*Compilation of House Price Index using Big Data Techniques**

This study aims to derive an alternative indicator for the residential property price in India adopting big data tools. Data on residential property prices from select web-portals are collected and processed using big data tools to compile a residential house price index. This index can be compiled almost immediately at the end of each quarter. It is found that there are statistically significant correlations between residential house prices based on web data and registration data (both index and growth), though there are divergences in terms of price levels. There is a bidirectional relationship between the indices based on registration data and web data.

House price is an essential indicator of the economy. It has a co-movement with the Gross Domestic Product (GDP), consumption, investment, inflation, current account balance, and the output gap (Ahearne *et al.*, 2005). Residential property is the largest single asset for most households around the world. The variations in residential property price affect households' long-term investment strategy, influence their spending and borrowing patterns. Change in the property price influences the banking and financial sectors of the economy through bank lending and mortgage channel.

In view of this, residential property price is an important piece of information for policymakers, especially for the central banks. In the Indian context, there are two established housing price indices, *viz.*, (a) House Price Index (HPI) compiled by the Reserve Bank of India and (b) RESIDEX, which is compiled by the National Housing Bank (NHB)[1]. These indices are typically available with a lag of (at least) one quarter.

In order to obtain an early signal on residential house prices, an attempt has been made to compile an alternate residential house price index based on data available on the web portals of select real estate agencies in India. Big data tools have been employed to collect, process, and store these data.

The rest of the article is organised as follows: Section 2 provides a brief review of the literature in this domain. Section 3 highlights data collection using dynamic chart scrapping methodology and processing unstructured data using big data tools. Section 4 describes the methodology for compiling the index on the basis of web data. Section 5 outlines stylised facts and empirical methodologies to study the index's usefulness for estimating the Reserve Bank's HPI. Section 6 focuses on empirical results, and Section 7 sets out the concluding observations.

2. Review of Literature

In the domain of big data analytics, the techniques involved in the programmatic collection of intended data from a specific website is referred to as web scraping. The web scraping as a data collection method was adopted in the past. However, the same was used in the core economic activity in the Billion Prices Project (BPP), an academic initiative at the MIT Sloan and Harvard Business School (Cavallo and Rigobon, 2016). This work describes how the alternative sources, like the data web scraped from online e-commerce in BBP, can be a potential candidate for the compilation of price index (of underlying commodities). In their work, big data techniques are adopted to capture, clean, and process data received from online e-commerce portals. They argued that these sources are often cheap and enable users to

* This article is prepared by Indranil Gayen and Sasanka Sekhar Maiti in the Big Data Analytics Division, Department of Statistics and Information Management. The views expressed in the article are those of the authors and do not necessarily represent the views of the Reserve Bank of India.

[1] For this study, we have focused on HPI compiled by the Reserve Bank of India (RBI) as there will be an extended time series for comparison. Further, HPI is reported to the international forums by the RBI, while NHB RESIDEX is not having an All India level index (and has non-mutually exclusive coverage).

