Measuring Property Price Variations Using Online Property Advertisements

Indranil Gayen

Reserve Bank of India, Mumbai, India – igayen@rbi.org.in

Abstract: Using big data techniques relating to the web scraping, an alternative measure of residential house price index is proposed in the context of India. The index is computed based on the data scraped from online property advertisement websites. The proposed index tracks well the underlying property price dynamics, at least to the extent of existing indices. This index shares the benefit of low compilation cost and time. The index can be compiled in near real-time, at least, one quarter ahead of standard indices.

JEL Classification: C32, C43, R30, C55, C81

Keywords: big data, web scraping, real-estate, housing price index

Introduction

Residential property is the largest single asset for most of the households around the world. The variations in residential property price affect households’ long-term investment strategy, leaving influences on their spending and borrowing habits. Real-estate asset price changes have an influence on the banking and financial sectors of the economy through bank lending and mortgage channel. The patterns and behaviours of residential house prices, therefore, influences not only financial business cycle dynamics but also the performance of the financial system.

1 The views and opinions expressed in the study are attributable to the author only and do not represent the views of the Reserve Bank of India.
At present, in India, there are two established channels present for the dissemination of housing price statistics in the form index. One named as NHB RESIDEX, which is compiled by National Housing Bank (NHB), a wholly owned subsidiary of Reserve Bank of India (RBI) and another is compiled by the Statistical Analysis Division of the Department of Statistics and Information Management, Reserve Bank of India. In addition to these two indices, there was the presence of another index termed Residential Property Price Index (RPPI) which came into its existence from an initiative of RBI referred as “Residential Asset Price Monitoring Survey”. The RPPI series is discontinued in recent times from public dissemination.

These indices are typically available in a lag of one quarter and sometimes suffer from the irregular release schedule. As a consequence sometimes this information does not play a useful role in the advance outlook of the economic affairs and forecasting. The technique of nowcasting is usually used by researchers, economist and data scientist taking advantage of timely, high-frequency data to estimate one or few key economic indicators. Despite the fact that the price indices are low frequency, these could have been a candidate for input in many nowcasting exercises provided the data comes in a timely manner. However, compiling indices from micro-data is never an easy task and as a result, these indices are always released with a lag of, at least, one quarter. This article explores an option of alternate housing price index which at least can be obtained more timely than others. This is achieved by utilizing the data web scraped from leading property advertisement websites in India.

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 the web scraping. The use of the web scraping as a data collection utility has been there for a long period of time. However, the use of the same in the core economic activity perhaps was carried out first in The Billion Prices Project (BPP) which is an academic initiative by MIT Sloan and Harvard Business School. Cavallo and Rigobon (2016) have described in their work, how the alternative sources, like data web scraped from online e-commerce in BBP, can be a potential candidate for the compilation of price index numbers. In their work big-data techniques got used for capturing, cleaning and processing data received from the online e-commerce portal. They
showed that these sources are often cheap and enable users near real-time calculation of price index numbers. In a recent work by Banerjee, Singhal, and Subramanian (2018), the same concept was adopted in the Indian context, where they argued that despite several limitations of online prices, the web-data based index successfully tracks both the direction and the magnitude of Indian official Consumer Price Index (CPI).

Use of big data techniques and the web scraping, in the context of the residential property price, is limited. Several studies in this domain mainly include usage of Google Trend data in a nowcasting framework. Following the seminal work by Choi and Varian (2012), the scope of research has expanded rapidly, taking advantage of publicly available Google Trend data. Particularly, in the field of real-estate market and housing price, the use of Google Trend data was considered initially by Wu and Brynjolfsson (2015). Recently the similar approach is adopted in the context of India by Mitra, Sanyal, and Choudhury (2017). In this context, a similar project was described by Kristiawardani and Sampe (2017) from Bank Indonesia where they showed how the online advertisement data for Indonesia was successfully tracking property price indicator.

Typical big data analysis relating to indicator construction from the web scraped data usually involves a considerable amount of time for data collection in order to create a time series. This article adopts a novel approach to bridge this gap employing a specific kind of web scraping technique, which may be termed as “dynamic chart scraping”.

Subsequent sections of this article are organized as follows: Section 2 provides an overview of internet penetration in India, a review on the property advertisement websites in India and describes the web scraping framework adopted in this study; Section 3 summarizes the web scraped data and discuss the data cleaning and preparation strategies; Section 4 defines the index formulation methodology; Section 5 describes a comparative study among the proposed and established housing price indices; The observations based on empirical study is described in Section 6. Finally, Section 7 adds the concluding remark highlighting further scope of research.
2. Background

I. Internet Penetration in India

Studies, where data originated from the internet, is being used to infer on the general behaviour of the economy, a natural issue of internet penetration is raised towards the validity of the empirical results. Naturally, in this study the context of internet penetration is inevitable.

