Quantifying the Reversibility Phenomenon for the Repeat-Sales Index

Author Arnaud Simon

Abstract

The reversibility phenomenon in the repeat-sales index (RSI) is a serious obstacle for derivatives products. This article provides a solution for this problem, using an informational reformulation of the RSI framework. A theoretical formula (simple, easy to interpret, and easy to handle) is presented. For the derivatives, the technique has strong implications for the choice of underlying index and contract settlement. Even if reversibility of the RSI is probably higher compared with the hedonic approach, this index remains a challenger because of the predictability and quantifiability of its revisions.

With the repeat-sales technique, the past seems to change, but actually it is not the past itself that is changing; it is only its knowledge (its representation). This phenomenon is the consequence of the arrival of new data in the estimation set, these new data being relevant to the past. This mechanism of revision is an obstacle to the introduction of derivatives written on the repeat-sales index (RSI), and more generally it is an undesirable characteristic for management of real estate risk. Thus, it would be profitable to have at one's disposal an empirical methodology that could allow anticipating the size of the potential fluctuations, as mentioned in Clapham, Englund, Quigley, and Redfearn (2006): ''If a futures market requires index stability, it would be useful to know how often revision—either period-by-period or cumulative—exceeds some level. Say, for example, that futures markets could tolerate 0.5 percent revision in any one quarter and 2 percent cumulative revision to the initial estimate.'' But at the present time, such a general methodology does not exist in the RSI literature. This paper provides a solution to this problem, using an informational reformulation of the RSI framework. Our methodology is robust in the sense that its conclusions are not conditioned by a single dataset; indeed in Clapham, et al. (2006) one can ask whether the empirical results are still valid for another sample. In this article, the authors also conclude by acknowledging the superiority of the hedonic indexes because their reversibility fluctuations are smaller; however, they do not provide a methodology that would make the anticipations of these variations conceivable. As will be seen, the RSI technique makes these estimations possible. Consequently, even if the reversibility of the RSI is probably higher, this index can still challenge the hedonic approach because of its forecasting feature.

The rest of this article is organized as follows. The second section presents the theoretical and informational reformulation of the RSI. The third section studies the reversibility problem, first with a literature review [the results of Clapp and Giaccotto (1999) are given particular attention], then with informational formalism applied to the revisions issue. Most previous literature views the revisions as a selectivity problem; here we adopt the point of view of the informational content of the data and we show that revisions are also an intrinsic and general feature of the RSI. The fourth section is devoted to empirical implementation. In this section, a simulation algorithm is presented in order to answer Clapham, Englund, Quigley, and Redfearn's (2006) problem, establishing a conditional law for the distributions of the reversibility percentages. The problem is examined unconditionally here to give some indications of the best settlement of derivatives contracts (current or initial indexes, delayed or not). It could be useful for the reader to refer regularly to Appendix A in which the mathematics are exemplified with a basic example.

An Informational Reformulation of the RSI

In Simon (2007), a theoretical reformulation of the classical weighted RPI is developed. From the optimization problem associated to the general least squares procedure, it was demonstrated that a RPI estimation could be realized using the algorithmic decomposition presented in Exhibit 1. The left side is related to informational concepts, whereas the right side is associated with price measures. The final values of the index come from the confrontation of these two parts. This approach does not aim to create a new index, because all the equations presented in Simon (2007) are strictly equivalent to the classical Case-Shiller index. If at first this way of thinking appears more sophisticated, it might help to solve or to study some crucial RSI problems, for instance, quantification of the revisions. The next paragraphs introduce the fundamental concepts that try to reduce the technical side to a minimum level. A basic example can be found in Appendix A that illustrates this formalism.

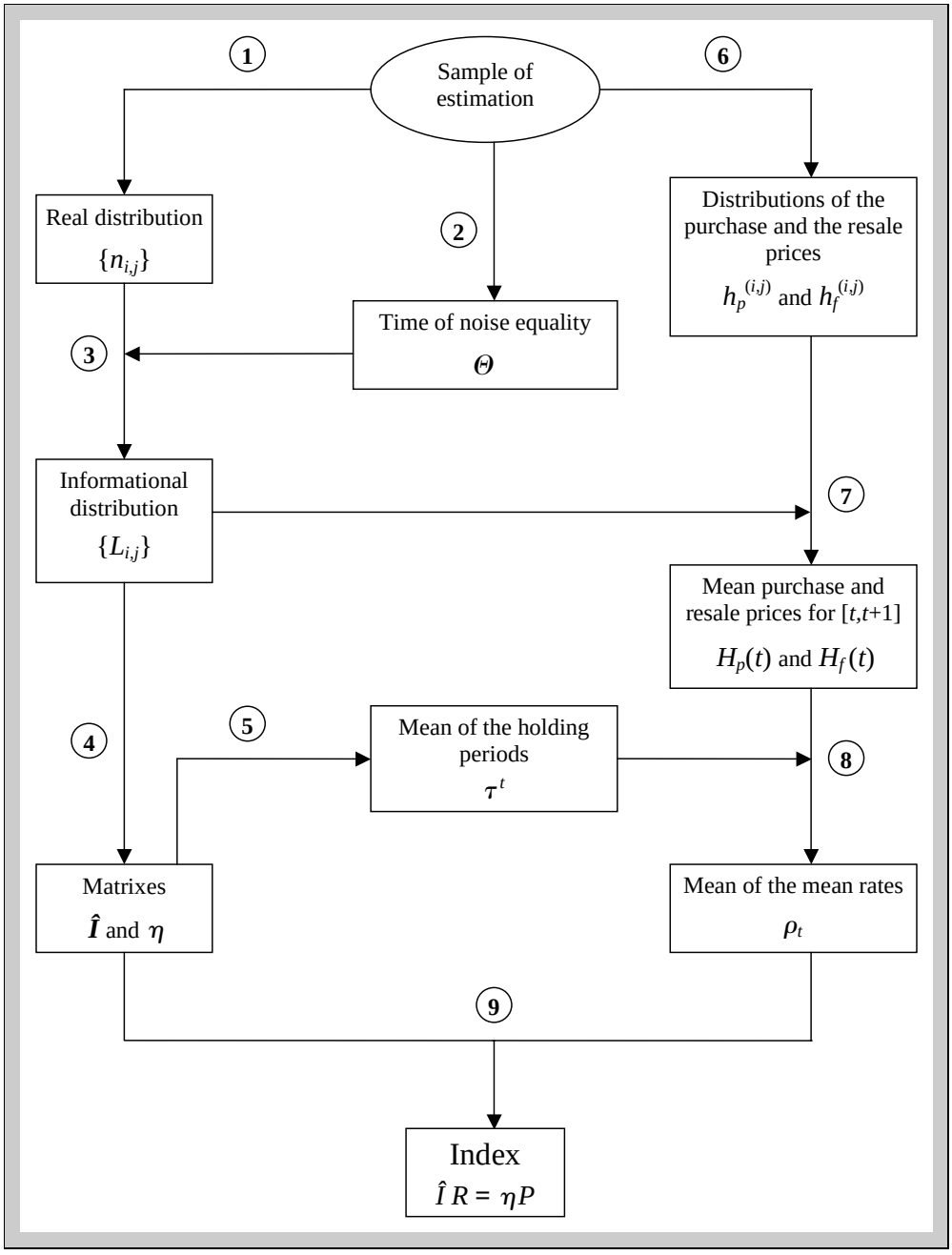
The Usual Estimation of RSI

In the repeat-sales approach, the price of a property k at time t comprises three parts:

$$\mathrm{Ln}(p_{k,t}) = \ln(Index_t) + G_{k,t} + N_{k,t}. \qquad (1)$$

$Index_t$ is the true index value, $G_{k,t}$ is a Gaussian random walk representing the asset's own trend, and $N_{k,t}$ is white noise associated with market imperfections. If $Rate = (rate_0, rate_1, ..., rate_{T-1})'$ the vector of the continuous rates for each elementary time interval $[t,t+1]$, becomes:

Exhibit 1 | Algorithmic Decomposition of the RPI



$$Index_t = \exp(rate_0 + rate_1 + ... + rate_{T-1}) \Leftrightarrow rate_t = \ln(Index_{t+1}/Index_t).$$

And when the formula at the purchase and resale dates is rewritten:

$$\text{Ln}(p_{k,j}/p_{k,i}) = \text{Ln}(Index_j/Index_i) + (G_{k,j} - G_{k,i}) + (N_{k,j} - N_{k,i}). \quad (2)$$

The return rate realized for property k is equal to the index return rate during the same period, plus the random walk and white noise increments. Each repeat-sale gives a relationship of that nature; in matrix form: $\mathbf{Y} = D^*LIndex + \varepsilon$. Here, $\mathbf{Y}$ is the column vector of the log return rates realized in the estimation dataset and $LIndex = (\ln(Index_1), ... , \ln(Index_T))'$. ε is the error term and D is a non singular matrix.[1] In the estimation process, the true values *LIndex* and *Rate* will be replaced with their estimators, respectively denoted $LInd = (\ln(Ind_1),... , \ln(Ind_T))'$ and $R = (r_0, r_1, ..., r_{t-1})'$. The usual estimation of $\mathbf{Y} = D^*LIndex + \varepsilon$ is carried out in three steps because of the heteroscedasticity of ε. Indeed, the specification of the error term leads to the relation $\text{Var}(\varepsilon_k) = 2\sigma_{N^2} + \sigma_{G^2}(j - i)$ in which the values σ_N and σ_G are the volatilities associated with $G_{k,t}$ and $N_{k,t}$, and $j - i$ is the holding period for the k^{th} repeat-sale. The first step consists of running an OLS that produces a series of residuals. These residuals are then regressed on a constant and on the length of the holding periods to estimate σ_N, σ_G, and the variance-covariance matrix[2] of ε, denoted Σ. Finally, the last step is an application of the generalized least squares procedure with estimated matrix Σ.

Central Concepts for the Reformulation

Time of Noise Equality. The variance of residual ε_k measures the quality of the approximation $\text{Ln}(p_{k,j}/p_{k,i}) \approx \text{Ln}(Ind_j/Ind_i)$ for the k^{th} repeat-sale. This quantity $2\sigma_{N^2} + \sigma_{G^2}(j - i)$ can be interpreted as a noise measure for each datum. As a repeat-sale is composed of two transactions (a purchase and a resale), the first noise source $N_{k,t}$ appears twice with $2\sigma_{N^2}$. The contribution of the second source $G_{k,t}$ depends on the time elapsed between these two transactions: $\sigma_{G^2}(j - i)$. Consequently, as time goes by, the above approximation becomes less and less reliable. Next, if factorized by σ_{G^2}: $2\sigma_{N^2} + \sigma_{G^2}(j - i) = \sigma_{G^2}[(2\sigma_{N^2}/\sigma_{G^2}) + (j - i)] = \sigma_{G^2}[\Theta + (j - i)]$. What does $\Theta = 2\sigma_{N^2}/\sigma_G^2$ represent? The first noise source provides a constant intensity ($2\sigma_{N^2}$), whereas the size of the second is time-varying ($\sigma_{G^2}(j - i)$). For a short holding period, the first one is louder than the second, but as the former is constant and the latter is increasing regularly with the length of the holding period, a time exists during which the two sources will reach the

same levels. Then, Gaussian noise $G_{k,t}$ will exceed the white noise. This time Θ, that we will call ''time of noise equality''[3], is the solution of the equation:

$$2\sigma_{N^2} = \sigma_{G^2} * time \Leftrightarrow time = 2\sigma_{N^2}/\sigma_{G^2} = \Theta. \quad (3)$$

The inverse of $\Theta + (j - i)$ can be interpreted as an information measure: if the noise is growing, that is, if the approximation $\text{Ln}(p_{k,j}/p_{k,i}) \approx \text{Ln}(Ind_j/Ind_i)$ is becoming less reliable, the inverse of $\Theta + (j - i)$ is also decreasing. Consequently, $(\Theta + (j - i))^{-1}$ is a direct measure[4] (for a repeat-sale with a purchase at t_i and a resale at t_j) of the quality of the approximation or, equivalently, of the quantity of information delivered. A noise measure is defined up to a constant. For example, $2\sigma_{N^2} + \sigma_{G^2}(j - i)$ could be used, but for ease of interpretation, $\Theta + (j - i)$ is preferred; this quantity is simply a time, unlike $2\sigma_{N^2} + \sigma_{G^2}(j - i)$. With this choice, the impact of holding duration becomes clearly explicit. Moreover, the entire estimation of the index can be realized with this single parameter Θ: when the classical Case-Shiller index is estimated, the size of the two random sources is not needed, just their relative sizes. A second parameter, for instance, σ_{G^2}, becomes useful only when calculating the variance-covariance matrix of the index values,[5] but not for simple estimation of the index.

Time Structure, Real and Informational Distributions, Subset Notations. The time is discretized from 0 to T (the present), and divided into T subintervals. We assume that transactions occur only at these moments, and not between two dates. Each observation gives a time couple $(t_i;t_j)$ with $0 \leq t_i < t_j \leq T$; thus there are $\frac{1}{2}T*(T + 1)$ possibilities for the holding periods. The set of repeat-sales with purchase at t_i and resale at t_j will be denoted by $C(i,j)$. The number of elements in $C(i,j)$ is $n_{i,j}$ and $N = \Sigma_{i<j}n_{ij}$ is the total number of repeat-sales in the dataset. As each element of $C(i,j)$ provides a quantity of information equal to $(\Theta + (j - i))^{-1}$, the total informational contribution of $n_{i,j}$ observations of $C(i,j)$ is:

$$n_{i,j}(\Theta + (j - i))^{-1} = n_{i,j}/(\Theta + (j - i)) = L_{i,j}. \quad (4)$$

Therefore, from real distribution $\{n_{i,j}\}$, there is informational distribution $\{L_{i,j}\}$, which is produced by simply dividing its elements by $\Theta + (j - i)$. The total quantity of information embedded in the dataset is: $I = \Sigma_{i<j}L_{ij}$. An observation is globally relevant for an interval $[t',t]$ if its holding period includes $[t',t]$; that is, if the purchase is at $t_i \leq t'$ and the resale at $t_j \geq t$. This subsample will be denoted $Spl^{[t',t]}$. For an elementary time-interval $[t,t+1]$, the simplified notation $Spl^{[t,t+1]} = Spl^t$ will be used. Exhibit 2 illustrates with a triangular upper table the repeat-sales associated with a given interval and the related quantity of information.