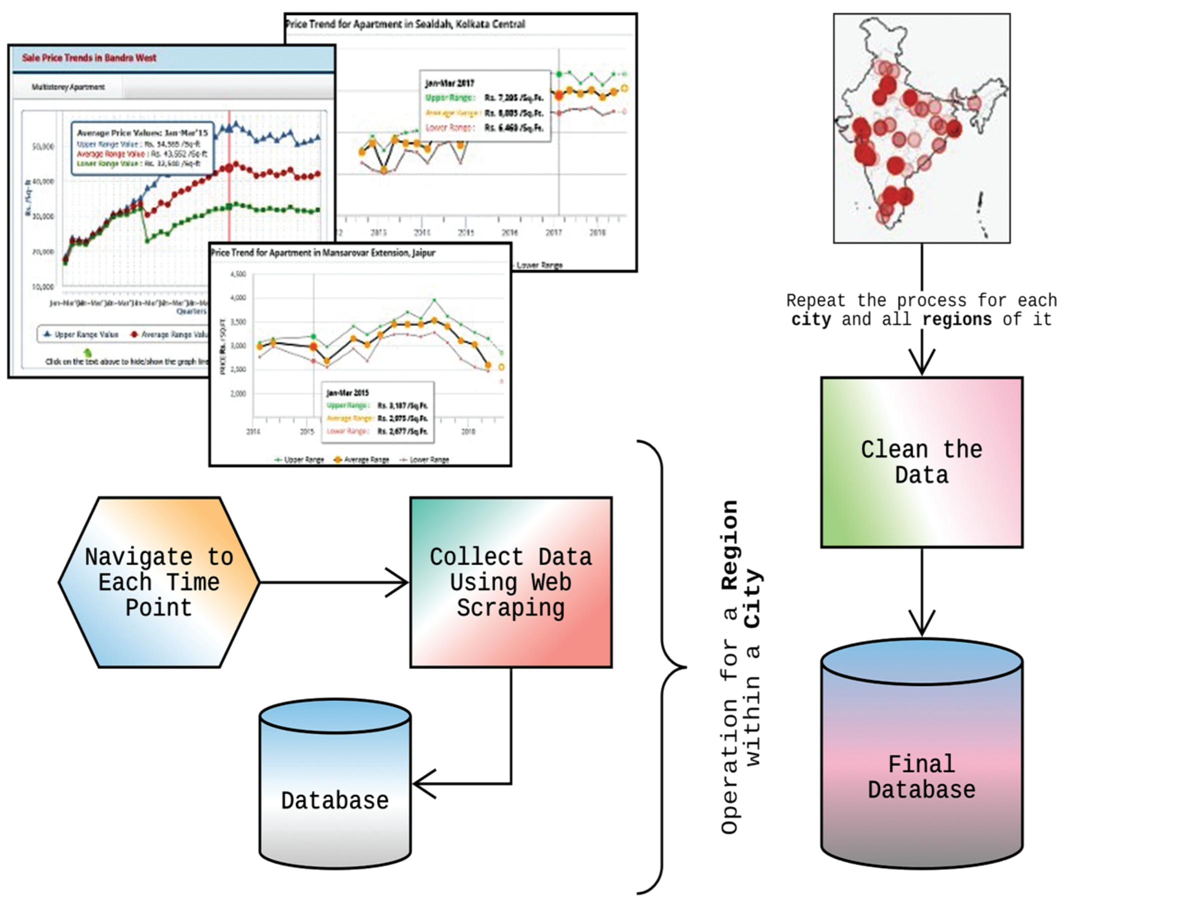
have near real-time price index numbers. The same concept was adopted in the Indian context, where it was argued that despite several limitations of online prices, the web-data based index successfully tracks both the direction and the magnitude of the Indian official Consumer Price Index (CPI) (Banerjee *et al.*, 2018).

The use of big data techniques and web scraping in the context of the residential property price has been minimal. However, several studies in this domain using Google Trend data include - predicting economic activities (Choi and Varian, 2012); assessment of real estate market and housing price (Wu and Brynjolfsson, 2015); and housing price in the Indian context (Mitra *et al.*, 2017). In this context, a similar study was carried out at the Bank Indonesia. It was argued that Indonesia's online advertisement data was successfully tracking property prices (Kristiawardani and Sampe, 2017).

3. Data Collection and Processing

The big data tools have been employed to collect data from websites[2] of select real estate advertising agencies in India. In order to have historical data on residential house price, a specific kind of web scraping technique, which may be termed as 'dynamic chart scraping', has been developed (Chart 1).

Chart 1: Data Collection and Processing Framework Using Big Data Tools



2 The leading real-estate advertisement websites in the Indian market.

Within the source website, the price trend is presented in the form of a dynamic chart. The web scraping technology was developed and employed on these dynamic charts in the web-portals to obtain desired housing price information. The data fields captured using the dynamic chart scrapping are source website, city, location within the city (termed as a region), property type, price of the property, type of price aggregation, unit of the price and the time point of price.

