Information and repeat-sales

Arnaud Simon

Information and repeat-sales

Arnaud Simon

Université Paris Dauphine

CEREG

Place du Maréchal de Lattre-de-Tassigny

75775 Paris Cedex 16

Arnaud.simon@dauphine.fr

Abstract

What is the informational structure of the repeat-sales index? In this article we answer it, deconstructing the global index in its building blocks. As by-products of this reformulation we establish very simple and intuitive formulas for the volatility of the index and the reversibility phenomenon. We study the formal link between the repeat-sales index and the price indexes (median, hedonic…). We introduce a methodology of data analysis that improves greatly the extraction of the information embedded in a dataset. At last, we give some elements for a more general theory of the informational repeat-sales indexes.

Key words: Repeat-sales index, information, algorithmic decomposition, volatility, reversibility

1. Introduction

Real estate or art markets are highly heterogeneous. In this kind of situation, the market price is usually given by a composite index; in other words a (sophisticated) weighted average of the observed prices or returns. Two natural questions come in mind: "*What is the structure of the information (the structure of the average)?"* and "H*ow should we choose the weights?"* The answer to the first question gives the various index methodologies, cf. Case, Pollakowski, Watcher (1991), while the second gives the variants within each class, cf. Wang, Zorn (1999). Among the traditional approaches we have of course the repeat-sales index (RSI) with the seminal articles of Bailey, Muth, Nourse (1963) and Case, Shiller (1987). Goetzmann (1992) studied the bias problem, Clapp and Giaccotto (1999) the role of the flips and the reversibility phenomenon. But we can also mention the hedonic technique developed by Rosen (1974), the appraisal approach examined critically in Geltner (1991) and the hybrid indexes, cf. Case, Quigley (1991) or Clapp, Giaccotto (1992). Here, in this article, we will only focus on the repeat-sales index. The main goal will be to make explicit its informational framework, that is the way information is defined and the way it is aggregated (paragraph 2). This approach will have some very substantial consequences presented in the paragraph 3. We will establish simple and intuitive formulas for the volatility of the index and for the reversibility phenomenon (an important point for the pricing of the derivatives written on a reversible index). A data analysis methodology will be developed from the algorithmic reformulation of the RSI, improving greatly the extraction of the information embedded in a dataset. We will also study the link between the price indexes and the RSI. At last in paragraph 4, we will suggest a way to generalize the repeat-sales index, transforming it in a family of informational repeat-sales indexes. This approach will allow unifying several article of the literature.

2. The theoretic relation Î R = η P

2.1. 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:

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

- $Index_t$ is the true index value
- $G_{k,t}$ a Gaussian random walk representing the asset's own trend
- $N_{k,t}$ a white noise associated to the market imperfections

The vector Rate = $(rate_0, rate_1, \ldots, rate_{T-1})'$ gives the instantaneous continuous rates for each elementary time interval [t,t+1] and 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, substracting : $Ln(p_{k,j}/p_{k,i}) = Ln(Index_j/Index_i) + (G_{k,j} - G_{k,i}) + (N_{k,j} - N_{k,i})$

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. Each repeat-sale gives a relation of that nature; we can thus express it under a matrix form:

$$Y = D*LIndex + \varepsilon \qquad (2)$$

- Y is the column vector of the log return rates realised in the estimation dataset
- $LIndex = (\ln(Index_1), \ldots , \ln(Index_T))'$
- ε is the error terms

- D is a matrix extracted from another matrix D'; the first column has been removed to avoid a singularity[1] 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 are corresponding to the different possible times for a trade. In each line -1 indicates the purchase date, +1 the resale date and the rest is completed with zeros.

Moreover, if we remark that there exists an invertible matrix[2] A, such that LIndex = A Rate, we can also write : $Y = (DA)\,(A^{-1}\,LIndex) + \varepsilon = (DA)\,Rate + \varepsilon$ **(3)**

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 occurred at t_i and the resale at t_j, we have:

$$\underset{1}{(0} \ \ldots \ \underset{t_i-1}{0} \ \ \underset{t_i}{1} \ \ \underset{t_i+1}{1} \ \ldots \ \underset{t_j-1}{1} \ \ \underset{t_j}{0} \ \ldots \ \underset{T}{0)}$$

Therefore, the relation **(3)** simply means that: $Log(return) = rate_i + \ldots + rate_{j-1} + \varepsilon$

In the estimation process, the true values Index and Rate will be replaced with their estimators, respectively denoted $LInd = (\ln(Ind_1),\ldots, \ln(Ind_T))'$ and $R = (r_0, r_1, \ldots, r_{T-1})'$. The estimation of **(2)** or **(3)** is carried out in three steps because of the heteroscedasticity of ε. The specification of the error term in **(1)** leads to the relation $Var(\varepsilon_k) = 2\sigma_N^2 + \sigma_G^2(j-i)$ where 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. The first step consists in running an OLS that produces a series of residuals. These residuals are then regressed on a constant and on the length of the holding period to get estimations for σ_N, σ_G and for the variance-covariance matrix[3] of ε, denoted Σ. The last step is an application of the generalised least squares procedure for the relation **(2)** with the estimated matrix Σ. Thus, we have to solve the minimisation problem:

---

[1] Cf. Baroni, Bathélémy, Mokrane (2004) for the details
[2] A is a triangular matrix whose values are equal to 1 on the principal diagonal and under it, 0 elsewhere.
[3] ∑ is a diagonal matrix whose dimension is equal to the size of the repeat sales sample

$$\text{Min}_{LInd}\,[(Y - D*LInd)'\,\Sigma^{-1}\,(Y - D*LInd)] \Leftrightarrow \text{Min}_{R}\,[(Y - (DA)\,R)'\,\Sigma^{-1}\,(Y - (DA)\,R)] \quad \textbf{(4)}$$

If we regroup the different time couples (t_i,t_j) in a first sum, and if in a second sum k' refers only to the trades inside this time-class, the Problem **(4)** becomes:

$$\text{Min}_{R}\,\Big[\sum_{i<j}\sum_{k'}\{\ln(p_{k',j}/p_{k',i}) - (r_i + \ldots + r_{j-1})\}^2 / \{(\sigma_G^2(j-i) + 2\sigma_N^2)\}\Big] \quad \textbf{(5)}$$

2.2. Notations, basic concepts and decomposition of the RSI

2.2.1. Time of noise equality

The variance of the residual ε_k actually measures the quality of the approximation $\text{Ln}(p_{k,j}/p_{k,i}) \approx \text{Ln}(\text{Ind}_j/\text{Ind}_i)$. Therefore, the 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 (purchase and 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. It can be useful to modify slightly the expression of the total noise, 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 source provides a constant intensity ($2\sigma_N^2$) and 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, there exists a duration where the two sources will reach the same levels, cf. **Figure 2**. Then, 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, we will call Θ the time of noise equality. In Case-Shiller (1987) four indexes are estimated, **Table 1** gives the different values of the parameter Θ for these cities. As we will see thereafter the function $G(x) = x/(x+\Theta)$ will sometimes appear in the equations. What does this function represent? For an holding period j-i we have

$G(j\text{-}i) = (j\text{-}i)/(\Theta+(j\text{-}i)) = \sigma_G^2(j\text{-}i)/[2\sigma_N^2+\sigma_G^2(j\text{-}i)]$. Actually $G(j\text{-}i)$ will be just the proportion of the time-varying noise in the total noise, these numbers will be used subsequently as a system of weights.

2.2.2. Three simplified situations

In the general algorithmic decomposition of the RSI, some of the mathematical expressions will be quite heavy. To keep the intuition alive, we will also work with three simplified situations. If $\Theta = +\infty$ we have $\sigma_G^2 << 2\sigma_N^2$, it means that the Gaussian noise $G_{k,t}$ is insignificant compared to the white noise $N_{k,t}$ ($\sigma_G^2 = 0$). This first situation is nothing else than the Bailey, Muth and Nourse framework, we will label it Situation-1. Situation-2 is associated to $\Theta = 0$; here the only active noise is the time-varying part ($\sigma_N^2 = 0$). For the great majority of the datasets, we will be between these two extremes cases : $0 < \Theta < +\infty$. The value of Θ indicates if the model is closer to Situation-1 or Situation-2. For instance, according to the **Table 1**, a BMN model is more appropriate for Atlanta ($\Theta = 12.89$) compared to San Fransisco ($\Theta = 4.20$). Finally, in Situation-3, we assume that all the transactions are made at the level of the true index values: $\{Index_t\}$. Here, the dispersal around the theoretical values is ignored. As mentioned above, the different formulas will be exemplified with these three examples. Thus, in the Situation-1 we have $G(x) = x$, in the Situation-2 $G(x) = 1$, and $G(x) = (j\text{-}i)/(\Theta+(j\text{-}i))$ for the Situation-3.

2.2.3. Quantity of information delivered by a repeat-sale

The theoretical reformulation developed in this article brings the concept of information at a crucial place. From where does it come from? If we factorise out σ_G^2 in the optimisation program (5), we get :

$$\text{Min}_R \left[\sum_{i<j} \sum_{k'} \left(\Theta + (j - i) \right)^{-1} \left\{ \ln(p_{k',j} / p_{k',i}) - (r_i + \ldots + r_{j-1}) \right\}^2 \right] \qquad \textbf{(6)}$$

For each observation the square error $\{\ln(p_{k',j}/p_{k',i}) - (r_i + \ldots + r_{j-1})\}^2$ is weighted by $(\Theta+(j-i))^{-1}$. If $\Theta + (j - i)$ 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[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. In the estimation process, the smaller weights for the long holding period make these observations less contributive to the index values. In Situation-1, the quantity of information for a repeat-sales realised between t_i and t_j is 1, and 1/ (j-i) in Situation-2.

2.2.4. Subsets notations

The set of the 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], we will say that an observation is relevant 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 sub-sample will be denoted $\text{Spl}^{[t',t]}$. For an elementary time-interval [t,t+1], we will also used the simplified notation $\text{Spl}^{[t,t+1]} = \text{Spl}^t$. If we organize the dataset in an triangular upper table, the sub-set $\text{Spl}^{[t',t]}$ will correspond to the cells indicated in **Table 2**.

2.2.5. The algorithmic decomposition of the RSI

From the optimization problem **(6)**, we demonstrate in **appendix A, B and D** that the repeat-sales index estimation can be realised using the algorithmic decomposition presented in

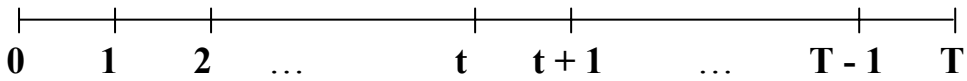
[4] These measures are relative ones. The matter is the relative sizes and not the absolute levels. They can be defined dividing by a constant in order to standardize the quantities.

Figure 1. This figure is the theoretical heart of this paper. The left-hand side is related to the informational concepts (for example the matrix Î, cf. **3.4**), whereas the right-hand side is associated to the price measures (for example the mean of the mean rates ρ_t, cf. 3.7). The final values of the index come from the confrontation of these two parts. The rest of this section define precisely all the building blocks that appear in this decomposition.

2.3. The real distribution and its informational equivalent

2.3.1. The real distribution

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 note $N = \sum_{i<j} n_{ij}$ the total number of the repeat-sales in the dataset. **Table 3a** is a representation of the real distribution of the $\{n_{ij}\}$.

