Rachman, Arief Noor

Article

An alternative hedonic residential property price index for Indonesia using big data: The case of Jakarta

Eurostat Review on National Accounts and Macroeconomic Indicators (EURONA)

Provided in Cooperation with:
Eurostat, Luxembourg

Suggested Citation: Rachman, Arief Noor (2019) : An alternative hedonic residential property price index for Indonesia using big data: The case of Jakarta, Eurostat Review on National Accounts and Macroeconomic Indicators (EURONA), ISSN 1977-978X, Publications Office of the European Union, Luxembourg, Iss. 2, pp. 73-93

This Version is available at:
https://hdl.handle.net/10419/309835

3 An alternative hedonic residential property price index for Indonesia using big data: the case of Jakarta ([1])

ARIEF NOOR RACHMAN ([2])

Abstract: Monitoring property price dynamics is a necessary task for central banks in order to maintain financial stability in the economy. Big data offers potential as a new source of data that might be used to produce official statistics on property. In this paper, we develop an alternative residential property price index (RPPI) for the secondary market for houses from online residential property listings using the time-dummy hedonic regression method. The dataset is based on residential property advertisement listings from Indonesia's major property web portals from January 2016 to September 2018. For this prototype index, the study initially focuses on Jakarta, the capital city of Indonesia. Our regression outputs generally show promising results and have the potential to become an official housing index. Future development will extend the index coverage to other large cities in the country and improve the characteristic variables in the model.

JEL codes: C43, E30, R31

Keywords: residential property prices index, big data, hedonic regression time-dummy method

([1]) Revised version of a paper that was presented to Eurostat's International Conference on Real Estate Statistics in Luxembourg, 20-22 February, 2019.

([2]) Statistics Department of Bank Indonesia.

1. Introduction

Most central banks monitor property prices as part of their work to maintain national financial stability. Therefore, Bank Indonesia (BI) has established a residential property prices survey and commercial property prices survey on a quarterly basis. The bank has been conducting two kinds of surveys for residential property statistics, both for the primary market (newly-built houses) and the secondary market (used houses or second-hand houses) with different methods to compute a residential property price index (RPPI). The primary market RPPI is computed using a chained index method based on the list price of new houses as provided by key property developers in 16 major cities, whereas appraisal methods are used to compute an RPPI for the secondary market in 10 major cities.

The compilation of an RPPI is a tricky process. A number of problems may arise across different stages of the process, from the identification of data sources to index calculation methods (Eurostat (2013)). The identification of a reliable data source is an issue that arises when computing an RPPI for Indonesia. Data on declared property transactions such as administrative data from the land registry or property tax records are difficult to acquire. The decentralisation of the administration for property taxes to local or municipal government makes it more difficult to collect data on transactions due to non-standard data records, wider coverage and reluctance on the part of local government to share data. These kinds of problem have led Bank Indonesia to conduct property price surveys using list prices from developers and appraisals of the activity of real estate agents as alternatives to the compilation of an RPPI. Timeliness is one of the main benefits when using asking (or listed) prices to construct property price indices. Nevertheless, this approach can potentially be a major weakness, insofar as differences between asking prices and actual transaction prices may result in misleading estimates. However, an RPPI based on asking prices may be considered a feasible solution for monitoring purposes, especially in the absence of data for actual transactions (IMF (2018)). Lyons (2019) also found that asking/listed prices can be an accurate indicator of actual transaction prices in Ireland's market for houses.

Recent developments of digital data known as 'big data' offer opportunities and benefits to official statistical institutions, such as: the ability to produce new indicators; the possibility to bridge time lags for existing official statistics; and the provision of an alternative source of data to produce official statistics. According to Hammer et al. (2017), big data is defined as a by-product of business and administrative systems, social networks and the internet of things and is often characterised by its high-volume, high-velocity and high-variety (3Vs) of data. However, big data also provides several challenges that need to be overcome, such as: i) concerns over data quality ([3]); ii) ensuring (legal) access to the data, considering big data is typically owned by private entities; and iii) developing advanced skills and making available the necessary technology to make use of such data (Das et al. (2014) and Hammer et al. (2017)). Therefore, statistical institutions need to be careful when using big data as a new source of official statistics.

This paper develops an alternative RPPI, making use of big data by drawing on online advertisements for property. The use of big data has the potential to improve the compilation of the RPPI and to challenge the existing practices employed to compute the existing RPPI.

([3]) This concerns having to deal with large volumes of data that need to be processed, cleaned/verified and then summarised without losing too much information.

Online data based on the listed price of properties as found in advertisements offers an immediate, inexpensive, and considerable amount of (alternative) data for constructing an RPPI. This study employs a direct hedonic approach to calculate robust property price indices based on the availability of data for various property characteristics. As a prototype index, the coverage for this study is limited to the Indonesian capital city, Jakarta.

This paper is organised as follows. The first section provides an introduction detailing the background to this study. Section two explains the data and the methodology used. A discussion of the results is presented in section three. Finally, conclusions and further work are presented in section four.

2. Data and methodology

2.1. Data sources

In the development of an alternative RPPI using big data, we collected monthly data from the two largest web portals for property advertisements in Indonesia, which together account for more than 50 % of the total market. Bank Indonesia secured the acquisition of these data through non-disclosure agreements (NDAs) with the two websites. The preparation and extraction of data from the two web portals was organised using virtual machines and Hadoop software (more details concerning the steps taken are presented at the end of this article in a *Data appendix*).

The data used were individual listings/advertisements for property with the following attributes: initial asking (or listed) price, offer type (for sale or rent), property type, lot size, dwelling size, number of bedrooms, number of bathrooms, address, and additional characteristics which are recorded as a 'free-text' description, such as the presence (or not) of a garage, gated property, swimming pool, its specific location or its distance from public facilities, and so on. For simplicity purposes, these 'free-text' characteristics were left aside in this study because the information was often incomplete and too granular to extract.

As it was initially an experiment, the study only focused on listings of houses that are 'for sale', while other types of residential property such as apartments/flats were excluded and left for future developments. Furthermore, the study only included the first instance of any advertisement/listing for each property; as such, only listings from the first month that an advertisement appeared were included, unless the listed price subsequently changed. Taking the listed price of each property on a monthly basis would implicitly give a larger weight to those properties which take longer to sell.

As mentioned above, the study was limited to computing an RPPI only for Jakarta, the capital city of Indonesia. Jakarta is one of the biggest cities in the world with a population of around 10 million people. As the nation's capital and the city with the largest population in Indonesia, Jakarta has the highest number of property transactions in Indonesia (when compared with other cities). Jakarta is believed to account for around one third of the national property market. Our dataset shows that Jakarta accounted for around 36 % of all online listings at a national level.

The study also used another dataset as a set of weights when aggregating the hedonic indices for each district into a composite property price index for houses across the whole of Jakarta. The total value of mortgage collateral by regions/district was used as a proxy for all transactions, in the absence of data covering both cash and mortgage-financed transactions. This dataset consists of individually appraised collateral values for mortgage loans that are derived from the centralised banking debtor information system which is jointly managed by Bank Indonesia and the Indonesian Financial Services Authority (OJK). We found that the share of each district in the total value of mortgage collateral value was relatively closely related to the share of each district in the total number of observations; in both cases South Jakarta had the biggest share among the five districts that compose the Indonesian capital. A table with this comparison is presented in the *Data appendix*.

