The Teranet–National Bank House Price Index TM

1. INTRODUCTION

The Teranet – National Bank House Price Index™ is an independently developed representation of the rate of change of home prices in six metropolitan areas, namely Ottawa, Toronto, Calgary Vancouver, Montreal and Halifax. The metropolitan areas are combined to form a composite national index. The indices are estimated on a monthly basis by tracking the sale prices of condominiums, row/town houses and single family homes within these six metropolitan areas.

The estimation of the indices is based on the assumption of constant level quality of the single family dwellings and that any price changes are driven only by market trends. Thus, the indices attempt to reflect market prices by minimizing or eliminating the influence of any changes in the physical characteristics (e.g., renovations) of the houses.

The estimation of the indices is based on the "repeat sales methodology". This methodology was originally developed by Bailey, Muth and Nourse as a method of avoiding the heterogeneity issues in housing markets. This methodology was extended by K. E. Case and R.J. Shiller, the foundation of which is that each property contributing to an estimation of aggregate home value change must have been sold at least twice in a particular time frame. The two sale prices are assumed to define a linear change of the value of the property between the two sales dates.

1.1 Index Estimation

The Teranet-National Bank House Price Indices**™** are estimated by tracking the registered home sale prices over time. At least two sales of the same property are required in the calculations. Such a "sales pair" measures the increase or decrease of the property value in the period between the sales in a linear fashion. The fundamental assumption of the constant level quality of each property makes possible the index calculation but imposes difficulties in filtering out those properties that do not satisfy it. This difficulty arises from the lack of information about the property, and only the amount of price fluctuation versus time may provide an indication on possible changes in the physical characteristics of the property or non-arms-length transaction. Such properties may not be included in the estimation process.

Any property that has been sold at least twice is considered in the calculation of the index, except those that may be related to influences from within the property (endogenous factors): a) non-arms-length sale, b) change of type of property, for example after renovations, c) data error, and d) high turnover frequency. Once the unqualified sales pairs have been eliminated, the estimation of the index in a certain jurisdiction can be initiated by casting all qualified sales pairs in a linear regression algorithm.

1.2 Weighting of the sales pairs

The most challenging procedure in the index estimation process is the assignment of weights to each sales pairs: not every sales pair should contribute the same to the index calculation. There are several factors that can be considered, each carrying different levels of uncertainly, primarily due to lack of information.

One attractive method of assigning weights is based on the statistical distribution of the sales pairs in the geographical area of interest. This can be done by calculating the average annual percentage change of the price of each property and then "tallying" all changes in different classes. This allows the determination of the experimental probability distribution of the sales pairs (also known as probability histogram). Properties whose annual percentage change is more frequent in the set are most probable to occur and thus are assigned higher weight than those that belong to a class with lower count of properties. The probability measure is thus the weight that can be applied. This methodology can also be used to filter out sales pairs that show highly improbable annual sale price change. In this case, a statistical methodology may be applied to eliminate outliers, based on the type of the statistical distribution of price changes. A null hypothesis for testing can be formulated for this purpose.

Sales pairs may also be weighted based on the time interval between the sales. A property with a long time interval is more likely to have experienced physical changes and thus its sale value may be reflecting non market trends. Such properties may be weighted less than the ones sold in a short time interval. This is known as *time interval weighting* and varies from one geographical area to another [*Case and Shiller*, 2007]. The need for "time interval weighting" can be identified by performing first level estimates of the indices and then examining whether the residuals (errors) of the linear regression increase with the time interval. The change of the residuals as a function of time is known as *heteroskedasticity* and indicates the presence of a trend in the errors as the time interval between the two sales increases.

It is noted that the above weighting schemes may not be exactly representative of the reality because there are many unknown factors. However, they comprise good initial estimates, which can be improved through iterations of the estimation process. The simplest case scenario is the use of equal weights in the first run of the estimator followed by an iteration that makes use of a new set of weights that are inversely proportional to the square residuals. It is clear that the residuals of any estimation process depend on the initial weights therefore, it is important to have as realistic weights as possible right from the start, rather than using equal weights and then iterating the solution using the inverse residuals as weights.

1.3 Index calculation methodology

The repeat sales index construction is based on a simple linear regression model, whose regression coefficient is the reciprocal of the desired index. The fact that there are more observations (sales pairs) than unknown regression coefficients, an overdetermined system of linear equations is formed from which one can obtain an optimal solution by using for example the method of least-squares (LS).

The sale prices are affected by factors, other than the property itself such as, financial market trends, consumer price index, and many others that are called *exogenous* factors. The set of sales values (observations) is likely correlated with the errors and the LS estimation procedure cannot be applied in principle. Other methods, such as the Instrumental Variables (IV) may provide consistent estimates of the indices. Details of the different estimation algorithms are provided in Section 3.

