Survey Data versus Administrative Data in the CPI Rent Price index, Challenges, Changes and Solutions – Case study of Iran

Mostafa Farrokhfal[1], Mohsen Saadatpour Moghaddam[2], Abdorahim Ira[3]

[1] - Price Indices Office, Statistical Centre of Iran, Dr. Fatemi Ave., Tehran, Iran, (E-mail: m.farokhfal@gmail.com)

[2] - Price Indices Office, Statistical Centre of Iran, Dr. Fatemi Ave., Tehran, Iran, (E-mail: m.saadatpour@aut.ac.ir)

[3] -Price Indices Office, Statistical Centre of Iran, Dr. Fatemi Ave., Tehran, Iran, (E-mail: ar_Ira@yahoo.com)

Abstract

Calculating the rent price index is always considered one of the most challenging topics in CPI, and there is no consensus on it. In this article, we calculated the rent price index for a two-year period using two approaches: Registry-Based Data and Survey-Based Data. The Registry-based approach uses the administrative data of the country's real estate system and the 12-Month Rolling Time Dummy Method (RTD) – a time dummy hedonic method with a rolling window – in R software. This registration data, which is regularly and monthly provided to the Statistical Center of Iran (SCI), includes all sales and rental transactions in the country, along with geographic address, age of building, area of building, type of building (apartment, villa, etc.), and the transaction price. In the survey approach, which is also the current approach of SCI, six symmetrical panels are used, from which prices are collected twice a year. Additionally, with the registration data, a more varied rent price index is obtained according to the characteristics of the dwelling, which can be useful for the policy makers of this sector.

1. Introduction

In some countries, rent is the largest household expenditure and CPI compilers encounter important problems when calculating the rent index in the CPI precisely. In Iran, Owner-Occupied Households (OOH) have the largest weight in the CPI. The OOH weight is obtained from the rental equivalence approach. For tenant rent, the Statistical Center of Iran (SCI) counts the amount of rent paid to the landlord for shelter. If a housing unit is occupied by the owners, the SCI computes what it would cost the owner to rent a similar place, known as the Owners' Equivalent Rent (OER). The SCI collects data on rent for about 30,000 residences in six panels through personal visits. Since rents do not change frequently, the rent of each panel is sampled every six months. The CPI measures price growth for the same baskets of goods and services over time, so the CPI adjusts for changes in the quality of the properties it observes. In this study, we intend to use registered data as an alternative for the rent price index survey and compare its results with the results of the current method.

1.1 Housing market in Iran

The rental market in Iran includes landlords, tenants, and real estate agencies. In both the selling and rental markets, special loans for housing constitute a very small share of total expenditure on it. Unfortunately, small share of rental transactions are registered as administrative data, and many of them are manipulated and not registered in the administrative system. However, in the selling market, usually transactions are most registered. The registered data recorded by real estate agencies in the rental market contains building area, age, mortgage value, monthly rent value, estate type (villa, apartment, etc.), skeleton type (concrete, metal, metal and concrete, brick, and no skeleton), address (including province and city name, street, plaque number, and floor), unique postal code, and registration date. Total monthly rent equals to:

$$\textit{Total monthly rent} = (0.03 * \textit{Mortgage Value}) + \textit{Monthly Rent} \qquad (1)$$

In fact, it is accepted in the rental market that the monthly interest rate is 3 percent and by this rate, tenants or landlords can change their asking or bidding price. For example, the landlord can reduce 1 million IR-Rial in mortgage value and instead increase 30,000 Rials in monthly rent.

1.2 Rent index in current method

In the "2021 base year" for the CPI index disseminated newly by the Statistical Center of Iran (SCI), we used the approach introduced in the 2020 manual. In this way, we used six symmetric panels with about 5000 households in each panel that were enumerated every six months. Furthermore, we visited every household twice a year. For example, households were the same in January and July, or the same in February and August, and so on. When the relative prices were calculated, we aggregated the data using actual rent weights for each city (obtained from the HIES survey in the base year) and calculated the change of the actual rent price in the province. Then, the sixth root of the relative prices was regarded as the monthly relative prices:

$$\triangle_{rent}^{t-1\rightarrow t} = \sqrt[6]{\prod_{i=1}^{n}\left(\frac{r_i^t}{r_i^{t-6}}\right)}$$

For calculating the OOH index, we do not have any other data source and we used calculated relative prices for the actual rent index at the city level and then aggregated them with their own weights - which differ from the actual rent weights and are obtained from the HIES survey - to aggregate and calculate the OOH index, so we have equal indexes at the city level and different indexes at the aggregate level (province level).