Figure 1, which is prepared with the World Bank data, represents the number of internet users as a percentage of the population, over different time regimes.

Developing countries are constructed based on the classification of the World Bank. The percentages in figure 1 corresponding to India for the recent years (2017 and 2018) is calculated by adjustments to the data as mentioned in the newspaper article (2018) and forecasted population estimate from IMF. The data in the former newspaper article was actually generated based on a report from IAMAI and Kantar IMRB. The adjustment is carried out on the forecasted series to match the World Bank percentages corresponding to the overlapping years.
It is evident from the Figure 1 that internet penetration level is maturing rapidly in the recent years.

Adoptability of internet is known to be concentrated in urban regions. However, since the real estate activities are heavily urbanized, the online advertised prices for housing are likely to represent the appropriate population.

II. Selection of Property Advertisement Sites

There are many websites in India for the listing of property advertisements. However, for preparing a list of top websites it is essential to know the web traffic shares in these websites. Many online services available which provide web traffic-related information for popular websites. This kind of services are generally referred to as web traffic analytics tool or online competitive intelligence tool. Two such services are considered in this study which are SimilarWeb, a private company founded in 2007 and Alexa Internet, a subsidiary of Amazon founded in 1996. The Table 1 has been prepared based on the data collected from these portals for various property sites in India.

<table>
<thead>
<tr>
<th>Websites</th>
<th>SimilarWeb</th>
<th>Alexa</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monthly Visits (in Millions)</td>
<td>Global Rank</td>
</tr>
<tr>
<td>99acres.com</td>
<td>3.90</td>
<td>3,407</td>
</tr>
<tr>
<td>magicbricks.com</td>
<td>3.30</td>
<td>5,450</td>
</tr>
<tr>
<td>housing.com</td>
<td>1.30</td>
<td>13,268</td>
</tr>
<tr>
<td>makaan.com</td>
<td>0.94</td>
<td>18,360</td>
</tr>
</tbody>
</table>

Table 1

Above table depicts the leading property websites in India. Websites magicbricks.com and 99acres.com are major competitors in this domain and holds maximum monthly traffic share among all.

---

3 The numbers are collected as of 31 May 2018
To verify the ranking provided by the web traffic analytics tools, a short analysis of Google Trend data is performed for these websites.

The Figure 2 (Google Trend analysis), depicts the similar ranking of websites as obtained from the web-traffic analytics tools.

As a nature of the competitive market, these leading property sites tend to provide similar online services. One of such services is property price trends for specific locations, provided for price tracking. Limited property websites among leading ones, provides this facility. Top two sites (99acres.com and magicbricks.com) provides them for the majority of the locations.

Figure 2

To reduce sampling noise in the data obtained from Google Trend platform, the same data is captured during many days. Additionally, a trend component was retrieved from the median time series (monthly), using X-13ARIMA-SEATS. The figure depicts the trends together with a sample raw series (as background) and overall search volume share. Each search terms here is considered as a search topic, instead of exact keywords.

---

4 To reduce sampling noise in the data obtained from Google Trend platform, the same data is captured during many days. Additionally, a trend component was retrieved from the median time series (monthly), using X-13ARIMA-SEATS. The figure depicts the trends together with a sample raw series (as background) and overall search volume share. Each search terms here is considered as a search topic, instead of exact keywords.
In this study, one of these two sites is picked randomly as a source website, for the web scraping activity.

II. Web Scraping Framework

Within the source website, the price trend is presented in a form of a dynamic chart. The web scraping technology was applied to these dynamic charts to obtain desired housing price information.

The data is represented in the source website, as different dynamic charts, for different property types and regions of cities within India. The whole workflow of scraping information from these charts is described in Figure 3.
3. Data Preparation

The housing price data, obtained using the dynamic chart scraping technology, are segregated for each region of different cities and property types. For each such location (a region of a city) and property types, there are quarterly time series corresponding to different aggregations (like minimum, maximum, average) of the raw prices (per unit property price, corresponding to various properties, listed in the website during that point of time). All these prices are given per standard unit cost (like per square feet, per square yards) which depends on property type only.