Mean Holding Period, Mean Prices, and Mean Rate for Spl^t. The repeat-sales that bring information on $[t,t+1]$ are the ones that satisfy $t_i \leq t < t + 1 \leq t_j$. The length of their holding periods can differ. Thus, τ^t is denoted as the harmonic

Exhibit 2 | Repeat Sales in $Spl^{[t',t+1]}$ and Quantity of Information Associated

	0	…	t'	…	t	$t+1$		T	Sum
0			$L_{0,t'}$		$L_{0,t}$	$L_{0,t+1}$		$L_{0,T}$	B_0^t
t'					$L_{t',t}$	$L_{t',t+1}$		$L_{t',T}$	$B_{t'}^t$
t						$L_{t,t+1}$		$L_{t,T}$	
T									
					Sum	$S_{t+1}^{t'}$	…	$S_T^{t'}$	$I^{[t',t+1]}$

$B_0^t = L_{0,t+1} + \ldots + L_{0,T} B_{t'}^t = L_{t',t+1} + \ldots + L_{t',T}$ sum of the rows (buy-side).
$S_{t+1}^{t'} = L_{0,t+1} + \ldots + L_{t',t+1}$ $S_T^{t'} = L_{0,T} + \ldots + L_{t',T}$ sum of the columns (sell-side).
$I^{[t',t+1]} = B_0^t + \ldots + B_{t'}^t = S_T^{t'} + \ldots + S_{t+1}^{t'}$.

average of these durations in Spl^t (see Appendix B for more details). Within each repeat-sales class $C(i,j)$, the geometric averages of purchase prices is calculated as: $h_p^{(i,j)} = (\Pi_k, p_{k',i})^{1/n_{i,j}}$, and of resale prices: $h_f^{(i,j)} = (\Pi_k, p_{k',j})^{1/n_{i,j}}$. With classes $C(i,j)$ that correspond to Spl^t, the average $H_p(t)$ of $h_p^{(i,j)}$. $H_p(t)$ can also be seen as the mean of the purchase prices, weighted by their informational contribution $1/(\Theta + (j - i))$, for investors who owned real estate during at least $[t,t+1]$. Similarly, mean resale price is defined as $H_f(t)$. For a given repeat-sales k' in $C(i,j)$, with a purchase price $p_{k',i}$ and a resale price $p_{k',j}$, the mean continuous rate realized on its holding period $j - i$ is $r_{k'}^{(i,j)} = \ln(p_{k',j}/p_{k',i})/(j - i)$. In subset Spl^t, the arithmetic weighted average ρ_t of these mean rates $r_{k'}^{(i,j)}$ is determined; this value is a measure of the mean profitability of the investment. It was demonstrated in Simon (2007) that $\rho_t = \ln[H_f(t)/H_p(t)]/\tau^t$. This relation is actually just the aggregated equivalent of $r_{k'}^{(i,j)} = \ln(p_{k',j}/p_{k',i})/(j - i)$ for subset Spl^t. All these averages $H_f(t)$, $H_p(t)$, and τ^t are weighted by the information. The way prices and duration appear in the formula (through a logarithm for prices and inverse function for durations) explains whether the pattern will be geometric or harmonic. The vector of these mean rates is denoted $\mathbf{P} = (\rho_0, \rho_1, ..., \rho_{T-1})$.

Informational Matrix. A repeat-sale is globally relevant for the interval $[t',t+1]$ if purchase is at t' or before and resale takes place at $t + 1$ or after. The quantity of information globally relevant[6] for $[t',t+1]$ is thus $I^{[t',t+1]} = \Sigma_{i \le t' \le t < j} L_{i,j}$ (cf. Exhibit 2). For an interval $[t,t+1]$, I^t is denoted for $I^{[t,t+1]}$. These quantities of information are calculated for all possible intervals included in $[0,T]$ and then arranged in a symmetric matrix $\hat{I}$.

$$\hat{I} = \begin{pmatrix} I^{[0,1]} & I^{[0,2]} & I^{[0,3]} & I^{[0,T]} \\ I^{[0,2]} & I^{[1,2]} & I^{[1,3]} & I^{[1,T]} \\ I^{[0,3]} & I^{[1,3]} & I^{[2,3]} & I^{[2,T]} \\ \vdots & & & \\ I^{[0,T]} & I^{[1,T]} & I^{[2,T]} & I^{[T-1,T]} \end{pmatrix}$$

A diagonal matrix η also needs to be introduced. Its diagonal values simply correspond to the sums on the lines of $\hat{I}$.

The Index and the Relation $\hat{I}\,R = \eta\,\mathbf{P}$. The estimation of the RSI can now be realized simply by solving the equation: $\hat{I}\mathbf{R} = \eta P \Leftrightarrow R = (\hat{I}^{-1}\eta)P$. The unknown is the vector $\mathbf{R} = (r_0, r_1, ..., r_{T-1})'$ of the monoperiodic growth rates of the index. The other three components of this equation ($\hat{I}$, η, and $\mathbf{P}$) are calculated directly from the dataset. The main advantages of this formalism are its interpretability and its flexibility: matrix $\hat{I}$ gives us the informational structure of the dataset and vector $\mathbf{P}$ indicates the levels of profitability of the investment for people who owned real estate at different dates.

Example and Comments

As this approach to RSI is not usual, the algorithm and the various concepts of Exhibit 1 are illustrated in Appendix A with a small sample. Here, the estimation interval is [0,2] and the dataset is assumed to have only three pairs: a first house ''a'' bought at 1 and sold at 2, a second house ''b'' bought at 0 and sold at 1, and a third house ''c'' bought at 0 and sold at 2. In order to simplify the formulas,[7] $\theta = 0$. Consequently, only $L_{i,j} = n_{i,j}/(j - i)$, but the central point is maintained: goods with a long holding period are less informative. From $\{L_{i,j}\}$, $\hat{I}$ is produced, and summing each line of $\hat{I}$ produces diagonal matrix η. The quantity of information for interval [0,1] is equal to 1,5: the related goods are houses a and c. As the first one is associated with a short holding period, its informational contribution (equal to 1) is greater than for c (equal to 0,5). Now, for interval [0,2], what does it means to be relevant for this interval? According to the definition, the only good satisfying this condition is house c, bought at 0 and sold at 2. Here, one can ask why either house a or house b is overlooked, because both bring information to a portion of interval [0,2]. The answer is that ''relevant for an interval'' means globally relevant for the entire considered interval, and not just for a part of it. However, in doing so, no information is removed, because these partial pieces of information associated with houses a and b will be taken into account in quantities $I^{[0,1]}$ and $I^{[1,2]}$ in matrix $\hat{I}$. In other words, there is simply a gradation of information levels. Thus, as the holding period for c is equal to 2, $I^{[0,2]} = 0{,}5$. After distribution $\{L_{i,j}\}$, mean holding periods $\tau^0 = 1{,}33$ and $\tau^1 = 1{,}33$ are produced for Spl^0 and Spl^1 by dividing the diagonal elements of η by those of $\hat{I}$. Spl^0 and Spl^1 both comprise two observations: the first with a holding period equal to 1 and the second equal to 2. But, as long possessions are less

informative, the mean period is closer to 1. Tables $h_p^{(i,j)}$ and $h_f^{(i,j)}$ are very simple in this example because in each class $C(i,j)$ there is just have one element. In a more complex situation, geometric equally[8] weighted averages within $C(i,j)$ would have to be calculated. For the mean purchase and resale prices in Spl^t: $H_p(0)$ and $H_f(t)$, the relevant pairs weighted by 0,5 or 1 are gotten back according to their informativity. The expressions for ρ_0 and ρ_1 come directly from that, with the same weight structure. Finally, using the index rates solves the system. It is true that this approach to the classical Case-Shiller index could appear to be an unnecessary development if it does not provide strong results. Fortunately, as will be seen below, this decomposition of the index in its building blocks is the key required to solve the reversibility problem and to get a very intuitive formula. Moreover, this way of thinking could be useful for analyzing some other features of the index. The various quantities that appear in the algorithm are intuitive and could be interesting to study in an empirical approach.

The Reversibility Phenomenon: State-of-the-Art and Theoretical Solution

One of the specificities of the RSI is its time dependence on the estimation horizon; a past value Ind_t is not fixed once and for all. When the horizon is extended from T_1 to T_2 ($T_2 > T_1$), the new repeat-sales will bring information not only to interval $[T_1,T_2]$ but also[9] to $[0,T_1]$; unfortunately, there is no reason the new value $Ind_t(T_2)$ should be equal to the old one $Ind_t(T_1)$. This phenomenon of retroactive volatility is called reversibility; the magnitude of variations can be substantial, up to 10% for Clapp and Giaccotto (1999).

Literature Review

The two seminal articles in repeat-sales technique are Bailey, Muth, and Nourse (1963), in a homoscedastic situation, and Case and Shiller (1987) in a heteroscedastic context. Since publication of these two papers, the repeat-sales approach has become one of the most popular indexes because of its quality and flexibility. It is used not only for residential but also for commercial real estate (cf. Gatzlaff and Geltner, 1998). One can also refer to Chau, Wong, Yiu, and Leung (2005) for a recent example of a multisectorial application of RSI and to Baroni, Barthélémy, and Mokrane (2004) for the French context. The reversibility phenomenon was analyzed more specifically by Hoesli, Giaccotto, and Favarger (1997), with a two-period model. This very simplified environment allows for rigorous study of the mathematics of the RSI; Meese and Wallace (1997), for example, chose the same model in their appendices. But when the number of dates increases, the RSI equations quickly become burdensome. Clapham, Englund, Quigley, and Redfearn (2006) tried to compare the sizes of the reversibility phenomenon in the various index methodologies. They concluded that the hedonic index was probably the least affected, but as this article was an empirical one, it

can be asked whether the conclusion was dependent on the sample. Generally, in the literature, a theoretical approach is not the most frequent situation. One example is Wang and Zorn (1997), but other examples are scarce. For the reversibility problem there is an exception, namely the article by Clapp and Giaccotto (1999).

Clapp and Giaccotto's Solution

Clapp and Giaccotto (1999) deal with a Bailey, Muth, and Nourse (1963) context, but their formula can be generalized to a Case and Shiller (1989) model. The first step consists of running, for interval $[0,T_1]$, regression $\mathbf{Y}(T_1) = D(T_1)LInd(T_1) + \varepsilon(T_1)$, in which the unknown is the vector of the logarithms of index: $LInd(T_1)$. In $\mathbf{Y}(T_1)$, there are the log-returns realized for the repeat-sales in the sample. The lines of matrix $D(T_1)$ correspond to the data. In each line, $+1$ indicates the resale date, -1 the purchase date, and the rest is made of zeros.[10] In a second step, the estimation interval is extended to $[0,T_2]$ and the regression becomes $\mathbf{Y}(T_2) = D(T_2)$ $LInd(T_2) + \varepsilon(T_2)$. The vector of the log-returns can be written $\mathbf{Y}(T_2)' = (\mathbf{Y}(T_1)'; \mathbf{Y}(T_2/T_1)')$: the old observations $\mathbf{Y}(T_1)$ completed with the new ones $\mathbf{Y}(T_2/T_1)$. Matrix $D(T_2)$ is a fourblock matrix:

$$D(T_2) = \begin{pmatrix} D(T_1) & 0 \\ D_1(T_2/T_1) & D_2(T_2/T_1) \end{pmatrix}$$

In the upper left corner is the old matrix $D(T_1)$. The lower part of $D(T_2)$ is associated with new repeat-sales. $D_1(T_2/T_1)$ corresponds to the transactions realized before T_1 (only purchases in that case), and $D_2(T_2/T_1)$ corresponds to the transactions realized after T_1 (purchases and resales). There are two kinds of new data: purchases before T_1 and resales after T_1, or purchases and resales after T_1. For the first case, -1 is registered in $D_1(T_2/T_1)$ and $+1$ in $D_2(T_2/T_1)$, whereas both are in $D_2(T_2/T_1)$ for the second. Denote $\Delta(T_2) = (D(T_1)'; D_1(T_2/T_1)')'$ as the left part of the matrix and $F(T_2) = (0'; D_2(T_2/T_1)')'$ as the right part. Vector $\mathbf{LInd}(T_2)$ gives the logarithms of the index values for the second estimation. This can be separated into two pieces; the first gives the levels of the index on $[0,T_1]$ and the second on $[T_1,T_2]$: $\mathbf{LInd}(T_2)' = (\mathbf{LInd}_1(T_2)'; \mathbf{LInd}_2(T_2)')$. Clapp and Giaccotto's formula establishes the link between vectors $\mathbf{LInd}(T_1)$ and $\mathbf{LInd}_1(T_2)$, which both give the index values on the interval $[0,T_1]$, but using only the information embedded in $\mathbf{Y}(T_1)$ for $\mathbf{LInd}(T_1)$, while $\mathbf{LInd}_1(T_2)$ uses completed dataset $\mathbf{Y}(T_2)$. This formula requires an auxiliary regression $\mathbf{Y}(T_2/T_1) = D_1(T_2/T_1)\text{AUX} + \varepsilon'$. But, even if it is similar to the previous regressions, ''AUX is not an index of any kind. It's just the vector of coefficients in the artificial regression of $\mathbf{Y}(T_2/T_1)$ on $D_1(T_2/T_1)$'' (Clapp and Giaccotto, 1999). A matrix Ω must be introduced that is quite hard to interpret: $\Omega = [D(T_1)'D(T_1) + D_1(T_2/T_1)'D_1(T_2/T_1)]^{-1}\ D(T_1)'\ D(T_1)$. With all these elements, the reversibility formula is:

$$\mathbf{LInd}_1(T_2) = \Omega\mathbf{LInd}(T_1) + (I - \Omega)\mathbf{AUX} + [\Delta(T_2)'\Delta(T_2)]^{-1}\Delta(T_2)'F(T_2)\mathbf{LInd}_2(T_2). \quad (5)$$

Informational Approach to Reversibility

Notations. This section deals with the reversibility phenomenon using the reformulation presented in the first section: the formulas will be simple and intuitive. Assume here the initial time horizon T_1 is extended to T_2, $T_2 > T_1$. The main idea of the reversibility formulas below consists of working with three repeat-sales samples. The first is the old sample, denoted by its horizon T_1. The second consists of new repeat-sales used in the re-estimation at T_2 but not used in the first estimation because their resale occurred after T_1; this sample is denoted as $T_2\backslash T_1$. The last sample, denoted T_2, is the entire sample of available repeat-sales at T_2. Thus: T_1 **U** $T_2\backslash T_1 = T_2$. With T_1, the index and its building blocks can be determined for the time interval $[0,T_1]$. With $T_2\backslash T_1$ and T_2, the same can be done for $[0,T_2]$. The notations will be the same as those presented in the first section; however, the considered sample will be added as a parameter. For example, $H_p(t)$ will be denoted $H_p(t;T_1)$, $H_p(t;T_2\backslash T_1)$, or $H_p(t;T_2)$ according to the associated dataset. The result is illustrated in Appendix A, adding two new transactions to the small sample. Assume that house d is bought at 2 and sold at 3, and house e is bought at 0 and sold at 3. Index $T_2\backslash T_1$ is first estimated with these two new repeat-sales and then index T_2 with the five observations available on [0,3]. The reversibility formula is just a linear dependence between the vectors $\mathbf{R}(T_1)$, $\mathbf{R}(T_2\backslash T_1)$ and $\mathbf{R}(T_2)$. The coefficients correspond to the associated quantities of information (cf. formulas *i* and *v* in the proposition below).