2.3.2. The informational distribution

Each elements of $\mathcal{C}(i,j)$ provide a quantity of information equals to $(\Theta+(j-i))^{-1}$. The total informational contribution of the $n_{i,j}$ observations of this class is $n_{i,j} \times 1/(\Theta + (j - i)) = n_{i,j} /($

$\Theta + (j - i)$), that will be denoted $L_{i,j}$. Therefore, from the real distribution $\{n_{i,j}\}$ of the **Table 3a** we get the informational distribution $\{L_{i,j}\}$, just dividing its elements by $\Theta + (j - i)$. The denominator of $L_{i,j}$ is the sum of the time of noise equality Θ plus the length of the holding period $(j - i)$, for the class under consideration (cf. **Table 3b**). The repeat-sales index being an informational index this distribution will appear frequently in the formulas. The total quantity of information embedded in a dataset will be denoted: $I = \sum_{i<j} L_{ij}$

2.4. 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 $j - i$. Using the function G, we can define the G-mean[5] 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: $F^t = (n^t G(\zeta^t))^{-1} \sum_{i \leq t < j} \sum_{k,} G(j-i)*(1/(j-i))$. Its inverse $\tau^t = (F^t)^{-1}$ is then the harmonic average[6] of the holding period $j-i$, 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, $\zeta^t = \tau^t$ (**cf. appendix B**).

---

[5] 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 $\{x_1, x_2, \ldots, x_n\}$, weighted by the $(\alpha_1, \alpha_2, \ldots, \alpha_n)$, is the number X such that: $\alpha G(X) = \alpha_1 G(x_1) + \alpha_2 G(x_2) + \ldots + \alpha_n G(x_n)$ with $\alpha = \sum_{i=1,\ldots,n} \alpha_i$. An arithmetic mean corresponds to $G(x) = x$, a geometric one to $G(x) = \ln(x)$ and the harmonic average to $G(x) = 1/x$

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

In Situation-1, as G(x) = x, the above formulas are very simple:

$\sum_{i \le t < j} \sum_{k'} (j-i) = n^t \zeta^t$ $\quad F^t = (n^t \zeta^t)^{-1} \sum_{i \le t < j} \sum_{k'} 1 = (\zeta^t)^{-1}$ $\quad (n^t \zeta^t) / \tau^t = \sum_{i \le t < j} \sum_{k'} 1 = n^t$

In Situation-2, as G is no longer increasing, ζ^t and $G(\zeta^t)$ are not relevant. We just have:

$F^t = (n^t)^{-1} \sum_{i \le t < j} \sum_{k'} 1/(j-i)$ $\quad n^t/\tau^t = \sum_{i \le t < j} \sum_{k'} 1/(j-i)$.

Here, the two means F^t and τ^t are simply equally weighted.

2.5. Two matrixes

2.5.1. The matrix η

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 $G(\zeta^i)$. In Situation-1 it becomes $n^i \zeta^i$, and simply n^i in Situation-2.

2.5.2. The informational matrix Î

If we are working with the interval [t',t+1], a 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 $I^{[t',t+1]} = \sum_{i \le t' \le t < j} L_{i,j}$. For an interval [t,t+1] we simply denote I^t for $I^{[t,t+1]}$. As exemplified in **Table 4**, $I^{[t',t+1]}$ can be calculated buy-side with the partial sums B_0^t , B_1^t , … , $B^t_{t'}$ or sell-side with $S^{t'}_T$, $S^{t'}_{T-1}$, … , $S^{t'}_{t+1}$; thus 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. It is the

most important object of this paper because it sums up in a simple table the informational distribution of the whole sample.

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

2.5.3. Mathematical properties of η and Î

As these matrixes play a crucial role in the subsequent developments we precise briefly some of their properties (**cf. Appendix C for the demonstrations**). The matrix elements are indexed with $1 \le p,q \le T$. For $p \le q$ the (p,q) element $\hat{I}_{p,q}$ of $\hat{I}$ is $I^{[p-1,q]}$, and for $p > q$ $I^{[q-1,p]}$.

Proposition 1

- The values are all positive ($\hat{I}_{p,q} \ge 0$) and symmetric ($\hat{I}_{p,q} = \hat{I}_{q,p}$)
- the terms are decreasing in line and column from the diagonal elements :

$\hat{I}_{p,p} \ge \hat{I}_{p,p+1} \ge \ldots \ge \hat{I}_{p,T}$ and $\hat{I}_{p,0} \le \hat{I}_{p,1} \le \ldots \le \hat{I}_{p,p}$ for $p = 0, \ldots, T$

$\hat{I}_{p,p} \ge \hat{I}_{p+1,p} \ge \ldots \ge \hat{I}_{T,p}$ and $\hat{I}_{0,p} \le \hat{I}_{1,p} \le \ldots \le \hat{I}_{p,p}$ for $p = 0, \ldots, T$

Using the definition of $I^{[p-1,\,q]}$ as a partial sum in the table of the $\{L_{i,j}\}$, we can also establish the following relations for the upper[7] side of the matrix:

[7] These relations can be easily adapted for the lower side.

Proposition 2

- $\hat{I}_{p,q} = \hat{I}_{p-1,q} + \hat{I}_{p,q+1} - \hat{I}_{p-1,q+1} + L_{p-1,q}$ for $1 < p \leq q < T$
- $\hat{I}_{1,q} = \hat{I}_{1,q+1} + L_{0,q}$ for $1 \leq q < T$
- $\hat{I}_{p,T} = \hat{I}_{p-1,T} + L_{p-1,T}$ for $1 < p \leq T$

And as $L_{i,j}$ is always positive or null we get as corollary the inequalities[8] :

- $\hat{I}_{p,q} \geq \hat{I}_{p-1,q} + \hat{I}_{p,q+1} - \hat{I}_{p-1,q+1}$ for $1 < p \leq q < T$

The trace of the matrix $\hat{I}$ is $Tr(\hat{I}) = I^0 + I^1 + \ldots + I^{T-1}$. We can also introduce the sum of the diagonal elements just above the principal diagonal: $Tr_{+1}(\hat{I}) = I^{[0,2]} + I^{[1,3]} + \ldots + I^{[T-2,T]}$. With these concepts we establish the following proposition:

Proposition 3

- $Tr(\hat{I}) = N\, G(\zeta)$
- $Tr(\hat{I}) - Tr_{+1}(\hat{I}) = I$
- $\sum_{t' \leq t} I^{[t', t+1]} + \sum_{t' > t} I^{[t, t'+1]} = n^t\, G(\zeta^t)$

What express the first two relations is that we can easily get back the central concepts[9] I (total quantity of information in the dataset) and N (number of repeat-sales in the whole dataset), just reading the matrix $\hat{I}$ diagonally. The third one indicates that the sum of each line (or column since $\hat{I}$ is symmetric) of $\hat{I}$ is equal to the corresponding diagonal elements of η.

[8] The inequalities for $\hat{I}_{1,q}$ and $\hat{I}_{p,T}$ are already known

[9] ζ is the equivalent of ζ^t for the whole sample: $\sum_{i<j} \sum_{k'} G(j - i) = N\, G(\zeta)$

2.6. The mean prices

2.6.1. For the class $\mathcal{C}(i,j)$

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 $h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}}$. With these means the well-known geometric pattern of the RSI appears clearly. In Situation-3, as $p_{k',i} = index_i$ and $p_{k',j} = index_j$, we just have $h_p^{(i,j)} = index_i$ and $h_f^{(i,j)} = index_j$.

2.6.2. For the sub-set Spl^t

For an elementary time-interval [t,t+1] the relevant classes $\mathcal{C}(i,j)$ are the ones that satisfy to the inequalities $i \leq t < j$. 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 $I^t = \sum_{i \leq t < j} L_{i,j}$):

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

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:

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

In Situation-1 (BMN), as the information is constant between all the repeat-sales, we simply have two equally weighted averages of the purchase and resale prices. In Situation-3, we get $H_p(t) = \left(\Pi_{i \leq t < j} (index_i)^{L_{i,j}} \right)^{1/I^t}$ and $H_f(t) = \left(\Pi_{i \leq t < j} (index_j)^{L_{i,j}} \right)^{1/I^t}$. Rearranging these expressions we also have $H_p(t) = \left[\prod_{i \leq t} (index_i)^{B_i^t} \right]^{1/I^t}$ and $H_f(t) = \left[\prod_{j > t} (index_j)^{S_j^t} \right]^{1/I^t}$. These formulas are interesting because they allow deepening the intuition of $H_p(t)$ and $H_f(t)$. The mean purchase

price $H_p(t)$ in Spl^t is actually an average of the past values of the theoretical index ($index_i$ for $i \leq t$). The weights are equal to the informational contributions of the repeat-sales of Spl^t with a purchase at the corresponding dates: B_i^t (cf. **Table 4**). Thus, $H_p(t)$ can be interpreted as a mean purchase price weighted the informational activity, buy-side, of the market. The interpretation is the same for $H_f(t)$ with the future values, that is the resale dates, and the informational activity of the market (sell-side).

2.7. 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 on 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 by the G(j-i) whose total mass is $n^t G(\zeta^t)$: $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \leq t < j} \sum_{k'} G(j-i) r_{k'}^{(i,j)}$. 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. In Situation-1 we have $\rho_t = (n^t \zeta^t)^{-1} \sum_{i \leq t < j} \sum_{k'} (j-i) r_{k'}^{(i,j)}$, in Situation-2 $\rho_t = (1/n^t) \sum_{i \leq t < j} \sum_{k'} r_{k'}^{(i,j)}$, and in Situation-3 $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \leq t < j} L_{i,j} \ln(index_j / index_i)$. We demonstrate in **appendix D** that ρ_t can be expressed in all the cases (specific and general), in a simplest way, with the following formula:

$$\rho_t = (1/ \tau^t) * (\ln H_f(t) - \ln H_p(t))$$

Within Spl^t, this relation is 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 by $P = (\rho_0, \rho_1, \ldots, \rho_{T-1})$.

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

As shown in **appendix A**, the estimation of the RSI can be realised solving the equation:

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

The only 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) can be calculated directly from the dataset. In the traditional procedure employed to calculate the RSI we use the regression $Y = (DA)\ Rate + \varepsilon$, with an heteroscedastic variance-covariance matrix $\sum$ for the residuals ε. The solution of this problem is well-known and can be written directly in a matrix form: $R = [(DA)'\ \sum^{-1}\ (DA)]^{-1}\ (DA)'\ \sum^{-1}\ Y$ or equivalently with the normal equations $[(DA)'\sum^{-1}(DA)]\ R = (DA)'\sum^{-1}\ Y$. This mathematical formulation can be understood as a projection on a vectorial space but its financial interpretation is not really obvious. For instance, the matrix $[(DA)'\ \sum^{-1}\ (DA)]$ is not very intuitive, in economical terms. However, if we look at this product more precisely we can notice that it is simply the informational matrix $\hat{I}$ introduced previously. On the other hand, the term $(DA)'\sum^{-1}\ Y$ is nothing else than $\eta\ P$. Consequently, the central relation of this paper, $\hat{I}\ R = \eta\ P$, is just another way to write the normal equations of the weighted least squares procedure. Therefore, one can ask whether the model developed above really brings something new?

Actually, the main advantage of this formalism is its interpretability: the matrix $\hat{I}$ gives us the informational structure of the dataset, the matrix η counts the relevant repeat-sales for each time interval [t,t+1] and the vector P indicates the levels of profitability of the investment for the people who are owning real estate at the different dates. These three components have a very clear economical meaning. The regression technique is a very general method which can be applied in a great number of situations. What we have done here is just to specify the model within our particular situation. It allowed integrating all the details that could not have been introduced if we chose to confine us to the traditional model. Simultaneously, we tried to

deconstruct the framework, opening the black box, in order to understand how it works inside. The repeat-sales technique now appears as a very simple, flexible and easy to handle model. Indeed, we will see in the rest of the article that this formalism can be used directly to study the volatility of the index and the reversibility phenomenon (two very classical issues of the repeat-sales literature). We will also explore the theoretical link between a price-index and the RSI. A methodology improving the extraction of the information embedded in a dataset will be presented on this basis, briefly. And finally we will generalize the RSI with the global concept of the informational repeat-sales indexes.

