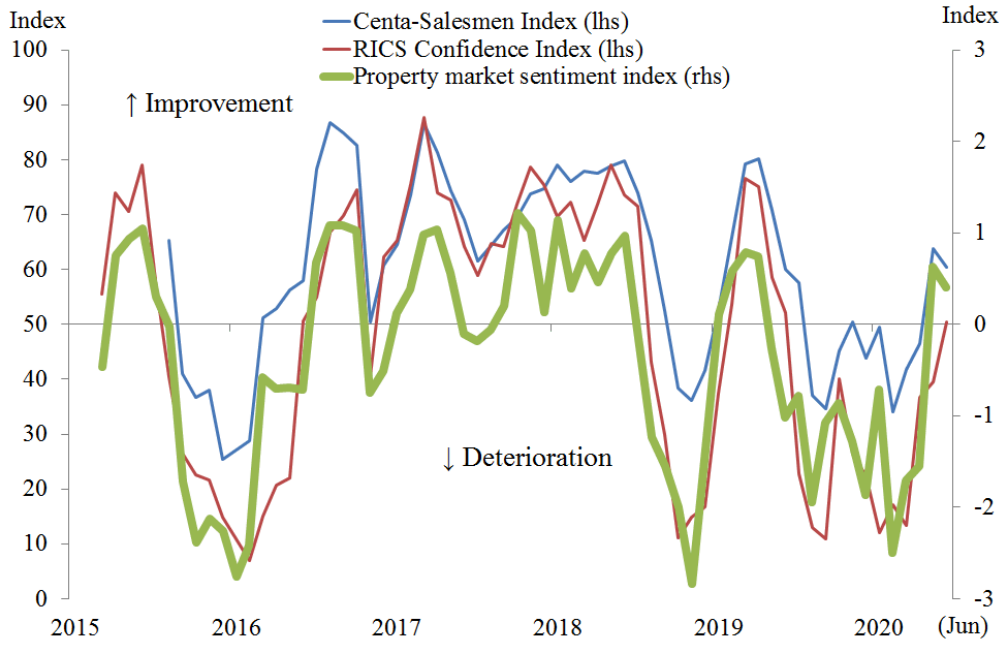
general tonal sentiment words could be irrelevant to identify market-specific sentiments and lead to noisy measures. To improve the efficacy of sentiments identification, we follow Soo (2018) method by introducing domain specific vocabulary. In particular we selected positive or negative tone keywords that Hong Kong media usually use in covering the property market, and some of them are related to a particular market segment only.

[7] Note that an article may display both positive sentiments and negative sentiments for a given market *m*. In this case, the article will not contribute to the aggregate sentiment S_{mt}. It is possible that the article has an uneven weighting on positive and negative sentiments in its content, but our indicator variables do not capture sentiments as a continuous scale. However, even with this relatively crude construction method, the resulting sentiment index is already quite informative and comparable to survey-based indices.

[8] Appendix A.1.1 reports the details of excluding non-domestic real estate markets information from the sentiment scores.

Figure 1 shows the news-based property market sentiment index (green line) together with other existing survey-based housing sentiment indices. Positive values mean market sentiments are improving, and vice versa. As shown in Figure 1, our news-based property market sentiment index fluctuates closely with the existing sentiment indices and is also highly correlated with the timing of domestic and external shocks, such as the start of the US rate hike cycle in 2015, regional market turbulence during 2015-2016, escalation of US-China tensions in the second half of 2018, and the domestic social incidents in 2019. During the periods, housing prices have experienced some corrections between 5% to 11%. The prolonged social incident in Hong Kong and the pandemic outbreak of COVID-19 have also suppressed the market sentiment between the second half of 2019 and the first half of 2020.

Figure 1: News-based property market sentiment index with other survey-based sentiment indices



Sources: RICS, Centaline property agency and staff estimates.

Figure 1 also compares our news-based property market sentiment index with two other indices, namely the Centa-Salesmen Index (CSI)[9] and the Royal Institution of Chartered Surveyors (RICS) Confidence Index[10]. For the period March 2015-June 2020, our index

[9] CSI is compiled from a weekly survey of Centaline property agency's salesman.
[10] RICS Confidence Index is compiled from a sentiments survey that collects and analyses the opinions of professionals in agency.

correlates with the CSI at 0.93 and correlates with the RICS Confidence Index at 0.86, implying that our index can reflect market sentiments. On the other hand, it is noticeable that there are periods (say, between 2017 and 2018) during which our index diverges from the survey-based sentiment indices. The divergence could reflect the difference in views expressed in the mass media versus those expressed by professionals.[11]

More importantly, our news-based index is more informative as it can distinguish the sentiments in the primary and secondary market, whereas the existing survey-based indicators cannot. This can help fill the information gap in the primary market, which was not transparent until in recent years[12]. As illustrated by Equation (2), the overall property market sentiment (S_t) can be decomposed into primary (S_{pt}) and secondary market (S_{st}). Figure 2 shows a decomposition of the dynamics of the news-based property market sentiment index into the contribution of the underlying market segments since April 1998. It is interesting to see that in general, the contribution of the primary and secondary markets is roughly the same, even though the secondary market dominates transaction volumes (over 70% during the same period). This is probably because newspapers usually have a disproportionate coverage on the primary market news. For example, they would concentrate on reporting new launches of some large-scale development projects, reflecting the marketing campaigns by property developers.

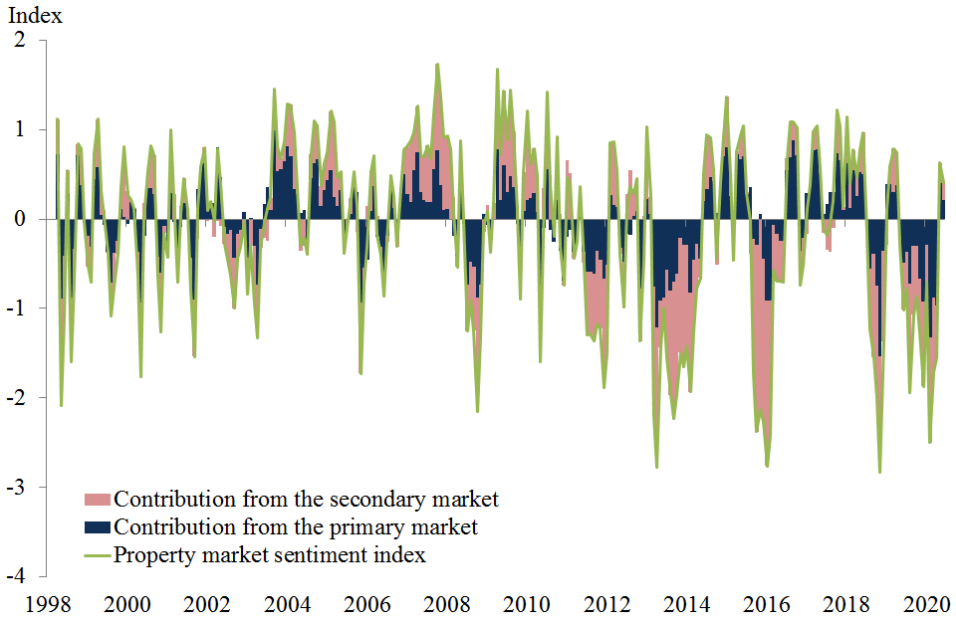
Another interesting observation is that the contribution of primary and secondary market sentiments is synchronised but asymmetric over the cycles. In particular, primary market sentiments usually contribute more to the up-cycle of the overall market sentiments rather than the down-cycle, such as the recent years since 2014 and the period between 2003 and 2005. One possible explanation is that property developers might adjust their sales pace to search for the best prices according to the market condition. For instance, they would try to raise their unit prices and accelerate the sale pace when the market is booming, while slowing

[11] Larsen, Thorsrud and Zhulanova (2020) have analysed such divergence in the context of inflation expectation forecasts by newspapers versus those made by professionals. They argue that these forecasts have different information contents.

[12] Before the implementation of Residential Properties (First-hand Sales) Ordinance in 2013, there was no standard record of primary market information (e.g. price list, quantity of sales in each launch) provided by property developers and the official sale and purchase records in Land Registry were only available after the transactions were completed.

the sale pace when the market is cooling.

Figure 2: Lexicon-based property market sentiment index with contributions

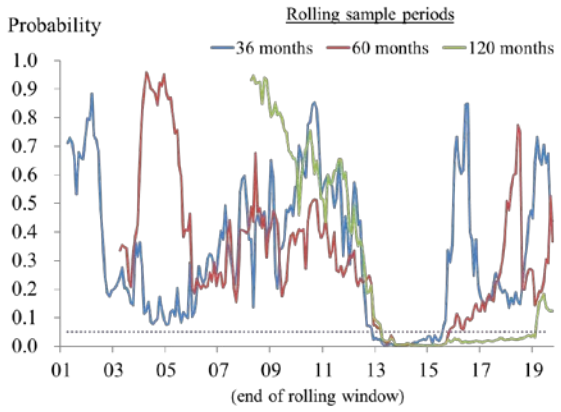


Source: Staff estimates.

Besides the contributions, our market-specific sentiment indices can also help shed some light on the lead-lag relationship between the primary and secondary market sentiments. We run a series of Pairwise Granger Causality tests with different rolling sample windows. Figure 3A illustrates the probabilities of primary market sentiments not Granger causing secondary market sentiments. We can see the probabilities went below 5% during 2009-2015 in different sample windows (Table 2). Meanwhile, the Granger Causality tests results of secondary market sentiments on the primary market sentiments are mixed (Figure 3B). These observations suggest that primary market sentiments tend to lead secondary market sentiments during the period of low housing supply (Figure 4). This can be explained by the anchoring effect between the two markets. Given the limited new launches, both buyers and sellers would naturally focus on the primary market condition. In addition, Leung and Tang (2015) and Leung et al. (2020a) argue that Hong Kong primary housing market can be considered as an oligopolistic market as it is dominated by a few real estate developers. This could form a relationship between price leaders (major developers in the primary market) and price followers (e.g. individual house-owners who try to sell their housing units in the secondary

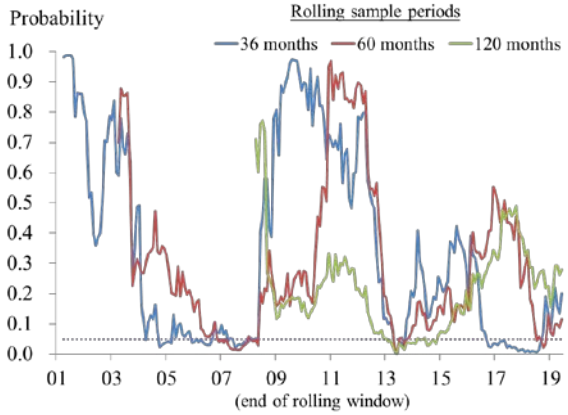
market). Therefore, primary market sentiments could be easily spilled over to secondary market sentiments. In recent years, the lead of primary market sentiments has weakened. One of the reasons is that the anchoring effect diminished gradually when the housing supply increased from 2014. The oligopolistic power of major developers might also have weakened as more small-and-medium developers entered the market.[13] Another possible explanation is that the primary market has become more regulated and transparent after the implementation of Residential Properties (First-hand Sales) Ordinance in 2013. Before 2013, newspapers were the only source for market participants to understand the primary market and the sentiments reported on the news would weigh heavily on their assessments. Since then, more information (e.g. prices and number of launches) on the primary market has become available, and this weakened the influence of primary market sentiments.