Our main challenge in the rent survey is that enumerators cannot collect data from tenant households and the unresponsive rate is very high, especially from wealthy households that are reluctant to cooperate with enumerators. As a result, we will miss changes in rent in the expensive area. An alternative method is to use administrative data that is registered by real estate agencies. In this article, we try to use register-based data as an alternative data source and compare their results to survey-based results.

2. Methodology

We access register-based data that exists in the Ministry of Roads and City Planning via web services regularly and monthly. This data includes all rental and selling data for the dwelling market in all provinces. In this database, we have much more data in comparison with survey-based. Because data is registered by real estate agencies and they don't have any legal compulsion to prohibit entering incorrect data, they may make mistakes intentionally or unintentionally. Therefore, due to these potential errors in entering data, the first step outlier data recognizing should be before any other process.

2.1 Outlier detection

Of course, it is not a universal approach that handles all possible cases. Since we do not have a strict outlier definition, it is always possible to find examples of results that are "incorrect" from the researcher's point of view for any outlier detection algorithm.

In the world of normal distributions, the typical approach for outlier detection is based on the standard deviation. It uses the following thresholds for outliers:

$$Lower = Mean - K \cdot StdDev, \qquad Upper = Mean + K \cdot StdDev, \qquad K_{default} = 3.$$

In the world of non-parametric distributions, there is another popular approach called Tukey’s fences (Tukey, 1997). It defines outlier thresholds as follows:

$$Lower = Q1 - K \cdot IQR, \qquad Upper = Q3 + K \cdot IQR, \qquad K_{default} = 1.5,$$

where IQR= Q3− Q1 is the interquartile range. we can use a more robust measure of statistical dispersion called the median absolute deviation (MAD) (Leys2013). It can be defined as follows:

$$MAD = C \cdot median(|X_i - median(X)|).$$

For the normally distributed values, there is a well-known relationship between the standard deviation and the median absolute deviation:

$$StdDev \approx 1.4826 \cdot median\ (|X_i - median(X)|).$$

Thus, we can use C=1.4826 (which is also known as the consistency constant) to make MAD a consistent estimator for the standard deviation estimation. With the MAD approach, we can define the outlier thresholds as follows:

$$Lower = Median - K \cdot MAD, \qquad Upper = Median + K \cdot MAD, \qquad K_{default} = 3.$$

The classic MAD approach defines a symmetric interval around the median. This does not work great for the left-skewed, right-skewed, and other kinds of non-symmetric distributions.

This problem can be resolved with the help of the Double MAD approach ([Rosenmai2013]). The idea is simple: for the obtained median value, we should calculate two median absolution deviations. One deviation should be calculated for the numbers below the median and one for the numbers above the median:

$$M = median(X), \qquad X^{(l)} = \{x|x∈X ∧ x≤M\}, \qquad X^{(u)} = \{x|x∈X ∧ x≥M\},$$

$$MAD^{(l)} = C·median(|Xi^{(l)}−M|), \qquad MAD^{(u)} = C·median(|Xi^{(u)}−M|).$$

Next, we can define outlier thresholds using $MAD^{(l)}$ for Lower and $MAD^{(u)}$ for Upper:

$$Lower = M−K.MAD^{(l)}, \qquad Upper = M+K·MAD^{(u).}$$

Unfortunately, the most common “straightforward” approach to calculate the median value is not always robust enough. However, it may not be the best estimation of the population median. This problem can be resolved with the help of the Harrell-Davis quantile estimator ([Harrell1982]). It defines estimation for the p^{th} quantile as follows:

$$Q_p = \sum_{i=1}^{n} W_{n,i}.X_i$$

$$W_{n,i} = I_{i/n}\{p(n+1)),(1-p)(n+1)\} - I_{(i-1)/n}\{p(n+1),(1-p)(n+1)\},$$

where $I_x\{a,b\}$ denotes the regularized incomplete beta function, this formula may look intimidating, but it provides a more robust median estimation than the straightforward approach.