The calculation of the indices is carried through by assigning a base period for which the index is set to unity (or 100). All sales pairs, prior to and including the base period, are retrieved from the database, then analyzed, filtered, and subsequently given an initial probability weight before they are cast into the linear regression estimator. We apply three different estimators, namely least squares (LS), instrumental variables (IV), and generalized method of moments (GGM). These estimators are designed to calculate all pre-base period indices simultaneously. The "simultaneous" nature of the calculation indicates that the interrelationship of the monthly indices is considered as it should. This is known as the pre-base index estimation process.

Once the pre-base regression coefficients have been estimated they are used in the post-base index calculation without any changes, i.e. they are kept constant. The post-base calculation is performed monthly, as new data come in. This implies that the post-base indices are dependent upon previous indices but they are independent from all future ones. In order to maintain accurate estimates of the market trends, the index of the current month is estimated simultaneously with the two previous monthly indices. Usually, but not exclusively, a simple average of the three monthly indices may be used as an estimate of the index of the current month whereas the indices of the two previous months are dropped.

1.4 Refinement of the weights

After their initial setting, the weights of the sales pairs can be refined by using for example the IV estimation process in an iterative approach. The initial weight matrix is used to run the IV estimator which produces residuals. The residuals (errors) represent the misfit of the data to the linear regression model. The square residuals may be taken to represent the variance of the misfit. A large variance indicates that the corresponding sales pair does not fit well to the model, or it is inconsistent with the rest of the sales pairs and it likely has a large error. In a first iteration of the solution, the initial weights are replaced by the reciprocals of the square residuals and new indices are calculated. This iteration can be repeated a few times until a desirable convergence has been achieved.

In order to test whether there is a need to apply interval weights, i.e., when there is heteroskedasticity present, a least-squares (LS) regression and its statistics can be used [*Baum et al.*, 2002; *Wallentin, and Ågren*, 2002]. First, the residuals from the LS estimator are tested for normality. If the statistical test shows that the residuals are normally distributed, it can be taken as a first crude indication that heteroskedasticity is not present. A more rigorous approach for heteroskedasticity testing is to use available statistical tests, e.g. Breusch-Pagan test, [*Breusch and Pagan*, 1979], White test [*White*, 1980], or similar tests that detect trends in the square residuals versus time interval. If such trends prove to be statistically significant, then heteroskedasticity is present and needs to be considered in the estimation process.

The methodology of heteroskedasticity testing can be further expanded to test for residual outliers i.e., residuals that are statistically inconsistent within the set of residuals. Sales pairs whose residuals are flagged as outliers can be down weighted before the iteration of the solution is attempted.

2. FILTERING AND WEIGHTING PROCEDURES

Filtering and weighting of the sales pairs are performed in different stages of the estimation process and are based on the statistics of the sales pairs and their internal consistency. Additional subjective filtering may also be applied by an experienced analyst who has a good knowledge of the housing market. Such subjective "algorithms" or "filtering parameters" can be selected before the start of the estimation process.

The filtering and weighting process is divided into two main stages, namely:

(a) Pre-analysis stage: sales pairs are examined and analyzed before they are cast into an estimation process. During this stage, the data (sale prices) and their statistics are examined in order to set initial threshold values for filtering and weighting.

(b) Post-analysis stage: after the data have been cast into a weighted least squares (WLS) and instrumental variables (IV) estimators and indices have been estimated, the covariance matrices of the estimated parameters and residuals are examined and analyzed. At this stage, rigorous statistical tests are performed to test the sales pairs against basic and simple hypotheses (also known as the Null hypotheses). Such tests reflect the internal consistency of the data from the point of view of how well they fit the regression model. Results from these analyses are fed back into the estimation process and final estimates of the indices along with their formal error estimates are obtained. On many occasions, iterations are required between the pre- and post-analysis stages.

What follows is a detailed description of the approaches and algorithms developed and implemented in relevant software. Several test cases are examined and pertinent numerical results are presented to accentuate the effectiveness of the methodology followed.

2.1 Pre-analysis of the sales pairs

The pre-analysis stage is entirely based on the original data set of sales pairs. Various statistical procedures of classification and testing are followed. Depending on the number of sales pairs in the original data set (sample size) one or more pre-analyses procedures as described below, may not be applicable since they may possibly become inconclusive with poor performance. Like in all statistical analyses approaches, care must be exercised when applying and interpreting statistical methodologies and results.

Property Type: The index calculation may only be based on one type of property, for example single homes, or condominiums or other, or a combination thereof. The decision to use a particular type of property or properties constitutes the first rudimentary but important filter of the sales pairs. In small jurisdictions, this filter may significantly decimate the data set. Statistically small number of sales pairs may result in highly inconsistent indices. Subsequent analyses may suggest the inclusion of additional types of properties to increase the number of sales pairs.

Sales Pair Time Interval: To avoid certain types of transactions, a sales pair whose corresponding dates of sale are six months or less apart are excluded from further consideration.

Extreme Sale Prices: In order to keep the index estimates consistent, a simple filter is applied to eliminate either very low or very high sale prices. This eliminates properties at both ends of the scale (shacks and mansions) that may not follow market trends. The price limits may vary from region to region. These values can either be set based on statistical analyses (usually not available) or on experience. The latter is most often followed.

