Quantifying the Reversibility Phenomenon for the Repeat-Sales Index

Arnaud Simon

Quantifying the reversibility phenomenon for the repeat-sales index

Arnaud Simon

Université Paris Dauphine

DRM-CEREG

Place du Maréchal de Lattre-de-Tassigny

75775 Paris Cedex 16

Arnaud.simon@dauphine.fr

Quantifying the reversibility phenomenon for the repeat-sales index

Abstract

The reversibility phenomenon for the repeat-sales index (RSI) is a serious obstacle for the derivatives products; it could hinder their introduction or their success. It is also an undesirable characteristic for the management of the real estate risk. This article provides a general solution for this problem, using an informational reformulation of the RSI framework. We present first a theoretical formula, easy to interpret and easy to handle, before implementing it. Our methodology is robust in the sense that its conclusions are not conditioned by any specific dataset; moreover, the numerical estimations of the reversibility percentages are reliable. For the derivatives our technique has strong implications for the choice of the underlying index. Indeed, even if the reversibility of the RSI is probably higher compared to the hedonic one, this index remains a challenger because of the predictability and the quantifiability of its revisions.

Key words: reversibility, quantification, information, Monte Carlo simulations, Markovian process

1. Introduction

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 the new data in the estimation set that are relevant for the past. This mechanism of revision is an obstacle to the introduction of the derivatives written on a RSI and more generally it is an undesirable characteristic for the management of the 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 et al. (2005) : "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 methodology does not exist in the RSI literature. This article provides a general solution for 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. (2005) one can ask if the empirical results are still valid for a non-Swedish sample. What's more the authors of this article also conclude to the superiority of the hedonic indexes because the reversibility fluctuations are smaller. However they do not provide a methodology which would make the anticipations of these variations conceivable. As we will see, the RSI technique makes possible these estimations. Consequently, even if the reversibility for 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. In the second paragraph we present more precisely the reversibility problem, with a literature review. The results of Clapp and Giaccotto (1999) are the subject of a particular attention. The third section presents firstly the theoretical reformulation of the RSI. Then, this new formalism is applied more

specifically to the reversibility problem; a simple and easy to handle formula is established. The fourth part is devoted to the empirical implementation. In this section, a simulation algorithm is presented and we will also answer to the above problem, mentioned in Clapham et al. (2005), establishing the law for the distributions of the reversibility percentages.

2. The reversibility phenomenon, state of the art

2.1. The phenomenon

One of the specificities of the RSI is its time dependence to 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 not only bring information on the interval $[T_1,T_2]$ but also[1] on $[0,T_1]$. And unfortunately, there is no reason why 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. Figure 1 is an illustration of this instability for the Los-Angles County (Clapp and Giaccotto (1999)) and Figure 2 for Paris (Baroni et al. (2004)). As we can see, the magnitude of the variations can be substantial, up to 10% in Clapp and Giaccotto for example.

2.2. Literature review

The two seminal articles for the repeat-sales technique are Bailey et al. (1963), in an homoscedastic situation, and Case, Shiller (1987) for the heteroscedasctic context. Since these two papers the repeat-sales approach has become of one most popular index because of its quality and its flexibility. It used not only for residential but also for commercial real estate, cf. Gatzlaff, Geltner (1998). One can also refer to Chau et al. (2005) for a recent example of a

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

multisectorial application of the RSI and to Baroni et al. (2004) for the French context. The reversibilty phenomenon had been analysed more specifically by Hoesli et al. (1997), in a two periods model. This very simplified environment allows studying rigorously the mathematics of the RSI; Meese and Wallace (1997) choose for instance the same model in their appendixes. But when the number of dates increases, the RSI equations become quickly burdensome. Clapham et al. (2005) tried to compare the sizes of the reversibility phenomenon for the various index methodologies. They conclude that the hedonic one was probably the less affected, but as their article is empirical their conclusion probably depends on the Swedish sample they used. And this is maybe the major concerns because a robust and general conclusion can only comes from theoretical arguments. Generally, in the literature, the theoretic approach is not the most frequent situation. We can mention the very interesting article of Wang and Zorn (1997), but others examples are globally scarce. For the reversibility problem there is an exception, namely the article of Clapp and Giaccotto (1999) ; their results are presented in the next paragraph.

2.3. The solution of Clapp and Giaccotto (1999)

This paper deals with a BMN[2] context, but its formulas can be generalised to a CS[3] model. The first step consists in running, for the interval $[0,T_1]$, the regression $Y(T_1) = D(T_1)LInd(T_1) + \varepsilon(T_1)$, where the unknown is the vector of the logarithms of the index: $LInd(T_1)$. Within $Y(T_1)$ we have the log-returns realised for the repeat-sales of the sample. The lines of the matrix $D(T_1)$ correspond to each data. In each line +1 indicates the resale date, -1 the purchase date and the rest is made of zeros[4]. In a second step, the estimation interval is extended to $[0,T_2]$ and the regression becomes $Y(T_2) = D(T_2)\ LInd(T_2) + \varepsilon(T_2)$. The

[2] Bailey, Muth and Nourse
[3] Case, Shiller
[4] The purchases at t = 0 are not included to avoid a singular matrix in the estimation.

vector of the log-returns can be written $Y(T_2)' = (Y(T_1)' ; Y(T_2/T_1)')$: the old observations $Y(T_1)$ completed with the new ones $Y(T_2/T_1)$. The matrix $D(T_2)$ is a four blocks 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 hand side corner we have the old matrix $D(T_1)$. The lower part of $D(T_2)$ is associated to the new repeat-sales. $D_1(T_2/T_1)$ is for the transactions realised before T_1 (only purchase in that case) and $D_2(T_2/T_1)$ for the transactions realised after T_1 (purchase and resale). The new data are of two types : purchase before T_1 and resale after T_1, or purchase and resale after T_1. For the first case, the -1 is registered in $D_1(T_2/T_1)$ and the +1 in $D_2(T_2/T_1)$, whereas both are in $D_2(T_2/T_1)$ for the second. We denote $\Delta(T_2) = (D(T_1)' ; D_1(T_2/T_1)')'$ the left part of the matrix and $F(T_2) = (0' ; D_2(T_2/T_1)')'$ its right part. The vector $LInd(T_2)$ gives the logarithms of the index values for the second estimation. It can be separated in two pieces, the first one gives the levels of the index on $[0,T_1]$ and the second on $]T_1,T_2]$: $LInd(T_2)' = (LInd_1(T_2)' ; LInd_2(T_2)')$. The Clapp and Giaccotto's formula establishes the link between the vectors $LInd(T_1)$ and $LInd_1(T_2)$, which both give the index values on the interval $[0,T_1]$, but using only the information embedded in $Y(T_1)$ for $LInd(T_1)$ while $LInd_1(T_2)$ uses the completed dataset $Y(T_2)$. This formula requires an auxiliary regression $Y(T_2/T_1) = D_1(T_2/T_1)AUX + \varepsilon'$. But, even if it looks like to the previous regressions, "AUX is not an index of any kind. It's just the vector of coefficients in the artificial regression of $Y(T_2/T_1)$ on $D_1(T_2/T_1)$" – Clapp,Giaccotto (1999). We also have to introduce a matrix Ω, quite hard to interpret: $\Omega = \left[D(T_1)' D(T_1) + D_1(T_2/T_1)' D_1(T_2/T_1) \right]^{-1} D(T_1)' D(T_1)$. With all these elements the relation for the reversibility is:

$$\mathbf{LInd_1(T_2)} = \Omega\, \mathbf{LInd(T_1)} + (I-\Omega)\, \mathbf{AUX} + \left[\Delta(T_2)'\Delta(T_2)\right]^{-1}\Delta(T_2)'F(T_2)\, \mathbf{LInd_2(T_2)}$$

3. The theoretical solution

3.1. An informational reformulation of the RSI

In this paragraph we summarize briefly the theoretical framework established in Simon(2007).

3.1.1. The classical estimation of the repeat-sales index

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

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

$Index_t$ is the true index value, $G_{k,t}$ is a Gaussian random walk representing the asset's own trend and $N_{k,t}$ is a white noise associated to the market imperfections. If we denote Rate = $(rate_0, rate_1, \ldots, rate_{T-1})$' the vector of the instantaneous continuous rates for each elementary time interval [t,t+1], we have $Index_t = \exp(rate_0 + rate_1 + \ldots + rate_{t-1})$, or equivalently $rate_t = \ln(Index_{t+1}/Index_t)$. For a repeat-sale we can write at the purchase time t_i : $Ln(p_{k,i}) = \ln(Index_i) + G_{k,I} + N_{k,i}$ and at the resale time t_j : $Ln(p_{k,j}) = \ln(Index_j) + G_{k,j} + N_{k,j}$. Thus, subtracting, we get $Ln(p_{k,j}/p_{k,i}) = Ln(Index_j/Index_i) + (G_{k,j} - G_{k,i}) + (N_{k,j} - N_{k,i})$. The return rate realised for the property k is equal to the index return rate during the same period, plus the random walk and the white noise variations. As each repeat-sales give a relation of that nature, we can express them under a matrix form $Y = D*LIndex + \varepsilon$. Here, Y is the column vector of the log return rates realised in the estimation dataset and $LIndex = (\ln(Index_1), \ldots , \ln(Index_T))'$. ε is the error term and D is a non singular matrix[5]. Moreover, if we remark that there exists an invertible matrix[6] A, such that LIndex = A Rate, we can also write[7] $Y = (DA) (A^{-1} LIndex) + \varepsilon$

[5] 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 the 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.

[6] A is a triangular matrix whose values are equal to 1 on the diagonal and under it, 0 elsewhere.

= (DA) Rate + ε. In the estimation process, the true values Index and Rate will be replaced with their estimators, respectively denoted Ind = (Ind_1,… , Ind_T)' and R = (r_0, r_1, …, r_{T-1})'. The usual estimation of Y = D*LIndex + ε or Y = (DA) Rate + ε is carried out in three steps because of the heteroscedasticity of ε. Indeed, the specification of the error term leads to the relation $Var(\varepsilon_k) = 2\sigma_N^2 + \sigma_G^2(j-i)$ 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 sales. Thus, the first step consists in running an OLS that produces a residuals series. These residuals are then regressed on a constant and on the length of the holding period to estimate σ_N, σ_G and the variance-covariance matrix[8] of ε denoted Σ. Finally the last step is an application of the generalised least squares procedure with the estimated matrix Σ. This approach is the traditional one. In Simon (2007) we established that it was equivalent to an algorithmic decomposition (cf. Figure 3), where the informational framework becomes explicit. The next paragraphs present briefly the mechanism of this algorithm.