In this study, we used the twice double MAD method for detecting outliers in data, once on total monthly rent and once again on total monthly rent per square meter, because some records had correct price data but incorrect area data. Using the "Hmisc" package in R software, outliers were detected and finally, database was constructed with clean data.

2.2 Preparing data set

In the second step, we used clean data to run a regression and estimate the corresponding coefficients. Our data set includes:

i) Total monthly rent is the in variable dependent logarithmic mode, obtained from (1)

ii) Building area in logarithm mode as integer data. Building area is limited between 30 and 1000 square meter.

iii) Building age; we categorized age of building in 5 groups as below:

Group	Interval ages
1	< 3 years
2	4 - 9 years
3	10 - 20 years
4	21 - 35 years

5	> 35 years

Due to the nature of the rental market, we specify these five categories and this age interval cannot be used for the selling market, age is considered as an integer data.

iv) Dwelling type; this independent variable contains apartments, villas and buildings.
v) Skeleton type; it could be concrete, metal, concrete and metal, brick and no skeleton.
vi) We used provinces as a dummy variable. Due to the fact that very few cities were in the final dataset and some of them had very few transactions, we used data only from the capital of the province and deleted the data from other cities.
vii) Time dummy variable

3.2 Rolling Time Dummy (RTD) method

With regard to the of above variables, the time dummy hedonic model is:

$$lnp_n^t = \beta_0 + \sum_{t=1}^{T} \delta^t D_n^t + \sum_{k=1}^{K} \beta_k Z_{nk}^t + \varepsilon_n^t$$

That the index for the current period (t) is derived as follows:

$$\mathrm{I_t} = \exp(\hat{\delta}_t) * 100$$

When data from a new period are added, the indices from the previous period may be changed because the estimated coefficients δ_t of the previous periods are revised based on the new observations. To avoid revisions to the indices, a rolling window approach is used; rolling window in this study is 12 months, with at least 700,000 records. Using the RTD method at the provincial level shows that the R-square is very low for some provinces. We do think not this is due to the number of data, as we have enough data for estimating (at least 20 times the survey data in each rolling window) the rent index, and it is likely that entering wrong data is causing such results. Therefore, we decided to aggregate data regardless of the provincial level for the final data set. Additionally, we know that most rent data exist from May to September each year, and in the other months, we have fewer data. Unfortunately, the outlier data rate has been increasing from 2 percent to 10 percent in rolling windows when it is near the current month.

3. Result

We used R software to calculated index with register-base data. On account of data limitation (lack of sufficient data), we consider data after Jan-2021 and all indexes converted to Jan-2021 = 100. We calculate calculated the rent index by register-based data at the country level.

As Table 1 shows, the CPI rent price index that we have nominated "Rent-All contract" is divergent from the register-based index that was calculated by the RTD method. This divergence is explainable: in register-based data, we have only households that finished their old contracts and changed their dwellings; because we know in Iran, households that stay in their primary dwellings don't update the previous contract in the real estate agency and they do it manually, so this contract doesn't exist in register-based data and we know the annual change of rent for these households is somewhat lower than other tenants that change dwellings. As a solution, we compared the "Rent-New Contract" index (R-NC) to the register-based index. This comparison is conceptually reliable, because both have only new contracts. The Rent-New Contract index shows the index for households that encounter new contracts and new prices. Since the relative prices for these households actually are year-on-year price changes, we converted them into a monthly index as follows:

In 2021, we suppose monthly rent change and therefore monthly index in the RTD method equal to the new contract index. For the following years, we forwarded the monthly index by the year-on-year change (for these cases, the new contract rent is equal to the annual percentage change of the rent contract).