2.2 Data treatment

Cleaning the data set was a crucial step when dealing with sources of big data due to concerns over data quality (as mentioned above). Since we obtained data from individual listings on web portals there were several issues regarding the data, including: (1) human error in data entry; (2) non-standard addresses — as a free-text field is employed; and (3) duplicate advertisements (which are mainly caused by the fact that one property can be advertised by more than one seller in a single portal as well as across different portals, and advertisements tend to be re-posted after their initial expiration date if the property has not been sold). When preparing the dataset, we removed all duplicate records and any data considered to be corrupt/incomplete.

The next step was to make statistical edits based on the assumption of a normally distributed dataset. We removed spurious price values using a median absolute deviation (MAD) test on price per unit of property size; the same method was also applied for building size. We removed those observations where the lot size was greater than 600 m^2, and also deleted any observations where the number of bedrooms was greater than 10 and the number of bathrooms was greater than eight, based on histogram (or bell-shaped curve) patterns. We trimmed any data cells which lay outside of the observed bell curve tails (outliers).

Finally, we ran a preliminary regression to identify further outliers using the Cook's distance method. This method identified outliers based on the combination of each observation's leverage and residual values, with the following formula:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(p+1)\hat{\sigma}^2}$$

Where D_i is Cook's distance for observation *i*, $\hat{Y}_j$ is the fitted response value, $\hat{Y}_{j(i)}$ is the fitted response value obtained when observation *i* is removed, $\hat{\sigma}^2$ is the mean squared error of the regression, and *p* is number of predictors. We removed those observations with value of D that was greater than $4/n$, a conventional standard (see O'Hanlon (2011)). Table 1 presents a record of the preparation steps applied for each district of Jakarta. On average we identified (and removed) around 5 % of records as outliers using Cook's distance method, while an additional 25 % of the data was removed due to statistical edits.

Table 1: Data records

Year	Steps	Districts				
		West Jakarta	Central Jakarta	South Jakarta	East Jakarta	North Jakarta
2016	# obs of active listings	618 858	110 293	852 252	457 734	369 972
	# obs of first instance listings (1)	77 031	14 527	103 925	55 930	42 378
	# obs raw data (2)	40 141	7 357	59 922	30 625	20 370
	# obs after stat edits	32 868	4 946	41 541	23 421	15 474
	# obs (3)	31 184	4 662	39 626	22 180	14 640
	# obs percentage from raw rata (%)	**77.7**	**63.4**	**66.1**	**72.4**	**71.9**
	# obs percentage from stat edits (%)	**94.9**	**94.3**	**95.4**	**94.7**	**94.6**
2017	# obs of active listings	579 507	106 175	905 145	450 882	319 099
	# obs of first instance listings (1)	82 150	14 159	130 380	67 901	39 626
	# obs raw data (2)	38 477	7 650	66 294	36 237	18 498
	# obs after stat edits	31 264	5 046	43 575	28 921	14 795
	# obs (3)	29 831	4 762	41 647	27 426	14 026
	# obs percentage from raw rata (%)	**77.5**	**62.2**	**62.8**	**75.7**	**75.8**
	# obs percentage from stat edits (%)	**95.4**	**94.4**	**95.6**	**94.8**	**94.8**
2018 (9 months)	# obs of active listings	403 232	78 728	657 112	348 836	225 357
	# obs of first instance listings (1)	64 024	14 666	151 420	88 366	45 423
	# obs raw data (2)	23 743	4 673	37 877	30 272	14 163
	# obs after stat edits	20 114	3 516	29 619	20 645	10 998
	# obs (3)	19 153	3 331	28 231	19 509	10 370
	# obs percentage from raw rata (%)	**80.7**	**71.3**	**74.5**	**64.4**	**73.2**
	# obs percentage from stat edits (%)	**95.2**	**94.7**	**95.3**	**94.5**	**94.3**

(1) Only new advertisements (ads) in each month, removing repeated advertisements.
(2) Raw data after removing the duplicated advertisements and corrupted data.
(3) Clean data after statistical edits and outliers are removed using the Cook's distance method.

2.3 Methodology

As mentioned in the Introduction, the calculation of an RPPI is a complex process because houses are infrequently sold and heterogeneous in terms of their structural characteristics such as location, size and facilities. This may lead to quality issues for price measurements since the differences in housing characteristics are hard to control, especially with a limited frequency of transactions. To identify factors for quality changes, quality-mix adjustments are needed to avoid misleading interpretations of the resulting indices. Silver (2016) identified several methods for making quality-mix adjustments such as hedonic methods, repeat sales, sales price appraisal ratios, and so on. The hedonic method is believed to be more preferable when compared with repeat sales due to its ability to use data on relevant property characteristics using regression techniques (Hülagü et. al (2015)). Furthermore, hedonic

regression analysis of house prices decomposes the overall price and provides estimates of marginal value for each of the characteristics. Li et al. (2006) highlighted three main approaches for hedonic methods: the time-dummy approach; the characteristics approach; and the hedonic (price) imputation approach. For a better explanation of these methods, see Diewert (2003), Hill (2012) and Silver (2016).

Silver (2016) indicated that both the characteristics and hedonic imputation approaches have major advantages over the time-dummy approach, but for simplicity we decided to continue with the time-dummy approach since it can immediately derive a price index from the estimated time-dummy coefficients. Li et al. (2006) also mention that the imputation method would likely give the same result as the hedonic method given the same dataset.

Our model specification used the semi-log regression model since the variable for house prices (in levels) was not normally distributed (there was a positively skewed distribution). The basic semi-log hedonic model is represented as follows:

$$\ln p_n^t = \beta_0^t + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + \varepsilon_n^t$$

Where:

p_n^t is the price of property *n* at time *t*;

z_{nk}^t is k characteristic variables of property *n* at time *t*;

β_0 and β_k are intercepts and house characteristic parameters; and

δ^τ is a dummy coefficient.

Our hedonic model only has quantitative characteristics such as building size, lot size, number of bedrooms and number of bathrooms. The number of bedrooms and the number of bathrooms are treated as dummy variables. We have three dummy variables for the number of bedrooms — one and two bedrooms, three bedrooms, and greater than four bedrooms (four bedrooms is used as a reference based on the highest frequency for the number of bedrooms). We also have three dummy variables for the number of bathrooms (with three bathrooms as a reference).

One of the shortcomings of this model specification is the lack of other characteristic variables that may be important (such as locational advantage, the condition of the property/house, or the age of the building). However, we had difficulties to identify the location advantage of property because the address and important information about the location were often incomplete, had less detail or were composed of 'free-text' information. The main work done so far in this area was centred on identifying property locations at the district level. On the other hand, information about building age or major renovation records were generally not disclosed in the advertisements. In order to identify the location advantage, we stratified the calculation of indices into five districts that together formed the metropolitan area of Jakarta, in other words, Central Jakarta, North Jakarta, East Jakarta, South Jakarta and West Jakarta.

