IMPUTATION OF VACANT UNITS IN THE RESIDENTIAL RENT INDEX OF THE U. S. CONSUMER PRICE INDEX

Robert M. Baskin, U.S. Bureau of Labor Statistics
2 Massachusetts Ave, N.E., Room 3655, Washington, D.C. 20212

Views in this paper do not constitute official Bureau of Labor Statistics policy.

KEY WORDS: empirical Bayes, primary sampling unit, variance components.

This is a report of an investigation into the effect of imputing the rents of vacant rental units on the current Residential Rent Index of the U.S. Consumer Price Index (CPI). This work is one task of the ongoing research into the Housing sample of the CPI in preparation for the revision of the sample design and estimation procedure which is currently scheduled for 1998. In this paper the ad hoc methodology used in the current system is investigated along with an empirical Bayes estimator.

In section one the sampling design and estimation considerations will be introduced. In section two the vacancy problem is described. Section three presents the model for the imputation of rents for vacant units. The estimation methodology will be explained in section four and findings will be presented in section five.

1. Introduction and 1987 Design Description

For a full discussion of the CPI the reader is referred to Chapter 19 of the *BLS Handbook of Methods*, (1992). However, the following features of the CPI methods of data collection and estimation are important for the present discussion.

Pricing for the CPI is conducted in 88 primary sampling units (PSU) in 85 geographic areas (New York consists of 3 PSUs and Los Angeles consists of 2 PSUs). In the CPI area design there is random selection of PSUs according to a stratified design in which one PSU is selected from each stratum. The four Census regions are used as the initial stratifying variable for PSUs. A description of the PSU selection can be found in Dippo and Jacobs (1983).

The PSU stage of sample selection is common to both Housing and Commodities and Services but in the Housing part of the CPI the next step is to divide each PSU into block clusters which are based on Census block groups. Block clusters include both Census enumeration districts and partial block groups. These are described in the *Handbook* p189. Segments are selected from block clusters and Housing units (HU) are systematically selected within segments. Thus, the design is a three stage stratified cluster sample.

The selected HUs are assigned to six panels. The six panels are collected on a rotating basis with rental units in panel one being collected in January and July, rental units in panel two being collected in February and August, etc., but owner units are only collected once every two years. Once a rental unit is initiated into the sample, attempts are made to collect data on the unit every six months until the unit is rotated out of the sample.

For each collection period, respondents for the rental units in the given panel are asked to give the current period rent and the previous period rent. The reason for the previous month's rent question is to attempt to get more timely data on rent change. However there are problems with using previous month's rent change which are not associated with the six month change in rent from the previous collection period. At this point it suffices to say that the problems with the one month rent change relate to bias and to the form of the index.

The CPI is a modified chained Laspeyres index, which is a ratio of the costs of purchasing a set of items of fixed quality and quantity in two different time periods. The Residential Rent index is the index of interest in this paper and it is estimated at the PSU level although not all PSUs are published. Let $I_{it,s}$ denote the index at time t, in pricing area i, relative to time period s. Then, at least conceptually,

$$I_{it,s} = 100{*}CW_{it} / CW_{is}$$

where CW_{it} and CW_{is} denote the aggregated weighted rents in Index Area i for times t and s respectively.

This conceptual form of the index is estimated by a composite index based on both one month rent change and six month rent change. The six month relative at time period t, denoted by $R_{t,t-6}$, is the ratio of aggregated weighted rents for the current period to the aggregated weighted rents for the six month ago period for the same set of units. Similarly, the one month relative at time period t, denoted by $R_{t,t-1}$, is the ratio of aggregated weighted rents for the current period to the aggregated weighted rents for the one month ago rents for the same set of units. Let $I_{it-6,s}$ denote the value of the index at time t-6 and $I_{it-1,s}$ denote the value of the index at time t-1. Then the form of the current composite index is

$$I_{it,s} = .35{*}R_{t,t-6}{*}I_{it-6,s} + .65{*}R_{t,t-1}{*}I_{it-1,s}$$

where the weights .35 and .65 were chosen to minimize variance.

2. The Vacancy Problem

Nonresponse is a major problem for any survey including the CPI. Whenever possible, in the CPI, nonresponse at the unit level is dealt with by noninterview adjustment. However, vacant units present a real problem both conceptually and practically. The purpose of the Residential Rent Index is to measure the changes in transaction rents on rental units in the housing stock. If a rental unit is vacant it has no transaction rent to be reported. A vacant unit is, in some sense, temporarily out of the defined universe even though it is still in the frame.

The process of imputation is to assign to a vacant unit the dollar value that the unit would rent for if it were occupied. The real goal of the imputation is to determine a change in price associated with units which are vacant.