There are missing values observed in the collected dataset. These missing values are estimated by interpolation using local polynomials[3].

Sometimes the price variation in the datasets is found to be very high. Such price variations are adjusted using three methods: (a) outlier detection rule as embedded in the X-13ARIMA-SEATS program; (b) additive outlier detection rule; and (c) heuristic approach involving moving interquartile ranges. Index of house price has been derived based on the processed final data.

4. Compilation of House Price Index

Denote, P^{a}_{ijkt} as price in the j[th] region of the i[th] city for the t[th] quarter and k[th] category of the property with a[th] type of price aggregation. Note that wherever the P^{a}_{ijkt} is not available the same is filled with $\widehat{P}^{a}_{ijkt}$ (Estimated P^{a}_{ijkt} where (i, j, k, t, a) is not present). The estimation is carried out using various machine learning and time series imputation techniques. The compilation of the index involves the following steps:

(a) Considering one specific time point as the base-period (let say we denote it by $t_0 \in \mathbf{T}$) price relative for each of these categories are calculated as $\pi^{a}_{ijkt} = \frac{P^{a}_{ijkt}}{P^{a}_{ijkt_0}}$. The weights for each of the price relatives (π^{a}_{ijkt}) are denoted using the following notation,

$$\omega^{a}_{ijk} = \text{Weight for} \begin{cases} \text{i}^{\text{th}} \text{ city, j}^{\text{th}} \text{ region, where } (i,j) \in \mathbf{L} \\ \text{k}^{\text{th}} \text{ category of property, where } k \in \mathbf{K} \\ \text{a}^{\text{th}} \text{ type of price aggregation, where } a \in A \\ \text{It is independent of quarter (time point) t} \end{cases}$$

(b) The weighted price relatives for different aggregate types, for the specific property type and location (location is a combination of *i and j*) ($\pi_{ijkt} = \frac{\sum_a \pi^{a}_{ijkt}\omega^{a}_{ijk}}{\sum_a \omega^{a}_{ijk}}$) are derived. The weighting pattern considered for aggregation levels[4] assigns (ω^{a}_{ijk}) highest weight to average aggregation type and lower weight on the other two ($\omega^{upper}_{ijk} < \omega^{average}_{ijk} > \omega^{lower}_{ijk}$). It is defined as $\omega^{a}_{ijk} = e^{-\left|\bar{P}^{a}_{t_0} - \bar{P}^{average}_{t_0}\right|}$. (Here $\bar{P}^{a}_{t_0}$ denote the appropriate average of the normalised unit free price at base period). (Final weights came as 1 for *average*, 0.913 for *upper*, 0.915 for *lower*)

(c) All of these weighted price relatives are then further aggregated to location level for each quarter as $\pi_{ijt} = \frac{\sum_k \pi_{ijkt}\omega_{ijk}}{\sum_k \omega_{ijk}}$. Choice of weights corresponding to each property types ($\{\omega_{ijk}\}$) is based on the count of regions *(j)* offering that specific property types in a city *(i)* at the base-period *(t_0)*.

(d) For each quarter, the weighted price relatives (π_{ijt}) corresponding to each location are further aggregated using region weights (ω_{ij}) to arrive at city level quarterly weighted price relatives ($\pi_{it} = \frac{\sum_j \pi_{ijt}\omega_{ij}}{\sum_j \omega_{ij}}$). The weights for regional locations $\{\omega_{ij}\}$ are constructed as $\omega_{ij} = e^{-\bar{P}_{ijt_0}}$[5]

[3] Lagrange polynomial (https://en.wikipedia.org/wiki/Polynomial_interpolation) is used in sliding window on time series data.

[4] The set A (of all possible aggregation types) has three elements which are upper, average and lower.