To calculate an RPPI, we followed the methodology/calculations employed for the Japanese residential property price index (JRPPI) (Land Economy and Construction Industries Bureau (2016)), using the rolling window technique to compute an RPPI from the hedonic regression time-dummy method. Estimated time-dummy coefficients ($\hat{\delta}^{\tau}$) were arranged as follows:

Table 2: Rolling window technique for the compilation of time-dummy coefficients

Regression r	*1*	*2*	*3*	*…*			*…*	$T-\tau+1$	*…*	T
1	$\hat{\delta}_1^1$	$\hat{\delta}_1^2$	$\hat{\delta}_1^3$	…	$\hat{\delta}_1^{\tau}$					
2		$\hat{\delta}_2^2$	$\hat{\delta}_2^3$	…	$\hat{\delta}_2^{\tau}$	$\hat{\delta}_2^{\tau+1}$				
3			$\hat{\delta}_3^3$	…	$\hat{\delta}_3^{\tau}$	$\hat{\delta}_3^{\tau+1}$	…			
…				…		…	…		…	
$T-\tau+1$								$\hat{\delta}_{T-\tau+1}^{T-\tau+1}$	…	$\hat{\delta}_{T-\tau+1}^{T}$

The index can be obtained by:

$$\frac{p^{\tau+1}}{p^1}=\exp\left(\hat{\delta}_1^{\tau}\right)\times\frac{\exp\left(\hat{\delta}_2^{\tau+1}\right)}{\exp\left(\hat{\delta}_2^{\tau}\right)}$$

Suppose the base period is the first period, then the price difference between the price at period 1 and the price at period $\tau+1$ can be obtained based on the time-dummy parameter calculated for the last period (time τ) of the first window time range and the time dummy parameters for the last period and the second to last periods of the next window time range (time τ and time $\tau+1$). By sequentially conducting the aforementioned calculations for all window time ranges, quality-adjusted price indices may be obtained for all time windows.

As done for the compilation of the JRPPI, the length of the window time was set to one year (12 months) which is common for analysis, as illustrated by Silver (2016). We assumed that this window would allow us to capture the seasonal dynamics of the market for houses. To compile a composite RPPI for the whole of Jakarta, we aggregated the indices for all five districts using total collateral mortgage values as weights; these data were provided by banks in 2017. Collateral mortgage values were used as a proxy for property transaction values in the absence of a more representative measure for the structure of the property market, such as tax revenues from property transactions.

3. Results

3.1. Regression results

The information used in this study indicated that house prices (in levels) were not normally distributed. Hence, we transformed the variable for house prices into a logarithmic format before running semi-log hedonic regression models. We ran a 12-month rolling windows regression from January 2016 to September 2018 for five different districts in Jakarta with a total of 110 regressions.

We ran two stages of regression, the first was to identify outliers using Cook's distance and the second regression (without outliers) to produce the index. We present a sample set of results below for a 12-month rolling window regression from January 2016 to December 2016 in North Jakarta and South Jakarta (see Table 3). These results show a relatively high degree of explanatory power as indicated by the adjusted R-square values. Given the limited availability of variables for house characteristics, such a high degree of explanatory power probably implies a relatively homogenous market for houses in these districts.

Table 3: Hedonic regression results for North Jakarta and South Jakarta

	North Jakarta			South Jakarta		
Dependent variable: Ln Price						
Independent variables	**Estimates**	**Robust standard error**		**Estimates**	**Robust standard error**	
Intercept	21.2900	0.00851	***	20.8540	0.00980	***
Building size	0.0010	0.00002	***	0.0023	0.00002	***
Lot size	0.0045	0.00003	***	0.0031	0.00002	***
Dum_# of bedroom 1-2	-0.2153	0.00824	***	-0.1042	0.00507	***
Dum_# of bedroom 3	-0.0440	0.00456	***	-0.4461	0.00983	***
Dum_# of bedroom >4	-0.0400	0.00536	***	-0.0929	0.00545	***
Dum_# of bathroom 1	-0.2225	0.00879	***	-0.2965	0.01033	***
Dum_# of bathroom 2	-0.1732	0.00478	***	-0.1389	0.00564	***
Dum_# of bathroom >3	0.0082	0.00492	*	-0.0095	0.00502	*
Dum_period 2016:2	0.0007	0.00867		-0.0132	0.01039	
Dum_period 2016:3	0.0006	0.00888		0.0164	0.01023	
Dum_period 2016:4	0.0034	0.00913		-0.0175	0.01053	*
Dum_period 2016:5	-0.0203	0.00765	***	0.0003	0.00921	
Dum_period 2016:6	-0.0123	0.00971		0.0390	0.01094	***
Dum_period 2016:7	-0.0132	0.00936		0.0015	0.01085	
Dum_period 2016:8	0.0022	0.00954		0.0148	0.01070	
Dum_period 2016:9	-0.0143	0.01008		0.0307	0.01086	***
Dum_period 2016:10	-0.0169	0.00829	**	0.0268	0.00974	***
Dum_period 2016:11	-0.0307	0.00916	***	-0.0158	0.00991	
Dum_period 2016:12	-0.0435	0.01056	***	-0.0030	0.01052	
Adjusted R-squared	0.863			0.783		
F-statistics	4 866			7 507		
Number of observations	14 640			39 626		

Note: *** significance at 1 %, ** significance at 5 % and * significance at 10 %.

Across all five districts of Jakarta, all characteristic variables in the model were statistically significant, stable, and in line with *a priori* expectations over time ([4]). On average, building size and lot size had a positive impact on house prices, while the number of bedrooms and the number of bathrooms had mixed results. Given no change in any other characteristics, it seemed that as the number of bedrooms increased beyond four this had a negative impact on house prices as it could reduce the living space available in the remainder of the house. The same argument applied to the number of bathrooms in South Jakarta, while for North Jakarta this result was in line with the *a priori* expectations. The results also implied that a house with an additional 10 m^2 of lot size would be 4.5 % more expensive than average (if other variables were kept constant).

Regression results for the South Jakarta district (based on a larger number of observations) provided similar findings. The explanatory power dropped slightly but still remained relatively high, with an adjusted R-squared value of 0.78. All of the house characteristics were significant, building size had twice the impact (compared with the results for North Jakarta) while the impact of lot size in South Jakarta was slightly less significant.

The Breusch-Pagan test was applied to detect heteroscedasticity. The result showed that heteroscedasticity was present in the model. Since heteroscedasticity only affects standard errors and the coefficients remained unbiased, we calculated robust standard errors to improve the value of the t-statistic.

3.2 Indices

Adopting the rolling window method, we estimated time-dummy regression coefficients for an RPPI for each of the five districts that compose the Indonesian capital city. The monthly indices suffered from short-term volatility, thus we employed three-month moving averages to smooth out the series ([5]). Thereafter, we compared the new indices with the existing RPPI (based on the appraisal method); note that the existing indices were expanded from a quarterly to a monthly frequency by simply putting the same index value for each month within a specific quarter.

([4]) We compared these regression results with the regression window for one year ahead (January 2017-December 2017) and had relatively consistent signs and coefficients for each explanatory variable.
([5]) The Central Statistics Office (CSO) of Ireland publishes a national house price index using the same technique to smooth out short-term volatility; see O'Hanlon (2011).

Figure 1 shows the index for North Jakarta. In this example, the hedonic index moves in a different direction to the existing RPPI for the sample horizon. The hedonic index generally revealed that prices were falling from 2016 to early 2017 before stabilising and remaining close to their new level. By contrast, the existing RPPI showed a generally upward trend for property prices during the sample horizon, with prices accelerating at a faster pace in the second half of 2017 and early 2018.

The situation in South Jakarta was different insofar as the hedonic index initially followed a similar pattern of development to that displayed for the existing RPPI; thereafter, the hedonic RPPI increased at a more rapid pace from the last quarter of 2017 to the third quarter of 2018 and therefore stood at a higher level than the existing RPPI (see Figure 2).