It might appear that reassigning the sample weights of vacant units to nonvacant units could be a solution. It has been found however, that vacant units act very differently from nonvacant units. This would imply that using the full set of nonvacant units to either reweight or impute the vacant units could bias the index. A second difficulty is that one month rent change and six month rent change behave differently with respect to errors in rent change values.

3. Imputation Models

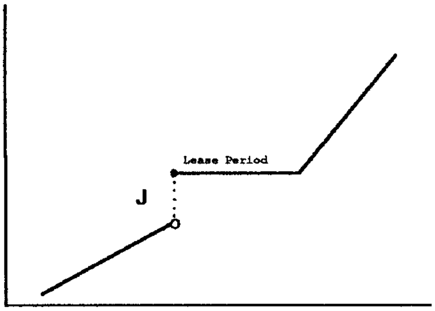
The three main assumptions underpinning the current vacancy imputation model along with some discussion follow:

1. Assumption: Rents tend to increase at a different rate for units that are vacant than for other units. Thus, including vacant units in the usual noninterview process results in a downward or upward bias in estimated one month price relatives.

Discussion: There have been several studies performed at BLS which indicate that this is a correct assumption. For a reference to earlier work see Sommers and Rivers (1983).

2. Assumption: Rents tend to increase at a steady rate of change, referred to as ROC, until the tenant moves out at which point a vacancy occurs and the rent jumps at some jump rate, J. After this, rents stay constant for the contract period of the lease.

A simple graph is used to illustrate these ideas.



Discussion: The major problem about the second assumption is that there appears to be an inconsistency in dealing with rents before and after a vacancy. Although aggregate rents may increase or decrease at some rate denoted ROC both before and after a vacancy, the rent for an individual unit may not change over time, at least an individual unit's rent should not be changing monthly. The current model assumes that a unit will have a tendency to have a lease after a vacancy since landlords wish to insure a minimum length of tenancy after a vacancy. However, prior to a vacancy, for a tenant who has occupied a unit for some time there may be no lease.

The question then arises as to whether the correct model should be that rents are constant before a vacancy and constant afterwards or should the model be that rents change before a vacancy and change afterward. The current housing data may or may not lend credence to either side of this debate. Presently the only information in our Housing database which can address this issue is the current (time t) and previous (time t-1) rent value for units in which there was a tenant change. Considering panels one and two only, the data show that over 70% of the units reported no change in rent in the collection period following a vacancy. This would seem to be partial support for the assumption that no changes occur in the six month period after a vacancy. However, over 20% of the units do report a change in rent after a vacancy which is partial support for the position that after a vacancy the rent should be assumed to increase at the rate ROC. This data is known to have some reporting problems which are surely reflected in the preceding numbers. There is also the problem of special deals on first month's rent, which may be part of the 20% changes. It should also be very clear that these numbers represent only part of the information which is actually desired. Arguments can be made for one form of the model or another based on mathematical or economical principles but, at this point, no model checking can be done without additional data. However, monthly changes in rents for an individual unit seem somewhat unreasonable and lead to a model which is impossible to estimate under the current panel structure.

Thus, the following concept has been proposed. For the aggregate of units the rents change at each time period by some rate known as ROC (which technically is a function of time) while for each individual unit the rents before and after a single vacancy are constant for the length of the preceding and following collection period.

3. Assumption: A set of BLS vacancies follows the Census distribution of vacancies gleaned from 1984 Annual Housing Survey data.

Implicit within the model is the assumption that the variable, length of occupancy in months, splits a given panel of the housing sample into i.i.d. groups which follow the Census distribution.

Discussion: The distribution of lengths of vacancies has been questioned as being not dynamic. Initial investigations indicate that the distribution may not be as stable over time as was initially hoped. However, the real problem with the distribution assumption is that our method of visiting a unit every six months censors the Census distribution. Even if the Census data were perfect and never changed over time we would not, on average, encounter that distribution in our sample.

The distribution of lengths of vacancies has been investigated by estimating the current cumulative distribution using BLS data. Using current BLS data it is possible, under the third assumption of the current vacancy imputation model to obtain *estimates* of the cdf of the distribution of length of vacancy. Note that BLS only collects length of occupancy of current tenant. Since each level of the cdf is estimated independently it is possible, although clearly undesirable, that the estimated cdf not be an increasing function. The data used in this estimation is from the six collection periods, January 1992 to June 1992, for each of the six panels.

From the Bureau of the Census "Vacancy Rates and Characteristics of Housing in the U.S.: Annual Statistics 1984" and from BLS data from 1992 the following distribution was gleaned.

