COMPOSITE AND MODEL BASED INDEX ESTIMATORS

John G. Kovar, Statistics Canada
11th floor, R.H.Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6

ABSTRACT

A problem of estimating monthly movements in rents based on data collected every four months is explored. Four alternative model based estimators and a class of composite estimators of the rent index will be derived and justified, both from an intuitive as well as theoretical point of view. Jackknife estimates of the precision of the index estimators will be presented. Empirical results using the Rent Survey Pilot Test data will be discussed.

KEY WORDS: Consumer Price Index; rotating samples; jackknife.

1. INTRODUCTION

The rent component of the Canadian Consumer Price Index is currently based on data collected on a six month rotating basis using a Labour Force Survey Supplement. Since changes in rents generally occur on an annual basis, the effective sample size of the Canadian Labour Force Survey (LFS) design is reduced. In other words, because the dwellings remain in the LFS sample for only six months, estimates of the magnitude of change can, on average, be obtained from only half the units. Furthermore, special annual benchmarks, obtained by reinterviewing the June sample of dwellings one year later, indicated that the rent component can suffer from various degrees of downward bias (Dolson, 1982). To ameliorate the situation, a new methodology of collecting rent data was proposed and has been tested using the Rent Survey Pilot Test.

Under the proposed methodology of the pilot test, the data are collected from the same units every four months over a period of thirteen months. The rotation scheme of the survey ensures that data on current rents as well as matched rents collected four months ago are available from exactly three rotation groups every month. For the purposes of the pilot test, the first two observations are the first and the fifth interviews of the regular rent component. A follow-up survey was conducted to obtain the other two observations in months nine and thirteen. The pilot test thus inherits the sample design of the Labour Force Survey.

In estimating the indices, the constraints of the Consumer Price Index publication policy must be kept in mind. In other words, it must be practically as well as technically possible to produce the indices on a monthly basis for each of the index cities. The estimates must be timely; produced no later than mid-month following the reference month. Furthermore, no revisions can be made once the indices are published. While not entirely essential, it would be desirable that any proposed estimator be able to reflect (real) sudden changes in trend very quickly. On the other hand, in order to remain credible, the indices must be relatively stable; volatile, saw-toothed indices are to be avoided.

In Section 2, four model based estimators will be presented, justified and compared on a theoretical basis. A class of composite estimators is proposed in Section 3. The associated precision estimators are discussed in Section 4. The results of applying the proposed methodology to the preliminary data from the Rent Survey Pilot Test are presented in Section 5. The conclusions and recommendations can be found in Section 6.

2. MODEL BASED INDEX ESTIMATORS

In this paper, only matched indices, indices based on matched samples, will be considered. While relative changes could easily be derived by comparing independent (unmatched) estimates of rent levels at distinct time points, such estimates of levels would have to be very reliable, necessitating prohibitively large sample sizes. Moreover, past studies indicate that such direct estimators tend to be volatile, upwardly biased and generally not practical in use (Szulc, 1983). In what follows, therefore, an estimate of relative change between two time points will be based only on matched units, that is, those units that report rents for both of these time points.

A rent index is customarily estimated by chaining one month relatives, that is, the ratios of changes in rents between two consecutive months, denoted by $I_{m/m-1}$. In other words, the index in month m over a base period zero, $I_{m/0}$, is estimated recursively by

$$\hat{I}_{m/0} = \hat{I}_{m-1/0} * \hat{I}_{m/m-1}$$
$$= \hat{I}_0 * \hat{I}_{1/0} * \hat{I}_{2/1} * \dots * \hat{I}_{m/m-1}, \qquad (2.1)$$

where $I_0=100$ is the (arbitrary) level of the index at time zero. The difficulty then rests only in estimating the monthly relatives.