The hedonic index for Central Jakarta showed a different pattern (see Figure 3). The series displayed a high degree of volatility and even when smoothed (three-month moving average) the high degree of volatility persisted. This was probably affected by the small number of observations that were available for estimation (as Central Jakarta accounted for only 4 % of the total number of observations in the whole of Jakarta).

Figure 4 (overleaf) presents a composite index for the whole of Jakarta (an aggregate covering all five districts together). This hedonic index provided promising results and less volatile results. It showed a smooth and increasing trend during the sample horizon and one which was in line with both of the existing indices — for the primary and secondary house markets — along the sample horizon.

Annual growth rates for the hedonic index were also seen to follow a similar pattern of development to that displayed for the RPPI for the secondary market for houses (based on appraisals of activity among estate agents (see Figure 5 overleaf)).

Figure 1: Comparison of indices for the secondary market for houses, North Jakarta, January 2016-September 2018
(January 2016 = 100)

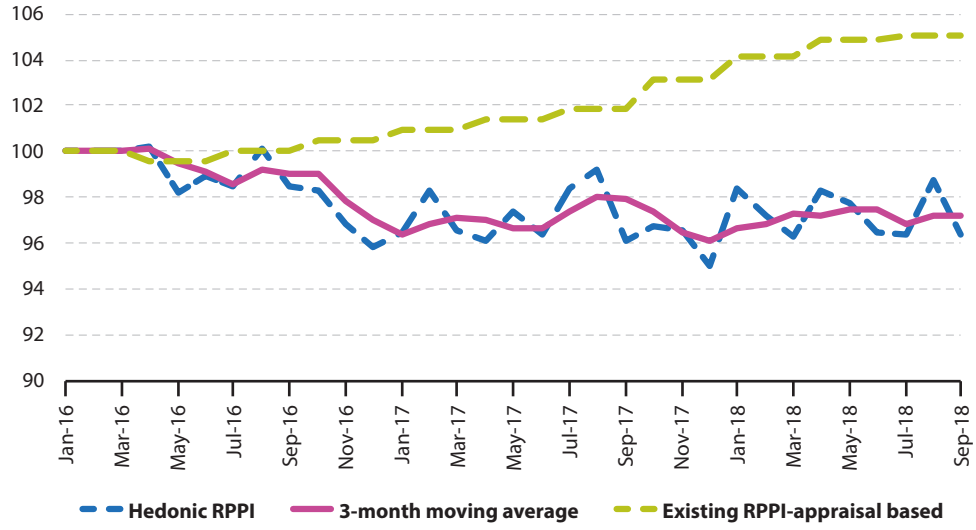


Figure 2: Comparison of indices for the secondary market for houses, South Jakarta, January 2016-September 2018
(January 2016 = 100)

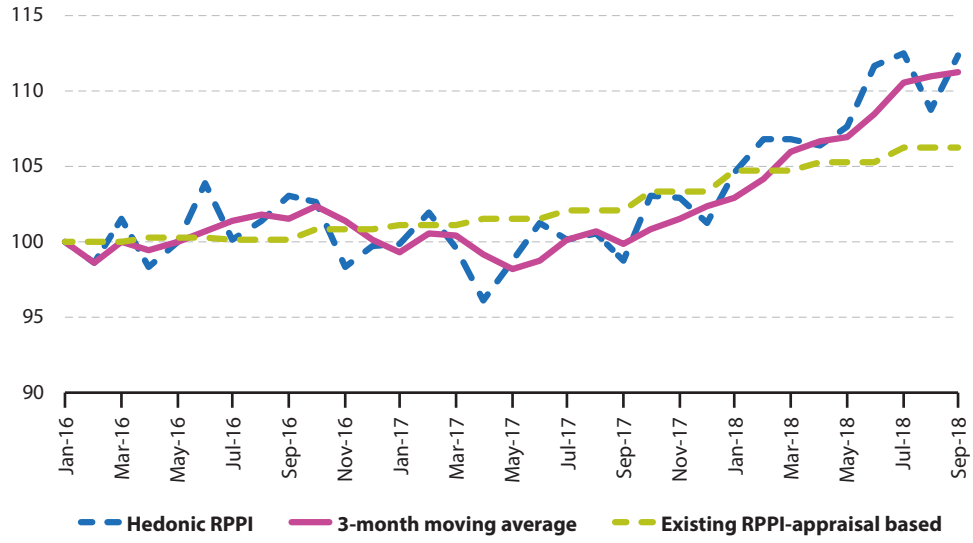


Figure 3: Comparison of indices for the secondary market for houses, Central Jakarta, January 2016-September 2018
(January 2016 = 100)

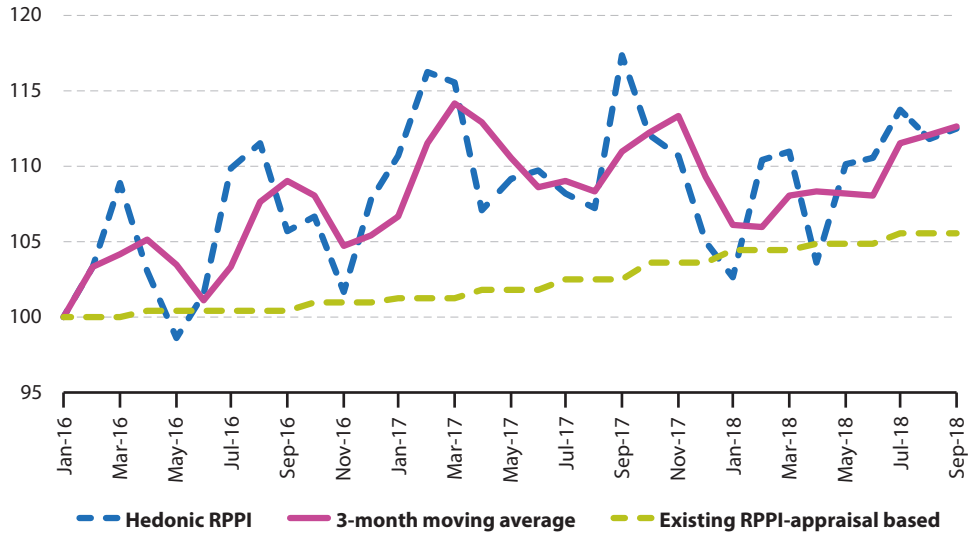


Figure 4: Comparison of indices for total housing markets, all of Jakarta, January 2016-September 2018
(January 2016 = 100)