Table 1
CDF of Vacancy Distribution

vacancy in months	Census percent	BLS percent
2 months or less	64.5%	73%
3 months or less	75.9%	79%
4 months or less	87.3%	83%
5 months or less	93.65%	85%

A comparison of the estimated cdf with the Census cdf indicates that the current BLS data is not following the Census data as closely as could be hoped. The Census distribution for two months or less shows a value of 64.5% but for all six panels the estimated value of this category is larger than 70%.

An alternative model is presented which is based on a simplified set of assumptions. This model will be the basis for the empirical Bayes estimator.

1. The model should be treated as a discrete time stochastic process.

2. The entire set of observed units is not representative of vacant units. In particular, with the data which is currently obtained, short tenure units are the best set of units to represent vacant units.

3. Clearly vacant units have no leases. However, for estimation purposes, it will be assumed that units which are vacant for a single collection period have only one change of rent from the last true observed rent before the vacancy until the next observed rent after the vacancy.

4. The ROC, as currently calculated, will be used to estimate the rent relative for the second consecutive vacancy for a unit.

This model does not really include any extra assumptions over the currently used model but does eliminate the need for the Census data.

4. Estimation Methodology

The current estimation is based on the jump rate model given in the previous section. It assumes that units which are *short tenure*, i.e., units with length of occupancy less than six months, are the best set of units for estimating rents for vacant units. It is also assumed that units do not experience multiple vacancies in the six month intervals between our visits. This is not perfectly true and currently there has been no field study to determine the extent to which multiple vacancies within the six month period are occurring.

For units which experience a vacancy between the previous pricing period (t-6) and the current pricing period (t), the assumed relationship between six month change of rent and the variables Rate of Change (ROC) and Jump rate (J) is presented below in Table 2. The table is presented by length of occupancy for the current pricing period (t) which must be less than six months. This relationship has several different values depending on the length of vacancy. For example, if the unit were occupied for five months, then the vacancy could only have been for one month. Thus, under the jump rate model, the current rent would be the jump rate (J) times the previous rent.

Table 2
Assumed Relationship of Six-Month Change to the Rate of Change (ROC) and the Jump rate(J)

Tenure (in months)	Six Month Change
(1) 1 month or less	ROC^5xJ or ROC^4xJ or ROC^3xJ or ROC^2xJ or ROCxJ or J
(2) between 1 and 2 months	ROC^4xJ or ROC^3xJ or ROC^2xJ or ROCxJ or J
(3) between 2 and 3 months	ROC^3xJ or ROC^2xJ or ROCxJ or J
(4) between 3 and 4 months	ROC^2xJ or ROCxJ or J
(5) between 4 and 5 months	ROCxJ or J
(6) between 5 and 6 months	J

These sets of possibilities occur because, while we know the length of occupancy of the current tenant, we do not know when the previous tenant moved out of the unit. The idea of using the Census data is to apply the distribution from the Census data to the above situations. Thus, for example, for the case of a length of tenancy of one month or less, we could expect 43% of the units to have a six month rent change of ROC^5xJ, 21.5% of the units should have a rent change of ROC^4xJ, etc.

Notationally, let $R^{t,i}$ denote the sum of weighted aggregate rents at time t for units with new tenants, which were occupied at time t-6, and have a length of occupancy of i, i = 1,...,6. If length of occupancy is i the unit has been occupied more than i-1 months and less than or equal to i months. The mathematical formula for the jump rate is given as

$$Jump\ Rate = \frac{\sum_{s=1}^{6} R^{t,s}\left(k_{1s} + \sum_{u=1}^{6-s} \frac{k_{2s}}{ROC^{u}}\right)}{\sum_{s=1}^{6} R^{t,s}}$$

The numbers k_{1S} and k_{2S} are from the Census distribution. The ROC value is calculated as the sixth root of six month change in weighted aggregate rents for units with length of occupancy greater than or equal to 6.

The housing economists have raised a point which indicates that the current methodology needs improving. The economists perceived the current methodology as performing more poorly in situations where a PSU or PSU replicate is deemed insufficient to estimate the current jump rate. The proposed methodology was designed with this problem in mind. The results are further addressed in the next section.

Under the simplified model proposed in the previous section, an average relative has to be calculated for each PSU replicate level combination.

Two alternatives are considered. First is the very simple idea of calculating the average relative for each PSU replicate combination as the six month change in weighted economic rents of the short tenure units. Under the proposed model this is all that is necessary. This has the advantage of extreme simplicity.

A second alternative is to use an empirical Bayes estimator to estimate the average relative. There is a short introduction to empirical Bayes estimation methodology in Casella (1985). The empirical Bayes estimator is calculated by region. First, the average relative is calculated for each PSU replicate combination exactly as in the previous method. The empirical Bayes estimator is then calculated as a weighted combination of the PSU replicate estimate and the region average of the PSU replicate estimates.