Reversibility formulas. The main results are summed up in the following proposition[11];

Proposition:

i. $I^t(T_2) = I^t(T_1) + I^t(T_2\backslash T_1)$

ii. $[H_p(t,T_2)]^{I^t(T_2)} = [H_p(t,T_1)]^{I^t(T_1)}[H_p(t,T_2\backslash T_1)]^{I^t(T_2\backslash T_1)}$
$[H_f(t,T_2)]^{I^t(T_2)} = [H_f(t,T_1)]^{I^t(T_1)}\,[H_f(t,\,T_2\backslash T_1)]^{I^t(T_2\backslash T_1)}$

iii. $\tau^t(T_2)$ is the harmonic weighted average of $\tau^t(T_1)$ and $\tau^t(T_2\backslash T_1)$

iv. $\rho_t(T_2) = [I^t(T_1)/I^t(T_2)][\tau^t(T_1)/\tau^t(T_2)]\rho_t(T_1) + [I^t(T_2\backslash T_1)/I^t(T_2)][\tau^t(T_2\backslash T_1)/\tau^t(T_2)]\rho_t(T_2\backslash T_1)$

v. $\hat{I}(T_2) = \hat{I}(T_1) + \hat{I}(T_2\backslash T_1)$.

vi. $\eta(T_2)\,P(T_2) = \eta(T_1)P(T_1) + \eta(T_2\backslash T_1)P(T_2\backslash T_1)$

vii. $\hat{I}(T_2)R(T_2) = \hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$

The first point (i) means that the quantity of relevant information I^t for a time interval $[t,t+1]$ for dataset T_2 is equal to the quantity of information provided by T_1 plus the quantity provided by $T_2\backslash T_1$. According to ii, the mean purchase and resale prices for T_2 are simply the weighted averages of the same quantities for T_1 and $T_2\backslash T_1$; the weights correspond to the informational contributions from the old sample and from new data. $\tau^t(T_2)$ is the weighted average of $\tau^t(T_1)$ and $\tau^t(T_2\backslash T_1)$ (cf. Appendix B for the weights). For $t < T_1$, iv means that $\rho_t(T_2)$ is the average of $\rho_t(T_1)$ and $\rho_t(T_2\backslash T_1)$. In this formula, quantity $I^t(T_1)/I^t(T_2)$ represents the percentage of the total information $I^t(T_2)$ already known when the horizon is T_1. $I^t(T_2\backslash T_1)/I^t(T_2)$ is the percentage of the information revealed between T_1 and T_2. The ratios $\tau^t(T_1)/\tau^t(T_2)$ and $\tau^t(T_2\backslash T_1)/\tau^t(T_2)$ measure the lengths of the holding periods for the old data and the new, relative to the lengths of the whole sample.[12] The scalar formula *i* can be generalized in a matrix[13] formula *v*. The matrix approach allows rewriting[14] formula iv under the synthetic form vi. Finally, relation vii gives the reversibility result for the index.

Comments. The logic of the reversibility phenomenon can be summarized as follows. First, estimate the RSI with old data on $[0,T_1]$; this gives an informational matrix $\hat{I}(T_1)$ and a vector $\mathbf{R}(T_1)$. Then, only with new data $T_2\backslash T_1$, the index is estimated on $[0,T_2]$; it gives $\hat{I}(T_2\backslash T_1)$ and $\mathbf{R}(T_2\backslash T_1)$. Finally, using the entire dataset (old data + new data), the RSI is calculated on $[0,T_2]$, with $\hat{I}(T_2)$ and $\mathbf{R}(T_2)$. What is expressed in the reversibility formula is simply that quantity $\hat{I}\mathbf{R}$ is additive if the horizon is extended from T_1 to T_2. As Clapp and Giaccotto (1999) have already proposed a formula to deal with this problem, how should these two different approaches be scrutinized? At the theoretical level, they are of course strictly equivalent because they are measuring the same phenomenon. But from a practical point of view, things are different. Clapp and Giaccotto's formula is rather complex and its financial interpretation is not obvious. For instance, what does matrix Ω represent? Moreover, as is pointed out in their article from 1999, the auxiliary regression is just an abstract estimation that does not correspond to an index of any kind. Taking a look at *formula vi*, it is simple, easy to handle, and easy to interpret. Just the informational matrixes and vectors of the monoperiodic growth rates of the indexes are needed; and these two notions are strongly intuitive. What is more, the equivalent of auxiliary regression AUX, namely $\mathbf{R}(T_2\backslash T_1)$, can be interpreted as the RSI for the interval $[0,T_2]$ if the estimation is run only with the new dataset, $T_2\backslash T_1$. Intuitively, this relation could be interpreted as a kind of ''equation of energy preservation'' for the datasets. Indeed, if product $\hat{I}\mathbf{R}$ measures the ''quantity of energy'' embedded in a dataset, the reversibility formula simply asserts that:

Energy of the whole dataset = Energy of the old data
+ Energy of the new data.

This idea of energy delivered by a sample also allows interpreting the relation $\hat{I}\mathbf{R} = \eta P$. The left side can be understood as the energy of the informational system of the index values, whereas the right side can be analyzed as the energy provided by the gross (real) dataset system. Here, also, there is a kind of equation of preservation:

Energy of the informational system

= Energy provided by the real system.

Predicting the Magnitude of the Revisions

A methodology is presented next that allows estimation of the magnitude of potential revisions that will be applied to settlement of the derivatives contracts.

The Exponential Benchmark

In order to simulate the behavior of repeat-sales between T_1 and T_2, a very simple model is introduced, based on an exponential distribution of resale decisions. More precisely, assume that:

1. The quantities of goods traded on the market at each date are constant and denoted K.
2. Purchase and resale decisions are independent between individuals.
3. The length of the holding period follows an exponential survival curve, with a parameter $\lambda > 0$ (the same for all owners).

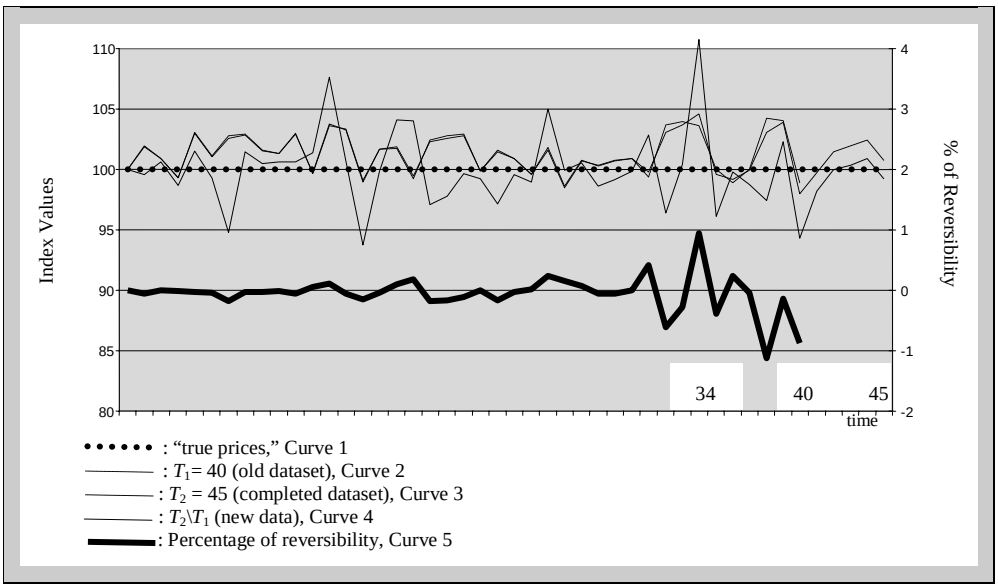
This last hypothesis means that, conditional to purchase at $t = 0$, the probability of not having sold the house at time t is equal to $e^{-\lambda t}$. This choice is unrealistic because it implies that the probability of selling the house in the next year is not influenced by the length of the holding period.[15] If the hazard rate is introduced,[16] which measures the instantaneous probability of a resale: $\lambda(t) = (1/\Delta t)^*$ Prob(resale $> t + \Delta t$|resale $\geq t$), it is well known that the choice of an exponential distribution is equivalent to the choice of a constant hazard rate. In the real world things are of course different. For the standard owner, the hazard rate is first low (quick resales are scarce). The second time, it increases progressively up to a stationary level, potentially modified by the economic context (residential time). Then, as time goes by, the possibility of moving because of retirement, or even death of the householder, would bring the hazard rate to a higher level (aging). However, even if the assumption is not entirely realistic, it generates a simple model. The aim of this benchmark is not to exactly describe reality; it is just trying to model a basic behavior. For an interval $[0,T]$, the benchmark dataset is

fully determined if the parameters K and $\alpha = e^{-\lambda}$ are known. This is established in Simon (2008), where the number of repeat-sales in an exponential sample is[17] $N = KT(1 - \pi)$ and the total quantity of information embedded in this dataset is[18] $I = K'[(T + \Theta + 1)\, u_T - T\pi]$. These two expressions will be useful in the calibration step.

An Example

For practical reasons, randomly generated artificial samples are used.[19] However, the methodology can be applied directly to real datasets, with no difficulties. Exhibit 3 presents the results of an estimation with $T_1 = 40$ and $T_2 = 45$. Curve 2 gives the index values on [0,40], for the old dataset, Curve 4 gives the index values on [0,45], using only the new data T_1/T_2, and Curve 3 is for the completed sample. Curve 5 shows the percentage of reversibility $(Ind_t(45)/Ind_t(40) - 1)$ for $t = 0,...,40$. The sample of the new data $T_2\backslash T_1$ is smaller than the two others; thus its curve logically presents a higher volatility. For the majority of the dates, the difference between the old index and the completed one is negligible; Curve 5 is close to zero. It is only in the last quarter of the interval [0,40] that the two curves can diverge (the spread can reach 1% with the simulated data). The direction of the variation is given by the new data. For instance, at $t = 34$ the index $T_2\backslash T_1$ is at 110, whereas the old index is around 104. Consequently, Curve 4 brings the old value (104) to a higher level (105). As can be seen, the reversibility phenomenon has a temporal pattern: it appears essentially for the nearest dates.

Exhibit 3 | An Example of Reversibility

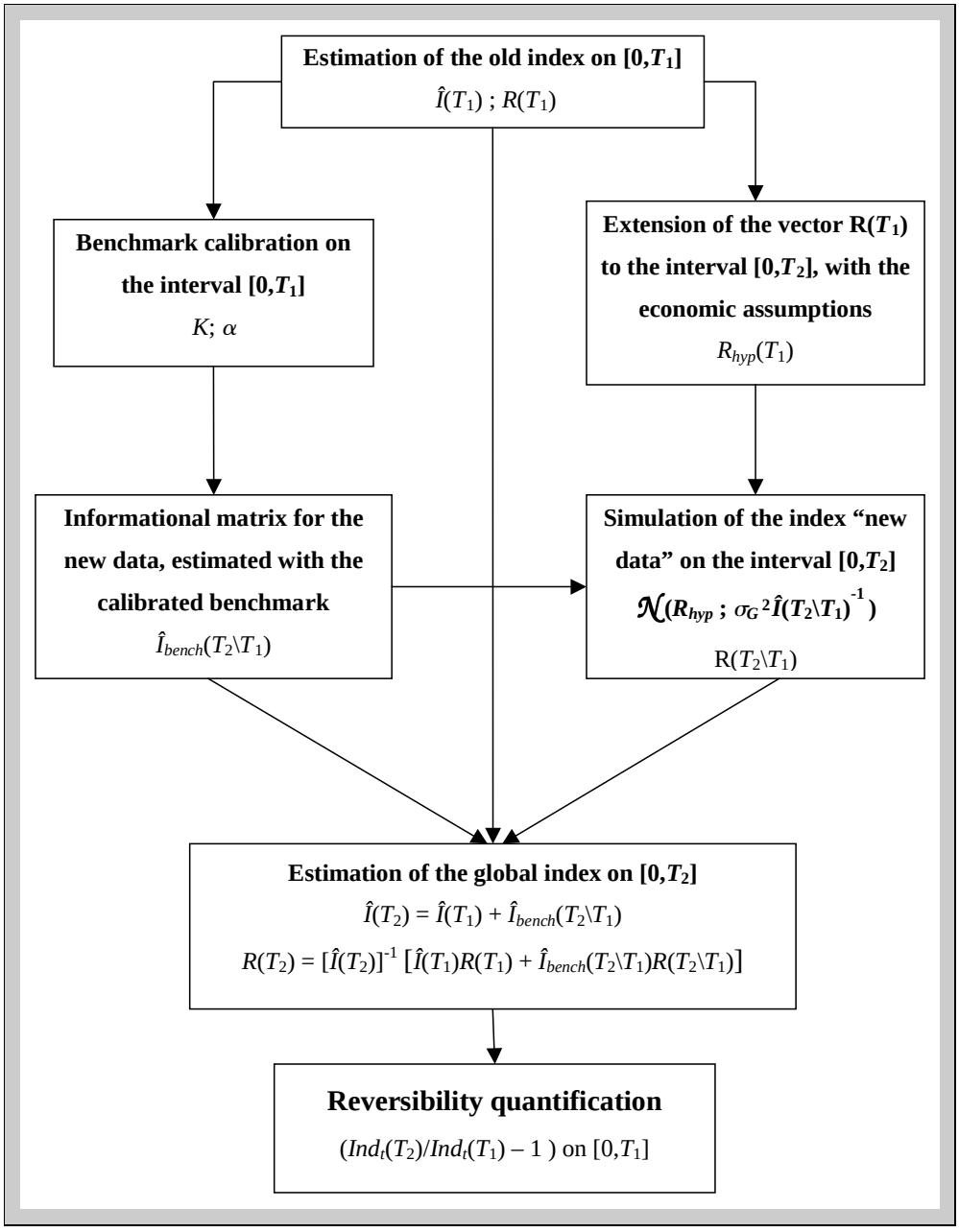


This phenomenon is also documented by Clapham, Englund, Quigley, and Redfearn (2006) and Deng and Quigley (2008). Unfortunately, from an investment point of view, these recent past values are in general the most important values. Therefore, it is crucial to elaborate a methodology able to indicate the level of reliability of the index values.