3. Four direct consequences

3.1. Volatility

We study here the volatility of the index. A very simple and intuitive formula will be established between the variance-covariance matrix of the vector R and the informational matrix Î. The idea of the demonstration is basic: we will use the algorithmic decomposition of the index (Figure 1) and we will establish the volatility formulas for the various building blocks, starting from the simplest one and finishing with the aggregative quantities (the method will be the same for the reversibility formula, **paragraph 4.4**).

3.1.1. Independency assumptions on $G_{k,t}$ and $N_{k,t}$

G_k is a Gaussian random walk; that is $G_{k,t} \sim \mathcal{N}(0 ; \sigma_G^2 t)$ and the increments of these processes are independent. We denote $G_{k,0}$ the starting point. Moreover, we assume that:

(H1) For a sufficiently large set of heterogeneous properties, the mean of the initial values $G_{k,0}$ is close to 0

(H2) The processes (G_k) are globally independent.

In the literature these assumptions are in general not specified. But here, in order to get our simple formula, they are essential. If (H1) is not satisfied, it would mean that for a large set of heterogeneous properties we have a common component. This part should then be captured by the index dynamic and not by the specific parts of the price processes. Thus, if (H1) is false, the index is not efficient. The assumption (H2) is often implicit in literature. Indeed, the estimation of the index is performed with the relation $Ln(p_{k,j}/p_{k,i}) = \ln(indice_j/indice_i) + (G_{k,j}-G_{k,i}) + (N_{k,j}-N_{k,i}) = \ln(indice_j/indice_i) + \varepsilon_k$ and the variance-covariance matrix of the residuals ε_k is diagonal. Even if correlation is a smaller requirement compared to independence, we can consider that (H2) is a classical assumption[10]. Regarding the white noises N_k we assume that they are Gaussian: $N_{k,t} \sim \mathcal{N}(0 ; \sigma_N^2)$ and for each k $N_{k,t}$ and $N_{k,t'}$ are independent ($t \neq t'$). Moreover, we will use the hypothesis (H3) :

(H3) The processes $(N_k)_{k=0,\dots,N}$ are globally independent of the $(G_k)_{k=0,\dots,N}$

3.1.2. The volatility formulas

The theoretical price decomposition is $\ln(p_{k,t}) = \ln(Index_t) + G_{k,t} + N_{k,t}$. In Situation-3, we assume that all the transactions are realised at the levels of the true index values: $\{Index_t\}$. In such a case, the dispersal around the theoretical values is null and the estimators of the index values $\{Ind_t\}$ are simply equal to the true values[11] $\{Index_t\}$. In the general situation the various estimators are random variables because their values depend on the dataset; however they will be centred on their theoretical values. Proposition 4 sums up the volatility behaviour of the different building blocks (demonstration : appendix E).

---

[10] We could mention an objection to (H2). If two properties are in the same street, their idiosyncratic components are probably correlated (cf. spatial correlation models). However, we choose not to deal with this issue here.

[11] The solution of the optimisation problem **(6)** is obvious: $r_t = rate_t$ for all t.

Proposition 4

- $p_{k,t} = \text{indice}_t \; C(p_{k,t})$ $\qquad C(p_{k,t}) \sim \mathcal{LN}\left(G_{k,0} ; \sigma_G^2 \, (\Theta/2 + t)\right)$
- $r_k^{(i,j)} \sim \mathcal{N}\left(r^{(I,j)} ; \sigma_G^2 / ((j-i)G(j-i)) \right)$ $\qquad r^{(i,j)} = \ln(\text{Indice}_j/\text{Indice}_i)/(j-i)$
- $h_p^{(i,j)} = \text{indice}_i \; C(h_p^{(I,j)})$ $\qquad C(h_p^{(i,j)}) \sim \mathcal{LN}\left(0 ; \sigma_G^2(\Theta/2 + i) / n_{i,j}\right)$
- $h_f^{(i,j)} = \text{indice}_j \; C(h_f^{(I,j)})$ $\qquad C(h_f^{(i,j)}) \sim \mathcal{LN}\left(0 ; \sigma_G^2(\Theta/2 + j) / n_{i,j}\right)$
- $\rho_{i,j} \sim \mathcal{N}\left(r^{(i,j)} ; \sigma_G^2 / (L_{i,j} (j - I)^2) \right)$ $\qquad \rho_{i,j} = \left(\sum_{k'} r_{k'}^{(i,j)} \right) / n_{i,j}$
- $H_p(t) = \left[\prod_{I \le t} \text{indice}_i^{B_i^t} \right]^{1/I^t} \; C(H_p(t))$ $\qquad C(H_p(t)) \sim \mathcal{LN}\left(0;\sigma_G^2 (1/I^t)^2 \sum_{i \le t < j} V_{i,j}(\Theta/2+ i)\right)$

  with $V_{i,j} = L_{i,j} / (\Theta+(j-i)) = n_{i,j} / (\Theta+(j-i))^2$
- $H_f(t) = \left[\prod_{j > t} \text{indice}_j^{S_j^t} \right]^{1/I^t} \; C(H_f(t))$ $\qquad C(H_f(t)) \sim \mathcal{LN}\left(0;\sigma_G^2 (1/I^t)^2 \sum_{i \le t < j} V_{i,j}(\Theta/2+ j)\right)$

  with $V_{i,j} = L_{i,j} / (\Theta+(j-i)) = n_{i,j} / (\Theta+(j-i))^2$
- $\rho_t \sim \mathcal{N}\left((n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j-i) \, n_{i,j} \, r^{(I,j)} ; \sigma^2_G / ((\tau^t)^2 I^t)\right)$ $\qquad \text{Cov}(\rho_t;\rho_{t'}) = \sigma_G^2 \, I^{[t',t+1]} / (\tau^t \tau^{t'} I^t I^{t'})$
- $R \sim \mathcal{N}\left(\text{Rate} ; \sigma_G^2 \, \hat{I}^{-1} \right)$ $\qquad$ Rate : vector of the true rates.
- $\text{Lind} \sim \mathcal{N}\left(\text{Lindex} ; \sigma_G^2 \, (\mathcal{T} - (\mathcal{L} + \mathcal{L}'))^{-1} \right)$

$$\mathcal{L} = \begin{pmatrix} 0 & L_{1,2} & L_{1,3} & \dots & L_{1,T} \\ 0 & 0 & L_{2,3} & & L_{2,T} \\ | & & & & | \\ 0 & 0 & 0 & \dots & L_{T-1,T} \\ 0 & 0 & 0 & & 0 \end{pmatrix} \qquad \mathcal{T} = \begin{pmatrix} B_1^1 + S_1^0 & 0 & \dots & 0 & 0 \\ 0 & B_2^2 + S_2^1 & \dots & 0 & 0 \\ | & & & & | \\ 0 & 0 & \dots & B_{T-1}^{T-1} + S_{T-1}^{T-2} & 0 \\ 0 & 0 & \dots & 0 & S_T^{T-1} \end{pmatrix}$$

and $\mathcal{V}(\text{Lind}) = \sigma_G^2 \, (\mathcal{T} - (\mathcal{L}+\mathcal{L}'))^{-1} = \sigma_G^2 \, \mathcal{T}^{-1} \sum_0^{+\infty} [(\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1}]^i$

3.1.3. Comments

R and Lind are unbiased estimators of the theoretical vectors Rate and Lindex. For R the result is very simple and intuitive: the variance-covariance matrix $\mathcal{V}I$ is equal to the inverse of the informational matrix $\hat{I}$ (with the coefficient σ_G^2). In other words, the estimation error for the index is equal to the inverse of the information. For a single repeat-sales we defined above the quantity of noise associated to an observation: $2\sigma_N^2 + \sigma_G^2(j-i)$, and its informational content: $(\Theta+(j-i))^{-1}$. Here also, we have the same kind of relation: error (noise) = $2\sigma_N^2+\sigma_G^2(j-i)$ = $\sigma_G^2 / (\Theta+(j-i)) = \sigma_G^2$ (information)$^{-1}$. Thus, the formula $\mathcal{V}(R) = \sigma_G^2 \hat{I}^{-1}$ is nothing else than the aggregative aspect of the relation between information and noise. The concepts are perfectly coherent from the elementary level of the goods to the index level. In the variance-covariance matrix of Lind two new matrixes appear: $\mathcal{L}$ and $\mathcal{T}$. $\mathcal{L}$ is equal to the informational table of the $\{L_{i,j}\}$, in which the first line had been removed, and completed with zeros. $\mathcal{T}$ is a diagonal matrix. Its elements give the quantities of information delivered by the transactions realised at each date, independently of the direction[12] of the transactions (purchase or resale). The second expression of $\mathcal{V}$(Lind) is interesting to study the sign of the covariances. As all the elements of the matrixes $\mathcal{T}^{-1}$ et $\mathcal{L} + \mathcal{L}'$ are positive, it also true for $[(\mathcal{L} + \mathcal{L}') \mathcal{T}^{-1}]^1$ for each i. Therefore, all the values within $\mathcal{V}$(Lind) are positive. The estimated vector Lind will be completely above or completely below Lindex more frequently than the situations where it will be oscillating around Lindex. However there is no systematic bias.

[12] For instance, for t = 2, B^2_2 is the quantity of information delivered by the repeat-sales with a purchase at t = 2, and S^1_2 the quantity of information for the repeat-sales with a resale at t = 2.

3.1.4. Index bias, examples

Lind is a Gaussian vector, centred on Lindex, with a variance-covariance matrix $\mathcal{V}$(Lind). Consequently, if we denote $diag_t[\mathcal{V}(Lind)]$ the t^{th} diagonal elements of $\mathcal{V}$(Lind), for each time t we have $Ind_t \sim \mathcal{LN}(Lindex_t ; diag_t[\mathcal{V}(Lind)])$. It is a well-known fact that the RSI presents a bias (Goetzmann 1992; Goetzmann, Peng 2002). With our notations the multiplicative bias can be computed exactly[13]: $\exp(\frac{1}{2} diag_t[\mathcal{V}(Lind)])$. In order to illustrate these results, we simulated a repeat-sales dataset ω_0 on the time interval [0,40] and we calculated its informational matrix Î. From this matrix, we computed the correlograms[14] for each r_t using the formula $\mathcal{V} = \sigma_G^2 \hat{I}^{-1}$. Figure 3a gives the correlogram for r_0, r_{13} and r_{39}. We have of course $Corr(r_0 ; r_0) = Corr(r_{13} ; r_{13}) = Corr(r_{39} ; r_{39}) = 1$ but the most interesting fact is that the coefficients of correlation of first order[15] are near -0,5. For the higher orders they are almost null. Figure 3b presents the theoretical biases in percentages of the true values[16]. As we can see the deviation is not uniform. It is smaller in the middle of the estimation interval and higher near the boundaries. Moreover it is not symmetric; the bias is more important for the right side than for the left side.