[5] It is inversely related to the average normalised price of a region at base period, which is taken as a proxy for the number of advertisements at a specific location. It is observed from data that usually the number of advertisements of the properties with higher relative price occur less frequently than that of lower relative price.

(e) Finally, to obtain quarterly All-India level figures, assuming the city and region have representative coverage, these price relatives (π_{it}) are used for the calculation of weighted average taking city weights as ω_i. This figure is donated as $\pi_t = \frac{\sum_i \pi_{it}\omega_i}{\sum_i \omega_i}$. The population size (using Census 2011 data) for each city is taken as respective city weight, following the approach as adopted in RBI's Reserve Bank's House Price Index (HPI) construction.[6]

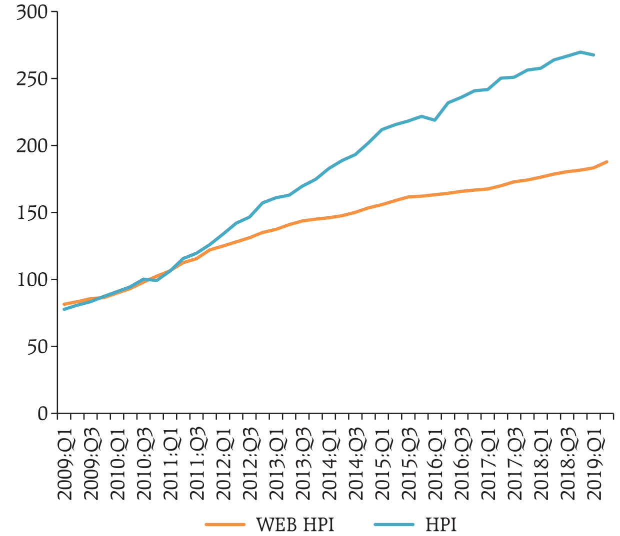
5. Stylised Facts and Empirical Methodologies

The Reserve Bank's HPI[7] is compiled every quarter based on the official data of property transactions collected from the registration authorities of respective state governments for ten major cities, *viz.*, Mumbai, Delhi, Chennai, Kolkata, Bengaluru, Lucknow, Ahmedabad, Jaipur, Kanpur and Kochi with the base year 2010-11 (RBI, 2014). The city, Kochi, has been excluded from this analysis from of HPI. The derived index based on web crawled data covers these cities except Kochi. It is termed as WEB HPI (for simplicity we have interchangeably used WEB to mean the same). The base year for WEB is considered the same as that of HPI.

Overall, the WEB Index is found to be lesser than HPI, implying that the price level based on registration data is higher than the price level indicated in the advertisements (Chart 2).

Further, the WEB Index graphs are smoother than the HPI. It means advertisement prices are not changed frequently, whereas, the HPI based on actual

Chart 2: House Price Index (HPI) *vis-à-vis* Web Index – All India



registration data is having higher price variation over time (as the actual sale varies significantly). Moreover, the registration prices are embedded with several underlying factors of the property such as floor rise, position, facing, and other factors, which may not be displayed in the advertisements. The empirical analysis of the divergence between HPI and WEB indices is beyond the scope of this study. However, it may be due to the following reasons:

(i) Advertised prices are changed at a longer time horizon, whereas the actual prices of properties are much higher as various underlying factors are incorporated in it;

(ii) Advertised prices are displayed at a discount to attract buyers whereas the final transaction takes place at a higher cost;

(iii) Several reports on real estate market research state that there is a declining trend in the sale of the residential properties from 2013-14 onwards[8]. As a result, the properties displayed in the advertisements may not be representative of the set of properties actually sold; and

6 All of the weighing patterns (except for All-India level aggregation) are based on the heuristic approach by the authors. There was limited, or no prior research work on this type of data. Details of which is omitted in this scope of the study.