In general, consider the relative change in rent in month m over month n, denoted by $I_{m/n}$. Denoting by x_m the total rent paid in the current month m by a certain subset s of dwellings in a given city, this "m over n relative" can be estimated by $\hat{I}_{m/n} = x_m/x_n$. However, for the proposed design, if one considers matched indices only, the only estimable relatives are those of the form $I_{m/m-4i}$, i=1,2,3. These relatives are estimated by $\hat{I}_{m/m-4i} = x_m/x_{m-4i}$, where the set s of dwellings consists of only those units that report rents both at time m and m-4i. Unfortunately, the interest lies in estimating monthly relatives, those of the form $I_{m/m-1}$. On the positive side, the rotation scheme ensures that a four-month relative will be available every month (Kovar, 1984). It is also assumed that units rotating out of the sample are replaced by equivalent units rotating into the sample. The set s of common dwellings therefore depends on the time m only, and any explicit future reference to it, while implicitly retained, can be suppressed in what follows. For a rigorous discussion of these assumptions and the effect on the index if the assumptions fail, the reader is invited to consult Szulc (1983) and Kovar (1984).

In the following paragraphs, four methods of estimating monthly relatives from four-month relatives will be described. Each will be justified intuitively as well as theoretically, and its advantages and disadvantages will be pointed out. The first three methods are derived by modeling the unobserved, previous month's rent, x_{m-1}, using the observed rents at time m and m-4. The fourth method attempts to exploit the rotation pattern of the survey. All four assume that at least a four month back history of data is available.

One way of estimating the relative $I_{m/m-1}$, is to estimate the previous month's total rent, x_{m-1}. This can be accomplished, among other methods, by linearly interpolating the observed rents at time m and m-4. Thus the previous month's total rent can be estimated by

$$x_{m-1} = .25\ x_{m-4} + .75\ x_m \qquad (2.2)$$

and consequently, the monthly relative for month m by

$$\hat{I}_{m/m-1} = x_m / x_{m-1} = 4x_m / (x_{m-4} + 3x_m) \qquad (2.3)$$

The <u>interpolated index</u> is derived by chaining these relatives as in (2.1) above. It can be shown that the interpolated index at time m depends on all the rents between time -4 and m. In other words, the index is susceptible to accumulating various biases over time. Note that the same index would be derived by assuming that the four-month increment, $x_m - x_{m-4}$, occurred in four additive steps, since then the previous month's rent would be estimated by $x_{m-1} = x_m - .25\ (x_m - x_{m-4})$, which is the same as (2.2).

In contrast to the above, we can attempt to estimate the monthly relative directly, by assuming that the four-month relative is due to four equal movements which act multiplicatively (Kosary, Branscome and Sommers, 1982). Under this assumption, the monthly relative can be estimated by

$$\hat{I}_{m/m-1} = (\ \hat{I}_{m/m-4}\ )^{.25} \qquad (2.4)$$

Assuming further that there are no sample changes or that units rotating out of the sample are replaced by equivalent units rotating into the sample, then the <u>geometric index</u> in month m over the base period zero becomes

$$\hat{I}_{m/0}=I_0*(I_{1/-3})^{.25}*(I_{2/-2})^{.25}*\ldots*(I_{m/m-4})^{.25}$$

$$=I_0*(x_{m-3}*x_{m-2}*x_{m-1}*x_m)^{.25} / (x_{-3}*x_{-2}*x_{-1}*x_0)^{.25} \qquad (2.5)$$

In other words, the index is a ratio of two geometric averages; hence the name geometric index. We note that at any time, the index depends on eight months worth of data only, and thus is independent of all movements between time 0 and m-4, and is therefore not susceptible to accumulating biases. However, it suffers from a mixture of one-month to three-month lags and will thus tend to dampen true sudden changes. These changes, on the other hand, will be reflected eventually, that is, the index will selfcorrect (Kovar, 1984). As a point of clarification, note also that the relationship in (2.4) can be expressed as

$$\log\ (x_{m-1}) = .25\ \log\ (x_{m-4}) + .75\ \log\ (x_m) \qquad (2.6)$$

The geometric index is therefore equivalent to an index derived by estimating the previous month's rent by linearly interpolating the logarithms of the observed rents at time m and m-4.