3.2. Is there a functional relation between a price index and the repeat-sales index?

3.2.1. Price-based index and return-based index

What is the formal relation between an index based on the returns, like the RSI, and an index based on the prices (hedonic, median…)? For a repeat-sale k' with a purchase at t_i and a resale

[13] If $X \sim \mathcal{LM}(p_1,p_2)$ then $E(X) = \exp(p_1 + p_2/2)$

[14] For r_t , the correlogram is the series of the coefficients of correlation between r_t and $r_{t'}$ (t' = 0,…, 40)

[15] Intuitively, the value -0,5 can be explained in the following way. We have $r_t = \ln(p_{t+1}) - \ln(p_t)$ et $r_{t+1} = \ln(p_{t+2}) - \ln(p_{t+1})$. In the simulation process the repeat-sales prices are generated independently. However for each data we will always have a common part between r_t and r_{t+1}: $\ln(p_{t+1})$. Thus, approximately half of the volatility of r_t and r_{t+1} comes from the same source.

[16] $100 * (E(Ind_t) / Indice_t - 1) = 100 * (\exp(\frac{1}{2} diag_t[\mathcal{V}(LInd)]) - 1)$

at t_j, the return is a function of the purchase and the resale prices: $r_{k,}^{(i,j)} = \ln(p_{k',j}/p_{k',i})/(j-i)$. The price $p_{k',I}$ contributes to the value of any price index at t_i: M_i. $p_{k',j}$ does the same for M_j at time t_j (Figure 4) and the value $r_{k,}^{(I,j)}$ contributes to the RSI. Thus, one can ask if the mathematical relation at the goods level has an equivalent at the index level. In other words, can we find a function F such that: $RSI = F(M_i, M_j)$?

3.2.2. A simplified situation

We examine first this issue with the Situation3. For each dates the transaction prices are equal for each observations: $p_{k',t} = Index_t$. Consequently any price index M_t is equal to $Index_t$. For the RSI we know that $H_p(t) = \left[\prod_{I \leq t}(index_i)^{B_i^t}\right]^{1/I^t}$ and $H_f(t) = \left[\prod_{j > t}(index_j)^{S_j^t}\right]^{1/I^t}$, that is $H_p(t)$ and $H_f(t)$ are geometric averages of the past and the future values of the price index M. As $\rho_t = (1/\tau^t)\,(\ln H_f(t) - \ln H_p(t))$, the vector P is also a function of these values and it is also true for the vector R with the relation $\hat{I}\,R = \eta\,P \Leftrightarrow R = (\hat{I}^{-1}\eta)\,P$. Of course, in Situation3 we do not really need to use these arguments to establish the formal relation because we already know that the RSI values are just equal to $\{Index_t\}$. However, this approach is interesting because it gives the intuition of the solution for the general situation.

3.2.3. The general solution

Here, the formulas for $H_p(t)$ and $H_f(t)$ are a little bit more complicated[17], but we are going to see that these two quantities are strongly related to a specific price index. If we want to

---

[17] $[H_p(t)]^{I^t} = \prod_{i \leq t < j}\left[\left(\prod_{k'} p_{k',i}\right)^{1/(\Theta+(j-i))}\right]$ and $[H_f(t)]^{I^t} = \prod_{i \leq t < j}\left[\left(\prod_{k'} p_{k',j}\right)^{1/(\Theta+(j-i))}\right]$

calculate a price index from the repeat-sales sample we have to use all the transaction prices observed at time t, whatever be their nature (purchase or resale). This subset, denoted E_t, is represented in **Figure 5a**. It includes all the repeat-sales with a resale at t or a purchase at t but we cannot use it directly. Indeed, if a good is bought at 0 and sold at t to a new house-owner, this one could perfectly choose to resale it at T-1. In that situation the price at t, is registered in the cell corresponding to $n_{0,t}$ but also in $n_{t,T-1}$ and there is no reason why we should count twice this value. Thus, we remove all the redundancies from E_t: it gives us a subset F_t of E_t. We denote $q_t(i)$ the transaction values (purchase or resale) in F_t and $inf_t(i)$ their informational weights.[18] We define now the price-index value at t, M_t, as the geometric average of the $q_t(i)$ weighted by the $inf_t(i)$: $M_t^{Inf_t} = \Pi_{F_t} [\, q_t(i) \,]^{inf_t(i)}$, with $Inf_t = \sum inf_t(i)$. The relation between $H_p(t)$, $H_f(t)$ and $\{M_t\}$ is established in appendix F. We have:

$$H_p(t) = [\, \Pi_{i=0,\ldots,t} M_i^{B^t_i} \,]^{1/I^t} \exp(\upsilon^t) \qquad H_f(t) = [\, \Pi_{j=t+1,\ldots,T} M_j^{S^t_j} \,]^{1/I^t} \exp(v^t)$$

with : $v^t \approx 0$, $\upsilon^t \approx 0$, $E[v^t] = 0$, $E[\upsilon^t] = 0$. We get back the same kind of expressions that we had in Situation3 but this time, the values $\{Index_t\}$ are replaced with the price-index values $\{M_t\}$. At first sight this price-index M is not completely natural: it is a geometric one and the weights (the informational contributions) are unequal. In fact, this geometric framework is not really a surprise; we know that the RSI is not an arithmetic index but a geometric one, cf. Shiller (1991). As regards to the weights, it has to be noticed that as they are decreasing with the holding period: the goods held for a long time are a little underweighted in this index. In the price dynamic associated to the RSI, the longer a property is held, the more it can deviate from the general trend. Therefore, the short detentions are more significant compared to the longer ones. These two features (geometric and underweighting) are then the natural consequences of the assumptions made for the repeat-sales model, for a price-index. The main difference with the Situation3 comes from the coefficients $\exp(v^t)$ and $\exp(\upsilon^t)$. They capture a

[18] $inf_t(i) = (\, \Theta + \text{length of detention} \,)^{-1}$

phenomenon of variability in the sub-samples. But if the dataset is large enough, they should be very close to 1. For the initial issue of this paragraph, we now have the answer in the general situation. If we define a price-index similar to M we have a functional relation between M and $H_p(t)$, $H_f(t)$. As the vector R of the RSI is a function of P, that is a function of $H_p(t)$ and $H_f(t)$, we also have a relation between the RSI and this specific price-index. In Situation3 the relation was deterministic; here the link is more stochastic because we have a certain degree of sample fluctuation. However this stochastic relation is centred on the deterministic case.

3.3. A methodology of data analysis

In this paragraph and in the next one we will not present the technical details because these two issues will be the subject of future articles. We just want to give an idea of the flexibility of the informational approach and its possibilities. Usually the computation of the RSI is realised with a basic regression. Instead of that procedure we propose to employ an algorithmic decomposition of the index (Figure 1), that is strictly equivalent. The main advantage of this method is that we get this way several indicators that could be very interesting to deepen our understanding of the real estate market situation and behaviour. We can mention for instance the mean purchase price $H_p(t)$ and the mean resale price $H_f(t)$ for the people who were owning real estate during at least [t,t+1] ; the mean return ρ_t is also interesting for the analysis. For the interpretation of the mean holding period τ^t and the informational activities in the market, S_t sell-side[19] and B_t buy-side[20], we will have to

[19] $S_t = L_{0,t} + L_{1,t} + \ldots + L_{t-1,t}$: quantity of information provided by the repeat-sales with a resale at t.
[20] $B_t = L_{t,t+1} + L_{t,t+2} + \ldots + L_{t,T}$: quantity of information provided by the repeat-sales with a purchase at t.

introduce a benchmark dataset[21] to have a reference. This benchmark sample will have to be a function of a small number of parameters. The calibration step will consist in choosing these parameters in order to have the same levels for some aggregative measures between the real dataset and the benchmark dataset. For example if the benchmark has two parameters, we could choose the ones that will provide the same level for N, total number of repeat-sales in the sample, and for I, total quantity of information in the sample. Then, comparing the values of B_t, S_t and τ^t between the two datasets, we could bring to the fore the moments when the informational activities will be far from their mean, and also the shortening and the lengthening in the mean holding periods. As we can see, this global approach allows increasing the extraction of the information embedded in a sample. We do not confine us to the calculation of a simple index curve; the market analysis is thorough.

3.4. Reversibility formula and numerical simulation

The last consequence of the informational reformulation of the RSI concerns the reversibility phenomenon. With the same technique used previously to establish the volatility formula (from the elementary building blocks to the aggregative quantities) we can demonstrate a very intuitive and simple formula to deal with this problem. The exact demonstration and the implementation of the result will be the subject of another article; here we just want to give an idea of the flexibility of the informational approach. In our model, the time horizon is extended from T_1 to T_2. First we estimate the RSI with the old dataset on $[0,T_1]$; we get an informational matrix $\hat{I}(T_1)$ and a vector $R(T_1)$. Then, only with the new dataset[22] $T_2\backslash T_1$, 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 sample

[21] An exponential benchmark for example: we assume that the volume of transactions K at each date is constant in the market and we model the resale decision with an exponential law with a parameter λ. This benchmark dataset is perfectly known as soon as the parameters K and λ are given.
[22] A new data is an observation with a resale after T_1.

(old data + new data), we calculate the RSI on $[0,T_2]$ with $\hat{I}(T_2)$ and $R(T_2)$. The reversibility formula[23] is then:

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

From this result we can develop a Monte-Carlo methodology which allows forecasting the size of the potential revisions. At T_1, $\hat{I}(T_1)$ and $R(T_1)$ are known. In order to simulate the repeat-sales that will arrive in the sample between T_1 and T_2 we introduce a simple model based on an exponential distribution of the resale decision, for instance. This benchmark sample is calibrated on the old dataset and we use it to have an approximation $\hat{I}_{bench}(T_2 \backslash T_1)$ of the matrix $\hat{I}(T_2 \backslash T_1)$. $R(T_2 \backslash T_1)$ gives the index evolution on the interval $[0,T_2]$ for the new dataset. We assume that on the first part $[0,T_1]$ it is centred on the best estimator that we have at T_1, that is $R(T_1)$. And for the rest of the interval $[T_1,T_2]$, we complete this central value in a T_2-vector $R_{hyp} = (R(T_1), R_{hyp}(T_1 ; T_2))$, making economical hypotheses on the future of the real estate prices. Theoretically, we know that $R(T_2 \backslash T_1) \sim \mathcal{N}(\text{ Rate}(T_2) ; \sigma_G^2\, \hat{I}(T_2 \backslash T_1)^{-1})$. However, because of the unobservability at T_1 of these parameters, we generate the vector randomly according to the distribution $\mathcal{N}(R_{hyp} ; \sigma_G^2\, \hat{I}_{bench}(T_2 \backslash T_1)^{-1})$. For the matrix $\hat{I}(T_2)$ we can demonstrate that we have $\hat{I}(T_2) = \hat{I}(T_1) + \hat{I}(T_2 \backslash T_1) \approx \hat{I}(T_1) + \hat{I}_{bench}(T_2 \backslash T_1)$. At this stage of the process, the last unknown in the reversibility equation is $R(T_2)$. Thus, if we solve it, we get the values of the RSI at T_2 and comparing them to the values of $R(T_1)$ we can estimate the reversibility percentages[24]. Iterating this process, we get the distribution of the reversibility fluctuations.

[23] The square matrix $\hat{I}(T_1)$ and the vector $R(T_1)$, size $T_1 * T_1$ and T_1, are completed with zeros to get $\hat{I}(T_1)$ and $R(T_1)$, size $T_2 * T_2$ and T_2.

[24] For the rates or for the index values.