The whole data points may be represented as below.

\[ P_{ijkt}^{a} = \text{Price for } \begin{cases} \text{i}^{\text{th}} \text{ city, } j^{\text{th}} \text{ region where } i \in I \text{ (set of cities)} \\ \text{and } j \in J_i \text{ (set of regions in a city)} \\ \text{in other words } (i,j) \in L \text{ (set of locations)} \\ k^{\text{th}} \text{ category of property where } k \in K \text{ (set of property categories)} \\ a^{\text{th}} \text{ type of price aggregation } (A \text{ being the set of these}) \\ t^{\text{th}} \text{ quarter (time point) where } t \in T \text{ (set of time points)} \end{cases} \]

This dataset has a large amount of missing values. To reduce the large amount of sparsity, raw scraped data, is considered for further processing where estimated figures are introduced. The following symbol is defined for estimated figures.

\[ \hat{P}_{ijkt}^{a} = \text{Estimated } P_{ijkt}^{a} \text{ where } (i,j,k,t,a) \text{ is not present} \]

During the estimation, several time series modelling techniques were considered. However, the method of interpolation using local polynomials is found to be adequate enough for this purpose.

After the step of estimating missing observations, to the extent possible, the whole data containing raw and estimated figures are considered for time series outlier adjustment. Three approaches are considered for outlier identification. These are, outlier detection rule as
embedded in the X-13ARIMA-SEATS program, additive outlier detection rule as described in Chen and Liu (1993) and lastly a heuristic approach involving moving interquartile ranges. Finally, the heuristic approach was used, as the other two are classifying the majority of the data as outliers, making the adjusted series too smooth. No observation is added or removed at this stage. Only outliers falling beyond the limits are adjusted to remain within the boundary. After this stage, the data is considered for index compilation.

The scraped objects contain geographic information (GI) for the majority of the locations. However, for a few locations, the latitude and longitude are missing. For these locations, Google Geocoding\textsuperscript{5} API is utilized through R as described by Kahle and Wickham (2013). This partly resolves the missing information on GI data. This geo-location information provides an important insight related to location coverage of the data. In the Figure 4 the geo-location information is plotted on the map of India\textsuperscript{6} over different time to describe how more and more locations are getting added to the directory of the property advertisement website.

![Web Data Locations Coverage Map: Variations over time](image)

---

\textsuperscript{5} Geocoding is the process of converting addresses into geographic coordinates. Reference: https://developers.google.com/maps/documentation/geocoding/

\textsuperscript{6} Note that the map of India is designed not to scale in all subsequent figures and are included in this article only for graphical illustration. The administrative boundary of India is constructed following all prevailing laws and resembles with Survey of India map. The map is made available by Data {Meet} Community Maps Project under the Creative Commons Attribution 2.5 India.
The base period for index construction is taken as 2013-Q2\textsuperscript{7}. For comparative purpose, new locations added after the base quarter is discarded.

The Figure 5 visualizes the data coverage graph over time, for all kinds of data obtained at different stages of the data cleaning process. The red shaded region corresponds to the data used for the Housing Price Index (HPI) calculation.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{Locations coverage in various data over time}
\end{figure}

In order to understand the location coverage of different HPI, Figure 6 is obtained using geo-location of the cities covered under different HPI. Common cities are marked in blue.

\textsuperscript{7} The quarter format used in this study follows usual calendar style as opposed to Indian fiscal year. Accordingly, quarter-I (Q1) refers to the period of January to March.

The base quarter should not be very old, as that will result in the omission of lot data. It should not be very recent also, as newly introduced locations may influence oddly into the calculation inducing more volatility.
City Coverage Map : Different indices

<table>
<thead>
<tr>
<th>HPI@RBI</th>
<th>HPI@WebData</th>
<th>RESIDEX@NHB</th>
<th>RPPI@RBI</th>
</tr>
</thead>
</table>

|| Map not to Scale || (Applicable for each maps. Administrative boundaries are only for graphical illustrations.)

4. Index Compilation Methodology

As described by Meese and Wallace (1997), several well-known methods (like repeat-sales, hedonic-regression) are available for computation of housing price index. In this study, since the access to the historical housing price is available through aggregates, option for computation of the index is limited to weighted average method only. The method is similar to what was explained in the occasional publication of Reserve Bank of India (2015)\(^8\) (the article explaining methodologies relating to RPPI@RBI) and a bulletin item by Statistical Analysis Division of the Department of Statistics and Information Management (Reserve Bank of India) (2014)\(^9\) (the article describing compilation methodology of HPI@RBI). The procedure is explained below using mathematical notations.