3.1.2. Notations, basic concepts and decomposition of the RSI

3.1.2.1. Time of noise equality

The variance of the residual ε_k measures the quality of the approximation $Ln(p_{k,j}/p_{k,i}) \approx Ln(Ind_j/Ind_i)$ for the k^{th} repeat-sales . This quantity $2\sigma_N^2 + \sigma_G^2(j-i)$ can be interpreted as a noise measure for each data. As a repeat-sales is compound 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. To make the interpretation easier it is useful to modify slightly the expression of the total noise,

---

[7] The basic rules of linear algebra imply that the matrix DA gets as many lines as the number of repeat sales in the sample, and that the columns correspond to the elementary time intervals. In each line of DA, if the purchase occurs at t_i and the resale at t_j, we have (0 … 0 $1(t_i)$ 1 … $1(t_{j-1})$ 0 … 0). Therefore, the relation Y = (DA) Rate + ε simply means that Log(return) = $rate_i$+ … + $rate_{j-1}$+ ε

[8] ∑ is a diagonal matrix with a dimension equal to the size of the repeat sales sample.

factorising 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, but as this one is constant and the second is increasing regularly with the length of the holding period, we can find a time where the two sources will reach the same levels. Thereafter, the Gaussian noise $G_{k,t}$ will exceed the white noise. This time is the solution of the equation: $2\sigma_N^2 = \sigma_G^2 * \text{time} \Leftrightarrow \text{time} = 2\sigma_N^2/\sigma_G^2 = \Theta$. For that reason, Θ will be called the "time of noise equality". In the below formula the function $G(x) = x/(x+\Theta)$ will sometimes appear. For an holding period j-i we have $G(j-i) = (j-i)/(\Theta+(j-i)) = \sigma_G^2(j-i)/[2\sigma_N^2+\sigma_G^2(j-i)]$. Actually, G(j-i) will represent the proportion of the time-varying noise in the total noise; these numbers will be used subsequently as a system of weights.

3.1.2.2. Quantity of information delivered by a repeat-sale

The theoretical reformulation developed in this article brings to the fore the concept of information. As $\Theta+(j-i)$ is a noise measure, its inverse can be interpreted as an information measure. Indeed, if the noise is growing, that is if the approximation $\text{Ln}(p_{k,j}/p_{k,i}) \approx \text{Ln}(\text{Ind}_j/\text{Ind}_i)$ is becoming less reliable, the inverse of $\Theta + (j-i)$ is decreasing. Consequently, $(\Theta+(j-i))^{-1}$ is a direct measure[9] (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. Within the estimation process, the smaller weights associated to the long holding periods make these observations less contributive to the index values.

3.1.2.3. Subsets and algorithmic decomposition of the RSI

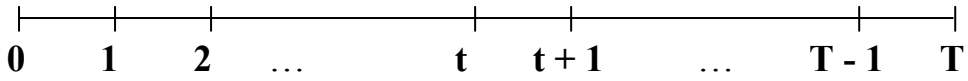
The set of repeat-sales with a purchase at t_i and a resale at t_j will be denoted by $\mathcal{C}(i,j)$. For a time interval [t',t], an observation is relevant only if its holding period includes [t',t] ; in other words, if the purchase occurs at $t_i \leq t'$ and the resale at $t_j \geq t$. This sub-sample will be denoted

[9] These measures are relative ones. The matter is the relative sizes and not the absolute levels. They can be defined dividing by a constant in order to standardize the quantities.

$Spl^{[t',t]}$. For an elementary time-interval [t,t+1], we will also used the simplified notation $Spl^{[t,t+1]} = Spl^{t}$. If we organize the dataset in an triangular upper table, the sub-set $Spl^{[t',t]}$ will correspond to the cells indicated in Table 1. From the optimization problem associated to the general least squares procedure, we demonstrated in Simon (2007) that the repeat-sales index estimation could be realised using the algorithmic decomposition presented in Figure 3. The left-hand side is related to the informational concepts (for example the matrix $\hat{I}$), whereas the right-hand side is associated to the price measures (for example the mean of the mean rates ρ_t). The final values of the index come from the confrontation of these two parts.

3.1.3. The real distribution and its informational equivalent

The time is discretized from 0 to T (the present), and divided in T sub-intervals.



We assume that the transactions occur only at these moments, and not between two dates (the step can be for example a month or a quarter, depending on the data quality). Each observations give a time couple $(t_i;t_j)$ with $0 \leq t_i < t_j \leq T$, thus we have T*(T+1) possibilities for the holding periods. The number of elements in $\mathcal{C}(i,j)$ is $n_{i,j}$, and we denote $N = \sum_{i<j} n_{ij}$ the total number of the repeat-sales in the dataset. Table 2a is a representation of the real distribution of the $\{n_{ij}\}$. As each element of $\mathcal{C}(i,j)$ provides a quantity of information equal to $(\Theta+(j-i))^{-1}$, the total informational contribution of the $n_{i,j}$ observations of $\mathcal{C}(i,j)$ is $n_{i,j}\,(\Theta+(j-i))^{-1} = n_{i,j}\,/(\,\Theta + (\,j - i\,)) = L_{i,\,j}$. Therefore, from the real distribution $\{n_{i,j}\}$ we get directly the informational distribution $\{L_{i,j}\}$ (cf. Table 2b), just dividing the $n_{i,j}$ by $\Theta+(j-i)$. The total quantity of information embedded in the dataset is then $I = \sum_{i<j} L_{ij}$.

3.1.4. Averages for the noise proportions, the periods and the frequencies

The number of repeat-sales included in Spl^t is $n^t = \sum_{i \leq t < j} n_{i,j}$. For an element of $\mathcal{C}(i,j)$, the length of the holding period is j – i. With the function G, we can define the G-mean[10] ζ^t of these lengths in Spl^t by $\sum_{i \leq t < j} \sum_{k,} G(j-i) = n^t G(\zeta^t)$. The first sum enumerates all the classes $\mathcal{C}(i,j)$ 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 $\mathcal{C}(i,j)$, the quantity $G(\zeta^t)$ can also be interpreted as the mean proportion of this Gaussian noise in the global one, for the whole sub-sample Spl^t. In the same spirit, we define the arithmetic average F^t of the holding frequencies 1/(j-i), weighted by the G(j-i), in Spl^t: $F^t = (n^t G(\zeta^t))^{-1} \sum_{i \leq t < j} \sum_{k,} G(j-i)*(1 /(j-i)) = I^t / (n^t G(\zeta^t))$. Its inverse $\tau^t = (F^t)^{-1}$ is then the harmonic average[11] of the holding periods j-i, weighted by the G(j-i), in Spl^t. If at first sight the two averages ζ^t and τ^t can appear as two different concepts, in fact it is nothing of the sort. We always have, for each sub-sample Spl^t , $\zeta^t = \tau^t$.

3.1.5. Two matrixes

The matrix η is a diagonal one, its T diagonal coefficients are: $n^0 G(\zeta^0)$, … , $n^{T-1} G(\zeta^{T-1})$. The i^{th} element gives the number of repeat-sales relevant for [i,i+1], multiplied by $G(\zeta^i)$. Now, if we are working with the interval [t',t+1], a given repeat-sales provides information on it if the purchase is at t' or before and if the resale takes place at t+1 or after. The quantity of information relevant for [t',t+1] is thus $I^{[t',t+1]} = \sum_{i \leq t' \leq t < j} L_{i,j}$ (for an interval [t,t+1] we

[10] We 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 $\{x_1 , x_2 , \dots , x_n\}$, weighted by the $(\alpha_1 , \alpha_2 , \dots , \alpha_n)$, is the number X such that: $\alpha G(X) = \alpha_1 G(x_1) + \alpha_2 G(x_2) + \dots + \alpha_n G(x_n)$ with $\alpha = \sum_{i=1,\dots,n} \alpha_i$. 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

[11] We have $(n^t G(\zeta^t)) / \tau^t = \sum_{i \leq t < j} \sum_{k,} G(j-i) * (1/(j-i)) = I^t$

simply denote I^t for $I^{[t,t+1]}$). As exemplified in Table 3, $I^{[t',t+1]}$ can be calculated buy-side with the partial sums $B_0{}^t$, $B_1{}^t$, … , $B^t_{t'}$ or sell-side with $S^{t'}{}_T$, $S^{t'}{}_{T-1}$, … , $S^{t'}{}_{t+1}$: we have $I^{[t',t+1]} = B_0{}^t + \ldots + B^t_{t'} = S^{t'}{}_T + \ldots + S^{t'}{}_{t+1}$. For all the intervals included in [0,T] we get this way the quantities of information related. These values are arranged in a symmetric matrix Î.

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

3.1.6. The mean prices

Within each repeat-sales class $\mathcal{C}(i,j)$, we calculate the geometric and equally weighted averages of the purchase prices $h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}}$ and of the resale prices $h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}}$. For an elementary time-interval [t,t+1], the relevant classes $\mathcal{C}(i,j)$ are the ones that satisfy to the inequalities $i \le t < j$. With these classes, we calculate the geometric average $H_p(t)$ of the $h_p^{(i,j)}$, weighted by the corresponding $L_{i,j}$ (the total mass of the weights is $I^t = \sum_{i \le t < j} L_{i,j}$):

$$H_p(t) = \left(\Pi_{i \le t < j} (h_p^{(i,j)})^{L_{i,j}} \right)^{1/I^t} = \left(\Pi_{i \le t < j} (\Pi_{k'} p_{k',i})^{1/(\Theta + (j-i))} \right)^{1/I^t}$$

As indicated in the second part, $H_p(t)$ is also the geometric mean of the purchase prices, weighted by their informational contribution $1/(\Theta+(j-i))$, for the investors who were owning real estate during at least [t,t+1]. Similarly, we also define the mean resale price $H_f(t)$:

$$H_f(t) = \left(\Pi_{i \le t < j} (h_f^{(i,j)})^{L_{i,j}} \right)^{1/I^t} = \left(\Pi_{i \le t < j} (\Pi_{k'} p_{k',j})^{1/(\Theta + (j-i))} \right)^{1/I^t}$$

As we can see, $H_p(t)$ can be interpreted as a mean purchase price weighted by the informational activity, buy-side, of the market. The interpretation is the same for $H_f(t)$, with the informational activity of the market, sell-side.