Extreme Price Changes: This step concerns the detection and elimination of sales pairs that show extreme price changes. It is based on the statistical distribution of the price changes in the set. It involves the following:

(a) Calculation of the percentage change per annum of the sale price of each pair. Properties whose values change more than a certain set level per annum are eliminated from the set. The cut-off value may vary and can be determined more effectively by a feedback loop (iteration procedure). This can be based primarily on statistical considerations (see below).

(b) The filtered pairs from step (a) are organized into classes based on their annual percentage change, by applying the usual statistical tools, e.g. the probability histogram. The number of classes in the histogram is analogous to the size of the sample data set but does not usually exceed 50 classes. The usual, low order experimental moments of the histogram (e.g. mean, variance, skewness and kurtosis) are then calculated.

(c) The null hypothesis H$_0$ of normality of the distribution of the sales pairs is invoked and subsequently tested by using the *chi-square goodness-of-fit* test. This test uses the following statistic

$$y = \sum_{\ell=1}^{N} \frac{(a_\ell - e_\ell)^2}{e_\ell} \, , \tag{1}$$

where index ℓ indicates the class of the histogram and N is the number of classes. Variable a_ℓ is the normalized class count of class ℓ in the histogram, and e_ℓ is its corresponding normal distribution ordinate value calculated from the normal distribution $n(\xi; \bar{l}, s^2)$ using the mean and the variance of the sample (set of annual percentage changes). The above statistic *y* follows the *chi-square* distribution $\chi^2(\xi; n-3)$ of *n*-3 degrees of freedom because the mean $\bar{l}$ and variance s^2 of the sample (here, the mean of the percentage changes of all properties along with their variance, respectively) are unknown and are estimated. The loss of the third degree of freedom is due to the selection of the number of classes in the histogram.

(d) If the *chi-square goodness-of-fit* test passes, the annual percentage changes of the pairs are considered to be normally distributed and confidence intervals are used to eliminate pairs whose annual percentage changes are outside these limits. If the *chi-square goodness-of-fit* test fails, the annual percentage changes of the sales pairs are assumed to follow an alternative distribution; in such case, we cannot set confidence intervals based on the normal distribution but we can repeat step (a) above by setting slightly tighter limits. However, the intent is not to render the distribution to be normal. Tests with many different data sets from different Canadian cities indicate that the annual percentage changes, follow the normal distribution right from the start. If in the end the statistics show that the sales pairs are not normally distributed, no further measure is taken to eliminate pairs.

(e) The new "clean" set of sales pairs (as achieved from steps (a) through (d) above) is used to generate, once again, the experimental probability histogram using the same number of classes as in step (b) above. Based on this histogram, each of the sales pairs is assigned an experimental probability that serves as the initial weight in the estimation process.

Step (a) above can be bypassed altogether. In this case, a first pass through process (b) and (c) can eliminate extreme values. Subsequently, iteration through the same steps can filter effectively the bad pairs and assign the first weights.

Heteroskedasticity in the Sales Pairs: Heteroskedasticity (autocorrelation) is the phenomenon of non-constant variance in the data. In the housing market, the time interval between the two sales may introduce heteroskedasticity: The longer the time interval between the two sales, the higher the error in the sales pairs may be expected. This means that properties with higher errors should be weighted less. Possible heteroskedastic models have been suggested [e.g., *Case and Shiller*, 2007] that provide a scale factor by which the initial weights may be multiplied to remove the effect of time. However, care must be taken when applying such rather arbitrary models.

In our estimation process we do not consider heteroskedasticity, at least in the first run of the data through the estimator(s), for the following reasons:

(a) We do not have any evidence about the degree of heteroskedasticity present in the data. The application of an arbitrary model at the initial stage may in fact at times introduce heteroskedasticity where it does not exist. Over- or under-correction of the weights is as undesirable as ignoring it.

(b) The experimental probabilities that are used as initial weights in the estimation process, in fact account, to a certain degree, for heteroskedasticity that may exist in the data. Even if this account is only partial, any

residual heteroskedasticity may be weak to significantly influence the final index estimates. By the same token, if residual heteroskedasticity is still present, it can be detected and accounted for in the estimation process by using appropriate statistical tests (see Section 3.4).

(c) The most important reason for not accounting for heteroskedasticity in the data in the pre-analysis stage is however, the fact that we can account for it more precisely through its detection and estimation after the first estimation process is completed (in the post-analysis stage). We follow rigorous statistical tests on the estimated residuals that allow the detection and estimation of any heteroskedasticity present. If present, the estimation process is repeated with refined weights derived from the actual heteroskedasticity present. This is indeed a rigorous approach and it is detailed below.

2.2 Post-analysis stage

The post-analysis stage is followed after the pre-processed or "clean" sales pairs have been cast into a linear regression estimation algorithm. The analysis is performed on the estimated regression coefficients (solution) and estimated residuals. The former provides the degree of trust we can place on the results, whereas the latter tests for residual outliers. Both statistical tests require the calculation of the covariance matrix of the regression coefficients and of the estimated residuals. We note here that we calculate and use only the diagonal part of the covariance matrices.