Analogous to the above geometric index, we can assume that the four consecutive monthly relative net increments are equal and acting additively. More precisely, we can write $I_{m/n}$ as $I_{m/n} = 1 + i_{m/n}$, where $i_{m/n}$ is the relative net increment in month m over month n. To estimate the monthly relative $I_{m/m-1}$, and consequently the index, we assume that the available four-month net increment $i_{m/m-4}$ is composed of four equal monthly increments $i_{m/m-1}$. The relative $I_{m/m-1}$ can then be estimated by

$$\hat{I}_{m/m-1} = 1 + i_{m/m-1} = (x_m + 3x_{m-4}) / 4x_{m-4} \qquad (2.7)$$

We note that (2.7) can be written as

$$1 / x_{m-1} = .25 / x_{m-4} + .75 / x_m \qquad (2.8)$$

In other words, the <u>incremental index</u> corresponds to one which would be derived by estimating the previous month's rent by linearly interpolating the reciprocals of the observed rents at time m and m-4. As is the case with the interpolated index, the incremental index will not be independent of the intermediate observations and be therefore susceptible to various accumulating biases.

It is interesting to note at this stage that the interpolated, geometric and incremental monthly relatives are respectively the weighted arithmetic, geometric and harmonic means of rent quotations collected four months apart. The standard relationship between these means explains the fact that the three indices are ordered in magnitude from smallest to largest in the above order. That is, in an inflationary situation the interpolated index will always be smaller in value then the geometric index which in turn will always be dominated by the incremental index. The reverse holds true when the trend is downward, that is, when prices are decreasing.

Finally, the <u>carried index</u> is constructed by taking advantage of the rotating sample at hand. Noting that all units reappear periodically in the sample, we construct the index by simply carrying each unit's rent value forward until a new observation is recorded. In this way all units on the file have a matching previous month's rent and thus the monthly relative can be constructed in a straight forward manner. The obvious drawback is that the rent increases (decreases) are not recorded until observed, this being the implicit, underlying model. However, since all changes are eventually recorded, the index will selfcorrect (Kovar, 1984), but will suffer from a mixture of one to three-month lags. Just as for all of the above indices, sudden (real) changes will be dampened but will be reflected eventually.

On the technical side, we note that in computing the carried index for any given month, one quarter of the observations on the file reflect a four-month movement, while three quarters of the observations are carried for one to three months and reflect no change. The monthly relative is therefore estimated by

$$\hat{I}_{m/m-1}=(x_m+x_{m-1}+x_{m-2}+x_{m-3}) \;/\; (x_{m-1}+x_{m-2}+x_{m-3}+x_{m-4}) \qquad (2.9)$$

Chaining the relatives as in (2.1), and assuming no sample changes take place, we obtain the index for month m over the base period zero as

$$\hat{I}_{m/0} = I_0*(x_{m-3}+x_{m-2}+x_{m-1}+x_m) \;/\; (x_{-3}+x_{-2}+x_{-1}+x_0). \qquad (2.10)$$

In other words, the index is a ratio of two arithmetic averages. Analogous to the geometric index, the carried index depends on eight months worth of data only, and thus is independent of the movements between time 0 and m-4.

3. COMPOSITE INDEX ESTIMATORS

While the Rent Survey Pilot Test was ongoing, it was recommended that the data for the three months between observations should be collected by recall, instead of basing the index on observations that are four months apart. Thus in month m the rent for that month, x_m, would be collected along with the previous three months' rents: x_{m-1}, x_{m-2}, and x_{m-3}. The four-month old rent, x_{m-4}, could be obtained from the previous interview. In order to estimate the monthly index at time m over the base period zero, $I_{m/0}$, a composite estimator was proposed, namely

$$I_{m/0} = \sum_i w_i \, I_{m-i/0} \, (x_m / x_{m-i}) , \quad \text{where } \sum_i w_i = 1 \qquad (3.1)$$

or, in order to preserve the underlying multiplicative nature of the index,

$$^*I_{m/0} = \Pi_i \, (I_{m-i/0} \, (x_m / x_{m-i}))^{w_i} , \quad \text{where } \sum_i w_i = 1. \qquad (3.2)$$

In other words, the index is formed by obtaining a weighted arithmetic or geometric average of four estimates of change in the current rent over one of the last four months' rents, chained to the existing index at the appropriate time points.