Figure 3A: Probability of primary market sentiments does not have Granger Causality on the secondary market



* Granger Causality tests with the lag of 3 months
Source: Staff estimates.

Figure 3B: Probability of secondary market sentiments does not have Granger Causality on the primary market



* Granger Causality tests with the lag of 3 months
Source: Staff estimates.

[13] According to the land sales record from Lands Department, more than 80 residential land sites were purchased by non-major property developers (i.e. not the groups of Cheung Kong, Henderson, New World Development, Sun Hung Kai, Sino and Wheelock) between 2013-2018, compared with around 30 land sites between 2007-2012.

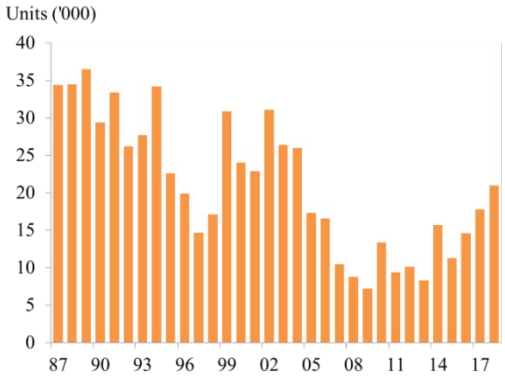
Table 2: Periods of primary market sentiment index significantly leads^ secondary market sentiment index

Rolling window of Granger Causality tests	Periods
36 months	Dec 2009 – Jul 2015
60 months	Apr 2008 – Oct 2015
120 months	Apr 2003 – Feb 2019

^Rejection of no causality at 5% significant level.

*Granger Causality tests with the lag of 3 months.
Source: Staff estimates.

Figure 4: Private housing completions



Source: Transport and Housing Bureau.

While our news-based sentiment index can generate new insights and compare favourably against the existing survey-based indices, our news-based textual analysis also has some limitations. As the word choice of the newspapers may evolve over time, it is possible that some trendy or outdated words are not captured by our dictionaries. Moreover, unlike machine learning, our lexicon-based identification may overlook the information hidden in the sentence structure, which could be essential in interpreting the content (e.g. a double negative statement[14]). Nevertheless, the potential bias is expected to be small, as the volume of our sample news articles is large enough, while local journalists usually prefer to use simple sentence structure in their news articles.

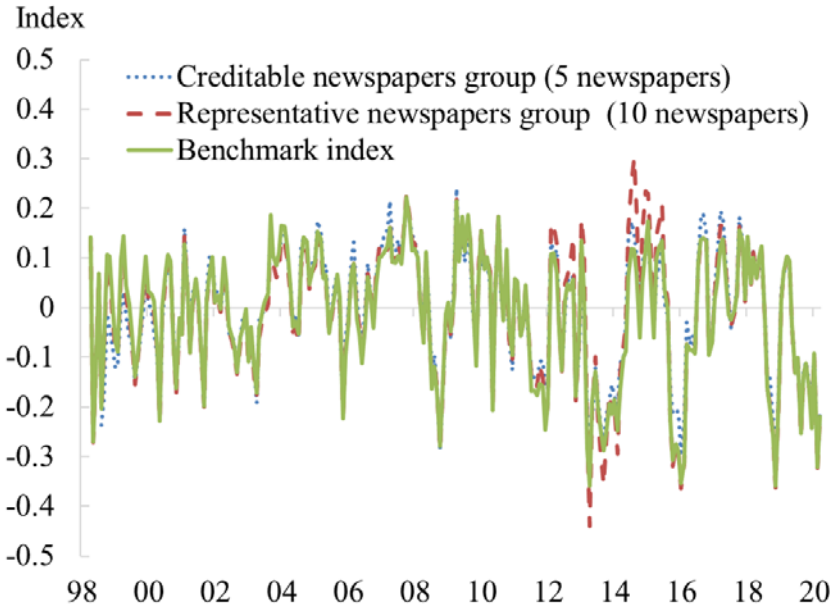
3.3 Robustness of the news-based property market sentiment index

To further test the robustness of our property market sentiment index, we conduct two validity checks on our benchmark sentiment index regarding our newspaper selection and the potential bias on word selection in our sentiment dictionary.

[14] For example, the word “bad” should be considered as a negative term. But if we add the word “not” before it, then the overall meaning of the sentence would be completely different.

Although all Chinese newspapers published in Hong Kong[15] are covered in our benchmark sentiment index, there is a concern whether some of the newspapers have credibility or readership issues in transmitting market sentiment. To show that our benchmark index does not distort by the component of publications, we follow Luk et al. (2020) method and re-compute the property market sentiment indices from two paid newspapers groups, i.e. creditable group[16] and representative group[17]. We can see that the sentiment indices from two newspapers groups are very similar with the benchmark index (Figure 5) and they have high correlation (0.97 and 0.98 respectively), indicating our benchmark index is robust to alternative sources of newspapers.

Figure 5: Benchmark index and indices from paid newspapers groups



Source: Staff estimates.

Next, we investigate whether or to what extent sentiment word selection may affect our index. As we did not perform an extensive audit study on our news searches like Baker et

[15] See Table A.1 for the list of Chinese newspapers included in our sample.

[16] We choose five most creditable newspapers (i.e. Sing Pao, Ming Pao, Hong Kong Economic Journal, Sing Tao Daily, and Hong Kong Economic Times) based on the long-term average rating in Public Evaluation on Media Credibility Survey. See the survey results of Public Evaluation on Media Credibility http://www.com.cuhk.edu.hk/ccpos/en/research/Credibility_Survey%20Results_2019_ENG.pdf.

[17] We choose ten paid newspapers, which are the constitutes of Hong Kong Economic Policy Uncertainty Index in Luk et al. (2020) paper, as a representative group. They are Wen Wei Po, Sing Pao, Ming Pao, Oriental Daily, Hong Kong Economic Journal, Sing Tao Daily, Hong Kong Economic Times, Apple Daily, Hong Kong Commercial Daily, and Ta Kung Pao.

al. (2016) did, it could raise a concern of sentiment identification. We resample the words randomly in different sizes (10%, 50% and 90% of total sentiment words in our dictionary) and draw 10,000 times to construct sentiment indices. Table 3 summarises the results in terms of correlations.

Table 3. Correlation between the benchmark and sample sentiment indices

Sample size (% of total sentiment words in the dictionary)	Correlation between the benchmark and sample sentiment indices (percentile)						
	1st	5th	10th	50th	90th	95th	99th
10%	0.824	0.861	0.881	0.978	0.793	0.732	0.653
50%	0.931	0.957	0.968	0.983	0.946	0.930	0.897
90%	0.979	0.983	0.984	0.982	0.976	0.973	0.966

In general, the results show the sample indices are highly correlated with the benchmark, even the number of sentiment words is small (e.g. 50th percentile of 10% sample of sentiment words). It suggests some omissions of sentiment words in our dictionary has a negligible impact on our index. Moreover, the improvement in correlation is small when we expand the sample size of dictionary from 50% to 90%, even at the extreme percentiles (e.g. 95th and 99th). It might indicate that expanding our current dictionary would not affect the benchmark sentiment index much as most of the words have been covered.

IV. MEASURING PROSPECTIVE BUYERS' INCENTIVES IN HONG KONG

In this section, we describe how textual analysis may help to measure prospective residential property buyer's incentives in Hong Kong. There is no existing measure of property buyer's incentives. We follow the literature (Kulkarni et al. (2009); Dietzel et al. (2014); Wu and Brynjolfsson (2015)) to measure buyers' incentives using Google Trends data.

Google Trends does not report the raw level of queries for a given search term. Rather, it reports a Google search volume index (GI), which uses an unbiased sample[18] of Google

[18] As Google Trends data is a random sample of all searches, the time series could deviate slightly each time it is extracted. To ensure the representativeness of our sample, we perform the data scrapping process repeatedly and compare results obtained from each sample. Robustness checks suggest that all of these individual samples yield similar results.

search data. Each data point is divided by the total searches of the geography and time range it represents to compare relative popularity. The resulting numbers are then scaled on a range of 0 to 100 based on the topic's proportion to all searches on all topics. GIs are available at a real-time basis.

The key identifying assumption is that a stable fraction of prospective property buyers would search for information on Google when they plan for property purchases. Indeed, Henderson and Cowart (2002) show that visitors of residential real estate agent websites make extensive use of the internet before making a purchase. Van Dijk and Francke (2018) provide a theoretical model and empirical analysis to relate internet search and housing market activities. The major advantage of Google Trends is that the search volume data is available on a real-time basis with no publication lag, which enables us to monitor the market more closely.

4.1 Methodology

A major challenge in constructing a keyword-based index is identifying the set of queries that are representative. A number of studies in the literature (e.g. Askitas and Zimmermann (2009); Fondeur and Karame (2013); Francesco (2009)) select queries in an ad hoc fashion, but the process is subjective and there is no guarantee that the set of selected queries captures the information in question better than alternative sets of queries. The index constructed based on ad hoc query selection is unlikely to be optimal in a statistical sense. Another disadvantage is that for some variables of interest, the researcher may not have a clear idea what the most relevant queries are. This shortcoming is particularly problematic in a bilingual city like Hong Kong, where people often mix in phrases from different languages when performing internet searches. This can lead to poor performance of the resulting index. With this in mind, we develop a robust algorithm to identify the optimal queries combination based on statistical criteria. The algorithm can be broken down into two steps:

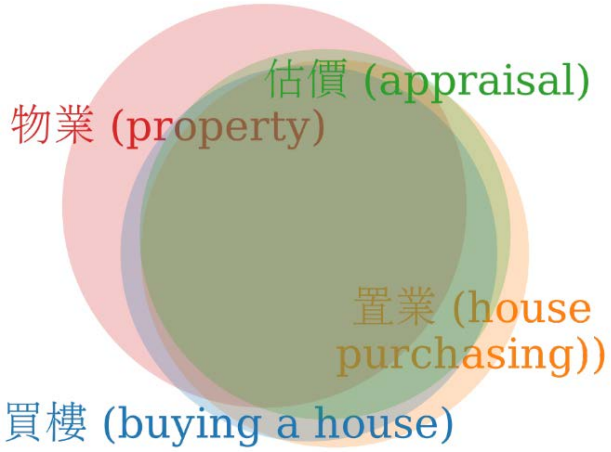
4.1.1 Queries selection

There are literally thousands of phrases a potential buyer may search when doing research for a property purchase. We select a representative subset of queries systematically using the 'related queries' feature provided by Google Trends (Yang et al. (2014)). 'Related

queries' lists what other queries users also search in the same search session, in descending order of average search intensity. This allows us to derive a list of popular phrases that are associated with property purchase using just one single seed word.

To evaluate how the choice of seed words may affect the outcome, we generate four lists of related queries using four generic seed words in Chinese – '置業' (house purchasing), '買樓' (buying a house), '估價' (appraisal) and '物業' (property). For each list, the top 100 related queries are retained. Figure 6 shows a Venn diagram of the four lists and reveals that the vast majority of the queries on the four lists are overlapping. We combine these lists and arrive at a single list with 131 queries.[19]

Figure 6: Venn Diagram of related queries derived from the four seed words



Source: Google.