Table 1- Register-Base index Vs survey-base index

	Register Base index	Survey Base index		RPPI index
	RTD	Rent-New Contract	Rent-All Contract	
Jan-21	100	100	100	100
Feb-21	110.3	110.3	100.7	104.7
Mar-21	110.2	110.2	102.3	108.1
Apr-21	103.4	103.4	103.8	108.2
May-21	112.2	112.2	104.7	109.4
Jun-21	118.2	118.2	106.1	112.0
Jul-21	122.7	122.7	110.2	113.2
Aug-21	121.8	121.8	112.5	116.3
Sep-21	125.6	125.6	116.4	118.6
Oct-21	130.0	130.0	120.1	119.3
Nov-21	135.1	135.1	121.8	120.8
Dec-21	143.9	143.9	123.1	124.0
Jan-22	151.6	148.2	128.1	124.8
Feb-22	157.3	162.4	129.1	128.8
Mar-22	158.5	173.4	130.0	129.5
Apr-22	166.2	150.6	132.5	133.8
May-22	172.6	162.3	135.3	141.8
Jun-22	183.3	181.1	138.3	151.8
Jul-22	192.3	195.5	142.6	164.3
Aug-22	196.3	195.0	146.8	167.4
Sep-22	202.2	199.1	151.6	172.8
Oct-22	202.8	183.6	157.8	172.6
Nov-22	206.6	204.7	163.5	181.2
Dec-22	213.1	213.9	168.5	188.5

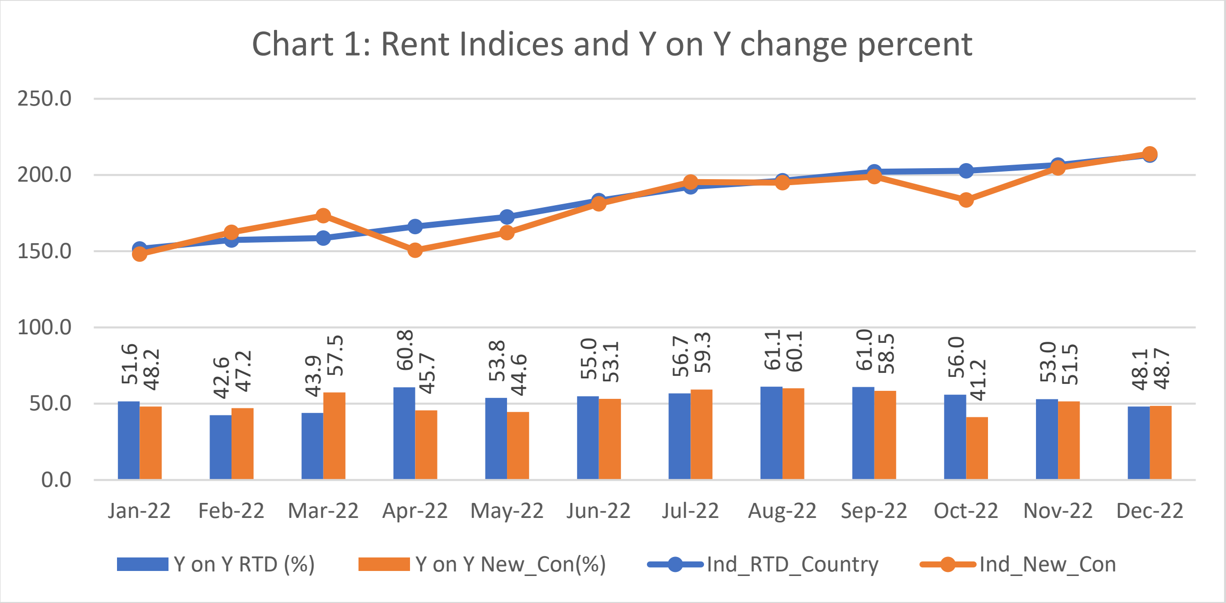

Chart 1: Rent Indices and Y on Y change percent

Chart (1) shows that the New-Contract index has decreased in April and October 2022. This is unexplainable theoretically, but some explanations may be consistent with this situation. From 20 March until 5 April, we have a holiday for the new year and housing transactions in both the rent and selling markets are very little. Also in October, schools and universities open a new educational year and households prefer not to change their dwellings. In both cases (April and October), New-Contracts are very few and this might have been affecting these months' indices.