3.1.7. The mean of the mean rates

For a given repeat-sales k' in $\mathcal{C}(i,j)$, with a purchase price $p_{k',i}$ and a resale price $p_{k',j}$, the mean continuous rate realised during its holding period j-i is $r_{k'}^{(i,j)} = \ln(p_{k',j}/p_{k',i}) / (j-i)$. In the subset Spl^t, we calculate the arithmetic mean of these mean rates $r_{k'}^{(i,j)}$, weighted[12] by the G(j-i) : $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} G(j-i) r_{k'}^{(i,j)}$. This value is a measure of the mean profitability of the investment for the people who were owning real estate during [t,t+1], independently of the length of the holding period. The weights in this average depend on the informational contribution of each data. We demonstrate in Simon (2007) that ρ_t can also be written, in a simpler way, with the following formula: $\rho_t = (1/\tau^t) * (\ln H_f(t) - \ln H_p(t))$. For Spl^t, this relation is actually the aggregated equivalent of $r_{k'}^{(i,j)} = \ln(p_{k',j}/p_{k',i}) / (j-i)$, with the harmonic mean of the holding periods τ^t, the mean purchase price $H_p(t)$ and the mean resale price $H_f(t)$. All these averages are weighted by the informational activity of the market. We denote the vector of these mean rates $P = (\rho_0, \rho_1, \ldots, \rho_{T-1})$.

3.1.8. The index and the relation $\hat{I}\, R = \eta\, P$

The global estimation of the RSI can now be realised just solving the equation: $\hat{I}R = \eta P \Leftrightarrow R = (\hat{I}^{-1}\eta)\, P$. The single unknown is the vector $R = (r_0, r_1, \ldots, r_{T-1})'$ of the monoperiodic growth rates of the index. The three others components of this equation ($\hat{I}$, η and P) are calculated directly from the dataset. The main advantages of this formalism are its interpretability and its flexibility: the matrix $\hat{I}$ gives us the informational structure of the

[12] The total mass of these weights is $n^t G(\zeta^t)$

dataset, the matrix η counts the relevant repeat-sales for each time interval [t,t+1] and the vector P provides the levels of profitability of the investment, for the people who are owning real estate at the different dates.

3.2. The reversibility formulas

3.2.1. Notations

We are now going to study how we can deal with the reversibility phenomenon using the above reformulation of the RSI. In all this section we assume that the initial horizon T_1 is extended to $T_2 > T_1$. Table 4 illustrates this extension for the informational distribution. We will keep the same notations, however the horizon will be added as a parameter; for example $H_p(t)$ will be denoted $H_p(t;T_1)$ or $H_p(t;T_2)$ according to the associated horizon of the estimation. There exists two kinds of new repeat sales: the ones with a purchase before T_1 and a resale after T_1 ($i < T_1 < j \leq T_2$), delimited by the continuous lines in Table 4, and those with a purchase and a resale realised between T_1 and T_2 ($T_1 \leq i < j \leq T_2$), delimited by the dotted lines. In Table 4 the relevant repeat sales for [t,t+1], if the horizon is T_1, are represented with a light grey. And if the horizon becomes T_2, the dark grey cells should also be included in this set.

3.2.2. Reversibility for I^t, n^t, $H_P(t)$ and $H_f(t)$

For an interval [t,t+1], $t < T_1$, the quantities of relevant information are $I^t(T_1) = \sum_{i \leq t < j \leq T_1} L_{i,j}$ for the first horizon and $I^t(T_2) = \sum_{i \leq t < j \leq T_2} L_{i,j} = I^t(T_1) + \sum_{i \leq t < T_1 < j \leq T_2} L_{i,j}$ for the second. The sum with $i \leq t < T_1 < j \leq T_2$ corresponds to the additional information (dark grey). If we denote it $I^t(T_2 \backslash T_1)$, we simply get the relation $I^t(T_2) = I^t(T_1) + I^t(T_2 \backslash T_1)$. Similarly, for the real

equivalents of $I^t(T_2)$, $I^t(T_1)$, $I^t(T_2\backslash T_1)$, that is $n^t(T_2)$, $n^t(T_1)$ and $n^t(T_2\backslash T_1)$, we have exactly the same kind of formula : $n^t(T_2) = n^t(T_1) + n^t(T_2\backslash T_1)$. In the following, the notation $T_2\backslash T_1$ will refer to the dataset of the new repeat sales that appear when the horizon is extended.

3.2.3. Reversibility for the mean prices $H_p(t)$ and $H_f(t)$

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

$$[H_p(t,T_1)]^{I^t(T_1)} = \prod_{i \le t < j \le T_1} \left(\prod_{k'} p_{k',i} \right)^{1/(\Theta + (j-i))} \qquad [H_p(t,T_2)]^{I^t(T_2)} = \prod_{i \le t < j \le T_2} \left(\prod_{k'} p_{k',i} \right)^{1/(\Theta + (j-i))}$$

Therefore we have : $[H_p(t,T_2)]^{I^t(T_2)} = [H_p(t,T_1)]^{I^t(T_1)} \prod_{i \le t < T_1 < j \le T_2} \left(\prod_{k'} p_{k',i} \right)^{1/(\Theta + (j-i))}$

Introducing $h_p^{(i,j)}$, this product becomes : $\prod_{i \le t < T_1 < j \le T_2} \left(\prod_{k'} p_{k',i} \right)^{1/(\Theta+(j-i))} = \prod_{i \le t < T_1 < j \le T_2} \left(h_p^{(i,j)} \right)^{L_{i,j}}$

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

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

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: $[H_p(t,T_2)]^{I^t(T_2)} = [H_p(t,T_1)]^{I^t(T_1)} [H_p(t,T_2\backslash T_1)]^{I^t(T_2\backslash T_1)}$. As we can see, the new value $H_p(t,T_2)$ is actually the geometric average between the old value $H_p(t,T_1)$ and a term which 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 we introduce $[H_f(t,T_2\backslash T_1)]^{I^t(T_2\backslash T_1)} = \prod_{i \le t < T_1 < j \le T_2} \left(h_f^{(i,j)} \right)^{L_{i,j}} = \prod_{i \le t < T_1 < j \le T_2} \left(\prod_{k'} p_{k',j} \right)^{1/(\Theta+(j-i))}$

we have: $[H_f(t,T_2)]^{I^t(T_2)} = [H_f(t,T_1)]^{I^t(T_1)} [H_f(t, T_2 \setminus T_1)]^{I^t(T_2 \setminus T_1)}$

3.2.4. Reversibility for τ^t

We study in this paragraph the link between the mean holding periods $\tau^t(T_1)$ and $\tau^t(T_2)$. We have: $I^t(T_2\backslash T_1) = \sum_{i \le t < T_1 < j \le T_2} L_{i,j} = \sum_{i \le t < T_1 < j \le T_2} \sum_{k,} G(j - i) * (1 / (j - i))$. Thus $I^t(T_2\backslash T_1)$ is almost the arithmetic average of the 1/(j-i) weighted by the G(j-i). It just lacks in this formula the total mass of the weights, that is $\sum_{i \le t < T_1 < j \le T_2} \sum_{k,} G(j-i) = n^t(T_2\backslash T_1)\, G(\zeta^t(T_2\backslash T_1))$, with $\zeta^t(T_2\backslash T_1)$ the G-average of the holding periods for the new repeat sales[13]. 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. We can now establish the formal link between $\tau^t(T_1)$ and $\tau^t(T_2)$ 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 $I^t(T_2) = I^t(T_1) + I^t(T_2\backslash T_1)$. We get the formula: $[n^t(T_2)G(\zeta^t(T_2))]/\tau^t(T_2) = [n^t(T_1)G(\zeta^t(T_1))]/\tau^t(T_1) + [n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))]/\tau^t(T_2\backslash T_1)$. And as we have $n^t(T_2)G(\zeta^t(T_2)) = n^t(T_1)G(\zeta^t(T_1)) + n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))$, we can assert that $\tau^t(T_2)$ is simply the harmonic weighted average of $\tau^t(T_1)$ and $\tau^t(T_2\backslash T_1)$.

3.2.5. Reversibility for ρ_t

3.2.5.1. Scalar formula for $t < T_1$

For $t < T_1$ we have[14]: $\rho_t(T_2) = [(1/\tau^t(T_1))*(\ln H_f(t,T_1) - \ln H_p(t,T_1))]*[(I^t(T_1)\tau^t(T_1)) / (I^t(T_2)\tau^t(T_2))]$

$+ [(1/\tau^t(T_2\backslash T_1)) * (\ln H_f(t, T_2\backslash T_1) - \ln H_p(t, T_2\backslash T_1))]*[(I^t(T_2\backslash T_1)\tau^t(T_2\backslash T_1)) / (I^t(T_2)\tau^t(T_2))]$

In the first square brackets we recognize $\rho_t(T_1)$. Moreover, we can easily prove that the third brackets are also equal to $[n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))]^{-1} \sum_{i \le t < T_1 < j \le T_2} \sum_{k,} G(j-i) r_{k,}^{(i,j)}$. This expression

[13] Here also, the quantity $G(\zeta^t(T_2\backslash T_1))$ can be interpreted as the mean proportion of the Gaussian noise in the whole noise, for the new data.

[14] $\rho_t(T_2) = [I^t(T_2)/(n^t(T_2)G(\zeta^t(T_2)))] * \ln[H_f(t,T_2)/H_p(t,T_2)] = [I^t(T_2)/(n^t(T_2)G(\zeta^t(T_2)))] * [\ln H_f(t,T_2) - \ln H_p(t,T_2)]$
$= [I^t(T_1)\ln H_f(t,T_1) + I^t(T_2\backslash T_1)\ln H_f(t,T_2\backslash T_1)] / [I^t(T_2)\tau^t(T_2)] - [I^t(T_1)\ln H_p(t,T_1) + I^t(T_2\backslash T_1)\ln H_p(t,T_2\backslash T_1)] / [I^t(T_2)\tau^t(T_2)]$

is simply the weighted mean of the mean rates $r_{k'}^{(i,j)}$ for the new repeat sales, and of course it will be denoted $\rho_t(T_2\backslash T_1)$. Thus, the reversibility formula for ρ_t , $t < T_1$, is:

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

The 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 for the new ones, relatively to the lengths of the whole sample. If we assume that the average holding periods are all equal, the relation simply becomes: $\rho_t(T_2) = [I^t(T_1)/I^t(T_2)]\rho_t(T_1) + [I^t(T_2\backslash T_1)/I^t(T_2)]\rho_t(T_2\backslash T_1)$.