Figure 7 shows these queries in a word cloud with their search intensities reflected by the size of the words. By inspection, the selected queries include names of major property agencies and banks in Hong Kong and other generic terms closely related to the property market. Given the high degree of overlap of the keywords, we believe that the use of these seed words are quite general, and that using other related seed words will give similar queries.

[19] In other words, only 10% of queries from each list is non-overlapping on average.

Figure 7: Word Cloud of property market-related queries derived

Note: Sizes of the queries reflect their average search intensities.
Source: Google.

4.1.2 Combination optimisation

The second step of the algorithm examines the forecasting power of the GIs on property price movement. The spirit of this optimisation is similar to Ginsberg (2009), in which the authors run linear regressions with each query and shortlist the ones with the best fits. We go one step further by exhausting as many combinations as possible using an iterative method.

In the first iteration, we run a benchmark model as follows:

$$g_{R\&VD,t} = \alpha_1 + \gamma_1 g_{R\&VD,t-1} + \varepsilon_{1,t} \qquad (3)$$

The dependent variable is the month on month growth of the domestic residential property price index, compiled by the R&VD.[20] The level of the R&VD index has a unit root, so its first difference is used instead.[21] This is just an AR(1) model. The inclusion of its first lag is

[20] The R&VD housing price index is based on the actual transaction prices reviewed by R&VD for collecting rent and rates purposes. It is released on a monthly basis with a 1-month public lag. Although the index does not include transactions of primary sales, it is representative of the overall market as secondary market transactions account for about 70% of total market transactions in Hong Kong. See https://www.landreg.gov.hk/en/monthly/agreement.htm.

[21] Similar to other commonly-used approaches for handling unit root, first differencing is subject to drawbacks. In particular, Baxter (1994) suggests that the approach could make uncovering correlations between variables at

based on the Bayesian information criterion (BIC).

Next, we add i ∈ {1,2,…131} queries one by one as follows:

$$g_{R\&VD,t} = \alpha_i + \gamma_i g_{R\&VD,t-1} + \beta_i g_{i,t} + \varepsilon_{i,t} \quad (4)$$

The dependent variable $g_{i,t}$ is the month on month growth of the three-month moving average of the Google search volume index (GI) of *query i.*[22] Since the R&VD Index is released with a month lag whereas the GI is available at real time, Equation (4) nowcasts the movement of the R&VD Index using (i) its lag value and (ii) the GI of *query i*, available at time *t.*[23] Each of these models is then compared against the benchmark model based on its explanatory power and only those that outperform the benchmark model are retained (63 queries in total).

We run the second iteration for each retained query $b \in \{1,2,\dots 63\}$ as follows: for each of the remaining 130 queries *i*, we run a regression using the first principal component of (i) its month on month growth (g_b), and (ii) the month on month growth of the remaining 130 queries (g_i):

$$g_{R\&VD,t} = \alpha_i + \gamma_{bi} g_{R\&VD,t-1} + \beta_{ij}\ PC_1(g_{b,t}\ \&\ g_{i,t}) + \varepsilon_{ij,t} \quad (5)$$

After running the above iteration for all $b \in \{1,2,\dots 63\}$ and $i \in \{1,2,\dots 131\}$, each model is compared against its own benchmark model (without the new query) and only those that outperform the benchmark are retained. After that, all models are pooled together, and only the top 1,000 combinations of queries that produce the highest explanatory power are retained for the next round of iteration.[24] (See Figure 8.) As the above process repeats, the marginal gain in explanatory power of including additional queries gradually diminishes. We stop the

low frequencies difficult. For this reason, the literature that focus on the medium/long-run relationship of housing market variables (e.g. Leung and Ng, 2019) often opt for the band-pass filter, Hodrick-Prescott filter or other detrending methods. As these methods are also subject to criticism (e.g. subjective smoothing parameter, potential spurious problem (Schüler, 2018), this study opts for the first-difference approach over others for its simplicity.

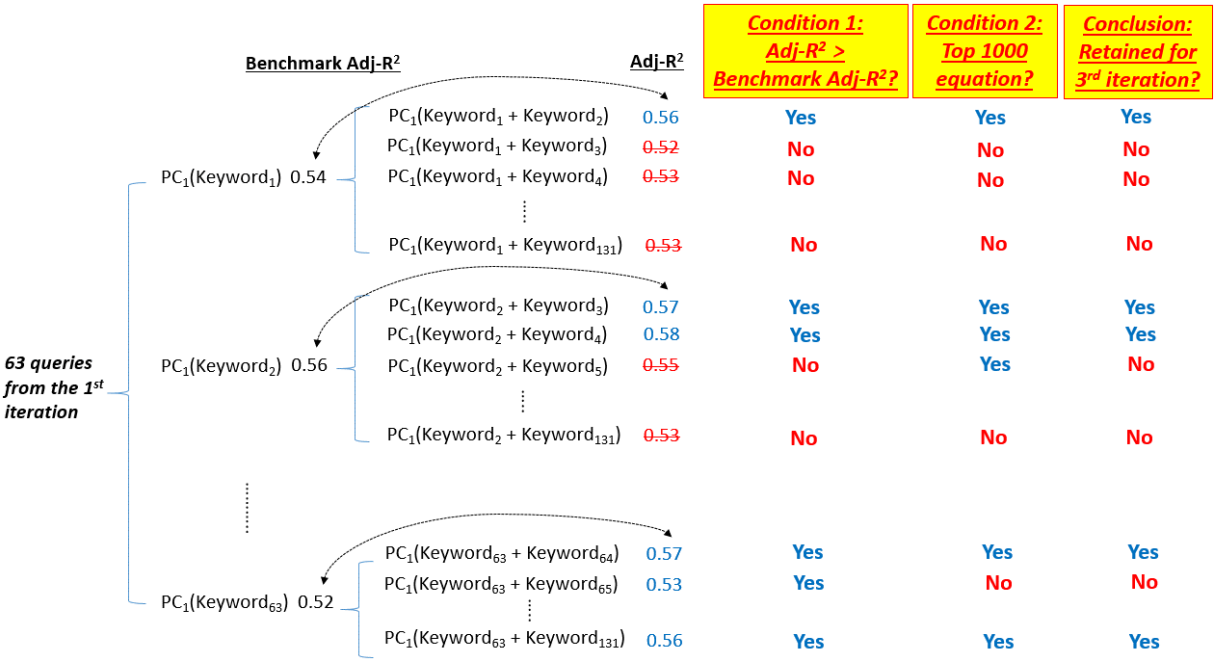
[22] As the raw Google Trends data is highly volatile, its 3-month moving average is used instead to reduce the noise.

[23] BIC is used to determine the optimal lag structure of the model. With the maximum lag length set at 6, BIC suggests that for 92% of the search queries, their models are optimal when only the contemporaneous GI is included as the explanatory variable.

[24] The restriction on number of combinations to 1,000 is imposed due to the bottleneck arising from computer processing power.

iteration process when the improvement is sufficiently small.

Figure 8: Selecting the optimal combination through iterations



4.2 Empirical results

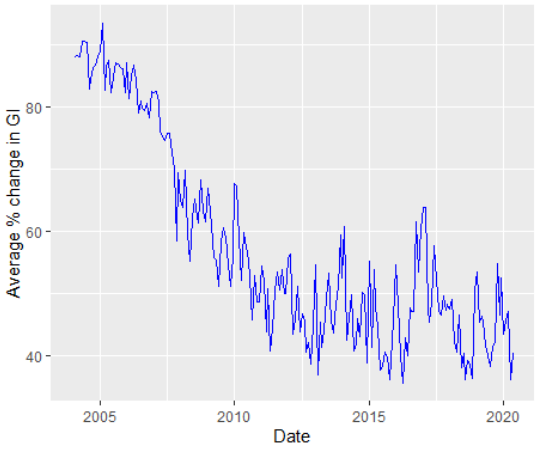
Before we run the algorithm, one decision needs to be made regarding the sample period. Since GI is based on a random sample of the search data, its value for certain query could be highly volatile if the underlying search volume is too small. This is particularly the case in earlier years where doing property research on internet was not as prevalent as we see today, which violates our key assumption. As the overall search volume is unavailable to us, we design a measure Vol_t to gauge the point where the volatility of GIs stabilised:

$$Vol_t = \frac{\sum_{i=1}^{n} vol_{i,t}}{n} \text{ , where } vol_{i,t} = \begin{cases} \left|\frac{\Delta GI_{i,t}}{GI_{i,t-1}}\right| & if \;\; GI_{i,t-1} \neq 0 \\ 1 & \text{otherwise.} \end{cases} ; \qquad (6)$$

By construction, vol_i captures the variance of GI_i across time, which theoretically varies inversely with the search volume of *query i*, ceteris paribus. Since a $GI_{i,t-1}$ of value 0 would

imply extreme low search volume and could turn $\frac{\Delta GI_{i,t}}{GI_{i,t-1}}$ into either 0 or infinity, we further assign a value of 1 to such scenario to avoid the overall volatility being understated.

Figure 9: Volatility of GIs across time



Visual inspection of Figure 9 suggests that the volatility of GIs declined sharply in the early years and then gradually stabilised around 2010. As such, we set 2010 as the start date of our sample period.[25]

The combination optimisation algorithm is run using adjusted-R-squared as the measure for explanatory power.[26] It stops when the improvement of adjusted-R-squared in an iteration is smaller than 0.001. Figure 10 shows the way adjusted-R-squared improves over the iterations. Our algorithm returns the optimal model after 17 rounds of iterations and the combination was formed by 17 search queries.[27] We set the first principal component of the corresponding queries as Google Buyer Incentive Index (GBII). We then compare GBII's performance with the baseline model, and an augmented model where Centa-City Leading Index (CCLI) is used as a regressor.[28]

[25] As a robustness check, we have also re-run the analysis using 2009 and 2011 as the start date. The results are similar to the one reported here.
[26] We also try other measures (RMSE, AIC, BIC) as a robustness check, and the results are broadly similar.
[27] Please refer to Table A.4 in the Appendix for the list of search queries used in compiling GBII. We also compare the queries selected by algorithm with those selected by a Bayesian variable selection model. See Appendix A.3 for details.
[28] The Centa-City Leading Index (CCLI) is a market index that captures secondary housing prices. It is a

$$g_{R\&VD,t} = \alpha_1 + \gamma_1 g_{R\&VD,t-1} + \varepsilon_{1,t} \quad \text{– Baseline model}$$

$$g_{R\&VD,t} = \alpha_2 + \gamma_2 g_{R\&VD,t-1} + \beta_1 g_{GBBI,t} + \varepsilon_{2,t} \quad \text{– GBII model}$$

$$g_{R\&VD,t} = \alpha_3 + \gamma_3 g_{R\&VD,t-1} + \beta_2 g_{CCLI,t} + \varepsilon_{3,t} \quad \text{– CCLI model}$$