The first estimator used is the weighted least-squares (WLS). Although the WLS may result in inconsistent estimates for the indices due to the correlation between the errors and the regressors, it serves us in three important ways:

(a) It is an efficient estimator by design.

(b) Any statistical tests applied on the WLS results are equally valid for other estimators, such as the IV estimator (*Wallendin and Ågren*, 2002; *Hansen, et al.*, 2006).

(c) The degree of bias of the results may be used to assess the extent to which *endogeneity* is present in the data and the robustness of the IV solution. From several numerical tests we performed, we found that the WLS estimator biases the results by a mere scale factor. The bias varies as a function of the quality of the data and the choice of the base period. For the latter, the more sales pairs contain the base period as one of the sales dates, the smaller the bias. The IV estimator is not very sensitive to such a choice.

What follows is a detailed description of the estimation processes, namely the WLS, the IV and the Generalized Method of Moments (GMM). WLS and IV may be considered as special cases of the EGMM estimator and as such it is important to examine it in more details.

3. THE LEAST-SQUARES ESTIMATOR

The estimation of the indices is achieved by following the weighted least-squares (WLS) and the weighted instrumental variables (WIV) approaches. In our case, since the regressors (sale prices) are considered as pairs in the regression equation they are most likely correlated with the errors. In this case the ordinary LS (no weights) and to a large extend the WLS will furnish inconsistent but efficient (smallest variance) estimation of the unknown parameters (regression coefficients), whereas the IV process will provide a consistent but inefficient estimates. The problem of inefficiency can be remedied by forming a realistic weight matrix (as realistic as possible) in the pre-analysis stage, and refining it during the estimation process. This can be done by using the WLS as a diagnostic tool before the WIV process is used for the final estimation. Therefore, the estimation process contributes to the refinement of the weights and renders the final estimation efficient.

3.1 The Regression Model

The weighted least-squares (WLS) estimator guarantees that the residuals are minimum thus offering a minimum variance estimate of the unknown parameters. The repeat sales index construction is based on a simple linear regression model. The index is the reciprocal of the regression coefficient, hereafter denoted as $\square$. The general equation of the linear regression model in matrix notation can be written as

$$Y = X\beta + U \, , \tag{2}$$

where X is the $n \times m$ matrix of *regressors* (design matrix), Y is the $n \times 1$ vector of independent variables, U is the $n \times 1$ vector of residuals, and β is the $m \times 1$ vector of unknown regression coefficients. In the above definitions, n is the total number of sales pairs considered with no endogenous regressors. We note that the quantity of interest to us is the index at certain epoch which is the inverse of the regression coefficient.

We denote the sale price of a property as P_{ij}, where subscript i is the property identifier and j refers to the epoch of the sale. For the purpose of this study, we consider monthly sales and thus, j will denote the month of transaction. $j=0$ signifies the base period with respect to which the index will be calculated. Eq. (2) can be written explicitly as:

$$P_{ij}\beta_j - P_{ik}\beta_k + u_i = 0 \, . \tag{3}$$

Eq. (3) is the most general form of the regression equation containing two unknown parameters, namely β_j and β_k that is, the regression coefficients of epochs j and k, respectively. If any of the regression coefficients is known, either from previous estimations or it corresponds to the base period (β_j =1 or β_k =1 for $j=0$ or $k=0$, respectively), then the corresponding product $P_{ij}\beta_j$ or $P_{ik}\beta_k$ becomes an independent variable and is moved to the right-hand-side (RHS) of Eq. (3). Note that these independent variables form vector Y (*cf.* Eq. (2)). If none of the indices is known, then the products $P_{ij}\beta_j$ and $P_{ik}\beta_k$ remain in the left-hand-side (LHS) of Eq. (3) as dependent variables, and the corresponding sale prices P_{ij} and P_{ik} (regressors) form the elements of matrix X.

3.2 The Weighted Least-Squares Estimator

When *a-priori* information on the weights of the pairs is available, then a diagonal $n \times n$ weight matrix P can be formed with, for example, elements proportional to the experimental probability. Then, the WLS estimator is given by [e.g., *Baum, et al.*, 2002]:

$$\hat{\beta}_{WLS} = \{X'PX\}^{-1} X'PY \, , \tag{4}$$

where the prime indicates matrix transposition. The covariance matrix of the WLS estimator is given by

$$C_{\hat{\beta}_{WLS}} = \sigma_0^2 \{ X'PX \}^{-1} , \tag{5}$$

where σ_0^2 is the *a-priori* variance factor that is usually taken as unity. However, in the case of the housing market, we do not really know the scale factor of the weight matrix **P** thus, taking σ_0^2 =1 will affect the scale of the covariance matrix given by (5). In such cases we use an estimate of σ_0^2, namely the *a-posteriori* variance factor $\hat{\sigma}_0^2$ that is given by:

$$\hat{\sigma}_0^2 = \frac{\hat{U}'P\hat{U}}{n-m} , \tag{6}$$

where $\hat{U}$ is the estimated vector of residuals (cf. Eq. (2)) given by

$$\hat{U}_{WLS} = Y - X\hat{\beta}_{WLS} \tag{7}$$

and *n-m* are the degrees of freedom of the system. Using $\hat{\sigma}_0^2$ in place of σ_0^2 in (5) we obtain the estimated covariance matrix of the unknown regression coefficients

$$\hat{C}_{\hat{\beta}_{WLS}} = \hat{\sigma}_0^2 \{ X'PX \}^{-1} . \tag{8}$$

The estimated residuals $\hat{U}$ have a covariance matrix given by

$$\hat{C}_{\hat{U}} = \hat{\sigma}_0^2 \{ P^{-1} - X(X'PX)^{-1} X' \} . \tag{9}$$

The WLS residuals $\hat{U}$ and their estimated covariance matrix $\hat{C}_{\hat{U}}$ are useful because:

(a) They can be used to flag any residual outliers that escaped the pre-analysis stage. Residual outliers in this case indicate misfit of a particular pair to the regression model, or equivalently, the pair whose residual is large is not consistent with the other pairs. The residual outlier can be detected by using rigorous statistical procedures (Hypothesis testing – see below).

(b) The Covariance matrix of the residuals reflects both, the initial weighting of the data (at the pre-analysis stage) and the level of consistency of the sales pairs when considered all together in the regression model.

3.3 Residual outlier detection

The WLS estimation provides solution for the unknown parameters (regression coefficients), and the residuals along with their associated covariance matrices. The covariance matrices carry information about the statistical properties of the unknowns and residuals and as such, they are used to statistically evaluate the solution.

The (estimated) covariance matrix of the residuals (*cf.* Eq. (9)) is used in the detection of residual outliers. This approach can be seen as an additional filter to the sales pairs in the post-analysis stage. Residual outliers can be traced back to their respective sales pairs, which can be either further down weighted or eliminated entirely from the set. Whatever the decision, the estimation process must be repeated (iterated) using the new set of sales pairs for the final solution.

The *chi-square goodness-of-fit test* is applied to verify whether the normalized residuals follow the standard normal distribution (H$_0$ hypothesis). This is an identical procedure to the one followed in Section 2.1, Eq. (1). The normalized residual $\tilde{u}_i$ corresponding to the sales pair *i* is given by [e.g. *Vanicek and Krakiwsky*, 1986]

$$\tilde{u}_i = \frac{\hat{u}_i}{\hat{\sigma}_i},\qquad\qquad(10)$$

where $\hat{u}_i$ is the estimated residual of sales pair i and $\hat{\sigma}_i$ is its corresponding estimated standard deviation (from $\hat{C}_{\hat{U}}$).

If the test passes, H$_0$ is true and the normalized residuals follow the standard normal distribution. If the test fails, the normalized residuals most probably follow an alternative but unknown distribution. The latter case does not allow us to test for residual outliers and we must return to the beginning of the screening process (pre-analysis) to make sure that there are no bad pairs in the set or whether the weighting scheme corresponds to the data at hand. In the latter case, we may need to test for the presence of heteroskedasticity; we will come to this case later. If, this second screening of the sales pair does not render the normalized residual normal, we proceed with the estimation process without taking any further action.

If the residuals pass the normality test we proceed to the next step, which is the detection of outliers. This is done by using the normalized residuals (*cf.* Eq. (10)) as the test statistics, and performing *in-context* testing of the residuals. The null Hypothesis H$_0$ states that the estimated residuals $\hat{u}_i \in \hat{U}$ follow the normal distribution (as tested previously). Then the statistic $\tilde{u}_i$ (*cf.* Eq. (10)) follows the *tau* distribution of *n-m* degrees of freedom. The statistical test examines whether each normalized residual falls into a predefined confidence interval (e.g. 95%) defined by the *tau* distribution. If outside the confidence interval, then the sales pair *i* is flagged as an outlier, and can be either eliminated or its weight can be reduced.

3.4 Heteroskedasticity Revisited

As discussed in the previous section, one of the reasons that the residuals may not be normally distributed is the presence of heteroskedasticity. Even if the residuals pass the normality test, which is not very sensitive to the presence of heteroskedasticity, there may be a mild heteroskedasticity present. In both cases it is prudent to test for heteroskedasticity by applying for instance the Breusch-Pagan, or the White test [*Baum, et al.,* 2002]. In fact, the WLS square residuals can be used for testing for heteroskedasticity [*Breusch and Pagan*, 1979; *Wallendin and Ågren*, 2002; *Hansen et al.*, 2006].