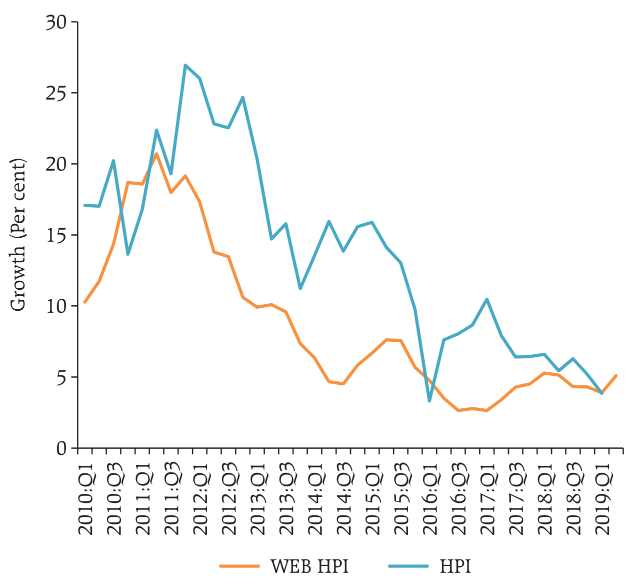
7 The data has been obtained from https://dbie.rbi.org.in/ (*Statistics > Real Sector > Prices & Wages > Quarterly*). The HPI series is available in two different base year with a common period. The index figures prior to 2010-Q2 (equivalently Q1 of 2010-11) are obtained using splicing with old base year (2008-09) indices.

(iv) As per the compilation methodology of HPI, the set of residential properties differ from one quarter to another, which is dependent on the actual sale during the underlying period.

The year-on-year growth figures on the basis of registration data was much higher than the growth based on advertisement data during January 2010 – September 2010, October 2011 – December 2015, and April 2016 – June 2017. However, there is a statistical co-movement of year-on-year growth of prices in both the series. The growth figures that fell below 5 per-cent during January – March 2019 could be due to the implementation of the Real Estate Regulation and Development (RERA) Act, 2016 (Chart 3).

In order to study the empirical relationship between the HPI and the WEB HPI, the correlation analysis has been carried out at levels and year-on-year growth.

Chart 3: Year-on-Year Growth in HPI and WEB HPI



Compute year-on-year growth of house prices using the formulas[9]

$$Y_t^{HPI} = \left(\frac{I_t^{HPI}}{I_{t-4}^{HPI}} - 1\right) * 100 \quad \ldots \quad (1)$$

$$Y_t^{WEB} = \left(\frac{I_t^{WEB}}{I_{t-4}^{WEB}} - 1\right) * 100 \quad \ldots \quad (2)$$

To analyse the usefulness of WEB HPI data for estimating the HPI, following steps have been used:

1. Carry out the stationary test of the variables Y_t^{HPI} and Y_t^{WEB}.
2. If the underlying variables are found to be non-stationary, then a statistical test is carried out to examine whether these variables are cointegrated.

   In case the variables are cointegrated, build a model based on the variables Y_t^{HPI} and Y_t^{WEB}; otherwise, build a Vector Auto Regression (VAR) model after transforming the variables Y_t^{HPI} and Y_t^{WEB} into stationary series taking a suitable number of differencing on both the variables (Asteriou & Hall, 2011).

   It is observed that the variables Y_t^{HPI} and Y_t^{WEB} are non-stationary and are not cointegrated. At first order differencing, the variables Y_t^{HPI} and Y_t^{WEB} are found to be stationary. Therefore, the underlying VAR model is

$$\begin{pmatrix} \Delta Y_t^{HPI} \\ \Delta Y_t^{WEB} \end{pmatrix} = \begin{pmatrix} \alpha^{HPI} \\ \alpha^{WEB} \end{pmatrix} + \sum_{j=1}^{p} \begin{pmatrix} \gamma_j^{HPI\sim HPI} & \gamma_j^{WEB\sim HPI} \\ \gamma_j^{HPI\sim WEB} & \gamma_j^{WEB\sim WEB} \end{pmatrix} \begin{pmatrix} \Delta Y_{t-j}^{HPI} \\ \Delta Y_{t-j}^{WEB} \end{pmatrix} + \begin{pmatrix} \epsilon_t^{HPI} \\ \epsilon_t^{WEB} \end{pmatrix} \quad \ldots \quad (3)$$