Figure 10: Maximum adjusted-R-squared and number of variables

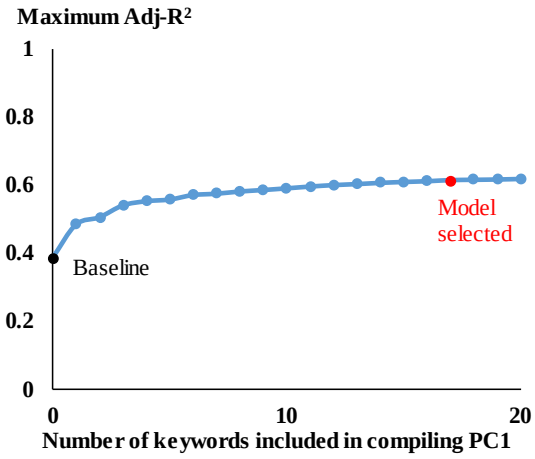


Table 4: Comparison of regression results

	Baseline	GBII	CCLI
t-stat (β_1 / β_2)	N/A	**8.5848**	8.4848
Adjusted-R-squared	0.3884	**0.6156**	0.6122
RMSE	0.0120	**0.0095**	0.0095
AIC	-745.43	**-802.50**	-801.41
BIC	-736.95	**-791.19**	-790.10
Observations	125	125	125

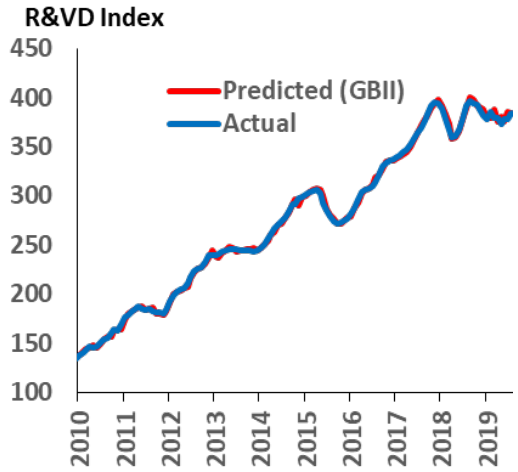
Note: Sample period runs from Jan 2010 to May 2020.

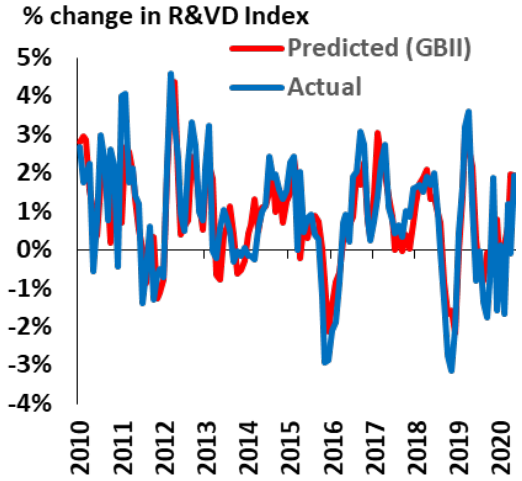
As shown in Table 4, the coefficient of GBII is significant, and the adjusted-R-squared is much higher compared with the baseline. Other performance measures (RMSE, AIC and BIC) also suggest notable improvements using GBII model. To our surprise, the performance of GBII is even on par with CCLI, a well-recognised price index that is widely used by the market in tracking housing price. Figure 11A and 11B show GBII tracks closely to the R&VD housing price index and its month-on-month changes. Given that CCLI is still subject to certain publication time lag, GBII could potentially help supplement the CCLI in tracking the turning points in Hong Kong housing cycle.

Figure 11A: R&VD housing prices index and predicted price index by GBII

Figure 11B: Monthly changes in R&VD Index and GBII

weekly index measuring the changes in market value of constituted estates, which is estimated by Centaline Property Agency Limited. As the index reflects the price changes on provisional sales records (rather than the transaction prices recorded in the Land Registry), it is often considered a leading indicator of R&VD Index. For details, see http://www1.centadata.com/cci/notes_e.htm.





To check the out-of-sample performance of the GBII, we further split the data into an in-sample period (90% of data) and an out-of-sample period (10% of data). After using the in-sample data as our training set, we put the fitted model into a test and see how well it performs in forecasting the out-of-sample data. Table 5 shows the result. The GBII model outperforms the baseline model by a significant margin, with MAE and RMSE decreasing by 23% and 20% respectively in the out-of-sample testing.

Table 5: Out-of-sample testing

	Baseline	GBII	CCLI	Improvement
MAE (out of sample)	0.0150	0.0116	0.0103	23%
RMSE (out of sample)	0.0172	0.0138	0.0124	20%

Note: In-sample period runs from Jan 2010 to May 2019; out-sample period runs from Jun 2019 to May 2020.

To sum up, we find that Google search queries provide valuable information about buyer sentiments in Hong Kong and can serve as a nowcasting tool in tracking the domestic property market. Nevertheless, our index is also subject to some caveats arising from using Google Trends. For example, it may not capture a demographically representative sample of prospective property buyers in Hong Kong, as internet usage varies widely across the population.

V. TRANSMISSION CHANNELS OF MARKET SENTIMENTS AND BUYERS' INCENTIVES TO THE HONG KONG HOUSING MARKET

Having constructed the news-based property market sentiment index and the Google Buyer Incentive Index, in this section we analyse the interactions between market sentiments and housing market variables in Hong Kong. In particular, we study the extent to which macroeconomic shocks transmit to the housing market through sentiment channels.

We adopt a Structural Vector Autoregressive (SVAR) model. A representation of the SVAR is:

$$B_0 X_t = c + B_1 X_{t-1} + B_2 X_{t-2} + \cdots + B_p X_{t-p} + \Xi S_t + \epsilon_t \qquad (7)$$

where c is a vector of constants, $B_0, B_1, \ldots B_p$ are coefficient matrices, and ϵ_t is a vector of structural innovations. The vector X_t contains the following endogenous variables: (1) Volatility Index of Hang Seng Index (*vhsi*); (2) Hang Seng Index (*hsi*); (3) average mortgage rates (*mort*); (4) Purchasing Manager Index (PMI) (*pmi*); (5) unemployment rate (*un*); (6) secondary market sentiments index (*sec*); (7) Google Buyer Incentive Index (*gbii*); (8) transaction volume in the secondary market (*tran*); and, (9) secondary market housing prices (*pp*).[29] We use housing prices and transaction volume of the secondary market only because these match the official definition of the housing price index. Moreover, we include an exogenous step function of policy variables, S_t, control for the policy effects from the demand management and macro-prudential measures of the Government and the HKMA during the sample period.[30] The SVAR model is estimated using monthly data from January 2010 – March 2020.[31] The lag length is set to one as suggested by the Schwarz Information Criterion.

Identification of the structural shocks is achieved using a standard Cholesky decomposition, with the ordering of the variables given above.[32] The stock market index and the volatility index are ordered first in the SVAR, given its responsiveness to all sorts of news including global financial market shocks. The average mortgage rate is ordered third, as it is

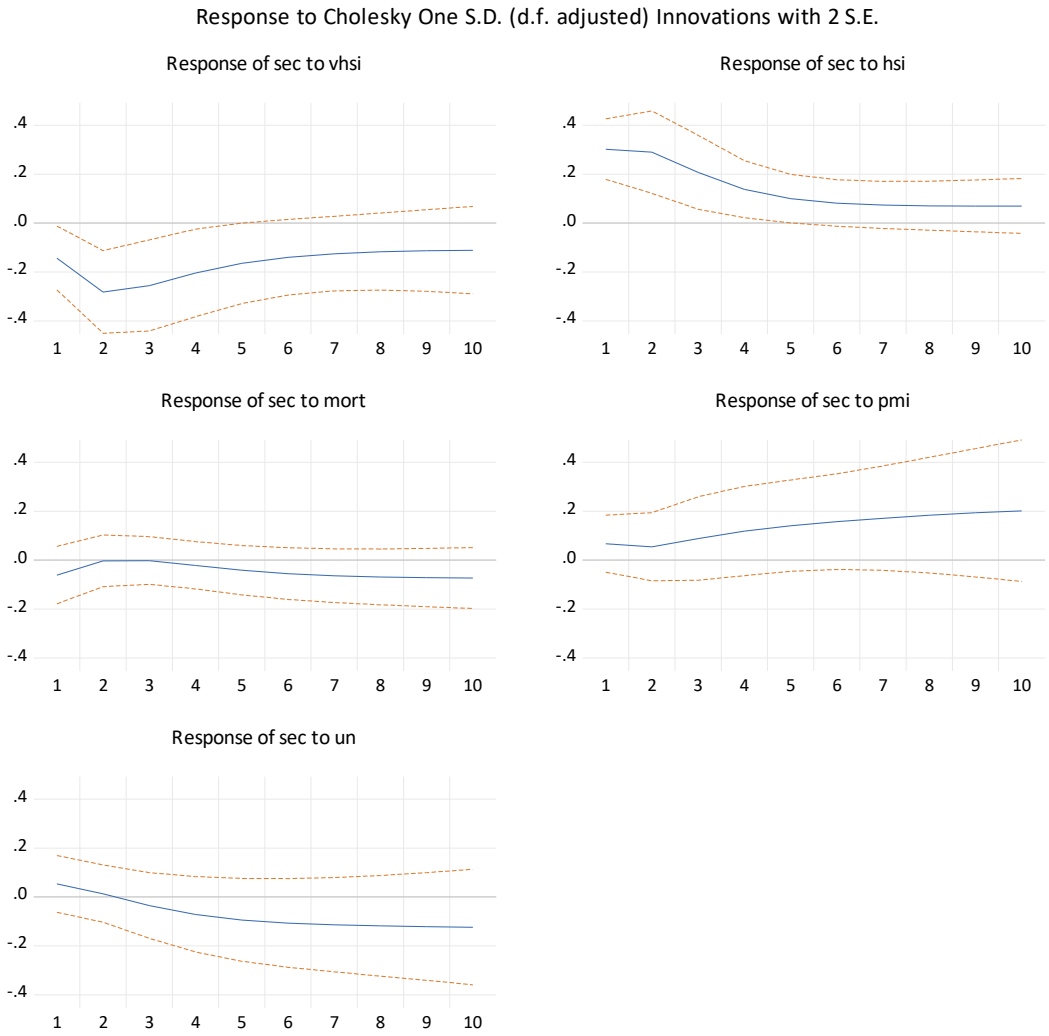
[29] The Hang Seng index, housing prices and transaction volume are stationarised by first-differencing.
[30] We follow HKMA (2014) and He (2014) to construct a step function of demand management and macro-prudential measures by increasing "count" for each round of new measures.
[31] As the pandemic of COVID-19 causes an unprecedented effect on the global and local economic condition, some possible structural changes in economic dynamics has been seen (e.g. semi-lockdown and social distancing measures) since March 2020. To avoid the distortion of our analysis, we end our estimation sample period at March 2020.
[32] As a robustness check, we have also tried generalised impulse response instead of Cholesky decomposition. The impulse responses are similar.

determined by both domestic banking liquidity conditions and the US policy rate under the Linked Exchange Rate System. Real economic indicators (PMI and unemployment rate) are ordered next and are included to represent the purchasing power of domestic households. After the macro-financial variables, we have our secondary property market sentiment index and our Google Buyer Incentive Index. Our rationale is that buying incentives are likely affected by property market news updates but not vice versa. Housing prices and transaction volumes are ordered last as they are influenced by the macro-financial conditions, market sentiments and buyers' incentives.