3.2.5.2. Vectorial formula

The above formulas are valid for $t < T_1$. However the expressions that define $I^t(T_2\backslash T_1)$, $\tau^t(T_2\backslash T_1)$, $\zeta^t(T_2\backslash T_1)$, $n^t(T_2 \setminus T_1)$ and $\rho_t(T_2\backslash T_1)$ can be generalized for $t \geq T_1$. Indeed, for these quantities the sums concern the classes $\mathcal{C}(i,j)$ such that $i \leq t < T_1 < j \leq T_2$, that is the new repeat-sales relevant for [t,t+1], with $t < T_1$. Now, if we choose $t \geq T_1$, the relevant cells will be the ones satisfying[15] to $i \leq t < j \leq T_2$. But, what we get this way is not 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 $I^t(T_2\backslash T_1) = \sum_{i \leq t < T_1 < j \leq T_2} L_{i,j}$ gives for $t \geq T_1$: $\sum_{i \leq t < j \leq T_2} L_{i,j} = I^t(T_2)$. We can now write the formulas in a more synthetic manner. We gather the values $\rho_t(T_2)$, for $0 \leq t < T_2$, in a T_2-vector $P(T_2)$ and the values $\rho_t(T_1)$, for $0 \leq t < T_1$, in a T_1-vector $P(T_1)$. From the vector $P(T_1)$, a T_2-vector is created adding at its end T_2-T_1 zeros; it will be noted in italics $P(T_1)$. The numbers $\rho_t(T_2\backslash T_1)$ are gathered in a T_2-vector $P(T_2\backslash T_1)$, and actually its last T_2-T_1 coordinates are simply equal to $\rho_t(T_2)$. Thus, for $t < T_1$ we have:

$$\tau^t(T_2)\, I^t(T_2)\, \rho_t(T_2) = I^t(T_1)\, \tau^t(T_1)\, \rho_t(T_1) + I^t(T_2\backslash T_1)\, \tau^t(T_2\backslash T_1)\, \rho_t(T_2\backslash T_1)$$

[15] $i \leq T_1 \leq t < j \leq T_2$ 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], we cannot forget it.

$\Leftrightarrow \quad n^t(T_2)\, G(\zeta^t(T_2))\, \rho_t(T_2) = n^t(T_1)\, G(\zeta^t(T_1))\, \rho_t(T_1) + n^t(T_2\backslash T_1)\, G(\zeta^t(T_2\backslash T_1))\, \rho_t(T_2\backslash T_1)$

And for $t \geq T_1$: $\quad n^t(T_2)\, G(\zeta^t(T_2))\, \rho_t(T_2) = n^t(T_2\backslash T_1)\, G(\zeta^t(T_2\backslash T_1))\, \rho_t(T_2\backslash T_1)$

The diagonal matrix $\eta(T_1)$ can be injected in a T_2-matrix, completing it with zeros, and denoted in italics by $\eta(T_1)$. $\eta(T_2)$ is the usual T_2-diagonal matrix and we denote $\eta(T_2\backslash T_1)$ the T_2-diagonal matrix built with $n^0(T_2\backslash T_1)\, G(\zeta^0(T_2\backslash T_1)),\ldots,n^{T_2-1}(T_2\backslash T_1)\, G(\zeta^{T_2-1}(T_2\backslash T_1))$. We can now write simultaneously these two kinds of equations (for $t < T_1$ and for $t \geq T_1$):

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

3.2.6. 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 computed only with the new $L_{i,j}$ (cf. Table 5), and for each interval $[t_i, t_j] \subset [0,T_2]$ they represent the additional quantity of information. $\hat{I}(T_2\backslash T_1)$ can be written with three sub-matrix $\hat{I}_a(T_2\backslash T_1)$, $\hat{I}_b(T_2\backslash T_1)$ and $\hat{I}_c(T_2\backslash T_1)$ (cf. Figure 4). $\hat{I}_a(T_2\backslash T_1)$ 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 $[t_i,t_j] \subset [0,T_1]$ and $[t_i,t_j] \subset [T_1,T_2]$, whereas $\hat{I}_b$ is for the intervals $[t_i,t_j] \subset [0,T_2]$ with $T_1 \subset]t_i,t_j[$; that is the ones which straddle the first horizon T_1. If we inject now the matrix $\hat{I}(T_1)$ in a T_2*T_2 matrix, completed it with zeros and denoted in italics $\hat{\mathit{I}}(T_1)$, the reversibility formula for the informational matrix is simply: $\hat{I}(T_2) = \hat{\mathit{I}}(T_1) + \hat{I}(T_2\backslash T_1)$.

3.2.7. Reversibility for the RSI

The last step consists in 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 $R(T_1)$[16]. 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 families of intermediate measures is known thanks to the 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 $\rho_t(T_2\backslash T_1)$. But if we examine precisely their definitions, we can notice that they are corresponding exactly to the quantities I^t, τ^t, ζ^t, n^t, ρ_t that we can get if the dataset is restricted only to the new repeat sales, as illustrated in Table 5. Thus, it suggests that it is useful to estimate the RSI on the interval $[0,T_2]$ just with the sample $T_2\backslash T_1$; we get this way a T_2-vector $R(T_2\backslash T_1)$[17]. If we now use the general relation $\hat{I}R = \eta P$ and the reversibility formula established for the vector P, $\eta(T_2)P(T_2) = \eta(T_1)P(T_1) + \eta(T_2\backslash T_1)P(T_2\backslash T_1)$, we finally get a very simple reversibility formula for the repeat-sales index :

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

3.3. Comments

The above formalism allows summing up the logic of the reversibility phenomenon as follows. First we estimate the RSI with the old data on $[0,T_1]$; we get an informational matrix $\hat{I}(T_1)$ and a vector $R(T_1)$. Then, only with the new data $T_2\backslash T_1$, we estimate the index on $[0,T_2]$; it gives $\hat{I}(T_2\backslash T_1)$ and $R(T_2\backslash T_1)$. At last, using the whole dataset (old data + new data), we calculate the RSI on $[0,T_2]$ with $\hat{I}(T_2)$ and $R(T_2)$. What is expressed in the reversibility formula is simply that the quantity $\hat{I}\, R$ is additive when the horizon is extended from T_1 to T_2. In their article of 1999, Clapp and Giaccotto proposed a formula to deal with this problem (cf. paragraph 2.3). How should we understand these two different approaches? At the theoretical

[16] In order to have a T_2-vector $R(T_1)$, we will sometimes complete the T_1-vector $R(T_1)$ with T_2-T_1 final zeros.

[17] We saw above that I^t, 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 relation is not true.

level they are of course equivalent because they are measuring the same phenomenon. But in an empirical point of view, things are quite different. The Clapp and Giaccotto's formula is rather complex and its financial interpretation is not obvious. For instance, what does the matrix Ω represent? Moreover, as it is pointed in this article of 1999, the auxiliary regression is just an abstract estimation which does not correspond to an index of any kind. On the other hand, the formula $\hat{I}(T_2)R(T_2) = \hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$ is simple, easy to handle and easy to interpret. The basic concepts are simply the informational matrix and the vector of the monoperiodic growth rates of the index; these two notions are strongly intuitive. What is more, the equivalent of the auxiliary regression AUX, namely $R(T_2\backslash T_1)$, can be interpreted as the RSI for the interval $[0,T_2]$ that we get if we run the estimation only with the new dataset $T_2\backslash T_1$. Thus, this new formula seems to be very interesting for the empirical applications. More generally, our relation could be understood as a kind of « equation of energy preservation » for the datasets. Indeed, if we consider that the product ÎR measures the quantity of energy embedded in a dataset, the reversibility formula simply asserts that:

Energy provided by the whole dataset		**Energy provided by the old data**		**Energy provided by the new data**
$\mathbf{\hat{I}(T_2)\,R(T_2)}$	=	$\mathbf{\hat{I}(T_1)\,R(T_1)}$	+	$\mathbf{\hat{I}(T_2\backslash T_1)\,R(T_2\backslash T_1)}$

This idea of energy delivered by a sample also allows interpreting the relation ÎR = η P. The left hand side can be understood as the energy of the informational system of the index values, whereas the right hand side can be analysed as the energy provided by the gross (real) dataset system. Here also, we have a kind of equation of preservation:

Energy of the informational system		**Energy provided by the real system**
$\mathbf{\hat{I}\,R}$	=	$\mathbf{\eta\,P}$

4. The empirical quantification methodology

Leaning on these theoretical results we are now going to implement a methodology which allows estimating the size of the potential variations due to the reversibility phenomenon.

4.1. The exponential benchmark

In order to simulate the behaviour of the repeat-sales between T_1 and T_2 we introduce a simple model based on an exponential distribution of the resale decision. More precisely, we assume that: 1) the quantities of goods traded on the market at each date are constant and denoted K 2) the buy decisions and the sell decisions are independent between the individuals 3) the length of the holding period follows an exponential distribution, with a parameter $\lambda > 0$ (the same for all the owners). This last hypothesis means that, conditionally to a purchase at t=0, the probability of not having sold the house at time t is $e^{-\lambda t}$. 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[18]. If we introduce the hazard rate[19] which measures the instantaneous probability of a resale: $\lambda(t) = (1/\Delta t)*\text{Prob}(\text{resale} > t+\Delta t \mid \text{resale} \geq t)$, we can demonstrate 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 (cf. Figure 5) we can reasonably think that the hazard rate is first low (quick resales are scarce). In a second time, it increases progressively to a stationary level, maybe modified by the economical context (residential time). Then, as time goes by, the possibility of a moving associated to the retirement or even the death of the householder would bring the hazard rate to a higher level (ageing). However, even if our assumption is not entirely realistic, we keep it because it generates a simple model in which the resale decision could be compared to a

[18] Prob (resale > t | resale ≥ t) = Prob (resale > s | resale ≥ s)

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

radioactive disintegration of an atom. The aim of this benchmark does not consist in describing precisely the reality; we just try to modelize a basic behaviour. For an interval [0,T], the benchmark dataset is fully determined if the parameters K and $\alpha = e^{-\lambda}$ are known. We demonstrated in Simon (2006) that the number of repeat-sales in an exponential sample is[20] $N = K\ T\ (\ 1 - \pi\ )$ and the total quantity of information embedded in this dataset is[21] $I = K' [\ (T+\Theta+1)\ u_T - T\ \pi]$. These two expressions will be useful for the calibration step.