4. Elements for an informational theory of the repeat-sales indexes

4.1. Repeat-sales and information

As we saw in the previous developments information is really the heart of the framework. Once it is known the estimation of the index is straightforward with the algorithm of **Figure 1**. In the Case-Shiller model the quantity of information was $(\Theta + (j - i))^{-1}$ for a repeat-sales of $\mathcal{C}(i,j)$, it was a constant in the BMN context and in Situation-2 we had $(j - i)^{-1}$. These informational choices were not really explicit; they were more a consequence of the assumptions made on the error term ε in the various regressions (white noise, random walk,…). From the traditional point of view, the issue of the informational quantification is, more or less, ignored. However, as in a heterogeneous context this question cannot be avoided, we often answer it without being aware of that. In a scientific angle this kind of things is never desirable but there is also another reason, much more important, to change our viewpoint. Indeed, if we could find a way to define directly the index with an explicit informational approach, we could adapt it to the various economical contexts easily and make the index more efficient and more flexible. The quantities of information would not be just $(j - i)^{-1}$ or $(\Theta + (j - i))^{-1}$; we could choose for each observation its level of representativity vis-à-vis the composite index just overweighting or underweighting it. If we take for example the real estate market of New-York, it seems reasonable to overweight the transactions realised after the 11th of September compared to the ones just before because their prices integrate the information. In a mathematical point of view, we will have three main changes with the inputs. First, within a class $\mathcal{C}(i,j)$ the quantities of information could vary between the goods. Secondly, for a repeat-sales of $\mathcal{C}(i,j)$ the rates on each elementary time interval could differ. Thirdly, an observation could contribute at different levels for each elementary time

interval of its holding period. The first point allows dealing with the situations like the 11th of September. The two next points can be useful to incorporate in the repeat-sales model the appraisal values. Indeed, if during the holding periods several appraisals are realised, we can divide the interval [i,j] in sub-intervals to take these valuations into account. On each of these sub-periods we will have a constant rate, and a specific level of information[25]. Consequently we also merge the repeat-sales model and a hybrid approach with our generalisation process. In the rest of this paragraph we will present the new theoretical definition of the generalised index. As it is no longer a consequence of a regression, we need to base the fundamental relation $\hat{I}R = \eta P$ on something else.

4.2. A global definition for the informational repeat-sales indexes

4.2.1. Generalization

For a repeat-sales k' of $\mathcal{C}(i,j)$ we assume that we have on each elementary time interval [s,s+1] of its holding periods [i,j] a rate $r_{k'}^{(i,j)}(s)$. Moreover, we also have a measure of the informational weight of this rate: $inf_{k'}(s)$. For the subset Spl^t, we define the mean informational profitability of the real estate investment[26] as:

$$\rho_t = (\eta(t))^{-1} \sum_{i\le t<j} \sum_{k'\varepsilon \mathcal{C}(i,j)} \sum_{s=i,\dots,j-1} inf_{k'}(s)\, r_{k'}(s) \quad \text{with} \quad \eta(t) = \sum_{i\le t<j} \sum_{k'\varepsilon \mathcal{C}(i,j)} \sum_{s=i,\dots,j-1} inf_{k'}(s)$$

These values are then gathered in a vector denoted $P_{real} = (\rho_0, \rho_1, \dots, \rho_{T-1})$. Now, if we assume

---

[25] On the sub-interval (purchase, appraisal1) we can think that the level of information is higher than for the sub-interval (appraisal1, appraisal2) because of the real transaction in the first couple of values.

[26] The mean profitability for the k' owner of $\mathcal{C}(i,j)$ is $\rho_{k'}^{(i,j)} = [\sum_{s=i,\dots,j-1} inf_{k'}(s)]^{-1} [\sum_{s=i,\dots,j-1} inf_{k'}(s)\, r_{k'}(s)]$ and it represents an informational weight of $inf_{k'}^{(i,j)} = \sum_{s=i,\dots,j-1} inf_{k'}(s)$. Within $\mathcal{C}(i,j)$, the mean profitability is $\rho^{(i,j)} = [\sum_{k'\varepsilon \mathcal{C}(i,j)} inf_{k'}^{(i,j)}]^{-1} [\sum_{k'\varepsilon \mathcal{C}(i,j)} inf_{k'}^{(i,j)} \rho_{k'}^{(i,j)}]$ and the informational associated weight is $inf^{(i,j)} = \sum_{k'\varepsilon \mathcal{C}(i,j)} inf_{k'}^{(i,j)}$. Note that we also have with these notations:

$$\rho_t = (\eta(t))^{-1} \sum_{i\le t<j} \sum_{k'\varepsilon \mathcal{C}(i,j)} inf_{k'}^{(i,j)} \rho_{k'}^{(i,j)} = (\eta(t))^{-1} \sum_{i\le t<j} inf^{(i,j)} \rho^{(i,j)} \quad \text{and } \eta(t) = \sum_{i\le t<j} \sum_{k'\varepsilon \mathcal{C}(i,j)} inf_{k'}^{(i,j)} = \sum_{i\le t<j} inf^{(i,j)}$$

that the particular values $r_{k'}^{(i,j)}(s)$ provided by the real sample are all replaced in the expression of ρ_t with a series of universal values $R = (r_0, r_1, \ldots, r_{T-1})'$ that depend only on the time interval, what does the quantity ρ_t become ? We get:

$$\rho_t = (\eta(t))^{-1} \sum_{i \le t < j} \sum_{k' \varepsilon\, C(i,j)} \sum_{s=i,\ldots,j-1} \inf_{k'}(s)\, r_s = (\eta(t))^{-1} \sum_{i \le t < j} \sum_{s=i,\ldots,j-1} \left[\sum_{k' \varepsilon\, C(i,j)} \inf_{k'}(s) \right] r_s$$

In this general situation, the variability of the information brings us to define a family of $\{L_{i,j}(s)\}$. More precisely, for a given $\mathcal{C}(i,j)$ and for $s = i,\ldots,j-1$, we denote $L_{i,j}(s) = \sum_{k' \varepsilon C(i,j)} \inf_{k'}(s)$. It gives $\rho_t = (\eta(t))^{-1} \sum_{i \le t < j} \sum_{s=i,\ldots,j-1} L_{i,j}(s)\, r_s$. The double sum is a linear combination of the $\{r_s\}_{s=0,\ldots,T}$. We can establish that the number of $r_{t'}$ is equal to:

$(L_{0,t+1}(t')+\ldots+L_{0,T}(t')) + (L_{1,t+1}(t')+\ldots+L_{1,T}(t')) + \ldots + (L_{t',t+1}(t')+\ldots+L_{t',T}(t'))$ for $t' \le t$

$(L_{0,T}(t')+\ldots+L_{t,T}(t')) + (L_{0,T-1}(t')+\ldots+L_{t,T-1}(t')) + \ldots + (L_{0,t'+1}(t')+\ldots+L_{t,t'+1}(t'))$ for $t' > t$

In these two sums, we just have the repeat-sales classes associated to the sub-sample Spl^t that are including the interval $[t',t'+1]$. In other words the coefficient in front of each $r_{t'}$ in the expression of ρ_t is simply equal to the quantity of information that Spl^t can provide for the interval $[t',t'+1]$. Now, for the given series of universal values $R = (r_0, r_1, \ldots, r_{T-1})'$, we denote P_{ind} the vector $(\rho_0, \rho_1, \ldots, \rho_{T-1})$ that we get. We put the values $\eta(t)$, for $t = 0,\ldots,T-1$, in a diagonal matrix η. And at last we introduce a square matrix $\hat{I}$ of dimension T. Its i^{th} line corresponds to Spl^i. The j^{th} element of the line i corresponds to the quantity of information that Spl^i provides for the interval $[j,j+1]$. With these notations we have $P_{ind} = \eta^{-1}\, \hat{I}\, R$; a very familiar formula.

4.2.2. Index definition

We can now give a rigorous definition of the index. It is no longer a consequence of a regression but rather an explicit choice of the informational convention. The BMN and the Case-Shiller situations are just two specific examples.

Definition

For a repeat-sales sample, we have for each goods a series of rates $\{r_{k'}(s)\}$ *and an associated informational series* $\{inf_{k'}(s)\}$. *With these data we calculate the vector of the mean profitability* P_{real}. *On the other side, for each universal series* $R = (r_0, r_1, \ldots, r_{T-1})'$ *we have a vector* P_{ind}. *The vector R of the monoperiodic growth rates of the informational index is the one that makes equal* P_{real} *and* P_{ind} *; in other words the one that summarizes with a series of constant values the mean returns observed in the real sample. That is:*

$$P_{ind} = P_{real} \quad \Leftrightarrow \quad \eta^{-1} \hat{I} R = P_{real} \quad \Leftrightarrow \hat{I} R = \eta P_{real}$$

4.3. Which definition for the information?

Real estate markets are illiquids and highly heterogeneous. Under these circumstances what is the meaning of the single number that we call "real estate market price" or "index value"? Is it really reasonable to summarize a very complex situation with just one number? However, if we have some difficulties to define in a very precise way the theoretical concept, the index is nevertheless essential for the practitioners (benchmark, economical indicator, underlying for derivatives…). The main problem, that is the heterogeneity, can be expressed with the following question: *How much real estate do we have in a specific house?* In other words, what is the degree of representativity of one good related to the global commodity called real estate. Answering to this question is equivalent to choose the informational level of the observation; that is to choose the values of the series $\{inf_k(s)\}$ in the generalised repeat-sales model. If in a formal view we brought to the fore the informational framework of the index, we still have this important question to solve: *How should we define the information?* Is the Case-Shiller model the best or should we choose others weights. According to the answer, we

will have different index curves. The literature already provides several alternatives. In these articles, the fundamental relation $\hat{I}R = \eta P$ remains unchanged, the modification just concerns the definition of the $L_{i,j}$. We can mention for example Peng (2002) or Shiller (1991) in which the weights are equally-weighted, price-weighted or value-weighted. In Dreiman, Pennington-Cross (2004) a quadratic term is introduced in the variance of the residuals ; the quantity of information for one repeat-sales of $\mathcal{C}(i,j)$ becomes $(\Theta + \psi (j - i) + (j - i)^2)^{-1}$ and not just $(\Theta + (j - i))^{-1}$ like in the Case-Shiller model. We can also mention Cannaday, Munneke, Yang (2005) in which the information is a function of the age of the property. We keep this issue for future researches. But whatever be our choice, it is clear that the informational approach allows unifying several papers of the literature.

5. Conclusion

In this article we tried to reformulate the repeat-sales index, making explicit its informational framework. We established that its estimation was equivalent to the fundamental relation $\hat{I} R = \eta P$ and we provided an algorithm in which the index is decomposed in elementary building blocks that can be used to improve the data analysis and the data interpretation. This approach is flexible, highly coherent and it allows establishing some very intuitive formulas (volatility, reversibility) thanks to the above mentioned decomposition. We also explored the formal link between the price indexes and the repeat-sales index. Finally, we generalized the informational approach to broaden the way we can define the information; unifying in the same time several article of the literature. The concepts introduced in this article are easy to handle, productive and there are many promising directions for the future researches like the implementation of the methodology of data analysis, the implementation of the reversibility

formula in order to forecast the magnitude of the potential fluctuations. We could also study the consequences of the informational conventions on the index curves. At last, it could be interesting to study the flip problem with these concepts. Usually the goods with the shortest holding periods are overrepresented in the sample. The natural idea is therefore to reduce them, but according which criterion? An informational criterion could be a solution.