Figure 12 first shows the impulse responses of market sentiments to exogenous innovations in different macro-financial variables. In general, market sentiments strengthened when stock prices increased or stock market volatility declined, while market sentiments were not responsive to other shocks[33]. It suggests that the equity market performance is the key driver of property market sentiments. The impulse response also shows that the transmission of equity price shock to property market sentiments was quick and the impact would last for at least one quarter.

[33] The insignificant impulse responses of unemployment rate and average mortgage rate might be due to the low volatility of those macro variables after the GFC.

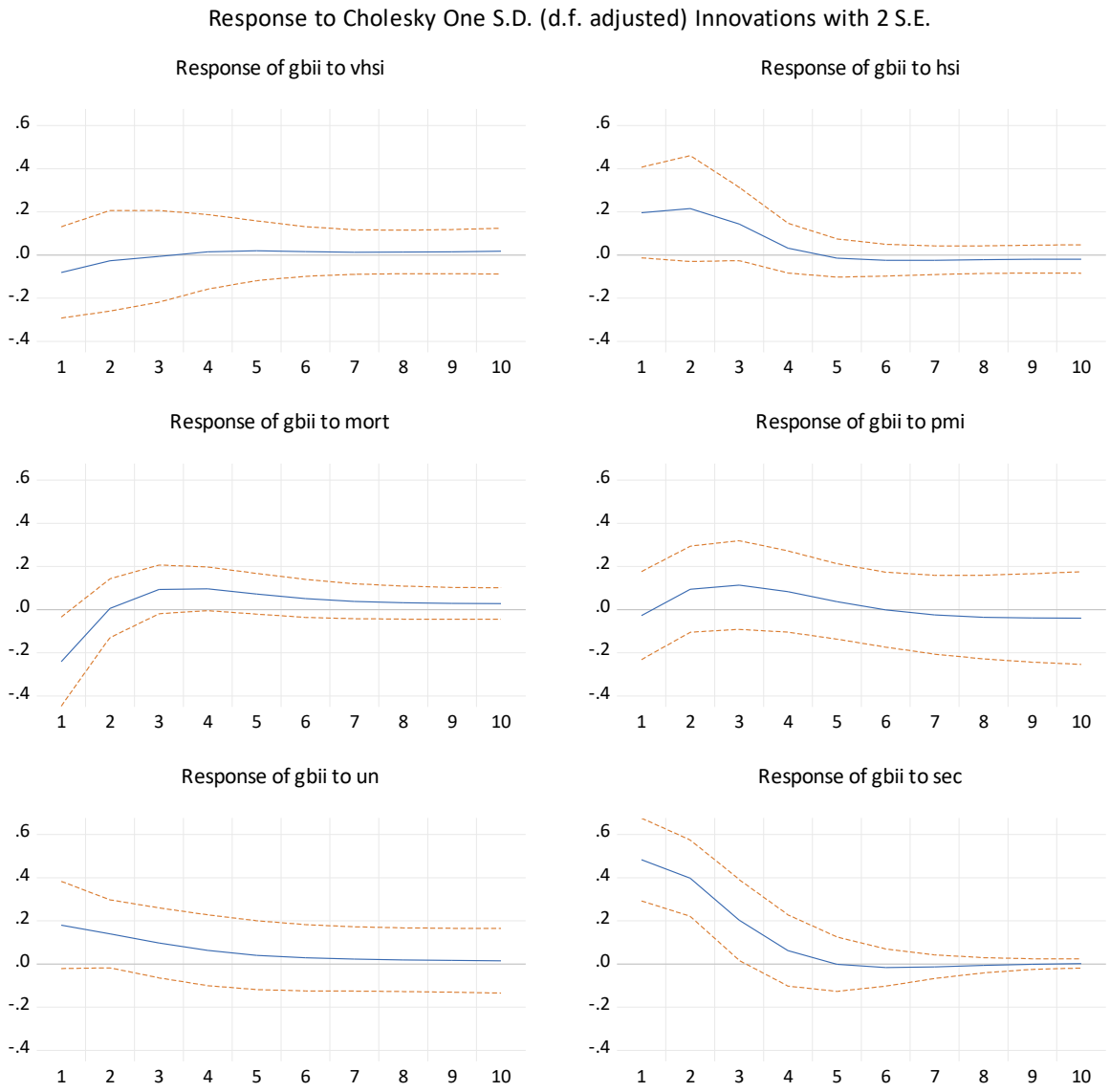
Figure 12: Cholesky impulse responses to one standard deviation innovation to the secondary property market sentiment index



Source: Staff estimates.

Figure 13 shows the impulse responses of buyers' incentives to innovations in macro-financial variables and market sentiments. In line with our expectation, buyers' incentives increased when market sentiments improved, and the impact would last for one quarter. The decline in mortgage rates would also stimulate buyers' incentives, whereas the other macro-financial variables do not have any significant impact. This might reflect that the mortgage rate is a common factor for people in considering to buy a flat.

Figure 13: Cholesky impulse responses to one standard deviation innovation to the Google Buyer Incentive Index

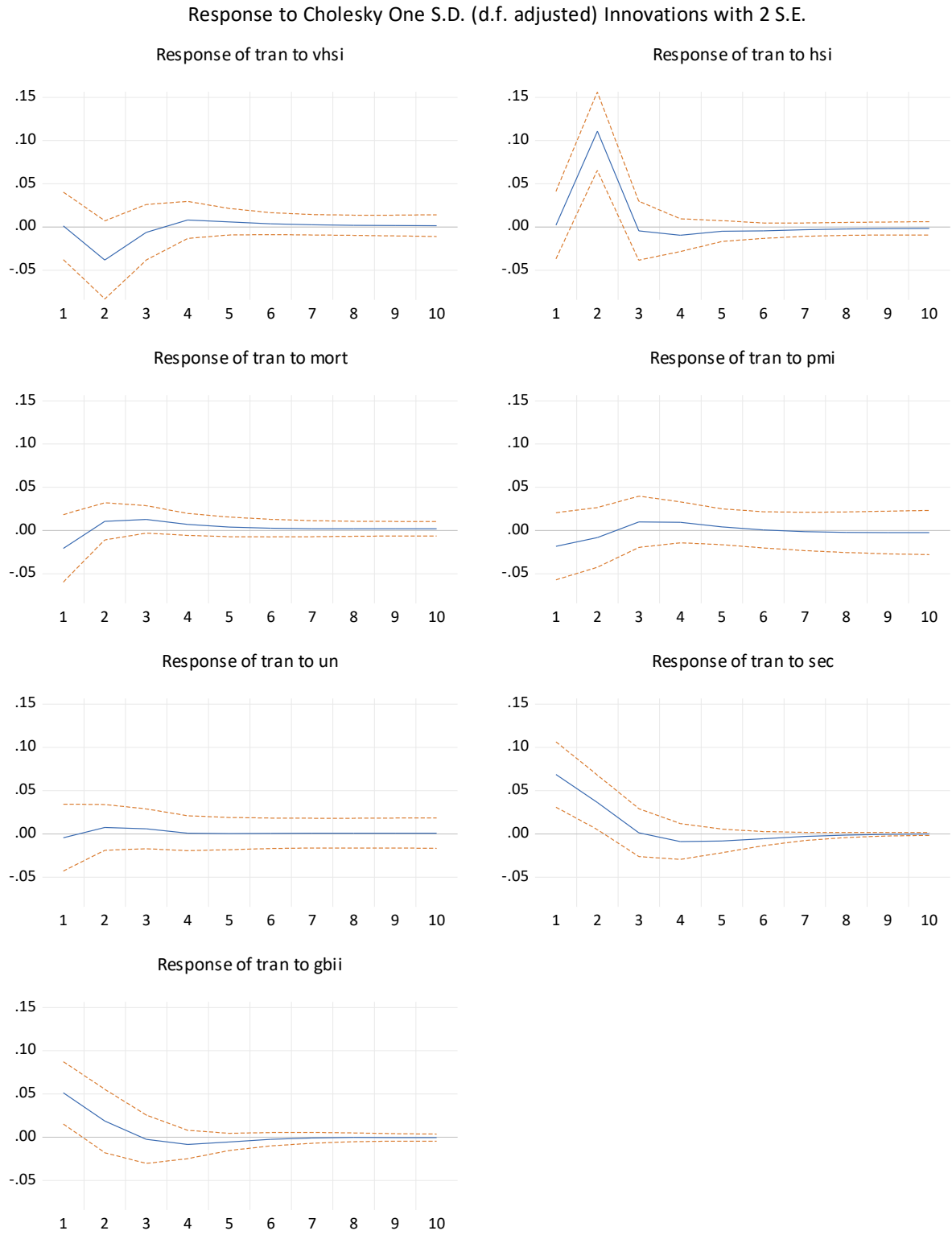


Source: Staff estimates.

To see the overall impacts of the housing market, Figures 14A and 14B show the impulse responses of transaction volumes and price to innovations in different variables. Similar to the previous studies, an improvement in the equity market would stimulate both housing prices and transactions. The volatility of the equity market could also have a negative impact on housing prices, but the impact was not significant for the transactions. On the other hand, both market sentiments and buyers' incentives have positive impacts on housing prices and transactions. These suggest that there is a sentiments channel of transmission in the housing market dynamics in Hong Kong, and this channel is additional to the standard macro fundamental channel.