The Simulation Process

The Monte Carlo technique is well adapted to our problem. The simulation algorithm is presented in Exhibit 4 (in several points some variations in the assumptions could be introduced according to the needs of the modeling; a basic version is presented here). From a repeat-sales sample on $[0,T_1]$, the associated index is calculated with $R(T_1)$ and $\hat{I}(T_1)$. These two quantities are fixed during the entire process. The present is time T_1, and the two estimations at T_1 and T_2 must satisfy the relation $\hat{I}(T_2)R(T_2) = \hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$. Variations of the index are attempted when the estimation will be renewed at T_2; in other words, comparing $R(T_1)$ with $R(T_2)$. Unfortunately, at point T_1 the quantities $\hat{I}(T_2)$, $\hat{I}(T_2\backslash T_1)$ and $R(T_2\backslash T_1)$ are unknown. The idea of the algorithm is to forecast these three measures to solve the equation with unknown $R(T_2)$, and then to compare $R(T_1)$ and $R(T_2)$. The first step consists of calibrating the exponential benchmark with the old data on $[0,T_1]$. More precisely, the search is for the values of parameters K_0 (constant flow on the market) and α_0 (resale speed) such that the total number of repeat-sales N and the total quantity of information I is equal between the real dataset and the benchmark sample.[20] Mathematically speaking, parameter α_0 can be estimated by working with quantity I/N, which does not depend on K (numerical resolution). When α_0 is known, K_0 can be calculated with $N = KT(1 - \pi)$. Once the benchmark is calibrated, the arrival of the information on the interval $[T_1,T_2]$ will occur according to the same rhythm as previously. This gives an approximation:[21] $\hat{I}_{bench}(T_2\backslash T_1)$ for matrix $\hat{I}(T_2\backslash T_1)$. At the same time, there is also an approximation for matrix $\hat{I}(T_2)$, adding $\hat{I}(T_1)$ and $\hat{I}_{bench}(T_2\backslash T_1)$ (cf. Proposition v). After the left side of Exhibit 4 devoted to informational matrixes, the focus turns to the right side, dedicated to growth rate vectors, and try to infer vector $\mathbf{R}(T_2\backslash T_1)$. $\mathbf{R}(T_1)$ gives the index evolution on interval $[0,T_1]$. For the rest of interval $[T_1,T_2]$, it is completed in a T_2-vector $\mathbf{R}_{\mathbf{hyp}} = (\mathbf{R}(T_1), \mathbf{R}_{\mathbf{hyp}}(T_1;T_2))$. $\mathbf{R}_{\mathbf{hyp}}(T_1;T_2)$ is a scenario for the future of real estate prices. Simon (2007) established that the vector $\mathbf{R}(T_2\backslash T_1)$ is Gaussian. It is centered on the growth rates of the theoretical index values,[22] and its variance-covariance matrix[23] is $\sigma_{G^2}\,\hat{I}(T_2\backslash T_1)^{-1}$. Because of its unobservability at T_1, $\mathbf{R}(T_2\backslash T_1)$ has to be generated randomly as a Gaussian vector $\mathfrak{N}(R_{hyp}\,;\,\sigma_{G^2}\,\hat{I}(T_2\backslash T_1)^{-1})$. The theoretical expectation (the true rates values) is replaced here with the best estimator that we have on $[0,T_1]$ at T_1, that is, $\mathbf{R}(T_1)$, and we complete it with the economic assumptions on $[T_2\backslash T_1]$ through vector $\mathbf{R}_{\mathbf{hyp}}(T_1;T_2)$. For the second parameter, the benchmark matrix is used as an approximation. At this stage of the process, $\hat{I}(T_1)$, $\hat{I}(T_2\backslash T_1)$, $\hat{I}(T_2)$, $R(T_1)$, and $R(T_2\backslash T_1)$. The final step consists simply of calculating vector $\mathbf{R}(T_2)$ with the

Exhibit 4 | Algorithm for the Quantification of the Reversibility Phenomenon



equation $\hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1) = \hat{I}(T_2)R(T_2)$. Once $R(T_2)$ is known, index values $Int_t(T_2)$ can be calculated on the interval $[0,T_1]$, and the size of the reversibility phenomenon can be measured for each simulation of $R(T_2\backslash T_1)$. Repeatedly running this procedure, produces an empirical distribution of the spreads.

The Theoretical Law of Reversibility in a Simplified Context

The Monte Carlo technique is interesting for the very complex situations in which closed formulas will be impossible to establish. For simple situations it is useful to deepen the mathematical analysis to get an idea of the dynamic of the revisions and to potentially model this phenomenon with a stochastic process. The process begins with an initial repeat-sales sample ω_0, associated with interval $[0,T_1]$. In the model, the benchmark is calibrated on this dataset and, using the corresponding parameters gives an estimation for matrix $\hat{I}(T_2\backslash T_1)$. The quantities $\hat{I}(T_1)$, $R(T_1)$, $\hat{I}_{bench}(T_2\backslash T_1)$, and $\hat{I}(T_2)$ are fixed and there is one random source, $R(T_2\backslash T_1)$. It is assumed that vector $\mathbf{R}_{\mathbf{hyp}}(T_1;T_2)$ is constant. Under these assumptions and with the formula $R(T_2) = [\hat{I}(T_2)]^{-1}\ [\hat{I}(T_1)R(T_1) + \hat{I}_{bench}(T_2\backslash T_1)R(T_2\backslash T_1)]$, it can be demonstrated that vector $R(T_2)$ is Gaussian:

$$E[R(T_2)] = \boldsymbol{R}(T_1) + [\hat{I}(T_2)^{-1}\hat{I}_{bench}(T_2/T_1)]\boldsymbol{R}_{\boldsymbol{hyp}}(T_1;T_2)$$
$$V[R(T_2)] = \sigma_{G^2}[\hat{I}(T_2)^{-1}\hat{I}_{bench}(T_2/T_1)][\hat{I}(T_2)]^{-1} \quad (6)$$

Matrix $\hat{I}_{bench}(T_2\backslash T_1)$ represents new information, $\hat{I}(T_2)$, the total information. Consequently, product $\hat{I}(T_2)^{-1}\ \hat{I}_{bench}(T_2\backslash T_1)$, which appears in these two formulas, can be interpreted as the (vectorial) proportion of the new information contained in the total. The first formula simply asserts that the expectation of $R(T_2)$ is equal to the old and constant vector $\mathbf{R}(T_1)$, plus a quantity that represents the influence of the economic hypotheses $R_{hyp}(T_1;T_2)$ on $[T_2\backslash T_1]$. This influence of $R_{hyp}(T_2\backslash T_1)$ is weighted by $[\hat{I}(T_2)]^{-1}\ \hat{I}_{bench}(T_2\backslash T_1)$; a relative measure of the informational weight of the new data. Regarding the variance formula, it has to be compared with the formula[24] $V[R(T_2)] = \sigma_{G^2}[\hat{I}(T_2)]^{-1}$ that would have to applied if the estimation for the index on $[0,T_2]$ was run directly with the entire dataset, without doing a halfway estimation at T_1. In a reversibility situation, part of the total sample is already known; therefore the resulting index is less volatile. What is expressed with the second formula is simply that the attenuation coefficient for the volatility is nothing other than $[\hat{I}(T_2)]^{-1}\ \hat{I}_{bench}(T_2\backslash T_1)$, once more. Now, if on $[T_1,T_2]$ real estate growth is null, in other words, $R_{hyp}(T_1;T_2) = 0$,[25] the following result can be demonstrated:

Reversibility Law

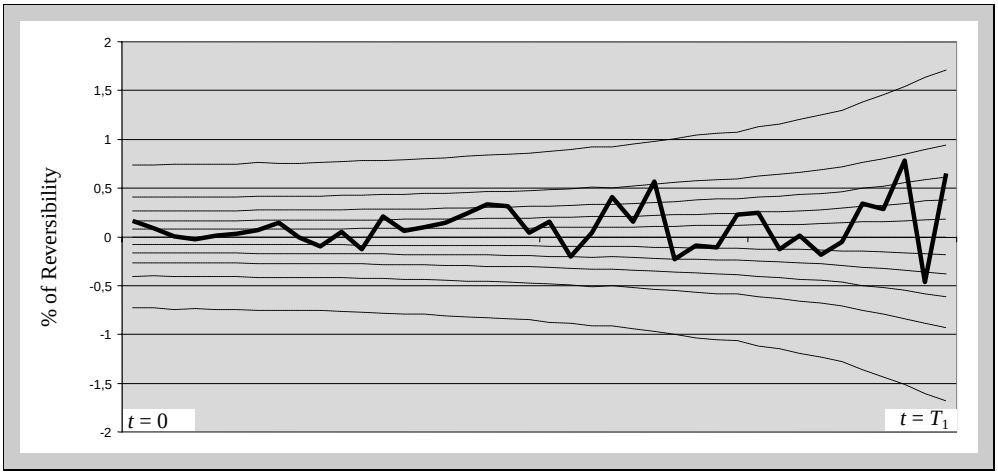
For $t = 1,...,T_1$ the ratio $Ind_t(T_2)/Ind_t(T_1)$ is log-normally distributed: $\mathfrak{LN}(0; v(t))$ $v(t)$ is the t^{th} diagonal element of the matrix[26] $\sigma_{G^2} A(T_2)[[\hat{I}(T_2)]^{-1}\hat{I}_{bench}(T_2\backslash T_1)][\hat{I}(T_2)]^{-1}[A(T_2)].'$

The reversibility percentage[27] for date t is a random variable that can be written as $100(Y_t - 1)$, with $Y_t \sim \mathfrak{LN}(0; v(t))$. Exhibit 5 represents the theoretical deciles, anticipated at T_1, using sample ω_0 on $[0,T_1]$. The bold curve gives the observed reversibility for this specific sample when the horizon is extended from T_1 to T_2. The theoretical curves are a good approximation of the empirical one. The magnitude of the potential revisions is small and approximately constant for the left side of the interval. But on the right side, things are different. Near T_1 the fluctuations are potentially more pronounced, as evidenced by the divergence of the theoretical curves in Exhibit 5. With the methodology developed in this paragraph, it now becomes possible to anticipate and to quantify reversibility effects in a reliable way.

Comments

In the above algorithm, randomness appears with simulation of the Gaussian vector[28] $R(T_2\backslash T_1)$. However, to deepen the simulation, two additional random sources could be introduced: for vector $\mathbf{R_{hyp}}(T_1;T_2)$ and for the couple (K, α).

Exhibit 5 | Deciles for Reversibility Percentages ($t = 1,..., T_1 = 40$)



The bold curve gives the observed empirical reversibility (at T_2) and the dotted ones give the theoretical deciles deduced from the reversibility law, just using the information known at T_1. The two extreme curves are the percentiles at 1% and 99%; the other curves give the deciles from 10% up to 90%.

Indeed, in order to estimate the expectation of vector $\mathbf{R}(T_2 \backslash T_1)$, vector $\mathbf{R}(T_1)$ could be completed with $\mathbf{R}_{\mathbf{hyp}}(T_1;T_2)$. This vector corresponds to a specific scenario for the evolution of real estate prices on $[T_1,T_2]$. But, as the future is uncertain, it could be reasonable to let these last coordinates fluctuate randomly, rather than restricting them to a single path. The second generalization concerns the couple (K, α). The first variable represents a constant level of liquidity in the market and the second the resale speed. With the calibration step on interval $[0,T_1]$, a mean couple (K_0,α_0) is found. However, for interval $[T_2 \backslash T_1]$, market conditions might be slightly different. To take this possibility into account, parameter K could be randomly chosen in an interval $[K_0 - \varepsilon; K_0 + \varepsilon]$ and α_0 in $[\alpha_0 - \varepsilon'; \alpha_0 + \varepsilon']$, for each Monte-Carlo simulation. This methodology could be extended to consider that the rhythm of transactions depends on the economic context and especially on future real estate prices. Here, a proportional hazard model would be calculated on $[0,T_1]$, like the one developed by Cheung, Yau, and Hui (2004), for instance. Then, according to the scenario simulated on $[T_1,T_2]$, the rhythm of resales could be deduced. The last point to examine is the direction of revisions. From Exhibit 5, the probability of having a positive revision is equal to the probability of a negative one. But, from Clapp and Giaccotto (1999) it is known that most of the time there is a downward revision, as is also documented by Clapham, Englund, Quigley, and Redfearn (2006). Actually, it seems that there are two sides to the reversibility problem. The first corresponds to the decrease in variance of the estimators when more data become available, as shown in Exhibit 5; thanks to the methodology developed in this paper, the magnitude of these revisions is readily apparent. The second aspect is a selectivity problem: the density of ''flips'' is higher near the right edge of interval $[0,T_1]$. When the estimation horizon is extended to T_2, the relative weights between the flips and the goods with a longer holding period come back to a more normal level for the dates near T_1. It would not be a problem if the financial features (trend and volatility) were the same across all types of goods, however, there is some evidence to suggest that this is not true [Clapham et al. (2006): ''This suggests *systematic differences in the relative appreciation* of those early entrants to the sample compared to those that arrive later'']. The simplest solution to avoid the revision problem would be to exclude all ''flips'' from the estimation sample; however, with this choice, some interesting information about market conditions would be removed. For an estimate of an index of the whole market, is it possible to deal with this problem using the formalism previously introduced? The answer is probably positive. For simplicity, assume that the initial sample can be divided into two subsamples: flips (holding period $<$ two years) and non-flips. For each subsample, for each elementary time interval $[t,t+1]$, the quantities of relevant information are calculated: $I_{flips}^{[t,t+1]}$ and $I_{non\text{-}flips}^{[t,t+1]}$. Near the right edge of the interval, the ratio $I_{flips}^{[t,t+1]}/I_{non\text{-}flips}^{[t,t+1]}$ increases automatically. The idea of a correction would be to remove just a portion of the flips from the global dataset in order to recover the same level for the ratio $I_{flips}^{[t,t+1]}/I_{non\text{-}flips}^{[t,t+1]}$ as the one observed in the middle of the interval. After this initial correction of the selectivity problem, the index is estimated on $[0,T_1]$. Then, the methodology developed above is applied to control for the revisions that do not depend on flips.