Considering one specific time point as base-period (base quarter, let say we denote it by \(t_0 \in \mathbf{T}\)) price relative for each of these categories are calculated as \(\pi_{ijk}^a = \frac{P_{ijk}}{P_{ijk0}}\). Wherever the price information is not available it may be replaced by corresponding estimates \(\hat{P}_{ijk}^a\). Now the weights for each of the price relatives \((\pi_{ijk}^a)\) are denoted using following notations

\(^8\) https://rbi.org.in/Scripts/PublicationsView.aspx?id=16223
\[ \omega_{ijk} = \text{Weight for} \begin{cases} 
  \text{i}^{\text{th}} \text{ city,} \ j^{\text{th}} \text{ region where} \ (i,j) \in L \\
  \text{k}^{\text{th}} \text{ category of property where} \ k \in K \\
  \text{a}^{\text{th}} \text{ type of price aggregation where} \ a \in A \\
  \text{It is independent of quarter (time point) t}
\end{cases} \]

All weights corresponding to different attributes, except weights for cities, are constructed from the cleaned data itself. To derive the all India level figure, population figures from 2011 Census Data\(^{10}\), of each city is considered.

To make the weights unit-free, weights which are linked to price values are calculated based on normalized price.

In the first step, the weighted price relative on different aggregate types, for specific property type and location \((\pi_{ijk} = \frac{\sum_a \pi_{ijk} \omega_{ijk}}{\sum_a \omega_{ijk}}\) are derived. All of these weighted price relatives are then further aggregated to location level for each quarter as \(\pi_{ij} = \frac{\sum_k \pi_{ijk} \omega_{ijk}}{\sum_k \omega_{ijk}}\). Here, \(\omega_{ijk}\) is the weight for \((i,j)^{\text{th}}\) location and \(k^{\text{th}}\) property type. For each quarter, the weighted price relatives \((\pi_{ij})\) corresponding to each location are further aggregated using region weights \((\omega_{ij})\) to achieve city level quarterly weighted price relatives \((\pi_{it} = \frac{\sum_j \pi_{ij} \omega_{ij}}{\sum_j \omega_{ij}}\). Finally, to obtain quarterly all India level figure, assuming the city and region have representative coverage, these price relatives \((\pi_{it})\) are used for the calculation of weighted average taking city weights as \(\omega_i\). This figure is donated as \(\pi_i = \frac{\sum_j \pi_{ij} \omega_i}{\sum_j \omega_i}\).

The weighting pattern considered for aggregation levels\(^{11}\) assigns \((\omega_{ijk})\) highest weight to average aggregation type and lower weight on the other two \((\omega_{ijk}^{\text{upper}} < \omega_{ijk}^{\text{average}} > \omega_{ijk}^{\text{lower}}\). It is defined as \(\omega_{ijk}^{a} = e^{-|\bar{P}_{ijk0} - \bar{P}_{ijk0}^{\text{average}}|}\). (Here \(\bar{P}_{ijk0}\) denote appropriate average of the normalized unit free price at base quarter). Choice of weights corresponding to each property types \(\{\omega_{ijk}\}\) is constructed based on the count of regions \((j)\) offering that specific property types in a city \((i)\) on base quarter \((t_0)\). The weights for regional locations \(\{\omega_{ij}\}\) are constructed as \(\omega_{ij} = e^{-\bar{P}_{ij0}}\) (inversely related to the average normalized price of a region at

\(^{10}\) http://censusindia.gov.in/

\(^{11}\) The set A (of all possible aggregation types) has three elements which are upper, average and lower.
base period). Finally, population size (using Census 2011 data) for each city is taken as respective city weight \( \{ \omega_j \} \), following the approach as adopted in HPI@RBI and RPPI@RBI construction.

5. Comparative Study

The proposed index (HPI@WebData) is compared with existing indices for the validation. As mentioned in earlier sections, HPI from different sources will be termed as a) HPI@RBI (HPI compiled by Statistical Analysis Division of the Department of Statistics and Information Management, Reserve Bank of India), b) HPI@WebData (The HPI compiled from dynamic chart scraped data and is proposed as an alternative index in this article) c) RESIDEX@NHB (NHB RESIDEX, India’s official housing price index from National Housing Bank, undertaken at the behest of the Government of India, Ministry of Finance. HPI at assessment price is taken here) and d) RPPI@RBI (index which was compiled by RBI based on transaction level data on housing loans).