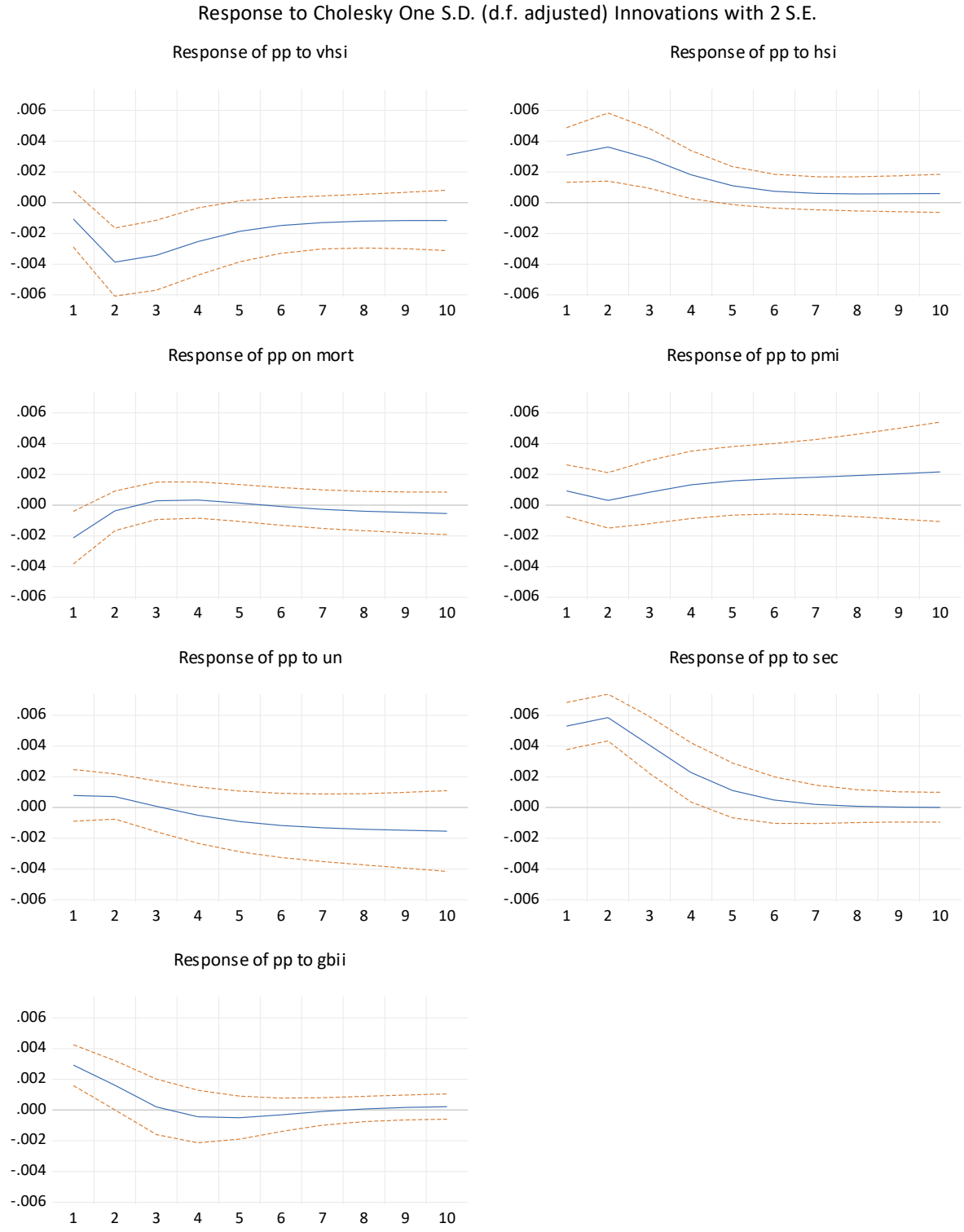
Figure 15 shows the variance decomposition of news-based secondary property market sentiments index, GBII, secondary transaction volumes and housing prices, illustrating the relative importance of different shocks. The result indicated that the market sentiments contributed 20% of the variation in GBII, more than 30% of the variation in prices and around 9% of the variation in transactions at a 10-month horizon. What this result implies is that, for example, exogenous changes in market sentiments could have a significant impact on individual buyer's incentive, as well as the housing prices through the sentiments channel. Meanwhile, the equity market performance (e.g. HSI and VHSI) is another contributor to short-term housing market dynamics, either directly on prices and transactions or indirectly through the sentiments channel. The buyer's incentive shock also contributes around 4-5% on housing prices and transactions. Given that home purchase is one of the biggest lifetime investment decisions, it is expected that buyers' incentives shock would account for some variations on housing prices and transactions.

Figure 14A: Cholesky impulse responses to one standard deviation innovation to the secondary housing transaction volumes



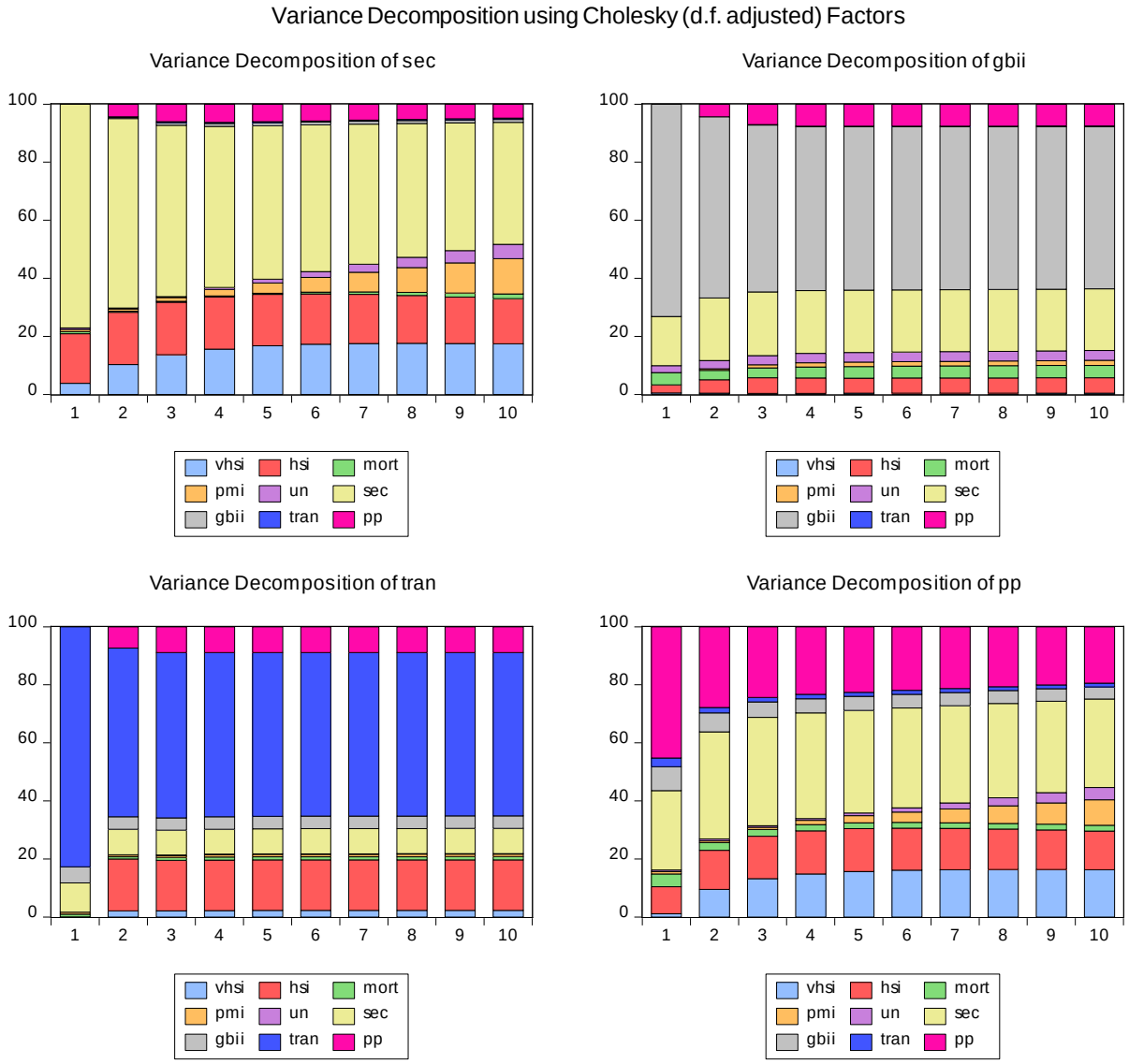
Source: Staff estimates.

Figure 14B: Cholesky impulse responses to one standard deviation innovation to the secondary housing prices



Source: Staff estimates.

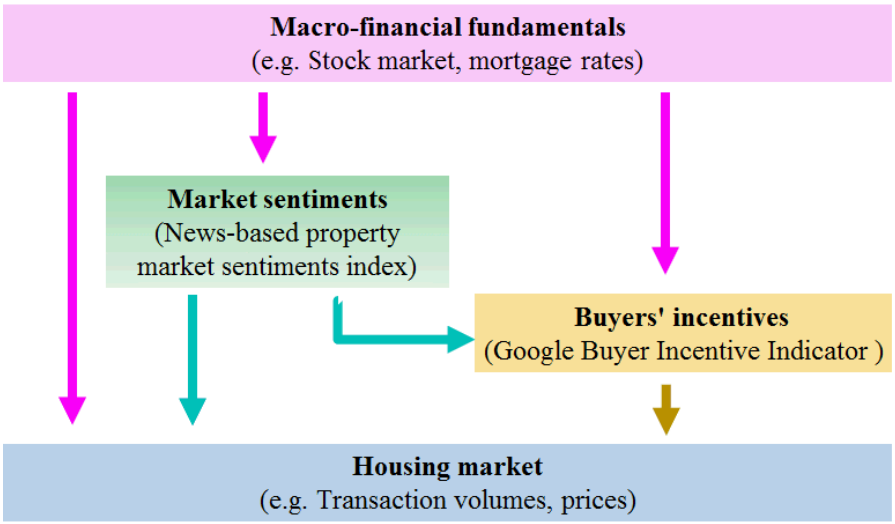
Figure 15: Contributions to the variations in secondary market sentiment index, GBII, secondary housing transactions and prices (percentage share)



Source: Staff estimates.

Based on the variance decomposition analysis, Figure 16 summarises the transmission mechanism in the Hong Kong housing market into a flow chart. Given the importance of the housing market to the Hong Kong economy, our findings suggest that market sentiments and buyers' incentives, in addition to macro-financial channels, can affect housing prices and transaction volumes through the sentiment channels. Therefore, for macro-financial surveillance purpose, it is important to track both of them closely.

Figure 16: Transmission mechanism with different channels in the housing market



VI. CONCLUDING REMARKS

This paper uses the novel approach of text mining to measure market sentiments and its transmission to the Hong Kong property market. We find that our news-based property market sentiment index is intuitive and is able to differentiate the sentiments in the primary and secondary markets, with primary market sentiments leading secondary market sentiments during the period of low housing supply. For our Google Buyer Incentive Index, we find that it has value-added in forecasting (or nowcasting) the official housing price index. In line with theories, we also find that negative property market sentiments would dampen buyers' incentives, which would then affect property prices and transaction volumes. Altogether, our paper suggests that market sentiments, as well as buyers' incentives can supplement existing property market indicators in identifying turning points in the property market cycle.

A. APPENDIX

A.1 Construction of the news-based property market sentiment index

This appendix describes in detail the construction of the news-based property market sentiment index. The universe of newspapers used in this exercise is all the Chinese news media in Hong Kong for the period April 1998 – June 2020, which is listed in Table A.1. Data is available in Wisers Information Portal, a digital newspaper archive of Chinese news media.