4.2. An example

For practical reasons, we are working in this article with artificial samples, randomly generated[22]. However, the methodology can be applied directly to the real datasets, without any difficulties. Figure 6 presents the results of the estimation when $T_1 = 40$ and $T_2 = 45$. The green curve gives the index values on [0,40], for the old dataset. The yellow curve gives the index values on [0,45], using only the new data $T_2 \backslash T_1$. And the red one is for the completed sample. As for the black curve, it gives 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; the black curve is near zero. It is only in the last quarter of the interval [0,40] that the two curves can diverge; the spread reach up to 1% with our 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, the yellow curve brings the old value (104) at a higher level (105). As we can see the reversibility phenomenon presents a strong temporal framework. It appears essentially for the nearest dates. But unfortunately, in an investment point of view, these recent past values are in

---

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

[21] $K' = K\ (1-\alpha)/\alpha$ $\quad \Theta = 2\sigma_N^2/\ \sigma_G^2$ $\quad u_n = \alpha\ /(\Theta + 1) + \alpha^2\ /(\Theta + 2)\ + \alpha^3\ /(\Theta + 3) + \ldots + \alpha^n\ /(\Theta + n)$

[22] First we first fix the numbers of transactions for each date, for the whole market. Then, the resale rates for each cohort are randomly generated. The estimation sample is made of the repeat-sales with a resale date before T.

general the most important ones. Therefore, it is really crucial to elaborate a methodology able to indicate the level of reliability of the old index values. In other words we are looking for a kind of confidence interval.

4.3. The simulation process

We will use for that purpose a Monte Carlo approach; the simulation algorithm is presented in Figure 7. From a repeat-sales sample on $[0,T_1]$, we calculate the associated index with $R(T_1)$ and $\hat{I}(T_1)$. These two quantities are fixed during all the process. The present time is T_1 and we are trying to infer what could be the variations of the index when the estimation will be renwed at T_2. The first step consists in calibrating the exponential benchmark with the old data on $[0,T_1]$. Precisely, we are looking for the values of the parameters K (constant flow on the market) and α (resale speed) such that the total number of repeat-sales N and the quantity of information I be equal between the real dataset and the benchmark sample[23]. Mathematically speaking, we can estimate the parameter α working with the quantity I/N which does not depend on K (numerical resolution). When α is known, we calculate K with the equation $N = K\,T\,(1 - \pi)$. Once the benchmark is calibrated, we assume that the arrival of the information on the interval $[T_1,T_2]$ will occur according to the same rhythm than previously. We get this way an approximation $\hat{I}_{bench}(T_2\backslash T_1)$ for the matrix $\hat{I}(T_2\backslash T_1)$ which represents the informational distribution of the new repeat-sales[24]. In the same time, we also get the matrix $\hat{I}(T_2)$ adding $\hat{I}(T_1)$ and $\hat{I}_{bench}(T_2\backslash T_1)$. After the left-hand side of Figure 7, devoted to the informational matrixes, we now focus on the right-hand side dedicated to the growth rates vectors. $R(T_1)$ gives the index evolution on the interval $[0,T_1]$. For the rest of the interval $[T_1,T_2]$, we complete it in a T_2-vector $R_{hyp} = (R(T_1), R_{hyp}(T_1;T_2))$ making economical

[23] Others choices are possible for this calibration step, according to the economical contexts.

[24] When K and α are known we demonstrated in Simon (2006) that $L_{i,j} = K'\alpha^{j-i}/(\Theta + j - i)$. We first build the informational distribution of the $\{L_{i,j}\}$ for the benchmark and for the interval $[0,T_2]$. Then, we just keep the columns between T_1 and T_2 (as in Table 5), which represent the new data for the exponential sample. From this partial table, adding its components, we get the matrix $\hat{I}_{bench}(T_2\backslash T_1)$.

hypotheses on the future of the real estate prices. In Simon (2007) we established that the vector $R(T_2\backslash T_1)$, is a Gaussian one. It is centred on the growth rates of the theoretical index values[25] and its variance-covariance matrix[26] is $\sigma_G^2\ \hat{I}(T_2\backslash T_1)^{-1}$. Because of its unobservability at T_1 we have to generate it randomly as a Gaussian vector $\mathcal{N}(R_{hyp}\ ;\ \sigma_G^2\ \hat{I}(T_2\backslash T_1)^{-1})$. The theoretical expectation is replaced with the best estimator that we have on $[0,T_1]$, that is $R(T_1)$, completed with the economical assumption on $[T_1,T_2]$. For the second parameter we simply use the benchmark matrix as an approximation. At this stage of the process we have $\hat{I}(T_1)$, $\hat{I}(T_2\backslash T_1)$, $\hat{I}(T_2)$, $R(T_1)$ and $R(T_2\backslash T_1)$. The final step consists in calculating the vector $R(T_2)$ with the 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 we can calculate the index values $Int_t(T_2)$ on the interval $[0,T_1]$ and we can measure the size of the reversibility phenomenon for this simulation. Running many times this procedure we finally get an empirical distribution for the spreads.

4.4. Comments

In the above process the randomness just appears in the generation of the Gaussian vector $R(T_2\backslash T_1)$. For its practical implementation we have to use the Cholesky factorization[27]. However, if we are interesting in deepening the simulation, we could introduce two additional random sources: the vector $R_{hyp}(T_1\ ;T_2)$ and the couple (K, α). Indeed, in order to estimate the expectation of the vector $R(T_2\backslash T_1)$, we completed the vector $R(T_1)$ with the economical assumptions associated to $R_{hyp}(T_1\ ;T_2)$, foreseeing a scenario for the evolution of the real estate prices on $[T_1,T_2]$. However, as the future is uncertain, it could be reasonable to let these

---

[25] $rate_t = \ln(Index_{t+1}/Index_t)$

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

[27] If Γ is a square matrix of dimension d, symmetric, positive, and with rank r
then we can find a matrix B, dimension d x r , rank r such that $\Gamma = B\ B'$ (Cholesky factorization)
Now, for a vector M of dimension d, and for a square matrix Γ of dimension d, symmetric, positive, rank r with its Cholesky factorization $\Gamma = B\ B'$: If $Y \sim \mathcal{N}(0\ ,\ I_d)$ then $M + B\ Y \sim \mathcal{N}(M\ ,\ \Gamma)$

$T_2 - T_1$ last coordinates fluctuate randomly, rather than restricting them to a single scenario. The second generalisation concerns the couple (K, α). The first variable represents a constant level of liquidity for the market and the second the resale speed. With the calibration step on the interval $[0,T_1]$ we found a mean couple (K_0,α_0). However, on the interval $[T_1,T_2]$ the market conditions could be slightly different. To take this possibility into account we can randomly choose the parameter K in an interval $[K_0 - \varepsilon ; K_0 + \varepsilon]$, and α_0 in $[\alpha_0 - \varepsilon' ; \alpha_0 + \varepsilon']$. We could even go further with this methodology, considering that the rhythm of the transactions depends on the economical context and especially on the future real estate prices. Thus, we should firstly calibrate a proportional hazard model on $[0,T_1]$, as the one developed by Cheung, Yau, Hui (2004). And then, according to the scenario simulated on $[T_1,T_2]$, the rhythm of the repeat-sales could be deduced.

4.5. The theoretical law of reversibility in a simplified context

Working in the simplified context with one random source (paragraph 4.3), we can deepen the mathematical analysis. As previously, a repeat-sales sample ω_0 is generated on the interval $[0,T_1]$. Then the benchmark is calibrated on this dataset, and using the corresponding parameters we get an estimation for the matrix $\hat{I}(T_2\backslash T_1)$. The quantities $\hat{I}(T_1)$, $R(T_1)$, $\hat{I}_{bench}(T_2\backslash T_1)$, $\hat{I}(T_2)$ are fixed and we just have one random source, that is $R(T_2\backslash T_1)$. The vector $R_{hyp}(T_1 ;T_2)$ of the economical assumptions on $[T_1,T_2]$ is also constant in all this paragraph. 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's easy to demonstrate that the vector $R(T_2)$ is Gaussian; we have:

$$E[R(T_2)] = \boldsymbol{R(T_1)} + [\hat{I}(T_2)^{-1} \hat{I}_{bench}(T_2\backslash T_1)] \boldsymbol{R_{hyp}(T_1;T_2)} \quad \text{and} \quad \mathcal{V}[R(T_2)] = \sigma_G^2 [\hat{I}(T_2)^{-1}\hat{I}_{bench}(T_2\backslash T_1)] \boldsymbol{[\hat{I}(T_2)]^{-1}}$$

The matrix $\hat{I}(T_2\backslash T_1)$ represents the new information, $\hat{I}(T_2)$ the total information. Consequently the 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 in the total one. The first formula simply asserts that the expectation of $R(T_2)$ is equal to the old and constant vector $R(T_1)$, plus a quantity

which represents the influence of the economical hypotheses $R_{hyp}(T_1;T_2)$ on $[T_1,T_2]$. This influence of $R_{hyp}(T_1;T_2)$ is weighted by $[\hat{I}(T_2)]^{-1}\ \hat{I}(T_2\backslash T_1)$; a relative measure of the informational weight of the new data. As for the variance expression, we have to compare it to the formula $\mathcal{V}[R(T_2)] = \sigma_G^2\ [\hat{I}(T_2)]^{-1}$ that we would have to apply if we wanted to run the estimation for the index on $[0,T_2]$, directly with the whole dataset, without doing an halfway estimation at T_1. In a reversibility situation we already know a part of the total sample; the resulting index is thus less volatile. What is expressed with the second formula is simply that the attenuation coefficient for the volatility is nothing else than $[\hat{I}(T_2)]^{-1}\ \hat{I}(T_2\backslash T_1)$, once more. Now, if we decide that on $[T_1,T_2]$ the real estate growth is null, in other words $R_{hyp}(T_1;T_2) = 0$, we can demonstrate[28] the following result:

Reversibility law

For $t = 1,\ldots,T_1$ the ratio $Ind_t(T_2) / Ind_t(T_1)$ is log-normally distributed: $\mathcal{LN}(0; v(t))$

$v(t)$ is the t^{th} diagonal element of the matrix[29] $\sigma_G^2 A(T_2)\left[[\hat{I}(T_2)]^{-1}\hat{I}(T_2\backslash T_1)\right]\ [\hat{I}(T_2)]^{-1}\ [A(T_2)]'$

Thus, the reversibility percentage[30] for the date t is a random variable that we can write $100*(Y_t - 1)$, with $Y_t \sim \mathcal{LN}(0; v(t))$. Figure 8 represents the theoretical deciles, anticipated at T_1, using the sample ω_0 on $[0,T_1]$. The black curve gives the observed reversibility for this specific sample when the horizon is extended from T_1 to T_2. As we can see, the theoretical curves are a good approximation of the empirical ones. The size of the potential revisions is small and approximatively constant for the left side of the interval. But for its right side, things are different. As we go closer to T_1 the fluctuations become more and more important, potentially, as testified by the divergence of the theoretical curves in Figure 8. With the

[28] For that purpose we just have to use the relations $LInd(T_2) = A(T_2)\ R(T_2)$ and $E[R(T_2)] = R(T_1)$

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

[30] $100*(Ind_t(T_2) / Ind_t(T_1) - 1)$

methodology developed in this paragraph, it now becomes possible to anticipate and to quantify the reversibility effects in a very reliable way.