4. THE INSTRUMENTAL VARIABLES ESTIMATOR – IV

Because there may be a correlation between the regressors **X** and the vector of residuals **U**, WLS (or simply LS) may give biased and inconsistent estimates for the regression coefficients. To remedy this, the method of instrumental variables (IV) can be used. IV is in reality a two-stage least squares estimator. Suppose that we have n sales pairs, and m unknown regressors out of which ℓ regressors $(\ell < m)$ are *exogenous*. We form the matrix of regressors (design matrix) **X** of dimensions $n \times m$ and the matrix of instrumental variables **Z** of dimensions $n \times \ell$, which is constructed from **X** by collapsing its columns that correspond to the endogenous regressors and then replacing its non-zero exogenous elements with 1 or -1, if the sale price belongs to the "current sale date" and "previous sale date," respectively. Note that the design matrix **X** is the same as in the LS case (*cf.* Eq. (2)).

4.1 The Ordinary IV Estimator

The consistent Instrumental Variables (IV) estimator is given by [e.g., *Baum et al.*, 2002]

$$\hat{\beta}_{IV} = \{X'Z(Z'Z)^{-1}Z'X\}^{-1}X'Z(Z'Z)^{-1}Z'Y ,$$ (11)

where all matrices have been defined previously. The estimated covariance matrix of the ordinary IV process is given by

$$\hat{C}_{\hat{\beta}_{IV}} = \hat{\sigma}_0^2\{X'Z(Z'Z)^{-1}Z'X\}^{-1} = \hat{\sigma}_0^2\{X'P_zX\}^{-1},$$ (12)

and

$$P_z = Z(Z'Z)^{-1}Z' .$$ (13)

The IV residual vector is given by (similar to (7))

$$\hat{U}_{IV} = Y - X\hat{\beta}_{IV} .$$ (14)

The *a-posteriori* variance factor is given by (6) using (14) and P=P_z. We notice that the ordinary IV estimator given by (11) is equivalent to the WLS estimator when its weight matrix $P=P_z$ (*cf.* Eqs. (4) and (13)). When all regressors are considered exogenous, then $\ell = m$ and (11) collapses to:

$$\hat{\beta}_{IV} = \{X'Z\}^{-1}Z'Y ,$$ (15)

and its covariance matrix again by (12).

4.2 The Weighted IV Estimator

We discuss here only the simple case when all regressors are considered exogenous i.e., Eq. (15) is valid. If a weight matrix **P** of dimension $n \times n$ for the sales pairs is available, then (15) and (12) become respectively:

$$\hat{\beta}_{WIV} = \{X'PZ\}^{-1}Z'PY ,$$ (16)

$$\hat{C}_{\hat{\beta}_{WIV}} = \hat{\sigma}_0^2[X'Z(Z'PZ)^{-1}Z'X]^{-1}$$ (17)

When there are endogenous regressors, the weighted IV estimator is equivalent to the Generalized Method of Moments (GMM) and it is discussed below.

5. THE GENERALIZED METHOD OF MOMENTS (GMM)

The Generalized Method of Moments (GMM) was introduced by Hansen [1982] to overcome, among others, the worrisome problem of heteroskedasticity. In fact, the GMM estimator allows a consistent and efficient (minimum variance) estimation of the regression coefficients even in the presence of heteroskedasticity of unknown form [*Hansen*, 1982; *Hansen et al.*, 2006]. The LS and IV estimators can be regarded as special cases of the GMM.

The use of a weight matrix **W** for the exogenous regressors (i.e., of dimensions $\ell \times \ell$) in the minimization of the weighted quadratic norm of the residuals **U** (*cf.* Eq. (2)) gives the GMM as follows [*Hansen at al.*, 2006]:

$$\hat{\boldsymbol{\beta}}_{GMM} = \{\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\}^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{Y} \,. \tag{18}$$

The optimal choice of the weight matrix **W** provides a minimum variance (efficient) GMM estimator. This optimal weight matrix that also eliminates heteroskedasticity is given by (ibid, 2006):

$$\hat{\boldsymbol{W}} = \hat{\boldsymbol{S}}^{-1} = (\boldsymbol{Z}'\hat{\boldsymbol{\Omega}}\boldsymbol{Z})^{-1}, \tag{19}$$

where $\hat{\boldsymbol{\Omega}}$ is a diagonal matrix of dimensions $n \times n$ comprising the inverse estimated square residuals $\hat{u}_i$ coming from a consistent estimator, e.g. from IV. Substituting (19) into (18) we obtain the efficient GMM estimator (EGMM) as follows:

$$\hat{\boldsymbol{\beta}}_{EGMM} = \{\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\hat{\boldsymbol{\Omega}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\}^{-1}\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\hat{\boldsymbol{\Omega}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \,, \tag{20}$$

with an estimated covariance matrix for the regression coefficients given by

$$\hat{\boldsymbol{C}}_{\hat{\boldsymbol{\beta}}_{EGMM}} = \hat{\sigma}_0^2 \{\boldsymbol{X}'\boldsymbol{Z}(\boldsymbol{Z}'\hat{\boldsymbol{\Omega}}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}\}^{-1} \,. \tag{21}$$

The estimated EGMM residuals and their covariance matrix can be written as:

$$\hat{\boldsymbol{U}}_{EGMM} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{EGMM} \,, \tag{22}$$

$$\hat{\boldsymbol{C}}_{\hat{U}} = \hat{\sigma}_0^2 [\hat{\boldsymbol{W}}^{-1} - \boldsymbol{X}(\boldsymbol{X}'\hat{\boldsymbol{W}}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'] \,. \tag{23}$$

The *a-posteriori* variance factor in (21) and (23) is again given by (6) in which $\hat{U}$ is from (22) and $\boldsymbol{P} = \hat{\boldsymbol{W}}$.