Since the Rent Survey Pilot Test was in progress when the above suggestions were made, it was impossible to modify the methodology in order to allow the collection of the intermediary rents. As such, in order to test the above estimators, this missing information had to be obtained by some other means. The intermediary information is needed at least for those units for which the current rent was not the same as the rent reported four months previously. For units which showed no change, it was assumed that the same rent prevailed throughout the four month period. Fortunately, the ongoing regular rent component could be used to estimate the probabilities of rent change in any given month in all cities. This information was used to simulate the date of change for units that in fact exhibited such change over any four month period. This partially simulated data was used to construct the indices defined in (3.1) and (3.2) above.

Since no appreciable difference was found between the two estimators, only the results for the additive estimator (3.1) are reported in what follows. It is realized, however, that the lack of difference between the two estimators may be due to the random nature of the data set induced by simulation. Therefore, both of the estimators will be considered when real data becomes available.

4. VARIANCE ESTIMATION

The indices described in the previous sections are functions of varying complexity depending on both current as well as past data. To further complicate the matter, they are based on partially rotating samples and thus data collected in periods which are close in time are very highly correlated. It is very difficult, therefore, to derive explicit expressions for the estimators of variances of these indices or even for their first order Taylor approximations. On the other hand, resampling methods, which recalculate the estimate of interest based on carefully selected subsets of the original data, can be of use here.

The precision of the proposed four-month estimators as well as the current monthly estimator can, for example, be obtained by jackknife repeated replication. In order to form the jackknife replicates however, the observations need to be linked in time. That is, units that rotate into the sample need to be associated with those units which rotate out of the sample and which they replace. Due to the design of the underlying Labour Force Survey, these linkages cannot be performed at the dwelling level, but can be accomplished with reasonable success at the primary sampling unit (PSU) level. In general, a PSU contains two to ten rented dwellings. The units that are linked in this manner thus form a constant number of panels that is unaffected by the rotation scheme. The panels of dwellings are then randomly grouped into larger panels of size g. The choice of g, the number of panels grouped together for the purpose of jackknifing, is quite arbitrary. Its main purpose here is to reduce the computing time.

For the purpose of variance estimation, a "pseudo-value" of the index is calculated from each jackknife replicate. A "pseudo-value", in the Tukey sense, is a weighted average of the original estimate and that obtained when data of one replicate is removed. One "pseudo-value" corresponding to each jackknife replicate can thus be calculated. The variability of these "pseudo-values" is then used to estimate the precision of the index. The confidence limits are obtained using the standard normal approximation assumptions. Alternatively, "pseudo-values" of relative changes could be used to obtain variances of relatives and subsequently those of the indices themselves (Kosary, Branscome and Sommers, 1982). This option has not as yet been investigated.

5. EMPIRICAL OBSERVATIONS

In September, 1985, the Prices Division of Statistics Canada initiated a Rent Survey Pilot Test in five urban centers according to the design described in Section 1. Under this design, each unit's rent is collected in months one, five, nine, and thirteen. The first two observations for each unit (month one and five data) are obtained from the ongoing regular rent component survey. Month nine and thirteen data are obtained using a vehicle similar to that used for the annual benchmarks. The Pilot Test was thus initiated in September, 1985 by collecting month nine data for units which had rotated into the regular component sample in January, 1985. As such, the first four-month relative could be obtained using May, 1985 data; the survey was fully rotated in as of January, 1986. Note that because of the collection method employed, the regular rent component and the pilot indices are not independent.

In this section we will describe the behaviour of the various index estimators as observed in application to the most recently available data, that is, the period of May, 1985 to December, 1986. It must be pointed out that the first four months of data in this period are based on one rotation group only, the following four months have two rotation groups available and only as of January, 1986 are the estimates based on all three rotation groups. For ease of comparison, all indices were initialized to their published value as of April, 1985. The raw regular rent component index and the published index are also presented.

The results for all the possible combinations of urban centers and indices are too numerous to include herein; they are available from the author. Some selected highlights will be put forth in the following paragraphs. While not exhaustive, they are hoped to be representative as well as indicative of the situation at hand.