![Comparison of different indices](image.png)

**Figure 7**
In HPI@RBI and HPI@RPPI the all India level index is computed and disseminated using city population as city weights (same Census data 2011 is used). Similar aggregation method is applied on RESIDEX@NHB\(^{12}\) to obtain all India figures (\(\pi_t = \frac{\sum x_t w_t}{\sum w_t}\)). The numbers corresponding to “common cities” is obtained based on the same formula including only common cities among all indices (Cities indicated by blue dots in the Figure 6).

The data corresponding to all indices was collected in May 2018. During that time only HPI@WebData was available for 2018-Q1 (January-March quarter of 2018).

The base quarter for each of these indices is shifted to match that of HPI@WebData (2013-Q2).

---

**Comparison of different indices**

All indices are normalized at same base quarter (2013 Q2)

The calculation for HPI@WebData is considered after 2011-Q1, as prior to that scraped data have huge noise and large volatility. This is possibly caused by low internet penetration

---

\(^{12}\) NHB does not disseminate RESIDEX at all-India level.
at that period and possibly because the online property advertisement platforms were still in the maturing phase.

Figure 7 and Figure 8 represents the comparison of index levels corresponding to aggregate times series based on all available locations (referred to as all India) and common cities respectively. In both figures, the proposed index (HPI@WebData) closely follow a similar trajectory with RESIDEX@NHB.

Figure 9 and Figure 10 are the year-on-year growth rates of the actual indices, respectively for all locations (all India) and common cities. The co-movement of the HPI@WebData series with RESIDEX@NHB, for common cities, show how the housing price dynamics in both of these indices match. Hence, price movement in HPI@WebData captures well, the patterns, represented by RESIDEX@NHB.

13 The sine curve like pattern, in all India figure, corresponding to RESIDEX@NHB is mainly induced by price variation in Delhi.
A heat map (Figure 11)\textsuperscript{14} has been generated to display the interrelationship among these indices (index levels are used in this analysis).

The heat map again shows that HPI@WebData have higher co-movement with both of the indices RESIDEX@NHB and HPI@RBI. Note that HPI@WebData and RPPI@RBI have 8 cities in common. In which all of the cities are having more than 70% co-movement. However, the mean $R^2$ is below 1.

\textsuperscript{14} For a pair of indices, the correlation (Kendall’s $\tau$, this will be referred as co-movement in subsequent sections) of time series for common cities are calculated. Then mean $R$-square is obtained from the average of square correlations corresponding to common cities. The heat map cell colour depth is proportional to mean $R$-square. Note that this list of common cities varies across different pairs of indices. Number of cities with co-movement higher than 70% are presented in brackets in each cell. The fraction of cities having co-movement higher than 70% is given outside the brackets in each cell. The diagonal cells of the symmetric heat map provide information about how many locations are present within an index.
6. Empirical Findings

A small overview of the regional price patterns based on HPI@WebData is given through a geo-tagged short time series plot in Figure 12. This diagram is used as an alternative to the commonly used Choropleth map. In this figure, each time series (provided for selected large cities) is given since 2015-Q1 (scales of each time series is different from others). The colour codes for each of these trends are determined by the direction of the latest trajectory (computed by the weighted mean of segment-wise co-movement with time when higher weights are assigned for recent years). More bluish tone indicates that the price for that location is recovering while the reddish tone indicates the residential price is falling in general.
Form the Figure 12 it is seen that residential house price in Mumbai and Delhi is in recovering phase. While cities in the south region (Chennai, Bangalore and Kochi) are facing diminishing pattern (the similar situation is observed for Jaipur and Ahmedabad). Prices in Kolkata, Hyderabad and Lucknow shows a potential recovering phase in recent time after a strong decline in the past.

7. Conclusion and the way forward

Compiling residential price indices has always been a challenge. Actual transaction in the diverse country like India is an information of rare availability. The data lack transparency
and get collected in a lagged manner. Thus, compiling a perfect indicator is a major overhead. In the presence of two established channels for dissemination of housing price index, this study introduces another indicator (HPI@WebData) utilizing the cutting edge big data and machine learning techniques which tracks the residential property price trajectory appropriately, at least to the extent of the existing indices. As it shares the benefit of almost no data lag, this index can be used as an advance estimate of residential property price movement. The added advantage in this method of data collection and index compilation is that it reduces the cost and improves the timeliness of the series.

The data collected has unexplored endless possibilities. Like the data contains, for each location, information about nearby points of interests, starting from the number of bus-stops to distance to the airport. This information may be utilized to evaluate the most important factor for price changes in each location. The data also contains commercial property price, for which currently not only in India, also in a considerable number of countries, there is no official indicator. Exploring the same should ideally lead to an interesting study. Thus, this work leaves ample opportunities for the future research work in this sector.
References