Recapitulating, in order to obtain the EGMM estimator the iterative procedure is applied as follows:

(a) Use the IV estimator given by (11) or (15) to calculate the regression coefficients $\hat{\boldsymbol{\beta}}_{IV}$ or $\hat{\boldsymbol{\beta}}_{WIV}$ and then the residual vector $\hat{\boldsymbol{U}}_{IV}$ given by (14).

(b) Use the inverse square residuals from the IV estimator to form $\hat{\boldsymbol{\Omega}}$,

(c) Use (20) to obtain the EGMM estimates along with their covariance matrix given by (21).

(d) If desired, calculate EGMM residuals (*cf.*, Eq. (22)) and then iterate EGMM estimator (go to step (b)).

The EGMM is consistent when arbitrary heteroskedasticity is present because of the presence of matrix $\hat{\Omega}$. Matrix $\hat{\Omega}$ is formed from the square residuals and therefore reflects the random character of heteroskedasticity. This implies that the data sample we have at our disposal must be sufficiently long. Small datasets are not adequate and cause poor performance. If heteroskedasticity is in fact not present then the standard IV process may be more preferable [*Baum et al.*, 2002].

6. STATISTICAL ASSESSMENT OF THE RESULTS

Once the WIV solution is obtained we can test it for reliability. This can be done in two stages:

(a) test the significance of the solution vector $\hat{\boldsymbol{\beta}}_{WIV}$ as one entity,

(b) test the significance of each individual member of vector $\hat{\boldsymbol{\beta}}_{WIV}$.

When testing vector $\hat{\boldsymbol{\beta}}_{WIV}$ as a single entity, the following statistic is formed [*Vanicek and Krakiwsky*, 1986]:

$$y = \frac{(\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{WIV})'\hat{C}_{\hat{\beta}}^{-1}(\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{WIV})}{m}, \qquad (24)$$

where the numerator is the quadratic norm of the difference between the estimated $\hat{\boldsymbol{\beta}}_{WIV}$ and a hypothesized set of values $\bar{\boldsymbol{\beta}}$ (see below), metricized (scaled) by the inverse estimated covariance matrix of the estimated regression coefficients. Statistic y is nothing but the average square distance of the regression coefficients from a hypothesized value and follows the $F(\xi; m, n-m)$ probability density function if $\bar{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{WIV}$ is normally distributed. Once again, a specific significance level (usually □=0.05) is invoked for the statistical test that defines parameter ξ of the *F* distribution. If $y > \xi_{F,1-\alpha}$ then the solution vector $\hat{\boldsymbol{\beta}}_{WIV}$ is statistically different from the hypothesized vector $\bar{\boldsymbol{\beta}}$ and in our case it should not be trusted. Otherwise, the solution is overall statistically very close to $\bar{\boldsymbol{\beta}}$ and should be trusted.

How do we actually select this hypothesized vector $\bar{\boldsymbol{\beta}}$? There are many possible ways to determine the hypothetical values keeping in mind that $\bar{\boldsymbol{\beta}}$ should reflect the smooth variation of the regression coefficients. It can be obtained by low pass filtering of vector $\hat{\boldsymbol{\beta}}_{WIV}$. The low pass filter can simply be a moving average process, or a more sophisticated one with appropriate cut-off frequency not to attenuate seasonal variations that are in fact part of the signal. A three-month moving average or a Parzen weighting function [*Jenkins and Watts*, 1968] can be used to obtain the smooth values. A seasonal cut-off frequency may not be adequate however to effectively filter out spikes in the results. Other approaches are also possible.

For the regression coefficients as individual members of $\hat{\boldsymbol{\beta}}_{WIV}$ it is important to test every one of them *in context* of the whole vector $\hat{\boldsymbol{\beta}}_{WIV}$. In simple terms, this test examines whether the difference of every individual coefficient from its hypothesized value taken from $\bar{\boldsymbol{\beta}}$ is small enough and therefore is to be trusted. The difference $\left|\bar{\beta}_j - \hat{\beta}_j\right|$ for every regression coefficient at epoch *j* is examined vis-à-vis the product of the estimated standard deviation of $\hat{\beta}_j$, namely $\hat{\sigma}_{\hat{\beta}_j}$ times an expansion factor $C_a(m)$ to account for the *in-context* nature of the statistical testing. The expansion factor is calculated from the *chi-square* distribution, and for large degrees of freedom (*n-m*> 40) and large number of unknown coefficients (*m*>40) the expansion factor approaches 3.56 for significance level □=0.05 when the *a-priori* variance factor is unknown. Mathematically this is expressed as follows [e.g., *Vanicek and Krakiwsky*, 1986]:

$$\left|\bar{\beta}_j - \hat{\beta}_j\right| \leq C_a(m)\hat{\sigma}_{\hat{\beta}_j} \qquad (25)$$

If inequality (25) holds, then the estimated regression coefficient $\hat{\beta}_j$ statistically close to $\overline{\beta}_j$ and is to be trusted statistically. Otherwise, $\hat{\beta}_j$ is very different from the hypothesized value and it is not to be trusted. In the latter case, coefficients that are not trustworthy may be eliminated from the set; they often appear as single spikes.