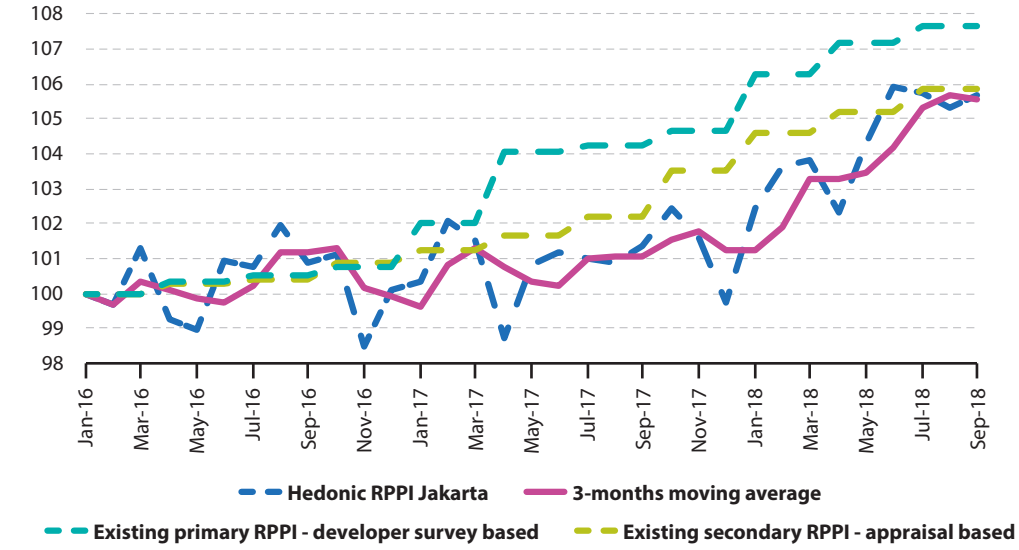
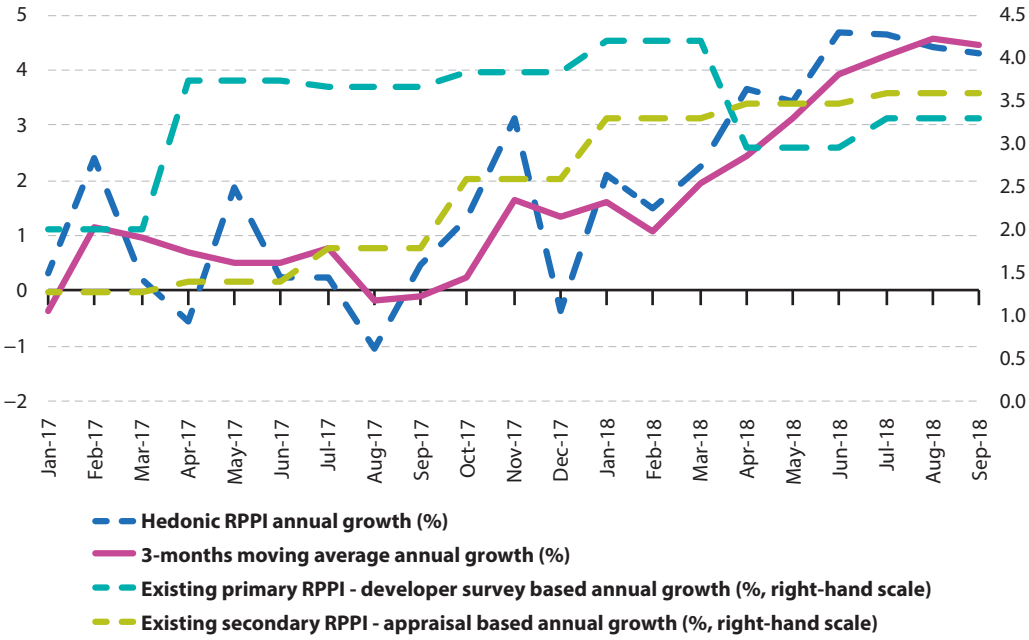


Figure 5: Annual growth rates for house markets, all of Jakarta, January 2017-September 2018
(%)



4. Conclusion

Using the time-dummy hedonic regression method we have computed alternative residential property price indices for the secondary house market in five districts of Jakarta based on property advertisements found on the web. These hedonic indices show promising results and have the potential in the future to replace the methods currently used to compile the RPPI for Indonesia. The regression outputs represent robust 'baseline' models for index compilation. Advertisement observations based on web listings seem more homogenous in nature, as indicated by the high degree of explanatory power, given the limited array of characteristic variables available. Smoothing may provide a better option for publishing an index based on these hedonic methods as it reduces short-term volatility.

For further developments, we will maintain these baseline models and extend coverage to other large cities in Indonesia. This extension will depend on the suitability of listings which may be available and the relative importance of different cities, as measured by their share in the national property market (derived from mortgage data). We need to ensure this new index remains representative of current market conditions by regularly reviewing the models' performance and updating the weights. We may also seek to enhance the models in the future by including a more granular spatial adjustment (location advantage) and other characteristics (such as the age of each property).

Acknowledgements

The author would like to thank Niall O'Hanlon (International Monetary Fund) for valuable technical assistance, Annisa Cynthia and Arinda Dwi Okfantia (both Bank Indonesia) for their helpful input. All views expressed are those of author and do not necessarily represent the views of the Bank Indonesia.

References

Daas, P. J. H., M. Puts, M. Tennekes and A. Priem (2014), 'Big Data as a Data Source for Official Statistics: experiences at Statistics Netherlands', proceedings of Statistics Canada Symposium 2014, Ottawa.

Diewert, W. E. (2003), 'Hedonic Regressions. A Consumer Theory Approach', National Bureau of Economic Research in *Scanner Data and Prices Indexes*, pp. 317-348.

Eurostat (2013), *Handbook on Residential Property Prices Indices (RPPIs)*, Publications Office of the European Union, Luxembourg.

Hammer, C., D. C. Kostroch, G. Quirós, and STA Internal Group (2017), 'Big Data: Potential, Challenges and Statistical Implications', Staff Discussion Note, SDN/17/06, International Monetary Fund, Washington D.C.

Hill, R. J. (2011), 'Hedonic Price Indexes for Housing', *OECD Statistics Working Papers*, Working Paper No. 36, OECD, Paris.

Hill, R. J. (2012), 'Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy', *Journal of Economic Surveys*, Volume 27, Issue 5, pp. 879-914.

Hülagü, T., E. Kizilkaya, A. G. Özbekler, and P. Tunar (2015), 'A Hedonic House Price Index for Turkey', presented to a Turkish Statistical Institute seminar and the European Real Estate Society 22nd Annual Conference, Istanbul.

Land Economy and Construction Industries Bureau (2016), 'Methodology of JRPPI: Japan Residential Property Price Index', Ministry of Land, Infrastructure, Transport and Tourism, Tokyo.

Li, W., M. Prud'homme and K. Yu (2006), 'Studies in Hedonic Resale Housing Price Indexes', presented to the Canadian Economic Association 40th Annual Meetings, Concordia University, Montréal.

Lyons, R. C. (2018), 'Can list prices accurately capture housing price trends? Insights from extreme markets conditions', *Finance Research Letters*, Volume 30, pp. 228-232.

Marsden, J. (2015), 'House prices in London — an economic analysis of London's housing market', *GLA Economics*, Working Paper 72, Greater London Authority.

O'Hanlon, N. (2011), 'Constructing a National House Price Index for Ireland', *Journal of the Statistical and Social Inquiry Society of Ireland*, Volume XL, pp. 167-196.

Radermacher, W. J. (2018), 'Official Statistics in the Era of Big Data Opportunities and Threats', *International Journal of Data Science and Analytics*, Volume 6, Issue 3, pp. 225-231.