References

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

Cannaday, Munneke, Yang. 2005. "A multivariate repeat-sales model for estimating house price indices" *Journal of urban economics* 57(2) : 320-342

Case, Pollakowski, Watcher. 1991. "On choosing among house price index methodologies" *AREUEA Journal* 19(3) : 286-307

Case, Quigley. 1991. “The dynamics of real estate prices”. *The review of economics and statistics* 73 (1) : 50-58

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

Clapp, Giaccotto. 1992. "Estimating price indices for residential property : a comparison of repeat sales and assessed value methods" *Journal of the American statistical association* 87(418) : 300-306

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

Dreiman, Pennington-Cross. 2004. "Alternative methods of increasing the precision of weighted repeat sales house prices indices" *Journal of real estate finance and economics* 28(4) : 299-317

Geltner. 1991. "Smoothing and appraisal-based returns" *Journal of real estate finance and economics* 4(3) : 327-345

Goetzmann. 1992. “The accuracy of real estate indices: repeat sales estimators”. *Journal of real estate finance and economics* 5 : 5-53

Goetzmann, Peng. 2002. "The bias of the RSR estimator and the accuracy of some alternatives" *Real estate economics* 30(1) : 13-39

Peng. 2002. "GMM repeat sales price indices" *Real estate economics* 30(2) : 239-261

Rosen. 1974. “Hedonic price and implicit markets : product differentiation in pure competition” *Journal of political economy* 1

Shiller. 1991. “Arithmetic repeat sales price estimators” *Journal of housing economics* 1 : 110-126

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

Appendix A: From the optimization problem to the algorithmic decomposition

The minimization problem is :

$$\text{Min}_R \left[\sum_{i<j} \sum_{k'} \left(\Theta + (j-i) \right)^{-1} \left\{ \ln(p_{k',j} / p_{k',i}) - (r_i + \ldots + r_{j-1}) \right\}^2 \right] \qquad (6)$$

If we develop the squares and if we keep only the non-constant terms, this problem is equivalent to the minimization of the function $\Phi(R)$:

$$\Phi(R) = \sum_{i<j} (\Theta + (j-i))^{-1} \sum_{k'} [(r_i + \ldots r_{j-1})^2 - 2\ln(p_{k',j}/p_{k',i}) (r_i + \ldots + r_{j-1})]$$

As k' is varying between 1 and $n_{i,j}$ for each repeat-sales class $\mathcal{C}(i,j)$, we also have :

$$\Phi(R) = \sum_{i<j} (\Theta + (j-i))^{-1} [n_{i,j}(r_i + \ldots + r_{j-1})^2 - 2(r_i + \ldots + r_{j-1}) \sum_{k'} \ln(p_{k',j}/p_{k',i})]$$

Introducing the notation $L_{i,j} = n_{i,j} / (\Theta + (j-i))$ we get:

$$\Phi(R) = \sum_{i<j} L_{i,j} [(r_i + \ldots + r_{j-1})^2 - 2\{(1/n_{i,j}) \sum_{k'} \ln(p_{k',j}/p_{k',i})\}(r_i + \ldots + r_{j-1})]$$

For each t = 0,…,T-1, we calculate the derivative of $\Phi(R)$ with respect to r_t. We use the notations $\check{r}_t$ for the sum of r_i to r_j, r_t excepted[27]. As r_t is present in the contribution of $\mathcal{C}(i,j)$, if and only if we have $i \le t < j$ we can write :

$$\partial\Phi(R) / \partial r_t = \sum_{i \le t < j} L_{i,j} [2 r_t + 2 \check{r}_t - 2 (1/n_{i,j}) \sum_{k'} \ln(p_{k',j}/p_{k',i})\}]$$

$$= 2 \sum_{i \le t < j} L_{i,j} (r_i + \ldots + r_{j-1}) - 2 \sum_{i \le t < j} L_{i,j} \{(1/n_{i,j}) \sum_{k'} \ln(p_{k',j}/p_{k',i})\}$$

Lemma 1 : $\sum_{i \le t < j} \mathbf{L_{i,j}} \{(1/n_{i,j}) \sum_{k'} \ln(p_{k',j}/p_{k',i})\} = \mathbf{n^t G(\zeta^t) \rho_t}$

For the k'th repeat sales in $\mathcal{C}(i,j)$, $\ln(p_{k',j}/p_{k',i})$ is just the log-return. If we introduce the notation $r_{k'}^{(i,j)} = \ln(p_{k',j} / p_{k',i}) / (j-i)$, this sum becomes:

$$\sum_{i \le t < j} L_{i,j} \{(1/n_{i,j}) \sum_{k'} \ln(p_{k',j}/p_{k',i})\} = \sum_{i \le t < j} (\Theta + (j-i))^{-1} \sum_{k'} r_{k'}^{(i,j)} (j-i)$$

$$= \sum_{i \le t < j} \sum_{k'} (j-i) / (\Theta + (j-i)) \, r_{k'}^{(i,j)}$$

[27] $r_i + \ldots r_{j-1} = \check{r}_t + r_t$ and thus $(r_i + \ldots r_{j-1})^2 = r_t^2 + \check{r}_t^2 + 2 \check{r}_t r_t$ that can be derived easily.

This expression is very similar to the arithmetic average of the $r_{k'}^{(i,j)}$ weighted by the coefficients $G(j-i) = (j-i)/(\Theta + (j-i))$. We can see these coefficients as the proportions of the noise coming from the Gaussian random walk $(j-i)$, in the total noise $\Theta + (j-i)$. However the total mass of the weights isn't specified yet.

On the time interval $[0; +\infty[$, the function $G(x) = x/(x+\Theta)$ is strictly increasing from 0 to 1 because as time goes by the time-varying noise component, measured by $x = j-i$, becomes the main noise source. The properties of G allow defining[28] a G-average ζ^t of the holding periods $j-i$ for all the repeat sales in Spl^t:

$$\sum_{i \le t < j} \sum_{k'} (j-i)/(\Theta + (j-i)) = \sum_{i \le t < j} \sum_{k'} G(j-i) = n^t\, G(\zeta^t) \qquad \left(\sum_{i \le t < j} \sum_{k'} 1 = n^t\right)$$

The total mass of the weights in the sum $\sum_{i \le t < j} \sum_{k'} G(j-i)\, r_{k'}^{(i,j)}$ leads us to define the quantity $\rho_t = (n^t\, G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} G(j-i)\, r_{k'}^{(i,j)}$. With this notation we get:

$$\sum_{i \le t < j} L_{i,j} \{(1/n_{i,j}) \sum_{k'} \ln(p_{k',j}/p_{k',i})\} = \sum_{i \le t < j} \sum_{k'} G(j-i)\, r_{k'}^{(i,j)} = n^t\, G(\zeta^t)\, \rho_t$$

Lemma 2: $\sum_{i \le t < j} \mathbf{L_{i,j}(r_i + \ldots + r_{j-1})} = \sum_{0 \le t' \le t} \mathbf{I}^{[t', t+1]}\, \mathbf{r_{t'}} + \sum_{t < t' < T} \mathbf{I}^{[t, t'+1]}\, \mathbf{r_{t'}}$

For this sum, it is useful to reorganize the calculation in the following way:

i \ j	t + 1	t + 2	…	T – 1	T
0	$(r_0+\ldots+r_t)$ $L_{0,\,t+1}$	$(r_0+\ldots+r_{t+1})$ $L_{0,\,t+2}$		$(r_0+\ldots+r_{T-2})$ $L_{0,\,T-1}$	$(r_0+\ldots+r_{T-1})$ $L_{0,\,T}$
1	$(r_1+\ldots+r_t)$ $L_{1,\,t+1}$	$(r_1+\ldots+r_{t+1})$ $L_{1,\,t+2}$		$(r_1+\ldots+r_{T-2})$ $L_{1,\,T-1}$	$(r_1+\ldots+r_{T-1})$ $L_{1,\,T}$
2	$(r_2+\ldots+r_t)$ $L_{2,\,t+1}$	$(r_2+\ldots+r_{t+1})$ $L_{2,\,t+2}$		$(r_2+\ldots+r_{T-2})$ $L_{2,\,T-1}$	$(r_2+\ldots+r_{T-1})$ $L_{2,\,T}$
⋮					
t – 1	$(r_{t-1}+r_t)$ $L_{t-1,\,t+1}$	$(r_{t-1}+\ldots+r_{t+1})$ $L_{t-1,\,t+2}$		$(r_{t-1}+\ldots+r_{T-2})$ $L_{t-1,\,T-1}$	$(r_{t-1}+\ldots+r_{T-1})$ $L_{t-1,\,T}$
t	r_t $L_{t,\,t+1}$	(r_t+r_{t+1}) $L_{t,\,t+2}$		$(r_t+\ldots+r_{T-2})$ $L_{t,\,T-1}$	$(r_t+\ldots+r_{T-1})$ $L_{t,\,T}$

[28] As regards to the quantity $G(\zeta^t) = \zeta^t/(\zeta^t + \Theta)$, it is simply the arithmetic mean of the proportions $(j-i)/(\Theta + (j-i))$, that is the mean contribution of the noise $G_{k,\,t}$ to the total noise, in Spl^t.

The "r_0" come from the first line, "r_1" from the first and the second, …, "r_t" from the t+1 lines. We get the contribution for "r_{T-1}" with the last column, for "r_{T-2}" with the two last columns, …, for "r_{t+1}" with all the columns the first one excepted.

Thus for $t' \leq t$, the quantity of $r_{t'}$ is :

$$\underbrace{(L_{0,\,t+1} + L_{0,\,t+2} + \ldots + L_{0,\,T})}_{\textit{(Line 0)}} + \underbrace{(L_{1,\,t+1} + L_{1,\,t+2} + \ldots + L_{1,\,T})}_{\textit{(Line 1)}} + \ldots + \underbrace{(L_{t',\,t+1} + L_{t',\,t+2} + \ldots + L_{t',\,T})}_{\textit{(Line t')}}$$

That is: $B_0^t + B_1^t + \ldots + B_{t'}^t$

And for $t' > t$, the quantity of $r_{t'}$ is :

$$\underbrace{(L_{0,\,T} + L_{1,T} + \ldots + L_{t,\,T})}_{\textit{(Column T)}} + \underbrace{(L_{0,\,T-1} + L_{1,T-1} + \ldots + L_{t,\,T-1})}_{\textit{(Column T-1)}} + \ldots + \underbrace{(L_{0,\,t'+1} + L_{1,t'+1} + \ldots + L_{t,\,t'+1})}_{\textit{(Column t'+1)}}$$

That is: $S^t_T + S^t_{T-1} + \ldots + S^t_{t'+1}$

If we refer to **Table 4**, the first sum corresponds to $I^{[t',\ t+1]}$ and the calculation is realised buy-side, whereas the second one corresponds to $I^{[t,\ t'+1]}$ with a sell-side calculation. The announced relation is now established.

With these two lemmas, the derivative of $\Phi(R)$ w.r.t. r_t becomes:

$$\partial\Phi(R) / \partial r_t \quad = 2 \sum_{0 \leq t' \leq t} I^{[\,t',\,t+1]}\, r_{t'} + 2 \sum_{t < t' < T} I^{[\,t,\,t'+1]}\, r_{t'} \; - 2\, n^t\, G(\zeta^t)\, \rho_t$$

And with the matrixes $\hat{I}$ and η, the vectors $R = (r_0, r_1, \ldots, r_{T-1})'$ and $P = (\rho_0, \rho_1, \ldots, \rho_{T-1})$, the solution of the optimization problem that is corresponding to the system of equations $\{\ \partial\Phi(R) / \partial r_t = 0\ ;\ t = 0,\ldots,T\text{-}1\}$ with unknown R, is actually the solution of $\hat{I}\, R = \eta\, P \Leftrightarrow\ R = (\hat{I}^{-1}\eta)\ P$.