   Note that the coefficient $\gamma_j^{HPI\sim WEB}$ in the right hand side of the equation means coefficient for ΔY_{t-j}^{HPI} while predicting ΔY_t^{WEB} and all other $\gamma_j^{*\sim *}$ are defined in the same line. For simplicity in the notation, we have subsequently used only γ_j^{*} in Table 3 first column (rest columns fully define $\gamma_j^{*\sim *}$).

The equation (3) is estimated based on the underlying data.

[8] Housing sales down 40 per cent in 2017 from 2013/14 levels: ANAROCK (Source: https://economictimes.indiatimes.com/industry/services/property-/-cstruction/housing-sales-down-40-per-cent-in-2017-from-2013/14-levels-anarock/articleshow/63352292.cms)

[9] We denote I_t^{HPI} and I_t^{WEB} as HPI and WEB HPI respectively at time $t \in \{2, 3, \ldots, T\}$.

6. Empirical Results

The empirical analysis has been carried out based on quarterly data on underlying indices, *viz.*, HPI and WEB HPI, available from January – March 2009 to January – March 2019 (41 quarters).

The estimated correlation coefficients between HPI and WEB HPI at both level and growth (year-on-year) are found to be statistically significant at 1 per-cent level (Table 1).

Unit root test of the underlying variables is carried out using the Phillips-Perron Test. It is observed that the first-order difference of both the variables Y_t^{HPI} and Y_t^{WEB} is found to be stationary (Table 2). The Johansen cointegration test between the variables Y_t^{HPI} and Y_t^{WEB} indicates that these variables are not cointegrated.

Accordingly, a VAR model (in line with Equation 3) has been estimated based on the sample data[10]. The VAR model provides an estimation framework for incremental growth of the house price.

The estimated VAR model reveals that the incremental growth of residential house price (both registration and online web) can be estimated using their lag values (past four quarters). In particular, residential house price available on the web could be useful in predicting the registration residential house price. There is a bi-directional causal relationship between the registration price and online web price (Table 3).

Table 1: Estimated Correlation Coefficient and Statistical Tests

Correlation Coefficient	Estimate	P-value	Sample Data Period
$\rho(I_t^{HPI}, I_t^{WEB})$	0.9869	0.00 (***)	January-March 2009 to January-March 2019
$\rho(Y_t^{HPI}, Y_t^{WEB})$	0.7844	0.00 (***)	January-March 2010 to January-March 2019
$\rho(\Delta Y_t^{HPI}, \Delta Y_t^{WEB})$	0.1296	0.23	April-June 2010 to January-March 2019

10 The lag order of the VAR model has been decided based on the Final Prediction Error (FPE) information criteria. The optimal lag order for the estimated VAR model is 4.

Table 2: Unit Root Test – Summary Results

Null Hypothesis	Summary Results of Phillips-Perron Test	Accept / Reject
Y_t^{HPI} has a unit root	Test statistic -0.9246, p-value 0.7688	Accept
ΔY_t^{HPI} has a unit root	Test statistic -7.2818, p-value 0.0000	Reject
Y_t^{WEB} has a unit root	Test statistic -1.0897, p-value 0.7096	Accept
ΔY_t^{WEB} has a unit root	Test statistic -4.4201, p-value 0.0012	Reject

7. Conclusion

With the help of big data tools, data on house price available in the web portals of real estate advertisers have been collected and processed. An alternative indicator of residential house price index has been constructed based on advertisement data available online. There is a divergence between the prices levels of HPI and WEB HPI. The level of residential house price based on advertisement data is found to be lower than the level of residential house price based on registration data. Also, the growth of price rise in registration data was much higher than the growth of price rise computed based on advertisements during January 2010 – September 2010, October 2011 – December 2015, and April 2016 – June 2017. However, the correlation between