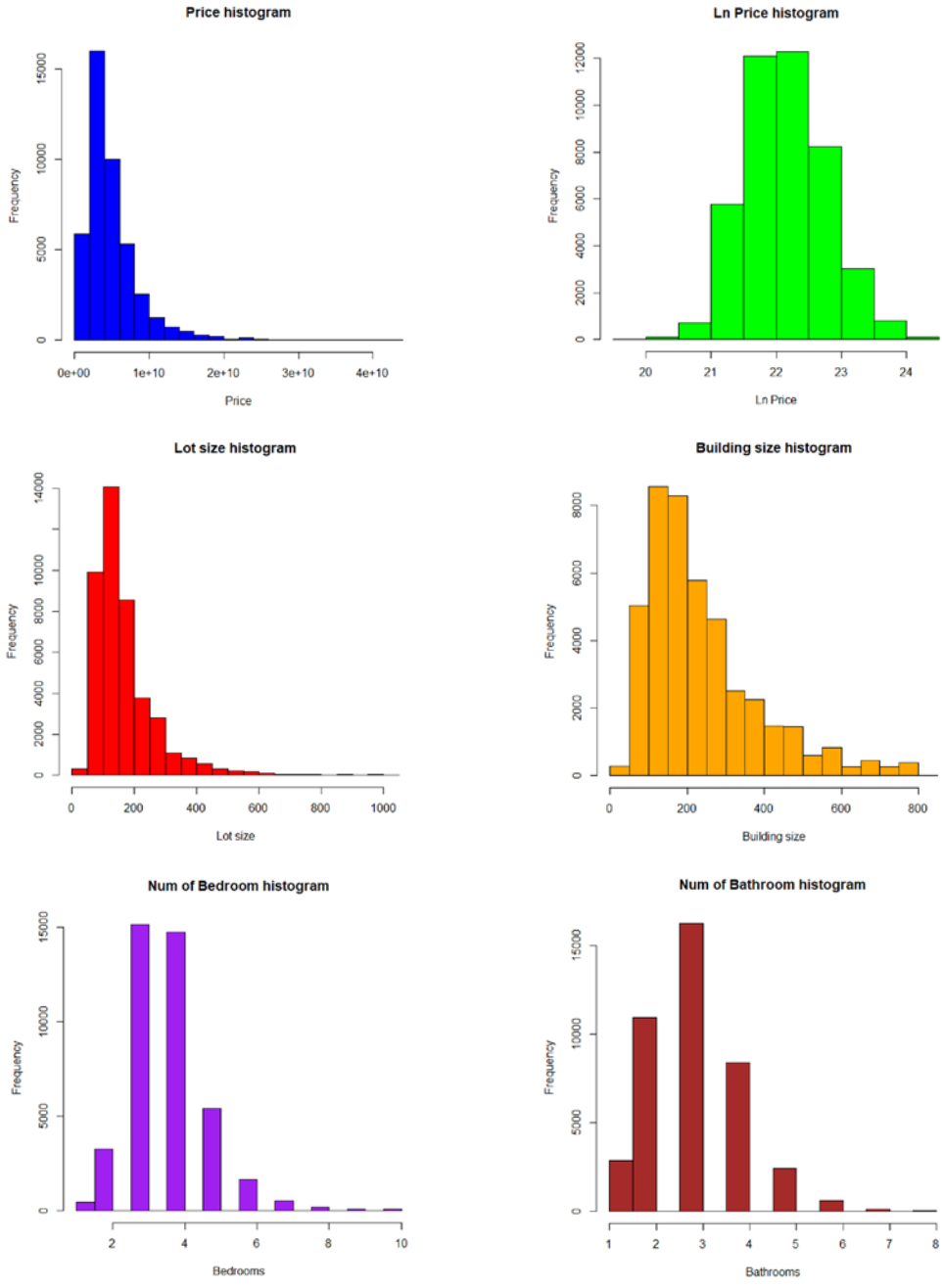
Silver, M. (2016), 'How to Better Measure Hedonic Residential Property Price Indexes', *IMF Working Papers*, WP/16/213, International Monetary Fund, Washington D.C.

Zwick, M. (2017), 'Introduction to Big Data in Official Statistics', Institute for Research and Development in Official Statistics, Federal Statistics Office Germany, Wiesbaden.

Data appendix

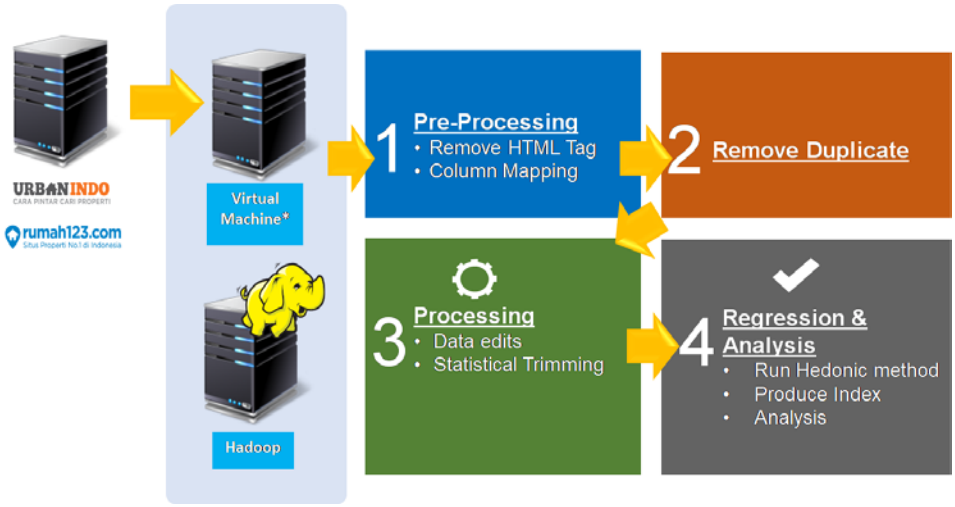
1. Data distribution for the sample, North Jakarta

Figure 6: Characteristic variables



2. Steps for data preparation

Figure 7: Workflow for the construction of an RPPI using big data



DATA SOURCE

- Online property advertisements from the two biggest property websites with approximately 9 000 new advertisements per month (for Jakarta only).
- Data available since 2015, listings only refer to the first instance that the price of a property was listed (a unique price).
- Data attributes:
  - title;
  - status of property : sell/rent;
  - type of property (house/apartment (flat)/villa/condotel/condominium);
  - advertising time start date and end date;
  - property price;
  - land and building size;
  - number of bedrooms and number of bathrooms;
  - address;
  - property description.

PRE-PROCESSING

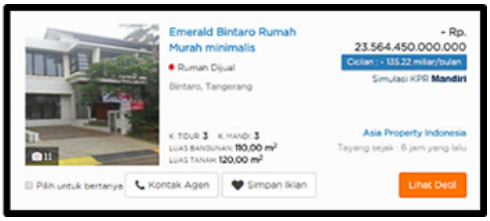
- **Cleaning**
  The cleaning process that formed part of the data pre-processing exercise included removing irrelevant characters such as HTML tags that formed part of the title and description for each advertisement. The removal of HTML tags was done through the Python programming language, using one of its libraries, HTMLParser.

- **City mapping**
  City mapping was conducted to standardise the addresses shown in the data to the city level. This process was carried out because some portals do not provide data pertaining to the city/district for each advertisement. City mapping used a list of city/districts and sub-districts obtained from Indonesian Statistics (BPS). If the address in the advertisement could not be found in the list of sub-districts, city mapping was carried out using the Geocoding API provided by Google Maps.

- **Column mapping**
  Data from different portals had distinct formats and column structures. For example, data from Rumah123 consisted of 21 attribute columns, using '|' (a bar) or '~' (a tilde) as a delimiter for the columns. In contrast, data from Urbanindo consisted of 13 attribute columns using a tab as the delimiter. Thus, column mapping was needed to standardise column structures, column names and the use of delimiters. Once completed, this allowed data from disparate sources to be compiled and processed simultaneously.