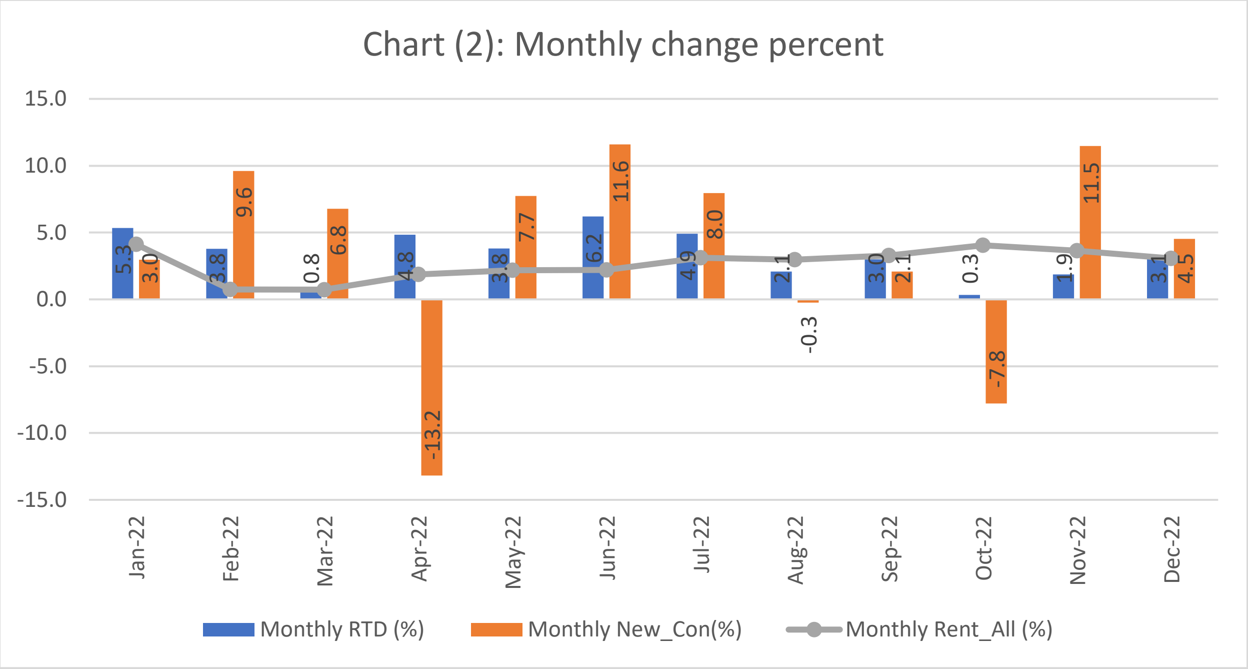

Chart (2): Monthly change percent

4. Conclusion

In this article, we try to use register-based data for indexation. In this way, we have faced some challenges, so that we couldn't build an index at the province level, despite the large amount of data we had available. Secondly, we obtained indexes that haven't had sensible monthly changes and seem incorrect. Since we want to use register-based data as an alternative data source, it seems we need some adjustments in our administrative data and should obligate real estate agencies to enter data correctly. Another proposal to improve the result is to aggregate some provinces' data instead of the whole country.

5. References

1. IMF (2020), RPPI Guide practical compilation guide.
2. IMF ILO Eurostat UN OECD WB (2020), consumer price index manual concepts and methods.
3. Leys, Christophe, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. “Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median.” Journal of Experimental Social Psychology 49, no. 4 (2013): 764-766.
4. Rosenmai, Peter. “Using the Median Absolute Deviation to Find Outliers.” Eureka Statistics. November 25, 2013. Accessed June 22, 2020.
5. Harrell, F.E. and Davis, C.E., 1982. A new distribution-free quantile estimator. Biometrika, 69(3), pp.635-640.
6. Consumer Price Index (CPI) monthly report, SCI
7. Residential Property Price Index (RPPI) monthly report, SCI