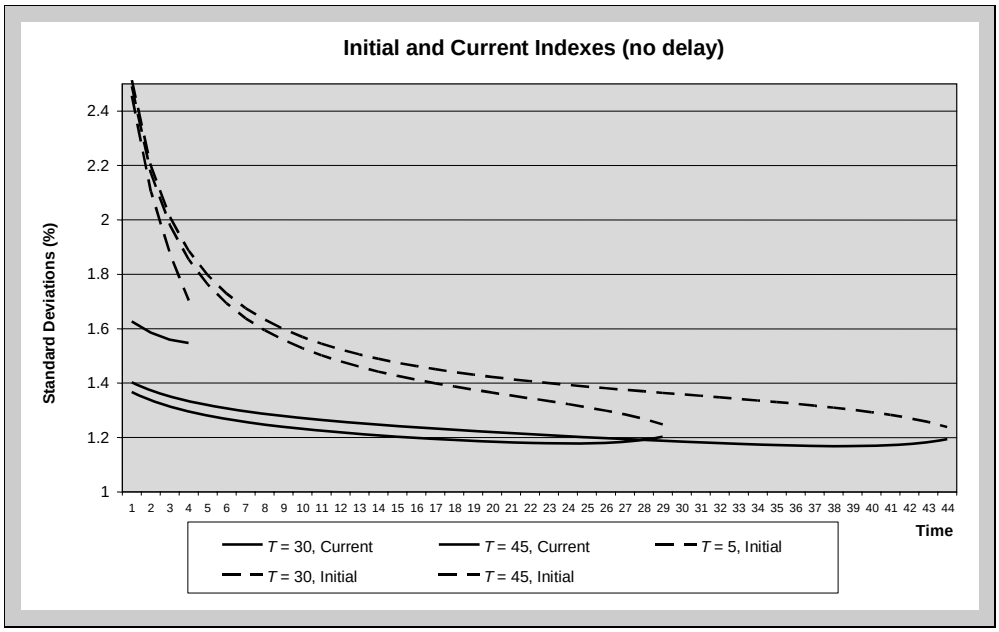
Derivatives and Reversibility

In the last section of Clapham, Englund, Quigley, and Redfearn (2006), several possibilities are investigated for the contract settlement of the derivatives. More precisely, the issue is when should cash flows be measured to give the best estimation of the true economic return realized between t and T, that is, $\text{Ln}(Index_T/Index_t)$? Four choices are examined on a specific dataset (from the least to the most effective):

1. $\text{Ln}(Ind_T(T)) - \text{Ln}(Ind_t(t))$: initial indexes
2. $\text{Ln}(Ind_T(T)) - \text{Ln}(Ind_t(T))$: current indexes
3. $\text{Ln}(Ind_T(T+d)) - \text{Ln}(Ind_t(t+d))$: initial indexes delayed
4. $\text{Ln}(Ind_T(T+d)) - \text{Ln}(Ind_t(T+d))$: current indexes delayed

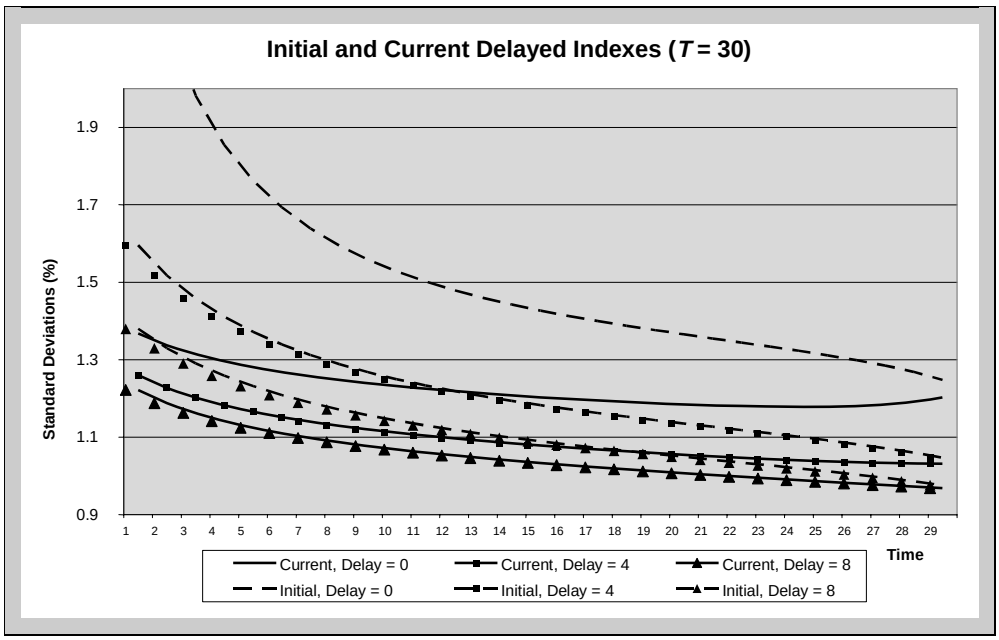
These four random variables are all centered on the true value: $\text{Ln}(Index_T) - \text{Ln}(Index_t)$. The levels of reliability for these approximations are given by their respective standard deviations. Formulas are given in Appendix C for variances. The formulas differ from those established in for the reversibility law. What is the difference? Actually, the reversibility law gives the conditional behavior for the revisions: at T_1, the value $Ind_t(T_1)$ is known, and an idea of the variation of this single index value is wanted when the index is re-estimated at T_2. Here, in this paragraph, things are different: there is an examination of the absolute behavior of the revisions. In others words, at 0 and what is the forecast of the error on the measure of the return, without knowing the index values at T_1. Can there be a ranking of Clapham, Englund, Quigley, and Redfearn with this theoretical approach? If there is to be an examination of this problem from a general point of view, a ''neutral'' sample must be employed. An exponential sample is selected, assuming that $K = 200$, $\lambda = 0.1$, $\sigma_{G^2} = 0.001$, and $\theta = 10$ (conservative choice). Exhibit 6a gives the results for the no-delay situations, with three choices for the resale date T (5, 30, and 45) and a purchase date t that varies between 0 and $T - 1$. It appears clearly that the current choice is always better than the initial one. This is especially true for transactions with a purchase near the left edge of the interval. For repeat-sales with a purchase date in the middle or at the end of the estimation interval, a difference also exists, but in a smaller proportion. With Exhibit 6b, there is an examination of the delay effect for a transaction realized between t and $T = 30$. The three kinds of curves give the results for the no-delay situation, for a delay equal to 4 (four quarters), and for a delay equal to 8 (two years). When the measure of the return is realized one year later, the global quality of the approximation improves significantly compared with the no-delay choice. But when the delay is extended up to two years, the additional improvement becomes lower (for instance, with the current choice and $t = 10$: no delay = 1.23%; delay one year = 1.11%; delay two years = 1.07%). From a cost/benefit point of view, this suggests that the optimal delay is maybe closer to one year than to two years. Note that the no-delay current index is a bit problematic for repeat-sales with a purchase near T; the quality of the measure deteriorates at

Exhibit 6a | Standard Deviations for $\text{Ln}(Ind_T(T)) - \text{Ln}(Ind_t(t))$ and $\text{Ln}(Ind_T(T)) - \text{Ln}(Ind_t(T))$



T is given by the curves; *t* corresponds to the horizontal axis. In all cases, the transaction date is *t* and the resale date is *T*.

Exhibit 6b | Standard Deviations for $\text{Ln}(Ind_T(T+d)) - \text{Ln}(Ind_t(t+d))$ and $\text{Ln}(Ind_T(T+d)) - \text{Ln}(Ind_t(T+d))$



the right of the interval. If these results are compared with those established in Clapham et al. (2006), the same ranking is found, except when dealing with delayed indexes. It seems that the current indexes are always better than the initial ones, according to this model. Moreover, the error strongly depends on the purchase date. The next step in the study of derivatives would be to choose a stochastic dynamic for the RSI in order to price the contingent claims. Unfortunately, things are rather complex because of the reversibility. If the basic assumption for stochastic processes in finance (related to the concept of market efficiency) is considered, that is their Markovian[29] behavior, a problem occurs. Is it really possible to describe the dynamic of the RSI with a single Markovian process? The answer is no. It is understood heuristically that there is a problem in just rewriting the reversibility formula: $\hat{I}(T_2)R(T_2) - \hat{I}(T_1)R(T_1) = \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$. The left side measures an increment between the present T_1 and the future T_2. If the Markovian assumption is satisfied, this variation cannot depend on the dates before T_1. But the right side $\hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$ is associated with new data arriving with the time extension, and adds information not only to $[T_1,T_2]$ but also to interval $[0,T_1]$. Consequently, RSI does not have a Markovian behavior. What follows from this result is the usual stochastic dynamics (geometric Brownian motion, Ornstein-Uhlenbeck...) cannot be used, at least not directly, to price a contingent claim. A solution might be to describe the reversibility process itself with a dynamic related to the reversibility law, and then to model the RSI as a noisy asset, as in Childs, Ott, and Riddiough (2001, 2002a, 2002b). Using this approach, the price discovery mechanism associated with the reversibility phenomenon could be captured. Even if the technical problems are important, the stakes are real and crucial for the finance industry. It is nothing less than the possibility of pricing the real estate derivatives written on the RSI.

Conclusion

An intuitive and easy to handle formula for the reversibility phenomenon was established using an informational reformulation of the RSI framework (cf. Appendix A for the example). Then, using an exponential benchmark for the resale decision and Monte-Carlo simulations, a methodology was developed for quantifying the size of the potential revisions, conditionally and unconditionally. In this way the problem[30] mentioned in Clapham, Englund, Quigley, and Redfearn (2006) was resolved for the repeat-sales index. For the moment, as there is not a similar technique for the hedonic indexes, it cannot be concluded that the RSI is not a suitable underlying for the derivatives contracts. Indeed, if its fluctuations are probably higher, they are nevertheless predictable, contrary to the hedonic approach. Regarding the best choice for settlement of the derivatives contracts, it seems that the current indexes with a delay equal to one year are the optimal choice. This article establishes that the reversibility phenomenon is not just a selectivity problem but also an inherent and intrinsic feature of the RSI, although the entire phenomenon cannot likely be reduced to a single sample effect. The natural question for future research is now to disentangle the two sources and to better understand their respective impacts.

Appendix A
Algebraic Examples

Old Index T_1 on [0,2]

Old Dataset

Time	0	1	2
House *a*		$p_{a,1}$	$p_{a,2}$
House *b*	$p_{b,0}$	$p_{b,1}$	
House *c*	$p_{c,0}$		$p_{c,2}$

Real Distribution

$n_{i,j}$	0	1	2
0		1	1
1			1
2			

Informational Distribution

$L_{i,j}$	0	1	2
0		1	0.5
1			1
2			

$$\begin{pmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{pmatrix}$$

Matrix $\hat{I}$

$$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

Matrix η

$$\tau^0(T_1) = \tau^1(T_1) = 2/1.5 \approx 1.33$$

Purchase Price in $C(i,j)$

$h_p^{(i,j)}$	0	1	2
0		$p_{b,0}$	$p_{c,0}$
1			$p_{a,1}$
2			

Resale Price in $C(i,j)$

$h_f^{(i,j)}$	0	1	2
0		$p_{b,1}$	$p_{c,2}$
1			$p_{a,2}$
2			

Mean Purchase and Resale Prices in *Spt*^{t}

$$H_p(0;T_1) = [(p_{b,0})^1\ (p_{c,0})^{0.5}]^{1/1.5} \qquad H_p(1;T_1) = [(p_{a,1})^1\ (p_{c,0})^{0.5}]^{1/1.5}$$

$$H_f(0;T_1) = [(p_{b,1})^1\ (p_{c,2})^{0.5}]^{1/1.5} \qquad H_f(1;T_1) = [(p_{a,2})^1\ (p_{c,2})^{0.5}]^{1/1.5}$$

Mean Return Rates in *Spt*^{t}

$$\rho_0(T_1) = (1/\tau^0)\ \ln[H_f(0;T_1)/H_p(0;T_1)] = 1/1.33$$

$$* [(1/1.5)\ \ln(p_{b,1}/p_{b,0}) + (0.5/1.5)\ \ln(p_{c,2}/p_{c,0})]$$

$$\rho_1(T_1) = (1/\tau^1) \ln[H_f(1;T_1)/H_p(1;T_1)] = 1/1.33$$
$$* [(1/1.5) \ln(p_{a,2}/p_{a,1}) + (0.5/1.5) \ln(p_{c,2}/p_{c,0})]$$

The term into brackets in the first equation represents the mean return for the repeat-sales that are relevant for [0,1], but not their mean rate. Indeed, these returns are realized on different time periods (one unit of time for the good *b* and two units for the good *c*). In order to get the mean rate, the brackets are divided by the associated mean holding period equal to 1.33.

Estimation of the index:
$$\begin{cases} 1.5r_0(T_1) + 0.5r_1(T_1) = 2\rho_0(T_1) \\ 0.5r_0(T_1) + 1.5r_1(T_1) = 2\rho_1(T_1) \end{cases}$$

These equations can be understood, for instance the first one, in the following way. The mean rate realized by the repeat-sales that are relevant for the interval [0,1] on their whole holding periods, in other words $\rho_0(T_1)$, depends on 75% of the first unitary index rate $r_0(T_1)$, and on 25% of the second unitary index rate $r_1(T_1)$. This set of transactions is more focused on the first time interval but not only because of the good *c*, bought at 0 and resold at 2. If the system is solved:

$$\begin{cases} r_0(T_1) = 1.5\rho_0(T_1) - 0.5\rho_1(T_1) \\ r_1(T_1) = 1.5\rho_1(T_1) - 0.5\rho_0(T_1) \end{cases}$$

$$\begin{cases} r_0(T_1) = 1/8 * [6 \ln(p_{b,1}/p_{b,0}) + 2 \ln(p_{c,2}/p_{c,0}) - 2 \ln(p_{a,2}/p_{a,1})] \\ r_1(T_1) = 1/8 * [6 \ln(p_{a,2}/p_{a,1}) + 2 \ln(p_{c,2}/p_{c,0}) - 2 \ln(p_{b,1}/p_{b,0})] \end{cases}$$

And if: $r_a = \ln(p_{a,2}/p_{a,1})$ $\quad r_b = \ln(p_{b,1}/p_{b,0})$ $\quad r_c = 1/2 \ln(p_{c,2}/p_{c,0})$

$$\begin{cases} r_0(T_1) = 1/8 * [6r_b + 4r_c - 2r_a] \\ r_1(T_1) = 1/8 * [6r_a + 4r_c - 2r_b] \end{cases}$$

For the first rate $r_0(T_1)$, the goods *b* and *c* contribute positively, whereas the good *a* that is informational just on the interval [1,2] contributes negatively in order to subtract to the return of the good *c* the part that is only depending on the interval [1,2]. This equation could be rewritten as: $r_0(T_1) = 1/8 * [(6r_b + 2r_c) + (2r_c - 2r_a)]$. The interpretation is the same for the second equation.