- **Removal of duplicates**
  After the initial processing, there were still many duplicate data entries. These included intra-portal and inter-portal duplicates. Intra-portal duplicates existed when the same property was advertised by different estate agents or when an advertisement was reposted by the same estate agent. Inter-portal duplications existed if the same property was advertised on different property portals. Property advertisements were considered the same if their price, land area, building area, number of bathrooms, number of bedrooms and city had the same value.

- **Removal of abnormal data**
  Aside from the issue of duplicate advertisements, there were also concerns regarding abnormal data. Below are some criteria which were used to identify abnormal data:
  a. Missing values — the advertisement did not provide data regarding the land area and/or the building's area.
  b. Unusual price — for example, land which cost Rp 0.00 or land of 100 m^2 which cost Rp 23 trillion.
  c. Unusual building or land area — for example, land area equal to 0 m^2 or a building area of 10 000 m^2.



  d. Unusual price due to location — for example, there was an advertisement for a house which was being sold at Rp 50 million in Jakarta Pusat.
  e. Unusual ratio of land area to building area — for example, an advertisement showed a house with a building area of 300 m^2 and a land area of 30 m^2, or a house with a building area of 30 m^2 and a land area of 1 000 m^2.

3. Property market share

Table 4: Comparisons of shares for weights and the number of observations
(%)

Districts	Mortgage collateral share	Number of observation share
West Jakarta	24.95	24.14
Central Jakarta	12.06	4.25
South Jakarta	31.92	37.34
East Jakarta	7.27	21.95
North Jakarta	23.80	12.31

4. Regression output

Table 5: Regression output — check for stability over time, sample for North Jakarta

Dependent variable: Ln Price				
Independent variables	**2016:1 - 2016:12 coefficients**		**2017:1 - 2017:12 coefficients**	
Intercept	21.2900	***	21.2900	***
Building size	0.0010	***	0.0011	***
Lot size	0.0045	***	0.0041	***
Dum_# of bedroom 1-2	-0.2153	***	-0.0310	***
Dum_# of bedroom 3	-0.0440	***	-0.2185	***
Dum_# of bedroom >4	-0.0400	***	-0.0326	***
Dum_# of bathroom 1	-0.2225	***	-0.1997	***
Dum_# of bathroom 2	-0.1732	***	-0.1853	***
Dum_# of bathroom >3	0.0082	*	0.0038	
Dum_period 2016 (2017):2	0.0007		0.0147	
Dum_period 2016 (2017):3	0.0006		0.0025	
Dum_period 2016 (2017):4	0.0034		-0.0077	
Dum_period 2016 (2017):5	-0.0203	***	-0.0012	
Dum_period 2016 (2017):6	-0.0123		-0.0049	
Dum_period 2016 (2017):7	-0.0132		0.0167	*
Dum_period 2016 (2017):8	0.0022		0.0244	***
Dum_period 2016 (2017):9	-0.0143		-0.0061	
Dum_period 2016 (2017):10	-0.0169	**	0.0009	
Dum_period 2016 (2017):11	-0.0307	***	-0.0023	
Dum_period 2016 (2017):12	-0.0435	***	-0.0195	*
Adjusted R-squared	0.863		0.862	
F-statistics	4 866		4 067	
Number of observations	14 640		14 026	

Note: *** significance at 1 %, ** significance at 5 % and * significance at 10 %.

Table 6: Testing for heteroscedasticity

Breusch-Pagan test for heteroscedasticity	
North Jakarta	BP = 204.1, df = 19, p-value < 2.2e-16
South Jakarta	BP = 366.96, df = 19, p-value < 2.2e-16
Central Jakarta	BP = 109.83, df = 19, p-value = 8.574e-15
West Jakarta	BP = 412.43, df = 19, p-value < 2.2e-16
East Jakarta	BP = 135.36, df = 19, p-value < 2.2e-16

Table 7: Regression output for three other districts of Jakarta

	Central Jakarta		West Jakarta		East Jakarta	
Dependent variable: Ln Price						
Independent variables	**Estimates**	**Robust standard error**	**Estimates**	**Robust standard error**	**Estimates**	**Robust standard error**
Intercept	21.0800	0.02999 ***	20.9310	0.00682***	20.5700	0.00906***
Building size	0.0014	0.00006 ***	0.0012	0.00002***	0.0013	0.00003***
Lot size	0.0038	0.00007 ***	0.0044	0.00003***	0.0033	0.00003***
Dum_# of bedroom 1-2	-0.0604	0.01616 ***	-0.0382	0.00351***	-0.0666	0.00482***
Dum_# of bedroom 3	-0.1352	0.02586 ***	-0.2171	0.00525***	-0.2564	0.00773***
Dum_# of bedroom >4	-0.1415	0.01474 ***	-0.0375	0.00483***	-0.0250	0.00590***
Dum_# of bathroom 1	-0.4678	0.02789 ***	-0.0968	0.00597***	-0.3393	0.00799***
Dum_# of bathroom 2	-0.1743	0.01564 ***	-0.0974	0.00336***	-0.1387	0.00456***
Dum_# of bathroom >3	0.0084	0.01505	0.0194	0.00436***	0.0293	0.00534***
Dum_period 2016:2	0.0185	0.03000	-0.0096	0.00676	0.0001	0.00895
Dum_period 2016:3	0.0752	0.02998 **	-0.0191	0.00667***	0.0414	0.00890***
Dum_period 2016:4	0.0275	0.02962	-0.0253	0.00668***	0.0064	0.00898
Dum_period 2016:5	-0.0306	0.02658	-0.0295	0.00578***	0.0326	0.00779***
Dum_period 2016:6	0.0164	0.03036	-0.0202	0.00706***	0.0383	0.00947***
Dum_period 2016:7	0.0816	0.03128 ***	-0.0103	0.00703	0.0202	0.00944**
Dum_period 2016:8	0.0966	0.03310 ***	-0.0146	0.00693**	0.0590	0.00910***
Dum_period 2016:9	0.0503	0.03384	-0.0405	0.00741***	0.0705	0.00950***
Dum_period 2016:10	0.0577	0.02918 **	-0.0164	0.00656**	0.0396	0.00839***
Dum_period 2016:11	0.0066	0.03123	-0.0259	0.00687***	0.0320	0.00890***
Dum_period 2016:12	0.0766	0.03284 **	-0.0098	0.00742	0.0606	0.00942***
Adjusted R-squared	0.731		0.813		0.835	
F-statistics	667		7 138		5 891	
Number of observations	4 662		31 184		22 180	

Note: *** significance at 1 %, ** significance at 5 % and * significance at 10 %.