Table 3: Estimated VAR Model

Coefficients	Endogenous Variables	
	ΔY_t^{HPI}	ΔY_t^{WEB}
α^{HPI} / α^{WEB}	-0.3850 *(0.44)*	-0.2737 *(0.18)*
γ_1^{HPI}	-0.0189 *(0.91)*	-0.0103 *(0.87)*
γ_2^{HPI}	-0.1922 *(0.21)*	0.0372 *(0.53)*
γ_3^{HPI}	0.2007 *(0.18)*	**0.1073** *(0.00)*
γ_4^{HPI}	**-0.4307** *(0.01)*	**-0.1366** *(0.04)*
γ_1^{WEB}	0.2033 *(0.63)*	0.2101 *(0.22)*
γ_2^{WEB}	0.0288 *(0.93)*	**0.4651** *(0.00)*
γ_3^{WEB}	-0.3672 *(0.27)*	-0.0591 *(0.65)*
γ_4^{WEB}	**0.9768** *(0.00)*	**-0.3212** *(0.02)*
$R^2(adj)$	0.32 *(0.03)*	0.40 *(0.01)*

Granger Causality Tests

Null Hypothesis	Summary Results
ΔY_t^{HPI} does not Granger cause ΔY_t^{WEB}	F-Test = 2.9509 (0.03)
ΔY_t^{WEB} does not Granger cause ΔY_t^{HPI}	F-Test = 2.8186 (0.04)

Source: Computations by authors.
Figures within the first bracket represent p-values.

indices (HPI and WEB) and their year-on-year growth is found to be positive and statistically significant. Therefore, the residential house price index based on advertisement data may be a useful indicator, and further research on its application may be attempted.

References

Alan G. Ahearne, John Ammer, Brian M. Doyle, Linda S. Kole, and Robert F. Martin, (2005). "House Prices and Monetary Policy: A Cross-Country Study", *Board of Governors of the Federal Reserve System.*

Asteriou, D. & Hall, S. G., (2011). "Vector Autoregressive (VAR) Models and Causality Tests". *Applied Econometrics (Second ed.). London: Palgrave MacMillan, p. 319–333.*

Banerjee, R., Singhal, N. & Subramanian, C., (2018). "Predicting Food Price Inflation through Online Prices in India". *Economic and Political Weekly, Volume LIII No 23, pp. 132-135.*

Cavallo, A., & Rigobon, R. (2016). "The Billion Prices Project: Using Online Prices for Measurement and Research". *Journal of Economic Perspectives, 30(2), 151–178. https://doi.org/10.1257/jep.30.2.151*

Choi, H., & Varian, H. (2012). "Predicting the Present with Google Trends". *Economic Record, 88(SUPPL.1), 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x*

Kristiawardani, K., & Sampe, I. (2017). "Using Online Property Advertisements Data as a Proxy for Property Market Indicators Bank Indonesia". *FC-Bank Indonesia Satellite Seminar on "Big Data" at the ISI Regional Statistics Conference 2017 Bali, Indonesia, March.*

Mitra, P., Sanyal, A., & Choudhury, S. (2017). "Nowcasting Real Estate Activity in India using Google Trend Data". *Reserve Bank of India - Occasional Papers, 38-No. 1. https://rbi.org.in/Scripts/bs_viewcontent.aspx?Id=3516.*

RBI, (2014). "House Price Index: 2010-11 to 2013-14". *RBI Monthly Bulletin, October, pp. 93-96.*

Wu, L., & Brynjolfsson, E. (2015). "The Future of Prediction : How Google Searches Foreshadow Housing Prices and Sales". *Economic Analysis of the Digital Economy: April 2015. https://doi.org/10.3386/w19549*