Appendix B: The two average holding periods are equal: $\zeta^t = \tau^t$

The relations that define the G-average ζ^t and the harmonic average τ^t are:

$n^t\, G(\zeta^t) = \sum_{i \le t < j} \sum_{k,} G(j - i)$ and $(n^t G(\zeta^t)) / \tau^t = \sum_{i \le t < j} \sum_{k,} G(j - i) * (1 / (j - i))$

With the second one we can write: $\tau^t = (n^t\, G(\zeta^t)) / \sum_{i \le t < j} \sum_{k,} G(j - i) * (1 / (j - i))$

Applying the function $G(x) = x/(x+\Theta)$ we get: $G(\tau^t) = G(\zeta^t)/[G(\zeta^t) + (\Theta/n^t)\sum_{i \le t < j} \sum_{k,} G(j-i)/(j-i))]$

Using the definition of ζ^t: $G(\tau^t) = G(\zeta^t) / [(1/n^t)\sum_{i \le t < j} \sum_{k,} G(j-i) + (\Theta/n^t)\sum_{i \le t < j} \sum_{k,} G(j-i)/(j-i))]$

That is : $G(\tau^t) = G(\zeta^t) / [(1/n^t) \sum_{i \le t < j} \sum_{k,} G(j-i) * (1 + \Theta / (j - i))]$

But, as: $1 + \Theta / (j - i) = (\Theta + (j - i)) / (j - i) = [G(j - i)]^{-1}$

It comes: $G(\tau^t) = G(\zeta^t) / [(1 / n^t) \sum_{i \le t < j} \sum_{k,} 1] = G(\zeta^t)$

The function G being strictly increasing, we get our identity $\tau^t = \zeta^t$.

Appendix C : Mathematical properties of the matrixes $\hat{I}$ and η

We clearly have $\hat{I}_{p,q} \geq 0$ and $\hat{I}_{p,q} = \hat{I}_{q,p}$. The inequalities in the Proposition 1 and the formulas in the Proposition 2 are just algebraic translations of very simple geometric relations. These geometric results can be illustrated easily with various areas of the Table 4. For Proposition 3, things are less obvious. We get the first formula about the matrix trace of $\hat{I}$, noticing that a given $L_{i,j}$ will appear (j – i) times in this sum:

$$\mathrm{Tr}\,(\hat{I}) = I^{0} + I^{1} + \ldots + I^{T-1} = \sum_{i<j} (j-i) L_{i,j} = \sum_{i<j} (j-i) \sum_{k,} (\Theta + (j-i))^{-1} = \sum_{i<j} \sum_{k,} G(j-i) = N\, G(\zeta)$$

For the second one we can write:

$$\mathrm{Tr}\,(\hat{I}) - \mathrm{Tr}_{+1}\,(\hat{I}) = \sum_{t=0,\ldots,T-1} I^{t} - \sum_{t=0,\ldots,T-2} I^{[t,\,t+2]} = \sum_{t=0,\ldots,T-2} (I^{t} - I^{[0,\,t+2]}) + I^{T-1}$$

$$= \sum_{t=0,\ldots,T-2} \sum_{i=0,\ldots,t} L_{i,\,t+1} + \sum_{i=0,\ldots,T-1} L_{i,\,T} = \sum_{t=0,\ldots,T-1} \sum_{i=0,\ldots,t} L_{i,\,t+1} = I$$

For the third one we can used the relation of lemma 2, with $r_0 = \ldots = r_{T-1} = 1$. It gives :

$$\sum_{t' \leq t} I^{[t',\,t+1]} + \sum_{t' > t} I^{[t,\,t'+1]} = \sum_{i \leq t < j} (j-i)\, L_{i,j} = \sum_{i \leq t < j} \sum_{k,} G(j-i) = n^{t}\, G(\zeta^{t})$$

Appendix D : Reformulation of ρ_t

The initial definition for ρ_t is $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} G(j-i)\, r_{k'}^{(i,j)}$ from where we get:

$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} (\Theta + (j-i))^{-1} \sum_{k'} \ln(p_{k',j}/p_{k',i}) = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} (\Theta + (j-i))^{-1} \ln(\Pi_{k'} p_{k',j} / \Pi_{k'} p_{k',i})$$

The geometric averages of the purchase prices and the resale prices[29] in $\mathcal{C}(i,j)$ are:

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

It gives: $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} L_{i,j} \ln(h_f^{(i,j)}/h_p^{(i,j)}) = (n^t G(\zeta^t))^{-1} \ln[(\Pi_{i \le t < j} (h_f^{(i,j)})^{L_{i,j}})/(\Pi_{i \le t < j} (h_p^{(i,j)})^{L_{i,j}})]$

If we now introduce the geometric averages of the $h_p^{(i,j)}$ and the $h_f^{(i,j)}$, weighted[30] by the $L_{i,j}$:

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

we get the formula : $\rho_t = (I^t / (n^t G(\zeta^t))) * \ln[H_f(t) / H_p(t)]$

We can interpret the coefficient $I^t/(n^t G(\zeta^t))$ with the two relations :

$$I^t = \sum_{i \le t < j} L_{i,j} = \sum_{i \le t < j} \sum_{k'} (\Theta + (j-i))^{-1} = \sum_{i \le t < j} \sum_{k'} G(j-i) * (1/(j-i))$$

$$n^t G(\zeta^t) = \sum_{i \le t < j} \sum_{k'} (j-i)/(\Theta + (j-i)) = \sum_{i \le t < j} \sum_{k'} G(j-i)$$

$I^t / (n^t G(\zeta^t))$ is nothing else than the arithmetic average F^t of the holding frequencies $1/(j-i)$, weighted by the $G(j-i)$. With the harmonic average $\tau^t = (F^t)^{-1}$ the final relation for ρ_t is then:

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

We can also notice that the geometric averages $H_p(t)$ and $H_f(t)$ can be expressed as:

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

That is the mean of the purchase and the resale prices for Spl^t, weighted by their informational individual contributions $1/(\Theta + (j-i))$.

---

[29] A purchase is a past event (p) and a resale a future event (f) in relation to the present [t, t+1].

[30] whose total mass is $I^t = \sum_{i \le t < j} L_{i,j}$

Appendix E : Volatility formulas

Lemma 3 : $p_{k,t} = indice_t\ C(p_{k,t})$

For a good k, at time t, the deviation of the price relatively to the index is $G_{k,t} + N_{k,t}$. With (H3) it comes that $G_{k,t} + N_{k,t}$ follows a normal law with a mean $G_{k,0}$ and a variance $\sigma_N^2 + \sigma_G^2 t = \sigma_G^2 (\Theta/2+t)$. The relation $\ln(p_{k,t}) = \ln(Index_t) + G_{k,t} + N_{k,t}$ gives $p_{k,t} = indice_t \exp(G_{k,t}+N_{k,t}) = indice_t\ C(p_{k,t})$ with $C(p_{k,t}) \sim \mathcal{LN}(G_{k,0} ; \sigma_G^2 (\Theta/2 + t))$.

Lemma 4 : $r_k^{(i,j)} \sim \mathcal{N}(r^{(i,j)} ; \sigma_G^2 / ((j - i) G(j - i)))$

For a good k, bought at t_i and sold at t_j, the return rate is $r_k^{(i,j)} = \ln(p_{k,j}/p_{k,i})/(j-i) = [\ln(Indice_j/Indice_i) + (G_{k,j} - G_{k,i}) + (N_{k,j} - N_{k,i})] / (j-i)$. The increment of the random walk follows a normal law $\mathcal{N}(0;\sigma_G^2 (j-i))$ and the one of the white noise a normal law $\mathcal{N}(0;2\sigma_N^2)$. With (H3) we get the law of $r_k^{(i,j)}$: $\mathcal{N}(\ln(Indice_j/Indice_i)/(j-i) ; (\sigma_G^2(j-i)+2\sigma_N^2)/(j-i)^2)$ or equivalently $\mathcal{N}(r^{(i,j)} ; \sigma_G^2 / ((j - i) G(j - i)))$ with $r^{(i,j)} = \ln(Indice_j/Indice_i)/(j-i)$.

Lemma 5 : $h_p^{(i,j)} = indice_i\ C(h_p^{(i,j)})$ $\qquad$ $h_f^{(i,j)} = indice_j\ C(h_f^{(i,j)})$

With the definitions of $h_p^{(i,j)}$ and $h_f^{(i,j)}$ we can write:

$$h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}} = indice_i \exp[(\Sigma_{k'} G_{k',i} + N_{k',i})/ n_{i,j}] = indice_i\ C(h_p^{(i,j)})$$

$$h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}} = indice_j \exp[(\Sigma_{k'} G_{k',j} + N_{k',j})/ n_{i,j}] = indice_j\ C(h_f^{(i,j)})$$

Using (H2) and (H3) we have $C(h_p^{(i,j)}) \sim \mathcal{LN}((\Sigma_{k'}G_{k',0}) / n_{i,j} ; \sigma_G^2 (\Theta/2+i)/ n_{i,j})$ and $C(h_f^{(i,j)}) \sim \mathcal{LN}((\Sigma_{k'}G_{k',0}) / n_{i,j} ; \sigma_G^2 (\Theta/2 + j)/n_{i,j})$. Moreover, if we use (H1) for each class $\mathcal{C}(i,j)$ we get respectively the law $\mathcal{LN}(0 ; \sigma_G^2(\Theta/2 + i) / n_{i,j})$ and $\mathcal{LN}(0 ; \sigma_G^2(\Theta/2 + j) / n_{i,j})$.

Lemma 6 : $\rho_{i,j} \sim \mathcal{N}\left(r^{(i,j)} ; \sigma_G^2 / (L_{i,j} (j - i)^2) \right)$

The mean return in $\mathcal{C}(i,j)$, $\rho_{i,j} = (\sum_{k,} r_{k'}^{(i,j)})/n_{i,j} = \ln(h_f^{(i,j)}/h_p^{(i,j)}) / (j-i)$, is a sum of $n_{i,j}$ independent and normal variables. With (H2) and (H3) we get immediately its law: $\mathcal{N}(r^{(i,j)}; \sigma_G^2/(L_{i,j}(j-i)^2))$.

Lemma 7 : $H_p(t) = \left[\prod_{i \le t} \text{indice}_i^{B_i^t}\right]^{1/I^t} C(H_p(t))$ $\quad$ $H_f(t) = \left[\prod_{j > t} \text{indice}_j^{S_j^t}\right]^{1/I^t} C(H_f(t))$

The mean purchase price and the mean resale price in Spl^t are:

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

In $H_p(t)$, if we replace $h_p^{(i,j)}$ with the product $\text{indice}_i * C(h_p^{(i,j)})$ we get:

$$H_p(t) = \left[\prod_{i \le t < j} (\text{indice}_i\, C(h_p^{(i,j)}))^{L_{i,j}}\right]^{1/I^t} = \left[\prod_{i \le t < j} \text{indice}_i^{L_{i,j}}\right]^{1/I^t} \left[\prod_{i \le t < j} C(h_p^{(i,j)})^{L_{i,j}}\right]^{1/I^t}$$

The first part corresponds to Situation-3. We already know that this expression is equal to the quantity $\left[\prod_{i \le t} \text{indice}_i^{B_i^t}\right]^{1/I^t}$; an average of the past values of the index weighted by their informational levels buy-side. In the second part, the laws of the random variables $C(h_p^{(i,j)})^{L_{i,j}}$ are log-normal with parameters 0 and $(L_{i,j})^2\sigma_G^2(\Theta/2+i)/n_{i,j}$. If we introduce the notation $V_{i,j} = L_{i,j}/(\Theta+(j-i)) = n_{i,j} / (\Theta + (j - i))^2$ this second parameter becomes $V_{i,j}\, \sigma_G^2(\Theta/2+ i)$. With (H2) and (H3) we can establish that the law of the second part, denoted $C(H_p(t))$, is lognormal: $\mathcal{LN}(0;\sigma_G^2(1/I^t)^2 \sum_{i \le t < j} V_{i,j}(\Theta/2+ i))$. Similarly $H_f(t) = \left[\prod_{j > t} \text{indice}_j^{S_j^t}\right]^{1/I^t} C(H_f(t))$ and the law of $C(H_f(t))$ is $\mathcal{LN}(0 ; \sigma_G^2 (1/I^t)^2 \sum_{i \le t < j} V_{i,j}(\Theta/2+ j))$.