There are several advantages to this methodology. It is not an ad hoc procedure. It is robust against several types of outliers. Even if a PSU replicate is insufficient, any information which does exist in the PSU replicate, can be used to improve the estimate. It is a simple formula which is easy to program. There is one weakness.

The problematic aspect of the empirical Bayes method is that the weights for the region and PSU ingredients depend on variance components that are difficult to estimate. In the current research ad hoc estimates of the weights indicated that the empirical Bayes methodology is at least as good as the current methodology without involving the complicated assumptions of the current model.

5. Findings

The three methods were compared for collection periods January of 1992 through June of 1992. The method of evaluation was to use the imputed rent from the given collection period, say t, and compare that result to an actual rent at time period t+6. According to both the current and proposed methodologies the rent at time t+6 is a gold standard although in reality it may be more like silver. Note that this implies that units which were still vacant at time t+6 had to be discarded from the analysis. For each of the three methodologies two variables were created. The proportion of the difference between the imputed value and the gold standard was created:

((imputed at t) / (actual at t+6)) - 1

Also the absolute value of this variable was used in the analysis. The mean and standard deviation of these variables were calculated for each region replicate level combination.

An example of the results for the set of units with a single vacancy by Region and for time period January 1992 to December 1992 are presented in Table 3.

TABLE 3 - 1992

VACANCY IMPUTATION ERROR
By Census Region
Mean of the Test Variable

Variable	North East	Mid West	South	West
Current	0.015	-0.002	-0.008	0.008
e Bayes	-0.003	0.001	-0.010	0.005
Simple	0.010	-0.002	-0.009	0.006
Absolute Value				
Current	0.125	0.083	0.081	0.096
e Bayes	0.116	0.084	0.081	0.096
Simple	0.125	0.081	0.081	0.095

At this point no one methodology is clearly consistently better in terms of the analysis. It is also clear that improvements must be made in the estimates. However, the empirical Bayes estimate does better in many cases, especially if a PSU is deemed insufficient in number of short tenure units.

Also simulations were run in a Unix environment to test the different techniques. In this set of simulations, a set of units in three PSUs was randomly generated. Then, a 10% random sample of the units was taken and the units were assigned to panel replicate combinations. Next rent changes were generated for the units. Some units were randomly assigned to be vacant independent of the sample and independent of the rent change. The rent index was calculated using each of the imputation methodologies as well as several experimental index forms.

The access to the entire set of units as well as the sample allowed a comparison of the calculated index to something dubbed *truth*. The theoretical concept of a cost of living index is to measure the change of transaction rents in a fixed set of housing stock. Since some units are vacant at any given time then in some sense the set of housing stock is shifting. Under this circumstance the concept of *truth* shifts a little also. At each point in time *truth* is measured as the aggregated set of rents for the nonvacant units from the current period divided by the aggregated set of rents for the same units in the initial time period.

There seemed to be an interaction of estimator with imputation method. Surprisingly, the combination of the current estimator with the current vacancy imputation performed about as well as any other estimator. This is somewhat difficult to explain given the perceived defects in using the Census distribution of vacancies. In situations where no PSUs were deemed insufficient in short tenure units the current estimator performed as well as the empirical Bayes estimator.

At this point in time the simulation has not been programmed to simulate PSUs with an insufficient number of short tenure units.

6. Conclusions

Thus far the results are mixed. The empirical Bayes estimator appeared to perform better on current units in the sample in terms of predicting future rents of currently vacant units. On the other hand in terms of effect on the index the current estimator together with the current imputation methodology seemed to perform as well as any other combination of estimator and imputation methodology. This result clearly needs more research.

7. Acknowledgments

The author would like to thank Bill Miller, Shawn Jacobson, Jim Branscome, and Steve Henderson for insight into the BLS Housing design. Thanks go to Ken Archer for help in accessing the Housing database. The author would like to thank Janet Williams for support on this project.

8. References

Bureau of Labor Statistics, *BLS Handbook of Methods* (1992), Washington. DC: U.S Government Printing Office, 176-235.

Casella, George (1985), "An Introduction to Empirical Bayes Data Analysis," *The American Statistician*, vol 39 #2, 83-87.

Dippo, C. S., and Jacobs, C. A. (1983), "Area Sample Redesign for the Consumer Price Index," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 118-123.

Fay, R.E. and Herriot R.A. (1979), "Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data," *JASA* 74 #366, 269-277.

Lane, W. F., and Sommers, J. P., (1984) "Improved Measures of Shelter Costs," *Proceedings of the Business and Economics Statistics Section,* American Statistical Association, 49-55.

Sommers, J. P., and Rivers, Joseph D., (1983) "Vacancy Imputation Methodology for Rent in the CPI," *Proceedings of the Business and Economics Statistics Section,* American Statistical Association, 201-205.