5. Conclusion

By means of an informational reformulation of the RSI framework we established first an intuitive and easy to handle formula for the reversibility phenomenon. Then, using an exponential benchmark for the resale decision and Monte-Carlo simulations, we developed a methodology to quantify the size of the potential revisions. In this way we answered to the problem[31] mentioned in Clapham et al. (2005) for the repeat-sales index. For the moment, as we do not have such a similar technique for the hedonic indexes, we cannot assert that the RSI is a bad underlying for the future contracts. Indeed, if its fluctuations are probably higher they are nevertheless predictable, contrary to the hedonic approach. Now, if we want to go further in the derivatives study, the next step would consist in choosing a stochastic dynamic for the RSI in order to price the contingent claims. Unfortunately things are rather complex because of the reversibility. If we consider the basic assumption related to the concept of market efficiency for the stochastic processes in finance, that is their Markovian[32] behaviour, a problem occurs. Is it really possible to describe the dynamic of the RSI with a single Markovian process? The answer is no. We can understand heuristically the problem 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 hand-side measures an increment between the present T_1 and the future T_2. If the Markovian

[31] "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?"

[32] A process is said Markovian if its future depends on its past only through its present. In others words, the path followed by the process to arrive at the level X_s, at the date s, will not influence the probability of realisation of its future X_t ($t > s$). 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 X_s. The market already integrated all the available and relevant information with the fixing of X_s.

assumption is satisfied this variation cannot depend on the dates before T_1. But we know that the right-hand side $\hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$, associated to the new data arrived with the time extension, not only brings information on $[T_1,T_2]$ but also on the interval $[0,T_1]$. Consequently, the RSI do not have a Markovian behaviour. What follows from this result is that we cannot use, at least directly, the usual stochastic dynamics (geometric Brownian motion, Ornstein-Uhlenbeck…) to price a contingent claim. A solution could consist in describing the reversibility risk itself with a dynamic, and then to model the RSI as a noisy asset like in Childs et al. (2001, 2002a, 2002b). Using this approach, we could catch the mechanism of price discovery associated to the reversibility phenomenon. But in spite of everything and even if the technical problems are important, the stakes are real and crucial for the finance industry. It is nothing less than the possibility to price the real estate derivatives written on a RSI. We now conclude this article with two small remarks. In Clapham et al. (2005) we can read: "This suggests that there is a link between the index revision and the sample selectivity of repeat-sales data". This affirmation could clarified because we saw previously that the reversibility phenomenon is inherent and intrinsic to the RSI framework. If the sample selectivity matters it's in a second time, through a specific information content of $\hat{I}(T_2\backslash T_1)$ or $R(T_2\backslash T_1)$. We probably cannot reduce the whole phenomenon to a single sample effect. The final remark concerns the title of the article Case, Shiller (1989): "The *efficiency* of the market for single-family homes". If the RSI is not Markovian, can we really study the efficiency with this index?

Réferences

Bailey, Muth, Nourse. 1963. “A regression method for real estate price index construction”. *Journal of the American Statistical Association* Vol 58

Baroni, Barthélémy, Mokrane. 2004. “Physical real estate : A Paris repeat sales residential index”. *ESSEC Working paper* DR 04007, ESSEC Research Center, ESSEC Business School

Case, Shiller. 1987. “Prices of single family homes since 1970: new indexes for four cities”. *New England Economic Review* September/October 1987 : 45-56

Case, Shiller. 1989. "The efficiency of the market for single-family homes". *The American economic review* 79(1) : 125-137

Chau, Wong, Yiu, Leung. 2005. "Real estate price indices in Hong-Kong" *Journal of real estate literature* 13(3) : 337-356

Cheung, Yau, Hui. 2004. "The effects of attributes on the repeat sales pattern of residential property in Hong-Kong" *Journal of real estate finance and economics* 29(3) : 321-

Childs, Ott, Riddiough. 2001. “ Valuation and information acquisition policy for claims written on noisy real assets”. *Financial Management* summer 2001 : 45-75

Childs, Ott, Riddiough. 2002a. “Optimal valuation of noisy real assets”. *Real estate economics* 30(3) : 385-414

Childs, Ott, Riddiough. 2002b. “Optimal valuation of claims on noisy real assets: theory and application”. *Real estate economics* 30(3) : 415-443

Clapham, Englund, Quigley, Redfearn. 2006. “Revisiting the Past and Settling the Score: Index Revision for House Price Derivatives”. *Real estate economics* 34(2) : 275-302

Clapp, Giaccotto. 1999. “Revisions in repeat-sales price indexes: Here today, gone tomorrow?” *Real estate economics* 27(1) : 79-104

Gatzlaff, Geltner. 1998. "A repeat-sales transaction-based index of commercial property" *A study for the real estate research institute*

Hoesli, Giaccotto, Favarger. 1997. "Three new real estate price indices for Geneva, Switzerland". *Journal of real estate finance and economics* 15(1) : 93-109

Meese, Wallace. 1997. “The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression and hybrid approaches”. *Journal of real estate finance and economics* 14 : 51-73

Simon. 2007. “An informational reformulation of the repeat-sales index”. Working paper, available on demand at arnaud.simon@dauphine.fr

Simon. 2006. “Is there a functional relation between the median index and the repeat-sales index”. Working paper, available on SSRN

Wang, Zorn. 1997. “Estimating house price growth with repeat sales data: What’s the aim of the game?” *Journal of housing economics* 6 : 93-118

Table 1: Relevant repeat-sales for the time interval [t',t]

Purchase date \ Resale date	0	…	t'	…	t	…	T
0	■						
⋮	■	■					
t'	■	■	■				
⋮	■	■	■	■			
t	■	■	■	■	■		
⋮	■	■	■	■	■	■	
T	■	■	■	■	■	■	■

Table 2a: Real distribution for the repeat-sales sample

	0	1	2	3	…	t	t + 1	…	T – 2	T – 1	T
0		$n_{0,1}$	$n_{0,2}$	$n_{0,3}$		$n_{0,t}$	$n_{0,t+1}$		$n_{0,T-2}$	$n_{0,T-1}$	$n_{0,T}$
1			$n_{1,2}$	$n_{1,3}$		$n_{1,t}$	$n_{1,t+1}$		$n_{1,T-2}$	$n_{1,T-1}$	$n_{1,T}$
2				$n_{2,3}$		$n_{2,t}$	$n_{2,t+1}$		$n_{2,T-2}$	$n_{2,T-1}$	$n_{2,T}$
3						$n_{3,t}$	$n_{3,t+1}$		$n_{3,T-2}$	$n_{3,T-1}$	$n_{3,T}$
⋮											
t							$n_{t,t+1}$		$n_{t,T-2}$	$n_{t,T-1}$	$n_{t,T}$
t + 1									$n_{t+1,T-2}$	$n_{t+1,T-1}$	$n_{t+1,T}$
⋮											
T – 2										$n_{T-2,T-1}$	$n_{T-2,T}$
T – 1											$n_{T-1,T}$
T											

Vertical axis: purchase date Horizontal axis: resale date

Table 2b: Informational distribution for the repeat-sales sample

	0	1	2	3	…	t	t + 1	…	T – 2	T – 1	T
0		$L_{0,1}$	$L_{0,2}$	$L_{0,3}$		$L_{0,t}$	$L_{0,t+1}$		$L_{0,T-2}$	$L_{0,T-1}$	$L_{0,T}$
1			$L_{1,2}$	$L_{1,3}$		$L_{1,t}$	$L_{1,t+1}$		$L_{1,T-2}$	$L_{1,T-1}$	$L_{1,T}$
2				$L_{2,3}$		$L_{2,t}$	$L_{2,t+1}$		$L_{2,T-2}$	$L_{2,T-1}$	$L_{2,T}$
3						$L_{3,t}$	$L_{3,t+1}$		$L_{3,T-2}$	$L_{3,T-1}$	$L_{3,T}$
⋮											
t							$L_{t,t+1}$		$L_{t,T-2}$	$L_{t,T-1}$	$L_{t,T}$
t + 1									$L_{t+1,T-2}$	$L_{t+1,T-1}$	$L_{t+1,T}$
⋮											
T – 2										$L_{T-2,T-}$	$L_{T-2,T}$
T – 1											$L_{T-1,T}$
T											

Vertical axis: purchase date Horizontal axis: resale date

Table 3: Relevant repeat-sales for [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	$S^{t'}_{t+1}$	…	$S^{t'}_T$	$I^{[t',t+1]}$

$B_0^t = L_{0,t+1} + \ldots + L_{0,T}$ $\quad B_{t'}^t = L_{t',t+1} + \ldots + L_{t',T}$ sum on the lines (buy-side)

$S^{t'}_{t+1} = L_{0,t+1} + \ldots + L_{t',t+1}$ $\quad S^{t'}_T = L_{0,T} + \ldots + L_{t',T}$ sum on the columns (sell-side)

$$I^{[t',t+1]} = B_0^t + \ldots + B_{t'}^t = S^{t'}_T + \ldots + S^{t'}_{t+1}$$

Table 4: 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									

Continuous lines: new repeat sales with a purchase before T_1 and a resale after T_1 ($i<T_1<j\leq T_2$)

Dotted lines : new repeat sales with a purchase and a resale between T_1 and T_2 ($T_1\leq i < j \leq T_2$)

Relevant repeat sales for [t,t+1] if the horizon is T_1 : light grey.

Relevant repeat sales for [t,t+1] if the horizon is T_2 : light grey + dark grey.