7. POST-BASE INDEX ESTIMATION

In this section we describe the methodology of calculating post-base regression coefficients that are based on the estimates of the historical regression coefficients (pre-base coefficients). Since the pre-base regression coefficients are to be used in the calculation of the post-base coefficients we continue to work in the domain of regression coefficients rather than in the index domain. The transformation of the regression coefficients into indices takes place only in the end, after all statistical assessments and filtering of the coefficients have been completed.

Usually, the historical regression coefficients contain high frequency noise that does not reflect market trends. High frequency noise means small rapid variations from month to month that need to be filtered out. The simplest low pass filter may be the three-month moving average, whether simple or weighted. In the latter case, the weights are the inverses of the variances of the regression coefficients (*cf.* Eq.(17)). The moving average process does not exhibit desirable characteristics in the frequency domain (ripple effects) [e.g. *Jenkins and Watts*, 1968; *Bendat and Piersol*, 1971] and other filtering algorithms such as the Parzen weighting function can be used. One important consideration when designing such a filter is its cut-off frequency. For instance, a seasonal cutoff frequency of 3-4 cycles per year will block monthly or bimonthly variations, whereas it will allow the slower volatile or seasonal variations that most probably reflect market trends to go through.

The previously estimated regression coefficients are subsequently used as constants in the calculation of the post-base regression coefficients. Post-base regression coefficients are calculated on a monthly basis as new sales pair data become available. When sales pairs are collected for the current month, for which the regression coefficient is to be estimated, then the same approach of repeat sales methodology is applied (*cf.* Sections 3, 4, and 5). However, in the case of monthly coefficient estimates, data from rolling three-month periods, namely the current month and the previous two months, are used to calculate the regression coefficient of all three coefficients in a simultaneous fashion. Subsequently, and in order to offset any delays in the flow of data in the registry, the current month's regression coefficient is taken as the average of the three months. The coefficients of the previous two months are not needed for further consideration and they are dropped. This moving average process provides a low pass filtering of the indices. In all the above averaging and/or filtering methods we can consider weighted estimates instead of simple averaging and/or filtering. In a weighted averaging scenario, weights are the inverses of the variances of the regression coefficients. Once the regression coefficients are final then their reciprocal value multiplied by 100 furnishes the final home price indices.

Mathematically, the methodology of estimating the post-base coefficients uses the same regression Eq. (3), to form all required matrices i.e., *X*, *Z*, *Y* and *P*. The characteristic of design matrix *X* is that it now has only three (3) columns, corresponding to the three coefficients to be estimated (current month and the previous two months). If all regressors are exogenous, then matrix *Z* also has three (3) columns. In Eq. (3) the term $P_{ik}\beta_k$ is the product of the previous sale price P_{ik} and the historical regression coefficient β_k. Coefficient β_k is a known "constant" estimated from pre-base calculations, which makes the term equal to $P_{ik}\hat{\beta}_k$ (independent variable), which is moved to the RHS of Eq. (3) to form the elements of matrix *Y*. Once all matrices are formed, they are cast into the regression coefficient estimator to calculate the coefficients of the current month plus the coefficients of the two previous months. It is clear that the monthly regression coefficients so calculated depend upon the regressions coefficients of the previous months only. This is in opposition to the pre-base calculations which provide coefficients that are dependent upon past and future coefficients.

8. CALCULATING THE NATIONAL COMPOSITE INDEX

The national composite index is the weighted average of all six metropolitan areas considered in this study. The weights are based on the aggregate dollar value of dwellings retrieved from the 2006 Statistics Canada Census. Table 1, summarizes the aggregate value of dwellings per metropolitan area and the weighting factor (in percent) that is the normalized aggregate value by the total value of all dwellings.

Table1: Summary of the 2006 Statistics Canada Census on the dwellings in the six metropolitan areas. The last column gives the normalized weight per city.

Metropolitan Area	Total Dwellings	Average Value of Dwellings	Aggregate Value	Weight (%)
Calgary	307,315	$ 381,866	$ 117,353,149,790	10.0
Halifax	99,200	$ 212,942	$ 21,123,846,400	1.8
Montreal	813,405	$ 244,417	$ 198,810,009,885	17.0
Ottawa*	221,690	$ 294,536	$ 65,295,685,840	5.6
Toronto	1,216,100	$ 403,112	$ 490,224,503,200	42.0
Vancouver	529,090	$ 520,937	$ 275,622,557,330	23.6
		Total	**$1,168,429,752,445**	

* Ottawa - Gatineau (Ontario part only)