5. Results for hedonic RPPI

Figure 8: Comparison of indices for the secondary market for houses, West Jakarta, January 2016-September 2018
(January 2016 = 100)

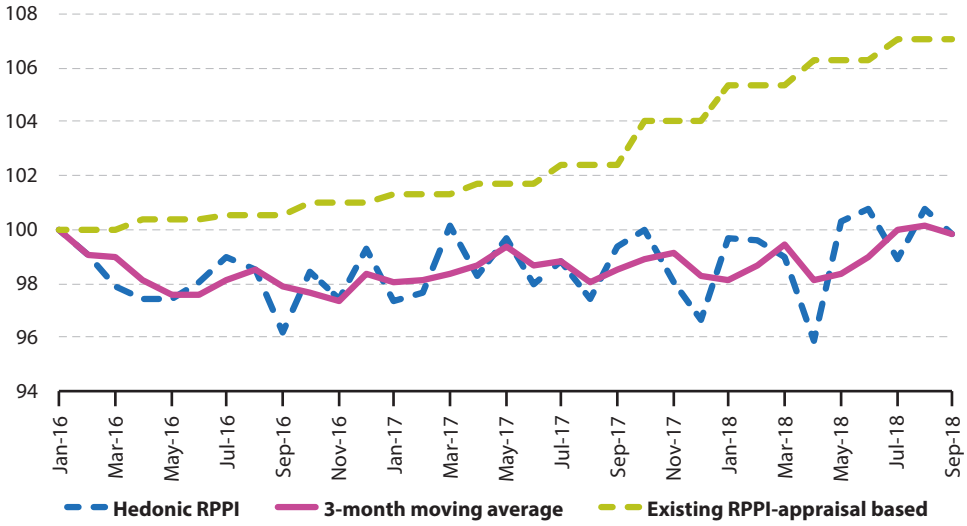


Figure 9: Comparison of indices for the secondary market for houses, East Jakarta, January 2016-September 2018
(January 2016 = 100)

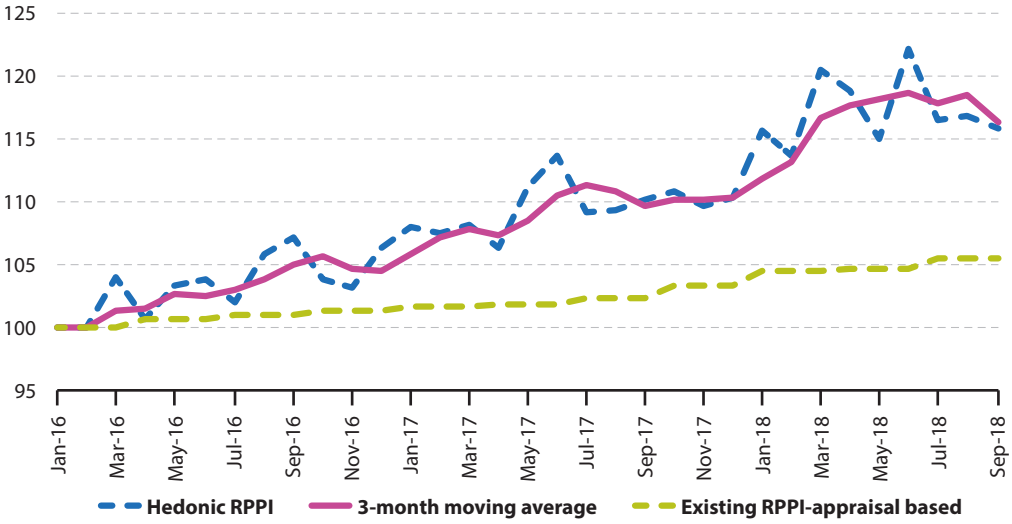


Table 8: Hedonic RPPI, January 2016-September 2018
(January 2016 = 100)

Period	West Jakarta		Central Jakarta		South Jakarta		East Jakarta		North Jakarta		All of Jakarta	
	TD Hedonic	Adjusted*	TD Hedonic	Adjusted*	TD Hedonic	Adjusted*	TD Hedonic	Adjusted*	TD Hedonic	Adjusted*	TD Hedonic	Adjusted*
Jan-16	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Feb-16	99.07	99.07	103.33	103.33	98.53	98.53	100.01	100.01	100.02	100.02	99.71	99.71
Mar-16	97.92	99.00	108.87	104.07	101.47	100.00	104.09	101.37	99.98	100.00	101.31	100.34
Apr-16	97.43	98.14	102.98	105.06	98.24	99.41	100.67	101.59	100.23	100.08	99.26	100.09
May-16	97.39	97.58	98.56	103.47	100.00	99.90	103.37	102.71	98.18	99.46	98.98	99.85
Jun-16	98.03	97.61	101.55	101.03	103.78	100.67	103.82	102.62	98.94	99.12	100.93	99.72
Jul-16	98.95	98.12	109.84	103.32	100.17	101.32	102.06	103.08	98.50	98.54	100.77	100.23
Aug-16	98.52	98.50	111.52	107.64	101.40	101.78	105.95	103.94	100.15	99.20	101.93	101.21
Sep-16	96.16	97.88	105.72	109.03	103.02	101.53	107.23	105.08	98.49	99.05	100.86	101.19
Oct-16	98.40	97.69	106.69	107.98	102.55	102.33	103.84	105.67	98.32	98.98	101.10	101.30
Nov-16	97.46	97.34	101.67	104.69	98.28	101.29	103.28	104.78	96.80	97.87	98.49	100.15
Dec-16	99.27	98.37	107.74	105.37	99.68	100.17	106.35	104.49	95.78	96.97	100.10	99.90
Jan-17	97.38	98.03	110.65	106.69	99.86	99.27	108.08	105.90	96.42	96.34	100.32	99.64
Feb-17	97.64	98.09	116.20	111.53	101.90	100.48	107.50	107.31	98.25	96.82	102.10	100.84
Mar-17	100.10	98.37	115.45	114.10	99.54	100.44	108.22	107.93	96.56	97.08	101.52	101.31
Apr-17	98.29	98.68	107.02	112.89	96.14	99.19	106.35	107.35	96.13	96.98	98.73	100.78
May-17	99.67	99.36	109.17	110.54	98.78	98.15	111.30	108.62	97.35	96.68	100.83	100.36
Jun-17	98.00	98.65	109.69	108.63	101.21	98.71	113.79	110.48	96.39	96.62	101.20	100.25
Jul-17	98.79	98.82	108.11	108.99	100.15	100.05	109.28	111.46	98.41	97.38	101.02	101.01
Aug-17	97.41	98.06	107.22	108.34	100.52	100.63	109.46	110.84	99.18	97.99	100.88	101.03
Sep-17	99.38	98.53	117.33	110.89	98.77	99.81	110.20	109.65	96.07	97.89	101.35	101.08
Oct-17	99.97	98.92	112.03	112.19	103.03	100.77	110.91	110.19	96.75	97.33	102.43	101.55
Nov-17	98.08	99.14	110.64	113.33	102.84	101.55	109.77	110.29	96.51	96.44	101.59	101.79
Dec-17	96.68	98.24	105.01	109.22	101.27	102.38	110.39	110.36	95.04	96.10	99.76	101.26
Jan-18	99.67	98.14	102.60	106.08	104.48	102.86	115.69	111.95	98.38	96.64	102.41	101.25
Feb-18	99.60	98.65	110.33	105.98	106.74	104.16	113.74	113.27	97.16	96.86	103.62	101.93
Mar-18	98.99	99.42	110.92	107.95	106.77	106.00	120.51	116.65	96.25	97.26	103.82	103.29
Apr-18	95.84	98.14	103.53	108.26	106.28	106.60	118.85	117.70	98.25	97.22	102.35	103.26
May-18	100.31	98.38	110.13	108.19	107.55	106.87	115.07	118.14	97.76	97.42	104.27	103.48
Jun-18	100.73	98.96	110.51	108.06	111.63	108.49	122.19	118.70	96.46	97.49	105.93	104.18
Jul-18	98.93	99.99	113.70	111.45	112.50	110.56	116.50	117.92	96.35	96.86	105.71	105.30
Aug-18	100.73	100.13	111.81	112.01	108.74	110.96	116.93	118.54	98.79	97.20	105.34	105.66
Sep-18	99.84	99.83	112.41	112.64	112.34	111.20	115.91	116.45	96.38	97.18	105.69	105.58

*Adjusted using 3-months moving average.