Table 5: Informational distribution for the 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							

Figure 1: Reversibility for the Los Angeles County, Clapp, Giaccotto (1999)

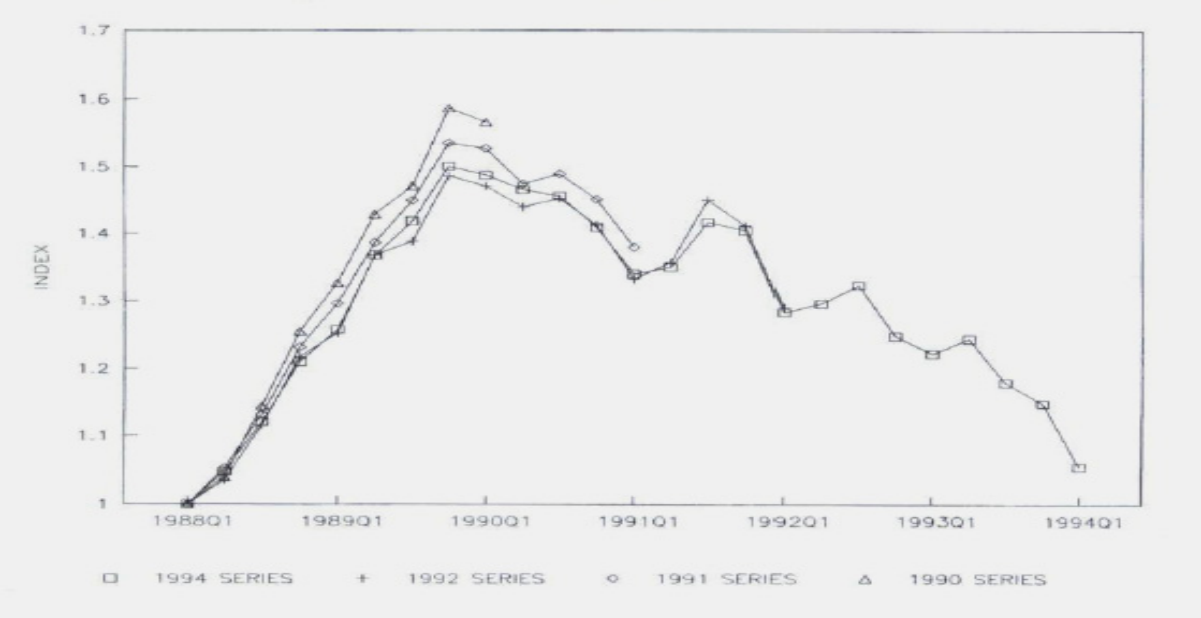


Figure 2: Reversibility for the Paris RSI (the after bubble period) ; Baroni et al (2004)

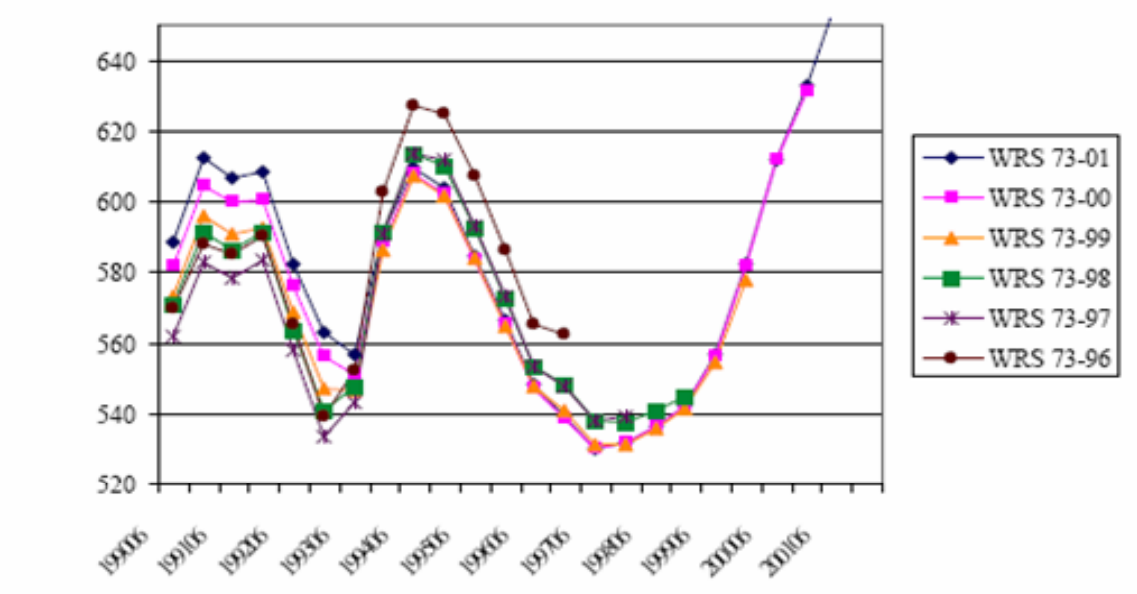
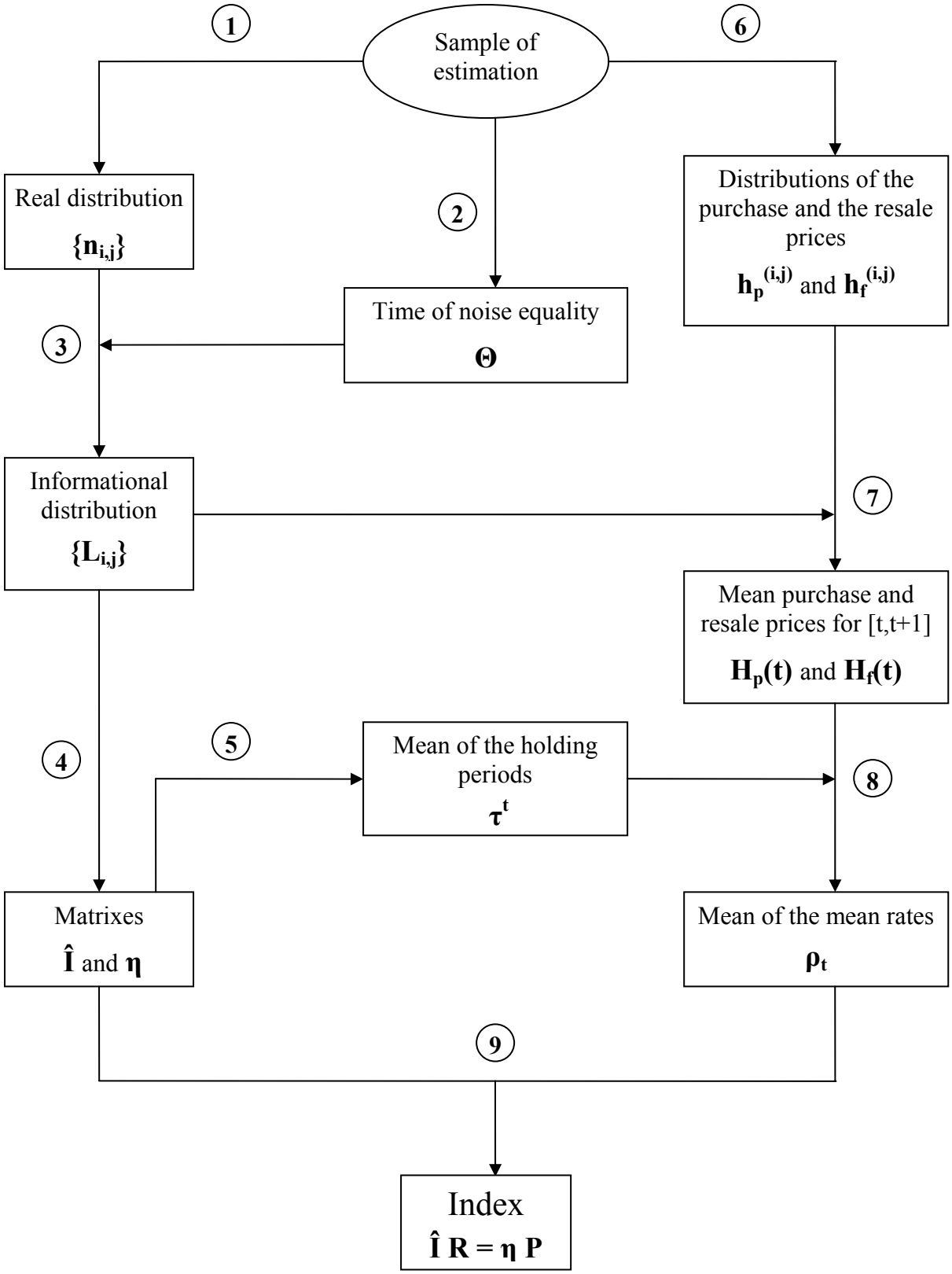


Figure 3: Algorithmic decomposition of the repeat-sales index



Legend of the Figure 3

(1) $\mathbf{n_{i,j}}$: Number of the repeat-sales with a purchase at t_i and a resale et 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 $\mathbf{\Theta} = 2\sigma_N^2 / \sigma_G^2$

(3) $\mathbf{L_{i,j} = n_{i,j} / (\Theta + j - i)}$: 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) We get the matrix $\hat{\mathbf{I}}$ 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 whose holding period is including [t,t']. The diagonal elements of the diagonal matrix $\boldsymbol{\eta}$ are equal to the sums (rows or columns indifferently) of the components of the matrix Î.

(5) Dividing the diagonal elements of Î by the diagonal elements of η we obtain directly the mean holding periods $\boldsymbol{\tau^t}$.

(6) For each repeat-sales class $C(i,j)$, the geometric averages of the purchase prices $\mathbf{h_p^{(i,j)}}$, and the resale prices $\mathbf{h_f^{(i,j)}}$ are:

$$\mathbf{h_p^{(i,j)} = \left(\prod_{k'} p_{k',i} \right)^{1/n_{i,j}} \qquad h_f^{(i,j)} = \left(\prod_{k'} p_{k',j} \right)^{1/n_{i,j}}}$$

(7) For the subset of the people who were owning real estate during [t,t+1], the mean purchase price $\mathbf{H_p(t)}$ (the mean resale price $\mathbf{H_f(t)}$) 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:

$$\mathbf{H_p(t) = \left(\prod_{i \le t < j} \left(h_p^{(i,j)} \right)^{L_{i,j}} \right)^{1/I^t} \qquad H_f(t) = \left(\prod_{i \le t < j} \left(h_f^{(i,j)} \right)^{L_{i,j}} \right)^{1/I^t}}$$

(8) The mean of the mean rates $\boldsymbol{\rho_t}$, realised by the people who were owning real estate during [t,t+1], can be calculated as a return rate with the fictitious prices $H_p(t)$ for the purchase $H_f(t)$ for the resale, and the fictitious holding period τ^t

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

(9) The vector R of the monoperiodic growth rates of the index is the solution of the equation:

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

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

Figure 4: Informational matrixes

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

$$\hat{I}_a(T_2\backslash T_1) = \begin{pmatrix} I^{[0,T_1+1]}(T_2) & I^{[0,T_1+1]}(T_2) & I^{[0,T_1+1]}(T_2) & \cdots & I^{[0,T_1+1]}(T_2) \\ I^{[0,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & \cdots & I^{[1,T_1+1]}(T_2) \\ I^{[0,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & I^{[2,T_1+1]}(T_2) & \cdots & I^{[2,T_1+1]}(T_2) \\ \vdots & \vdots & \vdots & & \vdots \\ I^{[0,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & I^{[2,T_1+1]}(T_2) & \cdots & I^{[T_1-1,T_1+1]}(T_2) \end{pmatrix}$$

$$\hat{I}_b(T_2\backslash T_1) = \begin{pmatrix} I^{[0,T_1+1]}(T_2) & \cdots & I^{[0,T_2]}(T_2) \\ I^{[1,T_1+1]}(T_2) & \cdots & I^{[1,T_2]}(T_2) \\ I^{[2,T_1+1]}(T_2) & \cdots & I^{[2,T_2]}(T_2) \\ \vdots & & \vdots \\ I^{[T_1-1,T_1+1]}(T_2) & \cdots & I^{[T_1-1,T_2]}(T_2) \end{pmatrix}$$

$$\hat{I}_c(T_2\backslash T_1) = \begin{pmatrix} I^{[T_1,T_1+1]}(T_2) & \cdots & I^{[T_1,T_2]}(T_2) \\ \vdots & & \vdots \\ I^{[T_1,T_2]}(T_2) & \cdots & I^{[T_2-1,T_2]}(T_2) \end{pmatrix}$$

Figure 5: standard hazard rate

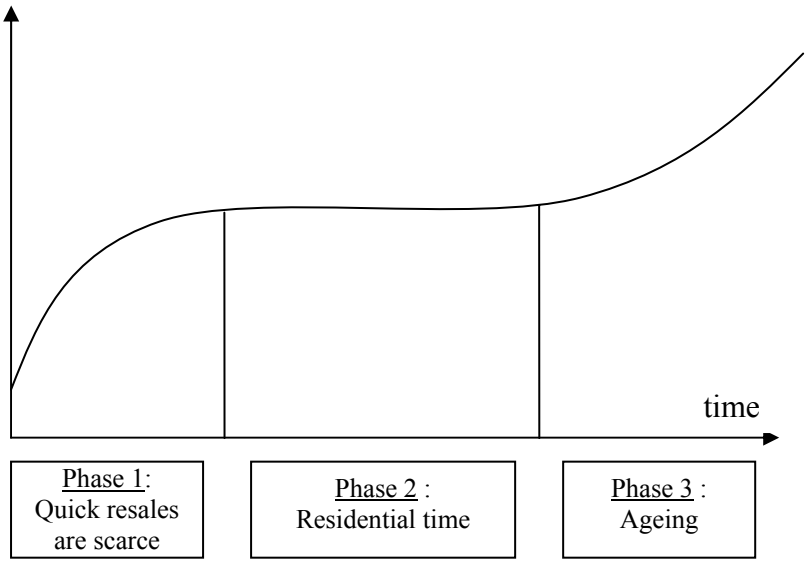
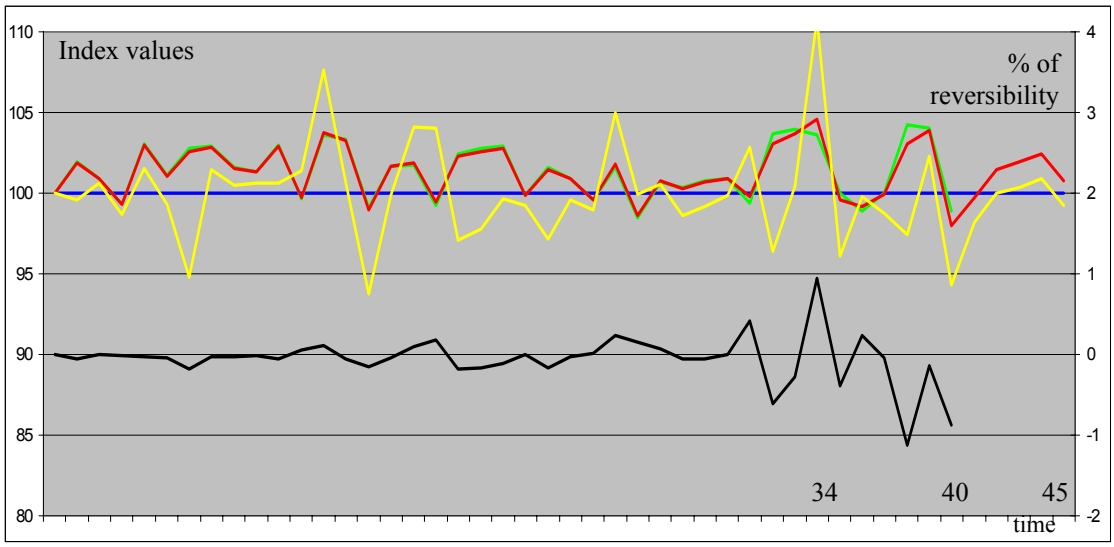


Figure 6: An example of reversibility



Left axis: index values Right axis: percentage of reversibility Blue curve: "true prices"
Green curve: T_1= 40 (old dataset) Red curve: T_2 = 45 (completed dataset) Yellow curve: $T_2 \backslash T_1$ (new data)
Black curve: percentage of reversibility between the red and the green curves

Figure 7: Algorithm for the quantification of the reversibility phenomenon

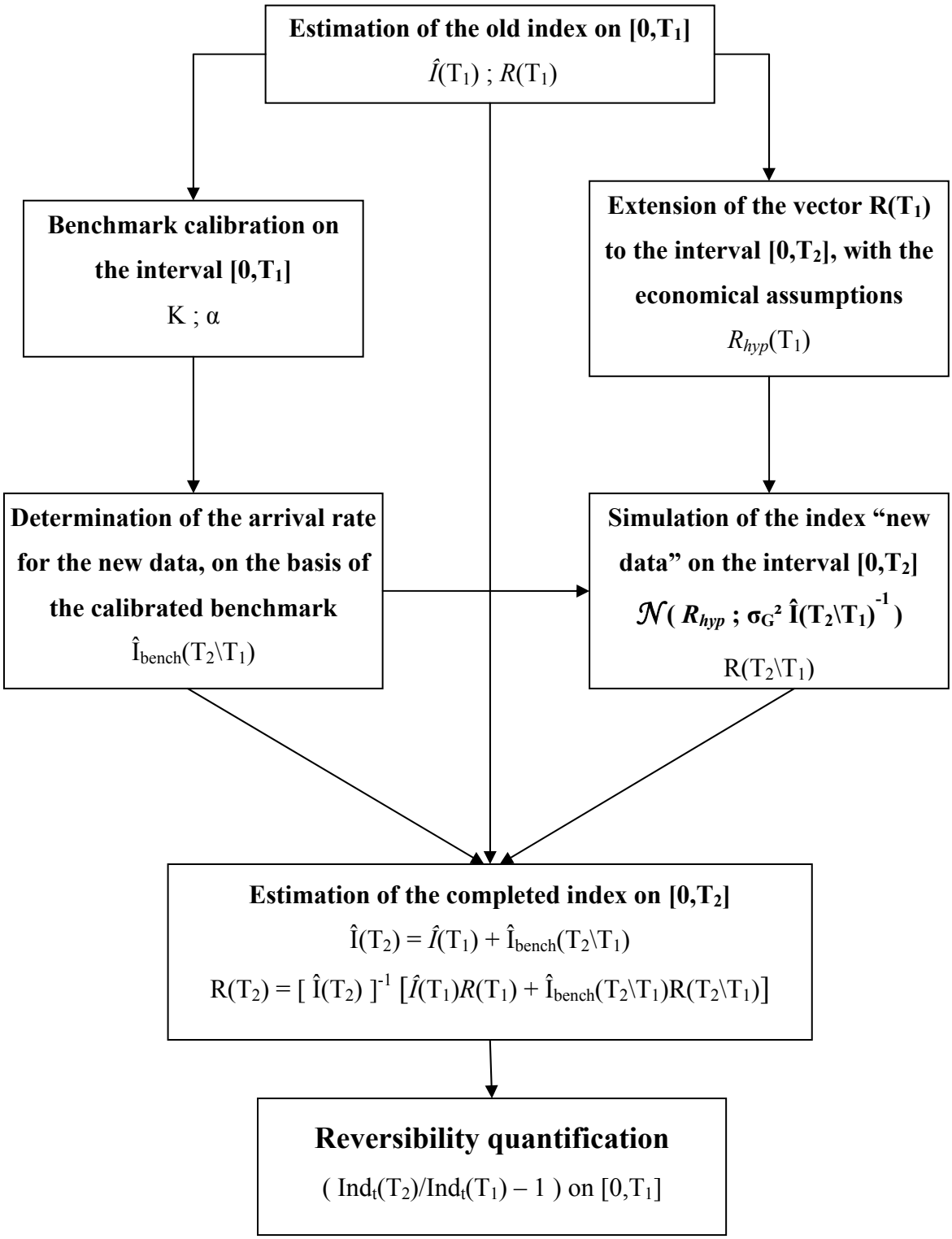
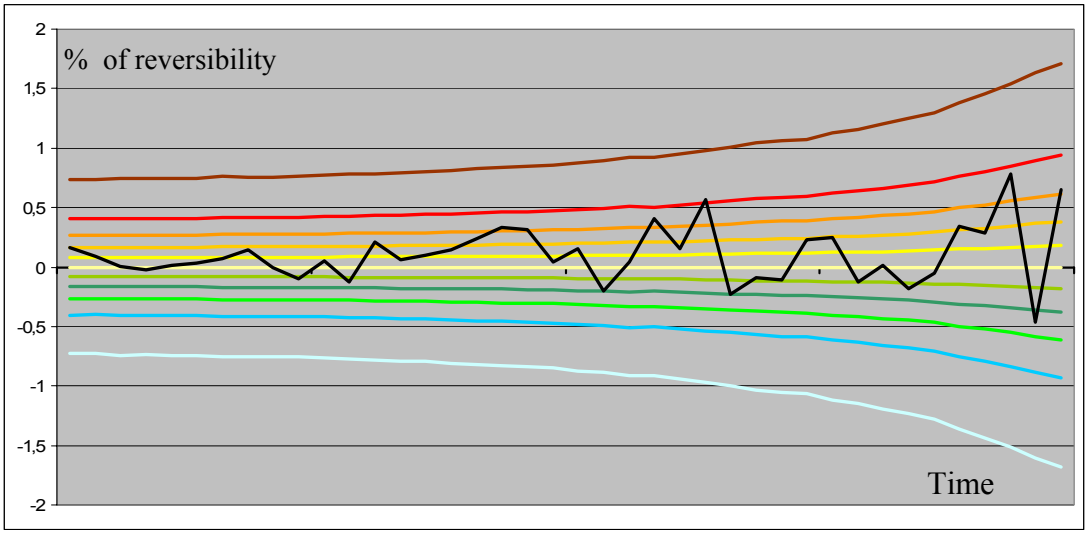


Figure 8: Deciles for the reversibility percentages (t = 1,…,40)



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