Index "New Data" $T_2\backslash T_1$ on [0,3]

New Data

Time	0	1	2	3
House *d*			$p_{d,2}$	$p_{d,3}$
House *e*	$p_{e,0}$			$p_{e,3}$

Mean Return Rates in *Spt*^{t}

$\rho_0(T_2\backslash T_1) = 1/3 * [\ln(p_{e,3}/p_{e,0})]$

$\rho_1(T_2\backslash T_1) = 1/3 * [\ln(p_{e,3}/p_{e,0})]$

$\rho_2(T_2\backslash T_1) = 1/1{,}5*[(0.33/1.33)\ln(p_{e,3}/p_{e,0}) + (1/1.33)\ln(p_{d,3}/p_{d,2})]$

Estimation of the index:

$$\begin{cases} 0.33r_0(T_2\backslash T_1) + 0.33r_1(T_2\backslash T_1) + 0.33r_2(T_2\backslash T_1) = 1\ \rho_0(T_2\backslash T_1) \\ 0.33r_0(T_2\backslash T_1) + 0.33r_1(T_2\backslash T_1) + 0.33r_2(T_2\backslash T_1) = 1\ \rho_1(T_2\backslash T_1) \\ 0.33r_0(T_2\backslash T_1) + 0.33r_1(T_2\backslash T_1) + 1.33r_2(T_2\backslash T_1) = 2\ \rho_2(T_2\backslash T_1) \end{cases}$$

Completed Index T_2 on [0,3]

Mean Return Rates in *Spt*^{t}

$\rho_0(T_2) = 1/1.63*[(1/1.83)\ln(p_{b,1}/p_{b,0}) + (0.5/1.83)\ln(p_{c,2}/p_{c,0}) + (0.33/1.83)\ln(p_{e,3}/p_{e,0})]$

$\rho_1(T_2) = 1/1.63*[(1/1.83)\ln(p_{a,2}/p_{a,1}) + (0.5/1.83)\ln(p_{c,2}/p_{c,0}) + (0.33/1.83)\ln(p_{e,3}/p_{e,3}/p_{e,0})]$

$\rho_2(T_2) = 1/1.5 *[(0.33/1.33)\ln(p_{e,3}/p_{e,0}) + (1/1.33)\ln(p_{d,3}/p_{d,2})]$

Estimation of the index:

$$\begin{cases} 1.83r_0(T_2) + 0.83r_1(T_2) + 0.33r_2(T_2) = 3\rho_0(T_2) \\ 0.83r_0(T_2) + 1.83r_1(T_2) + 0.33r_2(T_2) = 3\rho_1(T_2) \\ 0.33r_0(T_2) + 0.33r_1(T_2) + 1.33r_2(T_2) = 2\rho_2(T_2) \end{cases}$$

Reversibility for vector **P** gives:

$$\begin{cases} 3\rho_0(T_2) = 2\rho_0(T_1) + 1\rho_0(T_2\backslash T_1) \\ 3\rho_1(T_2) = 2\rho_1(T_1) + 1\rho_1(T_2\backslash T_1) \\ 2\rho_2(T_2) = 2\rho_2(T_2) \end{cases}$$

And equivalently for vector **R**:

$$\begin{cases} 1.83r_0(T_2)+0.83r_1(T_2) + 0.33r_2(T_2) = [1.5r_0(T_1) + 0.5r_1(T_1)] \\ \quad + [0.33r_0(T_2\backslash T_1) + 0.33r_1(T_2\backslash T_1) + 0.33r_2(T_2\backslash T_1)] \\ 0.83r_0(T_2) + 1.83r_1(T_2) + 0.33r_2(T_2) = [0.5r_0(T_1) + 1.5r_1(T_1)] \\ \quad + [0.33r_0(T_2\backslash T_1) + 0.33r_1(T_2\backslash T_1) + 0.33r_2(T_2\backslash T_1)] \\ 0.33r_0(T_2) + 0.33r_1(T_2) + 1.33r_2(T_2) = 0 + [0.33r_0(T_2\backslash T_1) \\ \quad + 0.33r_1(T_2\backslash T_1) + 1.33r_2(T_2\backslash T_1)] \end{cases}$$

Appendix B

Demonstration of the Informational Reversibility Formulas

Further Details (estimation on $[0, T_1]$)

The number of repeat-sales in Spl^t is $n^t = \Sigma_{i \le t < j} n_{i,j}$. For an element of $C(i,j)$, the length of the holding period is $j - i$. Using function G, the G-mean[31] ζ^t of these lengths in Spl^t can be defined by $\Sigma_{i \le t < j}\ \Sigma_k,\ G(j - i) = n^t G(\zeta^t)$. The first sum enumerates all the classes $C(i,j)$ that belong to Spl^t, the second, all the elements in each of these classes. Moreover, as $G(j - i)$ measures the proportion of the time varying-noise $G_{k,t}$ in the total noise for a repeat-sales of $C(i,j)$, the quantity $G(\zeta^t)$ can also be interpreted as the mean proportion of this Gaussian noise in the global one, for the entire subsample Spl^t. In the same spirit, the arithmetic average F^t of the holding frequencies $1/(j - i)$, weighted by $G(j - i)$, is defined in Spl^t: $F^t = (n^t G(\zeta^t))^{-1}\ \Sigma_{i \le t < j}\ \Sigma_k,\ G(j - i)^*(1/(j - i)) = I^t/(n^t G(\zeta^t))$. Its inverse $\tau^t = (F^t)^{-1}$ is then the harmonic average[32] of holding periods $j - i$, weighted by $G(j - i)$, in Spl^t. If at first the two averages ζ^t and τ^t appear to be two different concepts, in fact they are nothing of the sort. These is always, for each subsample Spl^t, $\zeta^t = \tau^t$ [cf. Simon (2007) for more details].

Reversibility for I^t and n^t

Table B1 exemplifies the extension of the horizon for the informational distribution. Two kinds of new repeat sales exist: those with a purchase before T_1

Table B1 | Informational Distribution When the Horizon is Extended from T_1 to T_2

	0	1	…	t	$t+1$	…	T_1	…	T_2
0		$L_{0,1}$	…	$L_{0,t}$	$L_{0,t+1}$	…	$L_{0\,T_1}$	…	$L_{0\,T_2}$
1			…	$L_{1,t}$	$L_{1,t+1}$	…	L_{1,T_1}	…	L_{1,T_2}
⋮				…	…	…	…	…	…
t					$L_{t,t+1}$	…	L_{t,T_1}	…	L_{t,T_2}
$t+1$						…	$L_{t+1\,T_1}$	…	$L_{t+1\,T_2}$
⋮							…	…	…
T_1								…	L_{T_1,T_2}
⋮									…
T_2									

Notes: Solid lines: new repeat sales with a purchase before T_1 and a resale after $T_1(l < T_1 < j \leq T_2)$. Dotted lines: new repeat sales with a purchase and a resale between T_1 and $T_2(T_1 \leq i < j \leq T_2)$.

and a resale after $T_1(i < T_1 < j \leq T_2)$, shown as solid lines, and those with a purchase and a resale between T_1 and $T_2(T_1 \leq i < j \leq T_2)$, shown as dotted lines. In this table, the relevant repeat sales for $[t,t+1]$, if the horizon is T_1, are in light gray. And if the horizon becomes T_2, the darker gray cells should also be included in this set.

For an interval $[t,t+1]$, $t < T_1$, the quantities of relevant information are $I^t(T_1) = \Sigma_{i\leq t<j\leq T_1}L_{i,j}$ for the first horizon and $I^t(T_2) = \Sigma_{i\leq t<j\leq T_2}L_{i,j} = I^t(T_1) + \Sigma_{i\leq t<T_1<j\leq T_2}L_{i,j}$ for the second. The sum with $i \leq t < T_1 < j \leq T_2$ corresponds to the additional information (darker gray). If denoted as $I^t(T_2\backslash T_1)$, the relation becomes $I^t(T_2) = I^t(T_1) + I^t(T_2\backslash T_1)$. Similarly, for the real equivalents of $I^t(T_2)$, $I^t(T_1)$, and $I^t(T_2\backslash T_1)$, that is, $n^t(T_2)$, $n^t(T_1)$, and $n^t(T_2\backslash T_1)$, gives exactly the same kind of formula: $n^t(T_2) = n^t(T_1) + n^t(T_2\backslash T_1)$. In what follows, the notation $T_2\backslash T_1$ will refer to the dataset of the new repeat sales that appear when the horizon is extended.

Reversibility for the Mean Prices $H_p(t)$ and $H_f(t)$

First calculate $H_p(t)$ with the purchase prices for the two horizons:

$$[H_p(t,T_1)]^{I^t(T_1)} = \Pi_{i\leq t<j\leq T_1}(\Pi_{k'}\, p_{k,,i})^{1/(\Theta+(j-i))}$$

$$[H_p(t,T_2)]^{I^t(T_2)} = \Pi_{i\leq t<j\leq T_2}(\Pi_{k'}\, p_{k,,i})^{1/(\Theta+(j-i))}.$$

Therefore: $[H_p(t,T_2)]^{I^t(T_2)} = [H_p(t,T_1)]^{I^t(T_1)}\Pi_{i\leq t<T_1<j\leq T_2}(\Pi_{k'}p_{k,,i})_{1/(\Theta+(j-i))}$.

Introducing $h_p^{(i,j)}$, the product becomes: $\Pi_{i\leq t<T_1<j\leq T_2}(\Pi_{k'}p_{k,,i})^{1/(\Theta+(j-i))} = \Pi_{i\leq t<T_1<j\leq T_2}(h_p^{(i,j)})^{L_{i,j}}$.

The total mass of these weights $L_{i,j}$ is equal to $I^t(T_2\backslash T_1)$. This geometric average is denoted:

$$[H_p(t,\ T_2\backslash T_1)]^{I^t(T_2\backslash T_1)} = \Pi_{i\le t<T_1<j\le T_2}(h_p^{(i,j)L_{i,j}}$$
$$= \Pi_{i\le t<T_1<j\le T_2}(\Pi_{k'}p_{k,,i})^{1/(\Theta+(j-i))}$$

For the interval $[t,t+1]$, $H_p(t,\ T_2\backslash T_1)$ represents the mean purchase price for the new relevant repeat sales. Thus, the reversibility formula is: $[H_p(t,T_2)]^{I^t(T_2)} = [H_p(t,T_1)]^{I^t(T_1)}[H_p(t,T_2\backslash T_1)]^{I^t(T_2\backslash T_1)}$. The new value $H_p(t,T_2)$ is the geometric average between the old value $H_p(t,T_1)$ and a term that represents the new data: $H_p(t,T_2\backslash T_1)$. Their respective contributions are measured by the informational weights $I^t(T_1)$ and $I^t(T_2\backslash T_1)$. Similarly, for the resale prices, if the following is introduced:

$$[H_f(t,T_2\backslash T_1)]^{I^t(T_2\backslash T_1)} = \Pi_{i\le t<T_1<j\le T_2}(h_f^{(i,j)})L_{i,j}$$
$$= \Pi_{i\le t<T_1<j\le T_2}(\Pi_{k'p_{k,,j}})^{1/(\Theta+(j-i)),}$$

gives:

$$[H_f(t,T_2)]^{I^t(T_2)} = [H_f(t,T_1)]^{I^t(T_1)}[H_f(t,T_2\backslash T_1)]^{I^t(T_2\backslash T_1)}.$$

Reversibility for τ^t

This section discusses the link between the mean holding periods $\tau^t(T_1)$ and $\tau^t(T_2)$. There is: $I^t(T_2\backslash T_1) = \Sigma_{i\le t<T_1<j\le T_2}L_{i,j} = \Sigma_{i\le t<T_1<j\le T_2}\Sigma_{k'}G(j-i) * (1/(j-i))$. Thus, $I^t(T_2\backslash T_1)$ is almost the arithmetic average of $1/(j-i)$ weighted by $G(j-i)$. This formula lacks only the total mass of the weights, that is $\Sigma_{i\le t<T_1<j\le T_2}\Sigma_{k'}G(j-i) = n^t(T_2\backslash T_1)\ G(\zeta^t(T_2\backslash T_1))$, with $\zeta^t(T_2\backslash T_1)$ the G-average of the holding periods for the new repeat sales. Therefore, as in the basic situation, $I^t(T_2\backslash T_1)/[n^t(T_2\backslash T_1)\ G(\zeta^t(T_2\backslash T_1))]$ is a mean frequency $F^t(T_2\backslash T_1)$, and its inverse a mean harmonic holding period $\tau^t(T_2\backslash T_1)$ for the new repeat-sales. Now the formal link between $\tau^t(T_1)$ and $\tau^t(T_2)$ can be established with the relations $I^t(T_2\backslash T_1) = [n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))]/\tau^t(T_2\backslash T_1)$ and $I^t(T_2) = I^t(T_1) + I^t(T_2\backslash T_1)$. Thus: $[n^t(T_2)G(\zeta^t(T_2))]/\tau^t(T_2) = [n^t(T_1)G(\zeta^t(T_1))]/t^t(T_1) + [n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))]/\tau^t(T_2\backslash T_1)$. And, as there is $n^t(T_2)G(\zeta^t(T_2)) = n^t(T_1)G(\zeta^t(T_1)) + n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))$, therefore $\tau^t(T_2)$ is simply the harmonic weighted average of $\tau^t(T_1)$ and $\tau^t(T_2\backslash T_1)$.