Table A.1: Chinese newspapers in Hong Kong

English Name	Chinese Name
A Daily	A 報
Apple Daily	蘋果日報
China Daily Hong Kong Edition	中國日報香港版
Express Post	快線周報
Goodnews	好報
HK iMail	香港郵報
HK01 Newspaper	香港 01 周報
Harbour Times	港報
Headline Daily	頭條日報
Headlinefinance	頭條財經報
Hong Kong Commercial Daily	香港商報
Hong Kong Daily News	新報
Hong Kong Economic Journal	信報財經新聞
Hong Kong Economic Times - Money Times	香港經濟日報 - 投資理財
Hong Kong Economic Times - Property Times	香港經濟日報 - 置業家居
Hong Kong Economic Times	香港經濟日報
Hong Kong Globe	公正報
Hong Kong Health Journal	香港健康報
I.T. times	資訊科技周刊
Invest Weekly	投資一週
Kowloon Post	龍週
Kung Kao Po	公教報
Lion Rock Daily	香港仔
Metro Daily	都市日報
Ming Pao Daily News	明報

Ming Pao Education	明報 - 教得樂
Ming Teens	明 teens
Oriental Daily News	東方日報
Sharp Daily	爽報
Sing Pao	成報
Sing Tao Daily	星島日報
Sky Post	晴報
Smarties	明報 Smarties
Ta Kung Pao	大公報
Ta Kung Pao Archive	大公報 - 歷史數據
Take me Home (Hong Kong Economic Times) - Hong Kong CWS	生活區報 (香港經濟日報) - 港島中西南
Take me Home (Hong Kong Economic Times) - Hong Kong East	生活區報 (香港經濟日報) - 港島東
Take me Home (Hong Kong Economic Times) - Kowloon East & Tseung Kwan O & Sai Kung	生活區報 (香港經濟日報) - 東九龍 將軍澳 西貢
Take me Home (Hong Kong Economic Times) - Kowloon West	生活區報 (香港經濟日報) - 西九龍
Take me Home (Hong Kong Economic Times) - New Territories East & Shatin & Ma On Shan	生活區報 (香港經濟日報) - 新界東 沙田 馬鞍山
Take me Home (Hong Kong Economic Times) - New Territories West	生活區報 (香港經濟日報) - 新界西
The Sun	太陽報
Tin Tin Daily News	天天日報
Wen Wei Po	文匯報
am730	

We limit the scope of newspaper search to the real estate section, which ensures that the sentiment index reflects development in the property market. Some newspapers in Table A.2 do not have a real estate section in part or whole of the sample period. We drop these observations.

To capture property market sentiments in a given month *t*, we build our index by counting the number of articles that display positive or negative sentiments for the property market $m \in \{p,s\}$, where *p* and *s* stand for the primary and secondary market respectively. Precisely, for each article *a* in newspaper *n* at time *t* and for property market *m*, we define two indicator variables

$pos_{a,n,m,t}$ and $neg_{a,n,m,t} \in \{0,1\}$, denoting the existence of positive and negative sentiments. A compound keyword-search approach assigns scores to these indicator variables with the keyword list given by Table A.2. For instance, $pos_{a,n,m=p,t}$ =1 is assigned to an article when there exists a paragraph in the article which has at least one keyword from the Primary Market List [P] and at least one keyword from the Positive sentiment list [$\#pos_{p,s}$ or $\#pos_{p}$]. Similarly, $neg_{a,n,m=p,t}$ =1 is assigned to an article when there exists a paragraph in the article which has at least one keyword from the Primary Market List [P] and at least one keyword from the Positive sentiment list [$\#neg_{p,s}$ or $\#neg_{p}$]. Positive and negative sentiments for the secondary market are assigned analogously.

Table A.2: Keywords of market segments and corresponding sentiments

Primary market [P]								
一手價單	新盤	發展商	貨尾	餘貨	推盤	成交紀錄冊	認購	推售
Secondary market [S]								
二手轉手	放盤	屋苑	銀主	易手	放售	原業主	易主	承接
Positive sentiment [#*pos* $_{p,s}$]								
回暖	見底	好轉	轉強	改善	轉旺	復甦	回勇	轉活
理想	熾熱	熱賣	升溫	活躍	報捷	暢旺	造好	不俗
睇好	樂觀	亮麗	大旺	凌厲	看俏	看好	出色	小陽春
旺場	旺市	高漲	踴躍	大熱	佳績	強勁	放心	受歡迎
追捧	熱烈	熱鬧	具信心	正面	百花齊放	如火如荼	向好	發力
白熱	轉好	熱捧	熱搶	好景	良好			
Negative sentiment [#*neg* $_{p,s}$]								
見頂	轉差	轉壞	轉弱	轉淡	軟化	回軟	受阻	減弱
平淡	撻定	撻訂	冷清	淡靜	淡風	降溫	睇淡	悲觀
低迷	疲態	萎縮	頹勢	淡市	遜色	受壓	看淡	看差
癱瘓	受挫	劣勢	疲弱	冷卻	慘淡	冰封	冷落	停頓
隱憂	淪陷	薄弱	勢危	爆煲	憂慮	疏落	欠佳	淡勢
黯淡	不景氣	冷淡	脆弱	陰影	倒退	驟減	尋底	擔憂
不利	寒冬	負面	嚴冬	零星	不振	乏力	唱淡	呆滯
陰霾白果	退卻	每況愈下	困難	乏人問津	急轉直下	停滯	惡化	低谷
Positive sentiment for primary market only [#*pos* $_{p}$]								
削優惠	爭購	人龍	熱銷	勁銷	打蛇餅	大排長龍		
Negative sentiment for primary market only [#*neg* $_{p}$]								
增優惠	加優惠	滯銷	停售	停賣	未有進展	終止買賣合約	自救	

The property market sentiment indices add up the market sentiments in newspaper articles and are defined in Equations (1) and (2) in the main text.

The sentiment scores are then scaled by the number of news articles in the local real estate section in the same period and standardised to have a unit standard deviation.

A1.1 Excluding regional real estate markets

The purpose of the sentiment index is to uncover the sentiments of the local real estate market. However, with the small-open economy nature of Hong Kong, local newspapers may also report about real estate market news in other regions (e.g. Pearl River Delta), although the proportion is small. To ensure our counts track the local property market sentiments, we exclude the counts of non-domestic real estate market through the following steps.

Recall that the newspaper articles used in this analysis are limited to those in the real estate section of the newspapers. We sub-divide the real estate section into different regions, as indicated in Table A.3. For each non-domestic real estate section, we repeat the above compound search procedures to obtain $S_{m,t}^{r}$, where r indicates the non-domestic region. Then, we can exclude non-domestic regions by:

$$S_{m,t}^{HK} = S_{m,t}^{overall} - \sum_{r} S_{m,t}^{r}.$$

The indices reported in the main text exclude non-domestic regions.

Table A.3: keywords of newspaper section

Real estate section (overall)

地產

Real estate section (non-domestic)

海外　　環球　　澳門地產　中國地產　中國房地產　內地樓市　內地房地產　　珠三角地產

A.2 Query selection

Using our algorithm, the queries selected are presented in Table A.4 below.

Table A.4: Queries used in compiling GBII

Category	Search queries
Real Estate Agency	centaline property hong kong
	midland property hk
	香港 屋 網
Mortgage	樓宇 按揭
	經絡 按揭
Appraisal	樓宇 估價
	物業 估價
	銀行 估價
Banks	hang seng bank
	hsbc
	中銀
Map	centamap
	google map
	中原 地產 地圖
Transaction information	中原 地產 成交
Others	property hk
	買樓

A.3 Robustness check on nowcasting algorithm

In the main text, we describe the algorithm to construct the GBII index. We show that the resulting index is intuitive and robust, and that it improves nowcasting of house prices in-sample and out-of-sample. However, one potential drawback of this algorithm is that its query selection is sequential, so it may miss out some queries, or combination of queries, that may be informative. This appendix uses a Bayesian algorithm as a robustness check and shows that our algorithm yields similar results, which demonstrate the robustness of our algorithm.

Specifically, we use a Bayesian variable selection model with a spike-and-slab prior, following closely George and McCulloch (1993), Ishwaran and Rao (2005), and Scott and Varian (2015). We consider a multivariate regression model as follows:

$$y_t \sim N\left(\sum_{j=1}^{m} z_{tj}\gamma_j + \sum_{i=1}^{p} x_{ti}\beta_i, \sigma^2\right),$$

where we have $t = 1,2,\dots,T$ is the time subscript. We regress y_t on z_{tj} where $j = 1,2,\dots,m$, and on x_{ti} where $i = 1,2,\dots,p$. The regressors z_{tj} are included unconditionally, but x_{ti} are not. We assume the following prior on β_i for all i:

$$\beta_i \sim (1-\pi_i)\delta_0 + \pi_i N(0, \sigma^2\tau^2),$$

where $\pi_i \in [0,1]$ is a mixture weight, and $\delta_0 \equiv 0$, which gives the 'spike'. The 'slab' is characterised by a Normal distribution with zero mean. We specify rest of the priors:

$$\begin{gathered}\tau^2 \sim \text{InverseGamma}(1/2, s^2/2),\\ \sigma^2 \sim \text{InverseGamma}(\alpha_1, \alpha_2),\\ \pi_i \sim Bernoulli(\theta),\\ \theta \sim Beta(a, b),\\ \gamma \sim N(\gamma^0, diag(\Sigma_1^{\gamma 0}, \Sigma_2^{\gamma 0}, \dots, \Sigma_m^{\gamma 0})).\end{gathered}$$

One needs to set the parameters $s, \alpha_1, \alpha_2, a, b$, γ^0 and $\left\{\Sigma_{\text{j}}^{\gamma 0}\right\}_{j=1}^{m}$. Then the parameters $\pi, \beta, \theta, \tau^2, \sigma^2$, γ are estimated given x_{ti}, z_{tj} and y_t using Gibbs sampling. The condition posteriors are now given by:

$$\theta|\pi \sim Beta\left(a + \sum_{i=1}^{p} \pi_i, b + \sum_{i=1}^{p} (1-\pi_i)\right),$$

$$\tau^2|\beta,\pi\sim\text{InverseGamma}\left(0.5(1+\textstyle\sum_{i=1}^{p}\pi_i),0.5\left(s^2+\frac{\beta'\beta}{\sigma^2}\right)\right),$$

$$\sigma^2|y,\beta,\gamma\sim\text{InverseGamma}\left(\alpha_1+\frac{N}{2},\alpha_2+\frac{(y-Z\gamma-X\beta)'(y-Z\gamma-X\beta)}{2}\right),$$

$$\beta|y,\pi,\tau^2,\sigma^2,\gamma\sim N\left(\left(\frac{1}{\sigma^2}X'X+\frac{1}{\sigma^2\tau^2}I\right)^{-1}X'(y-Z\gamma)\frac{1}{\sigma^2},\left(\frac{1}{\sigma^2}X'X+\frac{1}{\sigma^2\tau^2}I\right)^{-1}\right),$$

for those β_i where π_i=1, and $\beta_i=0$ when $\pi_i=0$,

$$\gamma|y,\beta,\sigma^2\sim N\left(\left(\Sigma_0^{-1}+\frac{1}{\sigma^2}Z'Z\right)^{-1}\left(\Sigma_0^{-1}\gamma^0+\frac{1}{\sigma^2}Z'(y-X\beta)\right),\left(\Sigma_0^{-1}+\frac{1}{\sigma^2}Z'Z\right)^{-1}\right),$$

$$\pi_j|y,\pi_{-j},\beta_{-j},\sigma^2,\tau^2,\theta,\gamma\sim Bernoulli(\xi_j),$$