Figures 1 and 2 present the indices for Montréal and Toronto. Similar graphs have been produced for the other three urban centers of Calgary, Regina, and Western Ontario, but are not presented here. Since the three interpolation based indices are too close together to be distinguishable, only the geometric index is presented. Detailed analysis shows, however, that as expected, their order is the same in all cases: interpolated, geometric, incremental. By contrast, the carried index is noticeably lower than the above three indices. For comparison purposes, the raw regular component index and the published index, that derived by imputing and otherwise cleaning up the raw regular index, are also plotted.

As can be seen in Figure 1, Montréal, the "smoothing" problem exhibited by the model based indices is corrected by the composite index. That is, in cities where a large proportion of units exhibits a change in rent in a given month, the model based indices will smooth the jump in the index over a period of four months. The composite index, however, attributes this change to where it in fact belongs. In other cities, where the index is relatively smooth, both the model based and the composite estimators perform relatively alike (see Figure 2).

We note that with the exception of Western Ontario, where no published index is produced, both the published and regular indices are very close. Secondly, they both tend to intertwine with the interpolation based indices, even in cities with very small sample sizes such as Regina. For the cities of Montréal, Calgary, and Regina, the carried index is substantially lower than both the regular component based indices and the interpolation pilot test based indices. In Toronto and Western Ontario, while the same relationship holds, it is not as clearly pronounced. The composite index is consistently higher than the regular series.

On the other hand, it was found that the composite index tends to be more volatile than the model based indices (Figures 2 and 3). While the stability of the model based indices is partly due to the strength borrowed from the underlying model, it must also be noted that the weights of the composite index were not optimized to minimize the variance of the index as could have been done. In this instance, all the weights were assumed equal since the dates of change were simulated. It is planned to optimize the weights once real data is available.

Confidence limits for all indices were produced as described in Section 4. In choosing a satisfactory value of g, the number of jackknife replicates, it became evident that all the variance estimators are very unstable. That is, the variance of the variance estimator is quite high. Some observations, however, can still be made.

Not surprisingly, the confidence bounds tend to get wider the further the index moves from its initial point (Apr.85), and as the sample size decreases (Figure 3). While in all instances the confidence bounds of the regular component monthly indices overlap with those of the model based pilot indices, this is not true in the case of the carried index or while the pilot based indices show a temporary lag due to a sudden jump in the regular series (Figure 3). The published series, however, remains within the bounds of the composite index at all times.

Finally, it must be noted that the regular component index is not necessarily synonymous with the true index. It is therefore difficult to decide which of the approximations are closer to the truth. In fact, in most cases, rather then suspecting the pilot interpolation indices of "running away", it is more likely that the regular component index is falling behind.

6. SUMMARY

Both the theoretical and the empirical observations suggest that the carried index demonstrates severe lags. While the interpolation based indices are not entirely free of these lags, they tend to recover more readily (Figure 1). The composite index, while more volatile, tends to pick up such changes at the appropriate time, and as such, is recommended here. Similar observations have been noted using simulated data (Kovar, 1986). When the new methodology is introduced, it is planned to produce the indices according to either one of the composite estimators discussed above. The geometric index will be produced for comparison purposes.

ACKNOWLEDGEMENT

Thanks are due to Patricia Whitridge for her technical assistance.

REFERENCES

Dolson, D.D. (1982). Rent Status Survey: Analysis. Statistics Canada technical report.

Kosary, C.L., Branscome, J.M. and Sommers, J.P. (1982). Evaluating alternatives to the rent estimator. Bureau of Labor Statistics technical report.

Kovar, J. (1984). Note on calculating the rent index. Statistics Canada technical report.

Kovar, J. (1986). Calculating monthly indices based on trimestrial data. Survey Methodology 12, 107-120.

Szulc, B. (1983). Linking price index numbers. Statistics Canada technical report.