Reversibility for ρ_t

Scalar formula for $t < T_1$

For $t < T_1$[33]: $\rho_t(T_2) = [(1/\tau^t(T_1))*(\ln H_f(t,T_1) - \ln H_p(t,T_1))]*[(I^t(T_1)\tau^t(T_1))/(I^t(T_2)\tau^t(T_2))] + [(1/\tau^t(T_2\backslash T_1)) * (\ln H_f(t, T_2\backslash T_1) - \ln H_p(t, T_2\backslash T_1))]*[(I^t(T_2\backslash T_1)\tau^t(T_2\backslash T_1))/(I^t(T_2)\tau^t(T_2))]$. In the first square brackets, it can be seen that $\rho_t(T_1)$. Moreover, it can easily be proved that the third brackets are also equal to $[n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))]^{-1}\Sigma_{i\le t<T_1<j\le T_2}\Sigma_{k'}G(j-i)r_{k'}^{(i,j)}$. This expression is simply the weighted mean of the mean rates $r_{k'}^{(i,j)}$, for the new repeat sales. And, of course, it is denoted: $\rho_t(T_2\backslash T_1)$. Thus, the reversibility formula for ρ_t, $t < T_1$, is:

$$\begin{aligned}\rho_t(T_2) &= [I^t(T_1)/I^t(T_2)][\tau^t(T_1)/\tau^t(T_2)]\rho_t(T_1)\\ &\quad + [I^t(T_2\backslash T_1)/I^t(T_2)][\tau^t(T_2\backslash T_1)/\tau^t(T_2)]\rho_t(T_2\backslash T_1).\end{aligned}$$

Vectorial Formula

The above formula is valid for $t < T_1$. However, the expressions that define $I^t(T_2\backslash T_1)$, $\tau^t(T_2\backslash T_1)$, $\zeta^t(T_2\backslash T_1)$, $n^t(T_2\backslash T_1)$, and $\rho_t(T_2\backslash T_1)$ can be generalized for $t \ge T_1$. Indeed, in these expressions the sums are for the classes $C(i,j)$ such that $i \le t < T_1 < j \le T_2$, that is the new repeat-sales relevant for $[t,t+1]$, with $t < T_1$. Now, if $t \ge T_1$ is selected, the relevant cells will be the ones that satisfy[34] to $i \le t < j \le T_2$. But what is produced is not really new; it is just $I^t(T_2)$, $\tau^t(T_2)$, $\zeta^t(T_2)$, $n^t(T_2)$ and $\rho_t(T_2)$. For instance $I^t(T_2\backslash T_1) = \Sigma_{i\le t<T_1<j\le T_2}L_{i,j}$ gives for $t \ge T_1$: $\Sigma_{i\le t<j\le T_2}L_{i,j} = I^t(T_2)$. Now the reversibility formula for ρ_t can be written in a more synthetic manner. The values $\rho_t(T_2)$ are regrouped, for $0 \le t < T_2$, in a T_2-vector $P(T_2)$ and the values $\rho_t(T_1)$, for $0 \le t < T_1$, in a T_1-vector $P(T_1)$. From vector $P(T_1)$, a T_2-vector is created, adding to its end $T_2 - T_1$ zeros; it will be denoted in italics $P(T_1)$. The numbers $\rho_t(T_2\backslash T_1)$ are regrouped in a T_2-vector $P(T_2\backslash T_1)$. Its last $T_2 - T_1$ coordinates are simply equal to $\rho_t(T_2)$. Thus, for $t < T_1$:

$$\begin{aligned}&\tau^t(T_2)I^t(T_2)\rho_t(T_2) = I^t(T_1)t^t(T_1)\rho_t(T_1)\\ &\qquad + I^t(T_2\backslash T_1)\tau^t(T_2\backslash T_1)\rho_t(T_2\backslash T_1)\\ \Leftrightarrow\quad &n^t(T_2)G(\zeta^t(T_2))\rho_t(T_2) = n^t(T_1)G(\zeta^t(T_1))\rho_t(T_1)\\ &\qquad + n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))\rho_t(T_2\backslash T_1).\end{aligned}$$

And for $t \ge T_1$: $n^t(T_2)G(\zeta^t(T_2))\rho_t(T_2) = n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))\rho_t(T_2\backslash T_1)$.

The diagonal matrix $\eta(T_1)$ can be included in a T_2-matrix, completing it with zeros, and denoted in italics: $\eta(T_1)$. $\eta(T_2)$ is the usual T_2-diagonal matrix. Denote $\eta(T_2\backslash T_1)$ the T_2-diagonal matrix that is achieved with $n^0(T_2\backslash T_1)G(\zeta^0(T_2\backslash T_1),...,n^{T_2-1}(T_2\backslash T_1)G(\zeta^{T_2-1}(T_2\backslash T_1))$. Now simultaneously these two kinds of equations can be written (for $t < T_1$ and for $t \geq T_1$):

$$\eta(T_2)P(T_2) = \eta(T_1)P(T_1) + \eta(T_2\backslash T_1)P(T_2\backslash T_1).$$

Reversibility for the Informational Matrix $\hat{I}$

For an interval $[t_i, t_j]$ the relevant information is denoted $I^{[t_i,t_i]}(T_1)$ or $I^{[t_i,t_i]}(T_2)$, according to the horizon. The associated informational matrixes are $\hat{I}(T_1)$ and $\hat{I}(T_2)$, dimension T_1 and T_2 respectively. A third matrix $\hat{I}(T_2\backslash T_1)$, dimension T_2, is the link between $\hat{I}(T_1)$ and $\hat{I}(T_2)$. Its values are calculated only with the new $L_{i,j}$ (cf. Table B2), and for each interval $[t_i, t_j] \in [0,T_2]$ they represent the additional quantity of information.

Now $\hat{I}(T_2\backslash T_1)$ can be written with three submatrixes $\hat{I}_a(T_2\backslash T_1)$, $\hat{I}_b(T_2\backslash T_1)$, and $\hat{I}_c(T_2\backslash T_1)$. $\hat{I}_a(T_2\backslash T_1)$ and $\hat{I}_c(T_2\backslash T_1)$ are two square matrixes of dimension T_1 and $T_2 - T_1$, whereas $\hat{I}_b(T_2\backslash T_1)$ is a $T_1*(T_2 - T_1)$ matrix and its transpose ${}^t\hat{I}_b(T_2\backslash T_1)$ a $(T_2 - T_1)*T_1$ matrix. $\hat{I}_a(T_2\backslash T_1)$ is symmetric and its diagonal elements correspond to the first column of $\hat{I}_b(T_2\backslash T_1)$; from one of these diagonal elements, the matrix values are the same on the right and below. The matrixes $\hat{I}_b(T_2\backslash T_1)$ and $\hat{I}_c(T_2\backslash T_1)$ are simply extracted from $\hat{I}(T_2)$. $\hat{I}_a$ and $\hat{I}_c$, respectively, represent the additional information for an interval $[t_i,t_j] \in [0,T_1]$ and $[t_i,t_j] \in [T_1,T_2]$. Whereas $\hat{I}_b$ is for the intervals $[t_i,t_j] \in [0,T_2]$ with $t_i < T_1 < t_j$. If the matrix $\hat{I}(T_1)$ is included in a T_2*T_2 matrix, completing it with zeros and denoting it in italics $\hat{I}(T_1)$, the reversibility formula for the informational matrix is simply: $\hat{I}(T_2) = \hat{I}(T_1) + \hat{I}(T_2\backslash T_1)$.

Table B2 | Informational Distribution for Dataset $T_2\backslash T_1$

	0	1	…	T_1	T_1+1	…	T_2
0		0	…	0	L_{0,T_1+1}	…	L_{0,T_2}
1			…	0	L_{1,T_1+1}	…	L_{1,T_2}
⋮				⋮	⋮	…	⋮
T_1					L_{T_1,T_1+1}	…	L_{T_1,T_2}
T_1+1						…	L_{T_1+1,T_2}
⋮							⋮
T_2							

$$\hat{I}(T_2\backslash T_1) = \begin{pmatrix} \hat{I}_a(T_2\backslash T_1) & \hat{I}_b(T_2\backslash T_1) \\ {}^t\hat{I}_b(T_2\backslash T_1) & \hat{I}_c(T_2\backslash T_1) \end{pmatrix}$$

Reversibility for the Index

The last step consists of establishing the reversibility formula for the index. For an horizon T_1 and $t < T_1$, the building blocks $I^t(T_1)$, $\tau^t(T_1)$, $\zeta^t(T_1)$, $n^t(T_1)$, and $\rho_t(T_1)$ give the repeat-sales index $R(T_1)$.[35] Similarly $I^t(T_2)$, $\tau^t(T_2)$, $\zeta^t(T_2)$, $n^t(T_2)$ and $\rho_t(T_2)$, calculated for $t < T_2$, give the repeat-sales index $R(T_2)$. The link between these two groups of intermediate measures is known, thanks to quantities $I^t(T_2\backslash T_1)$, $\tau^t(T_2\backslash T_1)$, $\zeta^t(T_2\backslash T_1)$, $n^t(T_2\backslash T_1)$, and $\rho_t(T_2\backslash T_1)$. Thus, it suggests that it is useful to estimate the RSI on the interval $[0,T_2]$, just with the sample $T_2\backslash T_1$. In this way, a T_2-vector $R(T_2\backslash T_1)$ is achieved.[36] If the general relation $\hat{I}R = \eta P$ and the reversibility formula established for vector P: $\eta(T_2)P(T_2) = \eta(T_1)P(T_1) + \eta(T_2\backslash T_1)P(T_2\backslash T_1)$ are used, a very simple reversibility formula for the repeat-sales index is:

$$\hat{I}(T_2)R(T_2) = \hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1).$$

Appendix C

Volatility and Contract Settlement

The fourth situations can be studied with one single formula, namely the variance of the difference: $Lind_{t1}(T_1) - LInd_{t2}(T_2)$, where $t_1 < T_1$ and $t_2 < T_2$. It is useful here to introduce the following notations: $e_i(L) = (0,..., 0, 1, 0,..., 0)'$. $e_i(L)$ is a column vector of dimension **L**, with all its components equal to zero, except the i^{th}, which is equal to 1. It is known that:

$$\begin{aligned} \mathrm{Var}(Lind_{t1}(T_1) - LInd_{t2}(T_2)) = {} & \mathrm{Var}(Lind_{t1}(T_1)) + \mathrm{Var}(LInd_{t2}(T_2)) \\ & - 2\mathrm{Cov}(Lind_{t1}(T_1); LInd_{t2}(T_2)). \end{aligned}$$

The first two variances with the matrix formula can be calculated: V(*LInd*) = $\sigma_{G^2} A\hat{I}^{-1}A'$ (cf. Simon, 2007). More precisely:

$$\mathrm{Var}(Lind_{t1}(T_1)) = \sigma_{G^2} e_{t1}(T_1)'A(T_1)\hat{I}(T_1)^{-1}A(T_1)'e_{t1}(T_1).$$
$$\mathrm{Var}(Lind_{t2}(T_2)) = \sigma_{G^2} e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}A(T_2)'e_{t2}(T_2).$$

For the third term, as there are estimations with different horizons, the reversibility formula has to be used: $\hat{I}(T_2)R(T_2) = \hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$. The random variables $R(T_1)$ and $R(T_2\backslash T_1)$ are independent because they do not have any repeat-sales classes $C(i,j)$ in common. As $Lind(T_1)$ and $Lind(T_2\backslash T_1)$ are linearly related to $R(T_1)$ and $R(T_2\backslash T_1)$ through the formula $Lind = AR$, these two vectors are also independent. The reversibility formula can be written:[37]

$$\begin{aligned}
\hat{I}(T_2)A(T_2)^{-1}A(T_2)R(T_2) &= \hat{I}(T_1)A(T_1)^{-1}A(T_1)R(T_1) \\
&\quad + \hat{I}(T_2\backslash T_1)A(T_2\backslash T_1)^{-1}A(T_2\backslash T_1)R(T_2\backslash T_1). \\
\hat{I}(T_2)A(T_2)^{-1}Lind(T_2) &= \hat{I}(T_1)A(T_1)^{-1}Lind(T_1) \\
&\quad + \hat{I}(T_2\backslash T_1)A(T_2\backslash T_1)^{-1}Lind(T_2\backslash T_1). \\
Lind(T_2) &= A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_1)A(T_1)^{-1}Lind(T_1) \\
&\quad + A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_2\backslash T_1)A(T_2\backslash T_1)^{-1}Lind(T_2\backslash T_1).
\end{aligned}$$

If the right this equation is multiplied by the vector $e_{t2}(T_2)'$:

$$\begin{aligned}
Lind_{t2}(T_2) &= e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_1)A(T_1)^{-1}Lind(T_1) \\
&\quad + e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_2\backslash T_1)A(T_2\backslash T_1)^{-1}Lind(T_2\backslash T_1).
\end{aligned}$$

And as $Lind(T_1)$ and $Lind(T_2\backslash T_1)$ are independent:

$$\begin{aligned}
&Cov(Lind_{t2}(T_2); Lind_{t1}(T_1)) \\
&\quad = Cov(e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_1)A(T_1)^{-1}Lind(T_1); Lind_{t1}(T_1)).
\end{aligned}$$

But as $Lind(T_1) = \Sigma_{i=1,\ldots,T_1} Lind_i(T_1)e_i(T_2)$:

$$\mathrm{Cov}(Lind_{t2}(T_2);\ Lind_{t1}(T_1)) = \Sigma_{i=1,\ldots,T_1} e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_1)A(T_1)^{-1}e_i(T_2)\ Cov(Lind_i(T_1);$$
$$Lind_{t1}(T_1)) = \sigma_{G^2}\Sigma_{i=1,\ldots,T_1} e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_1)A(T_1)^{-1}$$
$$e_i(T_2)e_i(T_1)'A(T_1)\hat{I}(T_1)^{-1}A'(T_1)e_{t1}(T_1).$$

If these results are summarized in a global formula:

$$Var(Lind_{t1}(T_1) - Lind_{t2}(T_2)) = \sigma_{G^2}[e_{t1}(T_1)'A(T_1)\hat{I}(T_1)^{-1}A(T_1)'e_{t1}(T_1) + e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}A(T_2)'e_{t2}(T_2)]$$
$$2\sigma_{G^2}\Sigma_{i=1,\ldots,T_1} e_{t2}(T_2)'A(T_2)\hat{I}(T_2)^{-1}\hat{I}(T_1)A(T_1)^{-1}$$
$$e_i(T_2)e_i(T_1)'A(T_1)\hat{I}(T_1)^{-1}A(T_1)'e_{t1}(T_1).$$

1. $n_{i,j}$: Number of repeat sales with purchase at t_i and resale at t_j, organized in an upper triangular table.
2. Estimation of the volatilities σ_N and σ_G for the white noise and the random-walk (step 1 and 2 of the Case-Shiller procedure). The time of noise equality is $\Theta = 2\sigma_{N^2}/\sigma_{G^2}$.
3. $L_{i,j} = n_{i,j}/(\Theta + j - i)$: Quantity of information delivered by the $n_{i,j}$ repeat sales of $C(i,j)$. These numbers are also organized in an upper triangular table.
4. Matrix $\hat{I}$ is derived from the informational distribution of the $\{L_{i,j}\}$ summing for each time interval $[t,t']$ the relevant $L_{i,j}$, that is, the ones with a holding period that includes $[t,t']$. The diagonal elements of the diagonal matrix η are equal to the sums (rows or columns indifferently) of the components of matrix $\hat{I}$.
5. Dividing the diagonal elements of $\hat{I}$ by the diagonal elements of η gives the mean holding periods τ^t.
6. For each repeat-sales class $C(i,j)$, the geometric averages of the purchase prices $h_p^{(i,j)}$, and the resale prices $h_f^{(i,j)}$ are:

$$h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}} \quad h_f^{(i,j)} = (\Pi_{k'} p_{k,j})^{1/n_{i,j}}.$$

7. For the subset of the people who owned real estate during $[t,t+1]$, that is Spl^t, the mean purchase price $H_p(t)$ (the mean resale price $H_f(t)$) is the

geometric average of the $h_p^{(i,j)}$ (respectively the $h_f^{(i,j)}$), weighted by the $L_{i,j}$, for all the relevant repeat-sales classes:

$$H_p(t) = (\Pi_{i \leq t < j}(h_p^{(i,j)})^{L_{i,j}})^{1/\Pi^t} \quad H_f(t) = (\Pi_{i \leq t < j}(h_f^{(i,j)})^{L_{i,j}})^{1/\Pi^t}.$$

8. The mean of the mean rates ρ_t realized by the people who owned real estate during $[t,t+1]$ can be calculated as a return rate with fictitious prices: $H_p(t)$ for the purchase, $H_f(t)$ for the resale, and fictitious holding period τ^t:

$$\rho_t = (1/\tau^t) * \ln[H_f(t)/H_p(t)].$$

9. Vector **R** of the monoperiodic growth rates of the index is the solution of the equation:

$$\hat{I}\mathbf{R} = \eta P \Leftrightarrow \mathbf{R} = (\hat{I}^{-1}\eta)P.$$

P is the vector $(\rho_0, \rho_1, \ldots, \rho_{T-1})$.