where

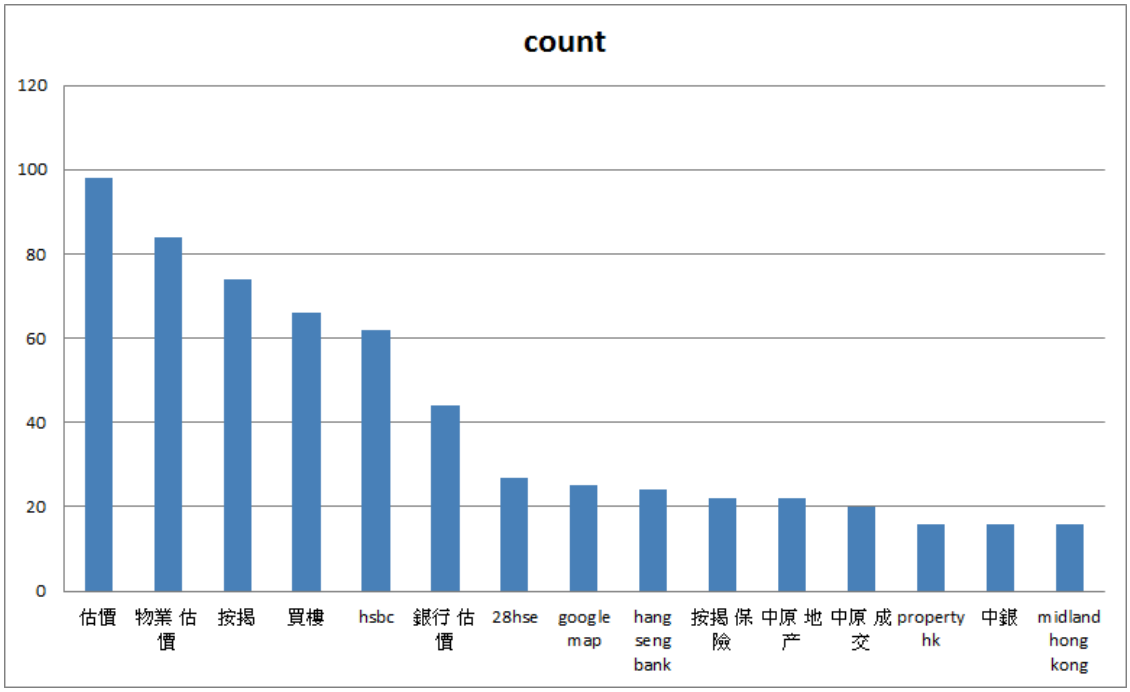
$$1-\xi_j=\frac{(1-\theta)}{(1-\theta)+\theta(\tau^2)^{-\frac{1}{2}}\left(x_j'\,x_j+\frac{1}{\tau^2}\right)^{-\frac{1}{2}}\exp\left[\frac{\left[\left(y-Z\gamma-X_{-j}\beta_{-j}\right)'x_j\right]^2}{2\sigma^2\left(x_j'\,x_j+\frac{1}{\tau^2}\right)}\right]},\qquad \forall j,$$

where β_{-j} denotes the vector β without the element β_j, and similarly for π_{-j}.

In our model, y_t is the month on month growth of *R&VD Index*; $z_{tj}=z_t$ is its lag (m=1), and x_{ti} are the 3-month moving average of the month on month growth of the Google Trends *query i*. In our estimation, we set $s=0.5,\alpha_1=1,\alpha_2=0.01,a=b=1$, $\gamma^0=0$ and $\Sigma^{\gamma 0}=0.01$. In each estimation, we make 5000 draws by Gibbs sampling and set the burn-in periods to be 3750. The mixture probability π_i is computed by dividing the number of non-zero β_i's by 1250.

We repeat this estimation 100 times, and in each estimation we record the 10 queries that are associated with the highest π_i. This allows us to compare the queries that are selected by this method with the ones that are selected by our algorithm. Figure A.1 reports the top queries returned using this model selection model. It turns out that the queries selected are quite similar to those selected by our preferred algorithm (See Table A.4). For instance, out of the top 10 keywords chosen by the Bayesian method, 6 appear in GGBI index.

Figure A.1: Top queries returned using spike-and-slab regression:


count
120
100
80
60
40
20
0
估價
物業 估價
按揭
買樓
hsbc
銀行 估價
28hse
google map
hang seng bank
按揭 保險
中原 地产
中原 成交
property hk
中銀
midland hong kong

REFERENCES

Askitas, N., & Zimmermann, K. F. (2009). "Google Econometrics and Unemployment Forecasting," *Applied Economics Quarterly*. 55: 107-20.

Baker, S. R., Bloom, N., & Davis, S. J. (2016), "Measuring economic policy uncertainty," *The Quarterly Journal of Economics*, 131(4): 1593-1636.

Bank for International Settlements (2019), "The use of big data analytics and artificial intelligence in central banking," *IFC Bulletin*, No 50.

Baxter, M. (1994). "Real exchange rates and real interest differentials: Have we missed the business-cycle relationship?" *Journal of Monetary Economics*, 33(1): 5-37.

Choi, H., & Varian, H. (2012) "Predicting the present with Google Trends," *Economic Record*.

Dietzel, M. A., Braun, N., & Schäfers, W. (2014), "Sentiment-based commercial real estate forecasting with Google search volume data," *Journal of Property Investment & Finance*, 32(6): 540-569.

Financial Stability Board (2017), "Artificial intelligence and machine learning in financial services - Market developments and financial stability implications," November.

Fondeur, Y., & Karamé, F. (2013). "Can Google data help predict French youth unemployment?" *Economic Modelling*. 30:117-25.

Francesco, D. A. (2009). "Predicting unemployment in short samples with internet job search query data," MPRA Paper 18403, University Library of Munich, Germany.

Galbraith, J. (1990), A Short History of Financial Euphoria. New York: Viking Press.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data," *Nature*. 457(7232): 1012-4.

George, E. I., & McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*. 88(423): 881-9.

Gao, Q., & Zhao, T. (2018), "The Influence of Home Buyer Sentiment on Chinese Housing Prices——Based on Media Text Mining," *International Journal of Economics and Finance*, 10(9).

He, D. (2014), "The Effects of Macroprudential Policies on Housing Market Risks: Evidence from Hong Kong," Banque de France Financial Stability Review No. 18.

Henderson, K. V. & Cowart, L. B. (2002), "Bucking e-commerce trends: A content analysis comparing commercial real estate brokerage and residential real estate brokerage websites," *Journal of Corporate Real Estate*, Vol. 4 No. 4, pp. 375-85.

Hong Kong Monetary Authority (2014), "Box 5: The Impact of Counter-Cyclical Prudential and Demand-Management Measures on Hong Kong's Housing Market," Half-yearly Monetary and Financial Stability Report September 2014, Hong Kong.

Huang, C., Simpson, S., Ulybina, D., & Roitman, A. (2019), "News-based sentiment indicators," IMF Working Paper No. 19/273.

Huang, D. J., Leung, C. K. Y.,& Tse, C-Y. (2018), "What Accounts for the Differences in Rent-Price Ratio and Turnover Rate? A Search-and-Matching Approach," *Journal of Real Estate Finance and Economics*, 57(3): 431-475.

Kwan, Y. K., Leung, C. K. Y., & Dong, J. (2015), "Comparing consumption based asset pricing models: The case of an Asian city," *Journal of Housing Economics*, 28(C): 18-41.

Iacoviello, M. (2005), "House prices, borrowing constraints, and monetary policy in the business cycle," *American Economic Review*, 95(3), 739-764.

Ishwaran, H., & Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*. 33(2), 730-73.

Kindleberger, C. P. (1978), Manias, Panics, and Crashes: A History of Financial Crises. First ed., John Wiley and Sons, Inc.

Kulkarni, R., Haynes, K. E., Stough, R. R. and Paelinck, J. H. P. (2009), "Forecasting Housing Prices with Google Econometrics," Research Paper School of Public Policy, George Mason University, No. 2009-10.

Larsen, V. H., Thorsrud, L. A., & Zhulanova, J. (2020), "News-driven inflation expectations and information rigidities," *Journal of Monetary Economics*, forthcoming.

Leung, C. K. Y., & Ng, C. Y. J. (2019). "Macroeconomic aspects of housing," Oxford Research Encyclopedia of Economics and Finance.

Leung, C. K. Y., & Tang, E. C. H. (2015), "Availability, Affordability and Volatility: The Case of the Hong Kong Housing Market," *International Real Estate Review*, 18(3): 383-428.

Leung, C. K. Y., Ng, J. C. Y., & Tang, E. C. H. (2020a), "What do we know about housing supply? The case of Hong Kong SAR," Economic and Political Studies, 8(1): 6-20.

Leung, C. K. Y., Ng, J. C. Y., & Tang, E. C. H (2020b), "Why is the Hong Kong Housing Market Unaffordable? Some Stylized Facts and Estimations," Globalization Institute Working Papers 380, Federal Reserve Bank of Dallas.

Leung, C. K. Y., & Tse, C-Y (2017), "Flipping in the housing market," *Journal of Economic Dynamics and Control*, 76(C): 232-263.

Liu, Z., Wang, P., & Zha, T. (2013), "Land-price dynamics and macroeconomic fluctuations," *Econometrica*, 81(3): 1147-1184.

Soo, C. K. (2018), "Quantifying Sentiment with News Media across Local Housing Markets," *The Review of Financial Studies*, 31(10): 3689–3719.

Tetlock, P. C. (2007), "Giving content to investor sentiment: The role of media in the stock market," *The Journal of Finance*, 62(3): 1139-1168.

Loughran, T., & McDonald, B. (2011), "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks," *The Journal of Finance*, 66(1): 35-65.

Luk, P., Cheng, M., Ng, P., & Wong, K. (2020), "Economic policy uncertainty spillovers in small open economies: The case of Hong Kong," *Pacific Economic Review*, 25(1): 21-46.

Schüler, Y. S. (2018). "Detrending and Financial Cycle Facts Across G7 Countries: Mind a Spurious Medium Term!" ECB Working Paper No. 2138.

Scott, S. L., & Varian, H. R. (2014), "Predicting the present with Bayesian structural time series," *International Journal of Mathematical Modeling and Numerical Optimisation*, 5(1/2), 4-23.

Scott, S. L., & Varian, H. R. (2015), "Bayesian Variable Selection for Nowcasting Economic Time Series," in *Economic Analysis of the Digital Economy*, edited by Goldfarb, A., Greenstein, S. M., and Tucker, C. E., pp. 119-135. University of Chicago Press.

Shapiro, A. H., Sudhof, M., & Wilson, D. (2020), "Measuring news sentiment," Federal Reserve Bank of San Francisco Working Paper.

Shiller, Robert J. (2009), Animal Spirits. Princeton, NJ: Princeton University Press.

Soo, C. K. (2018), "Quantifying sentiment with news media across local housing markets," *The Review of Financial Studies*, 31(10): 3689-3719.

Tang, C. H. (2019), "Speculate a lot," forthcoming in *Pacific Economic Review*.

Van Dijk, D.W. & Francke, M.K. (2018), "Internet search behavior, liquidity, and prices in the housing market," *Real Estate Economics*, 46(2): 368-403.

Walker, C. B. (2014), "Housing booms and media coverage," *Applied Economics* 46(32): 3954-3967.

Wu, L. & Brynjolfsson, E. (2015), "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales", Economic analysis of the digital economy. University of Chicago Press, 89-118.

Wu, T., Cheng, M., & Wong, K. (2017), "Bayesian analysis of Hong Kong's housing price dynamics", *Pacific Economic Review*, 22: 312– 331.

Wong, K., Ng, P., Cheng, M., & Luk, P. (2017), “Measuring Economic Uncertainty and its Effect on the Hong Kong Economy,” HKMA Research Memorandum, 11/2017

Yang, X., Pan, B., Evans, J. A., & Lv, B. (2015). “Forecasting Chinese tourist volume with search engine data,” Tourism Management, 46: 386-97.