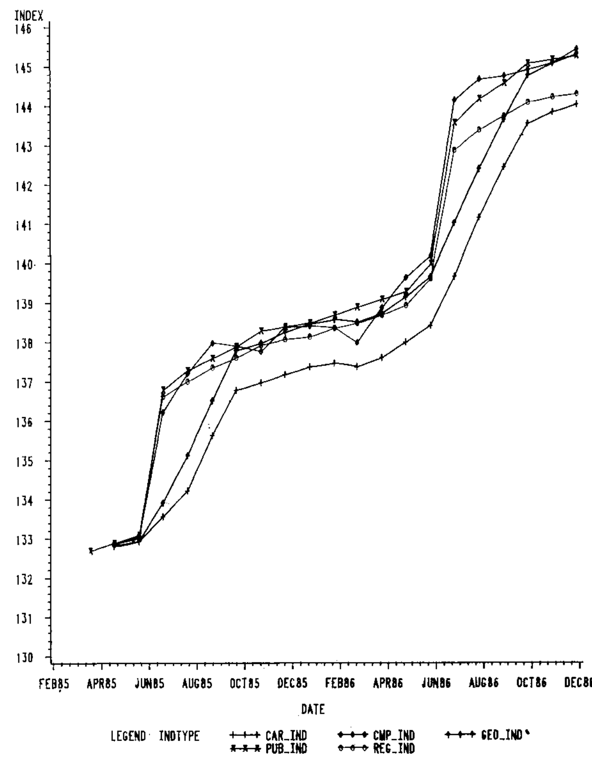
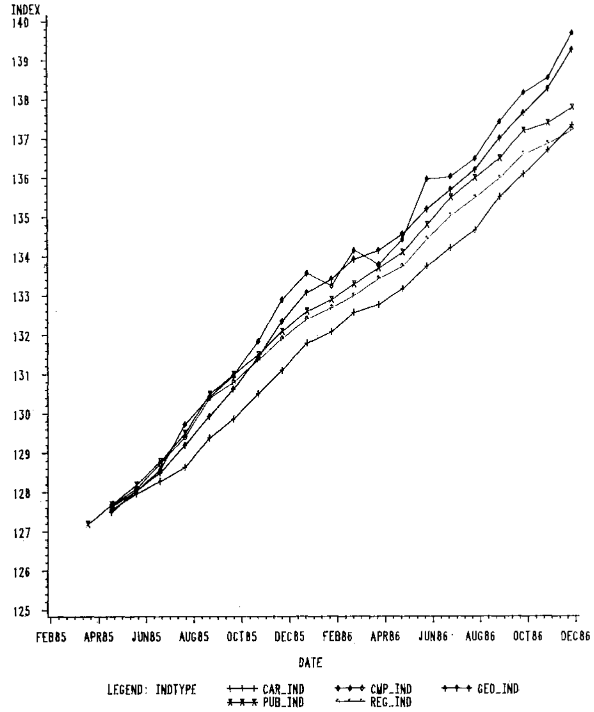
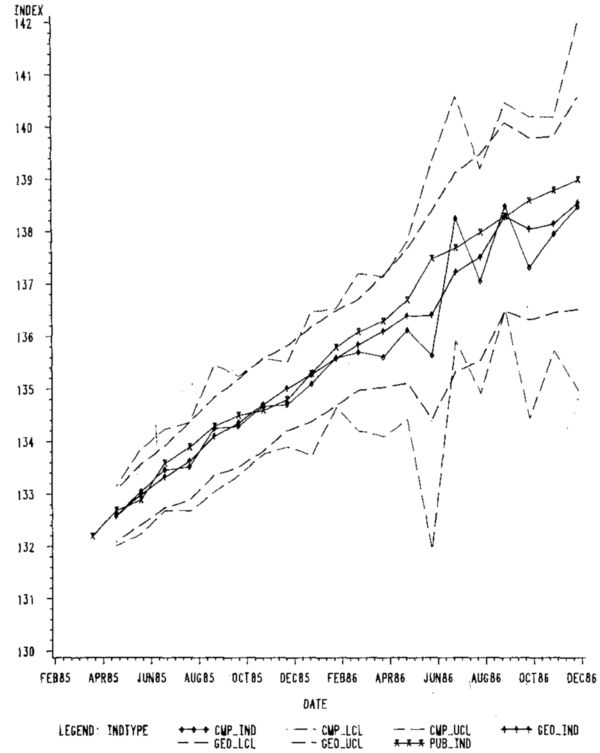


Figure 1: The Rent Index for the City of Montréal



<u>Figure 2:</u> The Rent Index for the City of Toronto



<u>Figure 3:</u> The Rent Index and its reliability for the City of Regina