Endnotes

[1] D is a matrix extracted from another matrix D'; the first column has been removed to avoid a singularity in the estimation process. The number of lines of D' is equal to the total number of repeat-sales in the dataset, and its $T+1$ columns correspond to the different possible times for the trades. In each line -1 indicates the purchase date, $+1$ the resale date, and the rest is completed with zeros.

[2] Σ is a diagonal matrix with a dimension equal to the size of the repeat-sales sample.

[3] In the Appendixes, the function $G(x) = x/(x+\Theta)$ will sometimes appear. For a holding period $j - i$, $G(j - i) = (j - i)/(\Theta + (j - i)) = \sigma_{G^2}(j - i)/[2\sigma N^2 + \sigma_{G^2}(j - i)]$. $G(j - i)$ is actually the proportion of the time-varying noise in the total noise; these numbers will be used as a system of weights.

[4] These measures are relative ones. What matters is their relative sizes and not their absolute levels. They can be defined up to a constant in order to standardize the measures.

[5] Simon (2007) establishes that the variance-covariance matrix of the vector of estimators R is equal to $\sigma_{G^2}\hat{I} - 1$ (cf. below for the definition of matrix $\hat{I}$).

[6] As exemplified in Exhibit 2, $I[t',t+1]$ can be calculated buy-side with the partial sums $\mathbf{B}0t, B1t, \ldots, Btt'$ or sell-side with $St'T, St'T-1, \ldots, St't+1$. But there is always $I[t',t+1] = B0t + \ldots + Btt' = St'T + \ldots + St't+1$.

[7] This simplification simply means that the noise coming from the white noise is null.

[8] Equally weighted because within a class $C(i,j)$ all the observations have the same degree of information.

[9] For instance, a datum with purchase at $t < T_1$ and resale at t' with $T_1 < t' < T_2$ will be informative for $[t,T_1]$. But, as the resale occurs after T_1, this repeat-sale cannot be used for the first index estimation.

[10] The purchases at $t = 0$ are not included to avoid a singular matrix in the estimation.

[11] More details are shown in Appendix B.

[12] If it is assumed that the average holding periods are all equal, the relation would simply become: $\rho t(T_2) = [It(T_1)/It(T_2)]\rho t(T_1) + [It(T_2/T_1)/It(T_2)]\rho t(T_2\backslash T_1)$.

[13] The vectorial quantities of dimension T_1, like $\mathbf{P}(T_1)$, $\mathbf{R}(T_1)$, $\boldsymbol{\eta}(T_1)$, and $\hat{\mathbf{I}}(T_1)$ can be injected in vectors and matrixes of dimension T_2, completing them with zeros. These equivalents will be denoted in italics in the formula: $\mathbf{P}(T_1)$, $\mathbf{R}(T_1)$, $\boldsymbol{\eta}(T_1)$, and $\hat{\mathbf{I}}(T_1)$.

[14] The informational matrixes $\hat{I}$ are used to get the diagonal matrixes η by just summing on the diagonals.

[15] Prob ($resale > t \mid resale \geq t$) = Prob ($resale > s \mid resale \geq s$).

[16] $\lambda(t)$ is a classical concept in the survival models [cf. Kalbfleisch and Prentice (2002)]. It appears, for example, in econometrical studies of prepayment and default options embedded in mortgages [cf. Deng, Quigley, and Van Order (2000)].

[17] $\pi = d(T) * (\alpha/T(1 - \alpha))$ $\quad d(k) = 1 - \alpha k$.

[18] $K' = K\,(1 - \alpha)/\alpha$ $\quad \Theta = 2\sigma N^2/\sigma G^2$ $\quad un = \alpha/(\Theta + 1) + \alpha^2/(\Theta + 2) + \alpha 3/(\Theta + 3) + ... + \alpha n/(\Theta + n)$.

[19] First the numbers of transactions realized at each date in the market are fixed. Then, the resale rates for each cohort are randomly generated. The estimation sample regroups the repeat-sales with a resale date observed before T. The prices are randomly generated around a ''true price'' curve; it is assumed that this curve is flat, for purposes of simplicity (Curve 1 in Exhibit 3).

[20] Other choices are possible for this calibration step, according to the economic contexts or the empirical issues.

[21] When K and α are known, it was demonstrated in Simon (2008) that $L_{i,j} = K'\alpha_j - i/(\Theta + j - i)$. First, the informational distribution of $\{L_{i,j}\}$ is calculated for the benchmark and for interval $[0,T_2]$. Then, we just keep the columns between T_1 and T_2, which represent the new data for the exponential sample. From this partial table, adding its components, gives the matrix $\hat{I}_{bench}(T_2\backslash T_1)$.

[22] $rate_t = \ln(Index_t + 1/Index_t)$.

[23] This formula is a general one. The variance-covariance matrix of vector $\mathbf{R}$, whatever the repeat-sales distribution, is always $V(\mathbf{R}) = \sigma_{G^2}\hat{I} - 1$.

[24] Cf. Simon (2007).

[25] For the purpose, the following was used: $LInd(T_2) = A(T_2)R(T_2)$ and $E[R(T_2)] = R(T_1)$.

[26] Matrix $A(T_2)$ is square and its dimension is T_2. It is composed of 1 on its diagonal and below, 0 elsewhere.

[27] $100*(Ind_t(T_2)/Ind_t(T_1) - 1)$.

[28] Here, the Cholesky factorization is used:
If Γ is a square matrix of dimension d, symmetric, positive, and with rank r then a

matrix B is found, dimension $d \times r$, rank r such that $\Gamma = BB'$ (Cholesky factorization) Now, for a vector $\mathbf{M}$ of dimension d, and for a square matrix Γ of dimension d, symmetric, positive, rank r, with its Cholesky factorization $\Gamma = BB'$: If $Y \sim N(0,Id)$ then $M + BY \sim N(M,\Gamma)$.

[29] A process is said to be Markovian if its future depends on its past only through its present. In others words, the path followed by the process to arrive at level Xs, on date s, will not influence the probability of realization of its future Xt ($t > s$). Financially, this mathematical assumption is one of the formulations for the concept of market efficiency. The present value incorporates all the past information; it is useless to study the past in order to get a better level for Xs. The market has already integrated all the available and relevant information with the fixing of Xs.

[30] ''If a futures market requires index stability, it would be useful to know how often revision—either period-by-period or cumulative—exceeds some level. Say, for example, that futures markets could tolerate 0.5 percent revision in any one quarter and 2 percent cumulative revision to the initial estimate—how often do the four indexes violate these criteria?''

[31] Recall here that the concept of average is a very general one. If a function G is strictly increasing or decreasing the G-mean of the numbers $\{x_1, x_2, \ldots, xn\}$, weighted by the $(\alpha_1, \alpha_2, \ldots, \alpha n)$, is the number X such that: $\alpha G(X) = \alpha_1 G(x_1) + \alpha_2 G(x_2) + \ldots + \alpha n G(xn)$ with $\alpha = \Sigma\, i = 1,\ldots,n\alpha i$. An arithmetic mean corresponds to $G(x) = x$, a geometric one to $G(x) = \ln(x)$ and the harmonic average to $G(x) = 1/x$.

[32] Thus $(tG(\zeta t))/\tau t = \Sigma i \leq t < j\Sigma k' G(j - i) * (1/(j - i)) = It$.

[33] $\rho t(T_2) = [It(T_2)/(nt(T_2)G(\zeta t(T_2)))] * \ln[Hf(t,T_2)/Hp(t,T_2)] = [It(T_2)/(nt(T_2)G(\zeta t(T_2)))] * [\ln Hf(t,T_2) - \ln Hp(t,T_2)] = [It(T_1)\ln Hf(t,T_1) + It(T_2\backslash T_1)\ln Hf(t,T_2\backslash T_1)]/[It(T_2)\tau t(T_2)] - [It(T_1)\ln Hp(t,T_1) + It(T_2\backslash T_1)\ln Hp(t,T_2\backslash T_1)]/[It(T_2)\tau t(T_2)]$.

[34] $i \leq T_1 \leq t < j \leq T_2$ is not correct because it would exclude the repeat-sales with a purchase at i such that $T_1 < i \leq t$. As these couples belong to the new data and are perfectly relevant for $[t,t+1]$, it cannot be removed.

[35] In order to have a T_2-vector $\mathbf{R}(T_1)$, the T_1-vector $\mathbf{R}(T_1)$ will sometimes be completed with T_2-T_1 final zeros.

[36] Above it was seen that It, τ_t, ζ_t, n_t, and ρ_t are equal for T_2 and $T_2\backslash T_1$ when $t \geq T_1$. Unfortunately, for the repeat-sales index this kind of relation is not true.

[37] For the matrix $A(T_1)$, its inverse $A(T_1)^{-1}$ is calculated first. The matrix $A(T_1)$ is equal to the matrix $A(T_1)$ completed with zeros, the same is don for $A(T_1)^{-1}$ and $A(T_1)^{-1}$.

References

Bailey, M.J., R.F. Muth, and H.O. Nourse. A Regression Method for Real Estate Price Index Construction. *Journal of the American Statistical Association*, 1963, 58, 933–42.

Baroni, M., F. Barthélémy, and M. Mokrane. Physical Real Estate: A Paris Repeat Sales Residential Index. ESSEC Working paper DR 04007, ESSEC Research Center, ESSEC Business School, 2004.

Case, K.E. and R.J. Shiller. Prices of Single Family Homes since 1970: New Indexes for Four Cities. *New England Economic Review*, 1987, September/October, 45–56.

——. The Efficiency of the Market for Single-Family Homes. *The American Economic Review*, 1989, 79:1, 125–37.

Chau, K.W., S.K. Wong, C.Y. Yiu, and H.F. Leung. Real Estate Price Indices in Hong-Kong. *Journal of Real Estate Literature*, 2005, 13:3, 337–56.

Cheung, S.L., K.W. Yau, and Y.V. Hui. The Effects of Attributes on the Repeat Sales Pattern of Residential Property in Hong-Kong. *Journal of Real Estate Finance and Economics*, 2004, 29:3, 321–39.

Childs, P.D., S.H. Ott, and T.J. Riddiough. Valuation and Information Acquisition Policy for Claims Written on Noisy Real Assets. *Financial Management*, 2001, Summer, 45–75.

——. Optimal Valuation of Noisy Real Assets. *Real Estate Economics*, 2002a, 30:3, 385–414.

——. Optimal Valuation of Claims on Noisy Real Assets: Theory and Application. *Real Estate Economics*, 2002b, 30:3, 415–43.

Clapham, E., P. Englund, J.M. Quigley, and C.L. Redfearn. Revisiting the Past and Settling the Score: Index Revision for House Price Derivatives. *Real Estate Economics*, 2006, 34: 2, 275–302.

Clapp, J.M. and C. Giaccotto. Revisions in Repeat-Sales Price Indexes: Here Today, Gone Tomorrow? *Real Estate Economics*, 1999, 27:1, 79–104.

Deng, Y. and J.M. Quigley. Index Revision, House Price Risk, and the Market for House Price Derivatives, *The Journal of Real Estate Finance and Economics*, 2008, 37, 191–209.

Deng, Y., J.M. Quigley, and R. Van Order. Mortgage Terminations, Heterogeneity and the Exercise of Mortgage Options. *Econometrica*, 2000, 68:2, 275–307.

Gatzlaff, D., and D. Geltner. A Repeat-Sales Transaction-Based Index of Commercial Property. A Study for the Real Estate Research Institute, 1998.

Hoesli, M., C. Giaccotto, and P. Favarger. Three New Real Estate Price Indices for Geneva, Switzerland. *Journal of Real Estate Finance and Economics*, 1997, 15:1, 93–109.

Kalbfleisch, J.D. and R.I. Prentice. *The Statistical Analysis of Failure Time Data*. Second edition. New York: John Wiley and Sons, 2002.

Meese, R.A. and N.E. Wallace. The Construction of Residential Housing Price Indices: A Comparison of Repeat-Sales, Hedonic-Regression and Hybrid Approaches. *Journal of Real Estate Finance and Economics*, 1997, 14, 51–73.

Simon, A. Information and Repeat-Sales. Working paper, available on SSRN, 2007.

——. Boundary Effects and Repeat-Sales. Working paper, available on SSRN, 2008.

Wang, F.T., and P.M. Zorn. Estimating House Price Growth with Repeat Sales Data: What's the Aim of the Game? *Journal of Housing Economics*, 1997, 6, 93–118.

Arnaud Simon, Université Paris Dauphine, 75775 Paris Cedex 16 or Arnaud.simon@dauphine.fr.