Lemma 8 : $\rho_t \sim \mathcal{N}\left((n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j-i) n_{i,j} r^{(i,j)} ; \sigma^2_G/((\tau^t)^2 I^t)\right)$ and $\text{Cov}(\rho_t;\rho_{t'}) = \sigma_G^2 I^{[t',t+1]}/(\tau^t \tau^{t'} I^t I^{t'})$

The formula for ρ_t can be expressed with the independent variables $\rho_{i,j}$:

$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j\text{-}i)\, n_{i,j}\ \left(\sum_{k'} r_{k'}^{(i,j)} / n_{i,j} \right) = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j\text{-}i)\, n_{i,j}\, \rho_{i,j}$$

With (H2) and (H3) its law is: $\mathcal{N}\left((n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j - i)\, n_{i,j}\, r^{(i,j)} ;\ \sigma^2_G / ((\tau^t)^2 I^t) \right)$. The expectation corresponds to Situation-3, ρ_t is centred. The variance of the law comes from the following calculation:

$$V(\rho_t) = (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} G^2(j\text{-}i)\, (n_{i,j})^2 \sigma_G^2 / (L_{i,j}(j\text{-}i)^2) = (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} (n_{i,j})^2\, \sigma_G^2 / (L_{i,j}(\Theta + j\text{-}i)^2)$$

$$V(\rho_t) = (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} (n_{i,j})^2\, \sigma_G^2 / (\, n_{i,j}\, (\Theta + j\text{-}i)\,) = (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} n_{i,j}\, \sigma_G^2 / (\Theta + j\text{-}i)$$

$$V(\rho_t) = (\sigma_G / (n^t G(\zeta^t)))^2 \sum_{i \le t < j} L_{i,j} = \sigma^2_G / ((\tau^t)^2 I^t)$$

Moreover if we write for t and t' :

$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j\text{-}i) n_{i,j}\, \rho_{i,j} \qquad \rho_{t'} = (n^{t'} G(\zeta^{t'}))^{-1} \sum_{i \le t' < j} G(j\text{-}i) n_{i,j}\, \rho_{i,j}$$

we can establish that $\mathrm{Cov}(\rho_t ; \rho_{t'}) = \sigma_G^2\, I^{[t', t+1]} / (\tau^t \tau^{t'} I^t I^{t'})$. Here, even if $t \neq t'$, the quantities ρ_t and $\rho_{t'}$ are not independent because a class $\mathcal{C}(i,j)$ can belong simultaneously to Spl^t and $\mathrm{Spl}^{t'}$. However the means $\rho_{i,j}$ are independent and for two distinct classes, $\mathcal{C}(i,j)$ and $\mathcal{C}(i',j')$, we have $\mathrm{Cov}(\rho_{i,j} ; \rho_{i',j'}) = 0$. If it is the same class we just have $\mathrm{Cov}(\rho_{i,j} ; \rho_{i,j}) = V(\rho_{i,j}) = \sigma_G^2/(L_{i,j}(j\text{-}i)^2)$. The announced result simply comes from the bilinearity of the covariance:

$$\mathrm{Cov}(\rho_t ; \rho_{t'}) = (n^{t'} n^t G(\zeta^{t'}) G(\zeta^t))^{-1} \sum_{i \le t' < t < j} (G(j\text{-}i)\, n_{i,j})^2\, \sigma_G^2 / (L_{i,j}\, (j\text{-}i)^2)$$

$$\mathrm{Cov}(\rho_t ; \rho_{t'}) = (n^{t'} n^t G(\zeta^{t'}) G(\zeta^t))^{-1}\, \sigma_G^2 \sum_{i \le t' < t < j} L_{i,j}$$

$$\mathrm{Cov}(\rho_t ; \rho_{t'}) = (n^{t'} n^t G(\zeta^{t'}) G(\zeta^t))^{-1}\, \sigma_G^2\, I^{[t', t+1]} = \sigma_G^2\, I^{[t', t+1]} / (\tau^t \tau^{t'} I^t I^{t'}).$$

And if we have $t = t'$, we get back the above formula for the variance of ρ_t : $\sigma^2_G / ((\tau^t)^2 I^t)$

<u>Lemma 9</u> : $\mathbf{R \sim \mathcal{N}(\mathrm{Rate} ;\ \sigma_G^2\, \hat{I}^{-1})}$

The coordinates of the expectation of the vector η P are: $\sum_{i \le t < j} G(j\text{-}i)\, n_{i,j}\, r^{(i,j)}$, for t = 0,…,T-1.

We also have $\mathrm{Cov}\,((\eta P)_t ; (\eta P)_{t'}) = \sigma_G^2\, I^{[t', t+1]}$ and $V\,((\eta P)_t) = \sigma_G^2 I^t$. Therefore, the variance-

covariance matrix of η P is nothing else than σ_G^2 $\hat{I}$. Now, in order to have the random behaviour of the vector R, we just have to use the fundamental formula of the RSI: $R = \hat{I}^{-1}(\eta P)$. Every linear combination Λ of the coordinates of R is a linear combination of the coordinates of P and consequently a linear combination of the Gaussian independent variables $r_{k'}^{(i,j)}$. Thus Λ follows a normal law and the vector R is a Gaussian one. The expectation vector E(R) is equal to $\hat{I}^{-1}$ E(ηP). But as we previously mentioned, this product corresponds to Situation-3 and in that case we know that we get back the true vector Rate = $(rate_0, rate_1, \ldots, rate_{T-1})'$. The vector R is thus an unbiased estimator of the theoretical vector, Rate. If two random vectors X and Y are linearly dependent (Y = BX), their variance-covariance matrixes $\mathcal{V}$X and $\mathcal{V}$Y are also dependent ($\mathcal{V}Y = B(\mathcal{V}X)B'$). For R we get: $\mathcal{V}(R) = \hat{I}^{-1}(\sigma_G^2 \hat{I})(\hat{I}^{-1})'$. With the simplification and the symmetry of $\hat{I}$ we finally have: $\mathcal{V}(R) = \sigma_G^2 \hat{I}^{-1}$.

<u>Lemma 10</u> : Lind ~ $\mathcal{N}$(LIndex ; σ_G^2 ($\mathcal{T}$ – ($\mathcal{L}$ + $\mathcal{L}$'))$^{-1}$)

The vectors Lind and R are linearly dependent: Lind = A R. A is a triangular matrix whose values are equal to 1 on the principal diagonal and under it, 0 elsewhere. As R is a Gaussian vector it is also true for Lind. The expectation of Lind is A E(R) = A Rate = LIndex. Its variance is $\mathcal{V}$(LInd) = A σ_G^2 $\hat{I}^{-1}$ A' = σ_G^2 A $\hat{I}^{-1}$ A' = σ_G^2 (A'^{-1} $\hat{I}$ A^{-1}) $^{-1}$. The left product, A'^{-1} time $\hat{I}$, means that to each line "i" of $\hat{I}$ we subtract the next line "i+1", except for the last one which stays unchanged. The result is then multiplied on the right by the matrix A^{-1} : the columns "j+1" are subtracted to the columns "j", the last one is unchanged. These operations on the lines and the columns of $\hat{I} = (\hat{I}_{p,q})_{1 \le p, q \le T}$ give four types of results for the coordinates of the matrix $A'^{-1} \hat{I} A^{-1} = (b_{p,q})_{1 \le p, q \le T}$:

- For $1 \le p, q < T$: $b_{p,q} = \hat{I}_{p,q} - (\hat{I}_{p+1,q} + \hat{I}_{p,q+1}) + \hat{I}_{p+1,q+1}$
- For $1 \le q < T$: $b_{T,q} = \hat{I}_{T,q} - \hat{I}_{T,q+1}$

- For $1 \leq p < T$: $b_{p,T} = \hat{I}_{p,T} - \hat{I}_{p+1,T}$
- For $p = q = T$: $b_{T,T} = \hat{I}_{T,T}$

With Proposition 2 we can simplify these expressions:

Case 1: $1 \leq p < q < T$

Changing slightly the indexes in Proposition 2 we get: $\hat{I}_{p,q} = \hat{I}_{p+1,q} + \hat{I}_{p,q+1} - \hat{I}_{p+1,q+1} - L_{p,q}$

Thus: $b_{p,q} = - L_{p,q}$

Case 2: $1 \leq q < p < T$

With the symmetry of $\hat{I}$ we have: $b_{p,q} = \hat{I}_{q,p} - (\hat{I}_{q+1,p} + \hat{I}_{q,p+1}) + \hat{I}_{q+1,p+1}$

Thus: $b_{p,q} = - L_{q,p}$ (similar to case 1)

Case 3: $1 \leq q = p < T$

Proposition 2 cannot be used here. However we can write:

$b_{p,p} = \hat{I}_{p,p} - (\hat{I}_{p+1,p} + \hat{I}_{p,p+1}) + \hat{I}_{p+1,p+1} = b_{p,p} = \hat{I}_{p,p} - 2\,\hat{I}_{p,p+1} + \hat{I}_{p+1,p+1}$

$b_{p,p} = (L_{p,p+1} + L_{p,p+2} + \ldots + L_{p,T}) + (L_{0,p} + L_{1,p} + \ldots + L_{p-1,p}) = B^{p}_{p} + S^{p-1}_{p}$

Case 4: $p = T$ and $1 \leq q < T$ $b_{T,q} = \hat{I}_{T,q} - \hat{I}_{T,q+1} = \hat{I}_{q,T} - \hat{I}_{q+1,T} = - L_{q,T}$

Case 5: $q = T$ and $1 \leq p < T$ $b_{p,T} = \hat{I}_{p,T} - \hat{I}_{p+1,T} = - L_{p,T}$

Case 6: $p = q = T$ $b_{T,T} = \hat{I}_{T,T} = L_{0,T} + L_{1,T} + \ldots + L_{T-1,T} = S^{T-1}_{T}$

As we can see, we have two kinds for the coefficients $b_{p,q}$ of the matrix $A'^{-1}\,\hat{I}\,A^{-1}$:

- For $p \neq q$ they are equal to the opposites of $L_{p,q}$ (or $L_{q,p}$ for $p > q$)
- For $p = q$ and $p < T$ (case 3), the values $B^{p}_{p} + S^{p-1}_{p}$ corresponds to the quantities of information delivered by the transactions realised at the dates p, independently of the type of the transactions (purchase or resale). And this interpretation is also valid for $p = q = T$ because at the date T we can only have resales (case 6 : S^{T-1}_{T}).

Therefore, if we introduce the matrixes $\mathcal{L}$ and $\mathcal{T}$, the variance-covariance matrix of Lind is equal to: $\mathcal{V}(\text{LInd}) = \sigma_G^2 \, (A'^{-1}\,\hat{I}\,A^{-1})^{-1} = \sigma_G^2 \, (\mathcal{T} - (\mathcal{L} + \mathcal{L}'))^{-1}$

$$\mathcal{L} = \begin{pmatrix} 0 & L_{1,2} & L_{1,3} & \dots & L_{1,T} \\ 0 & 0 & L_{2,3} & & L_{2,T} \\ | & & & & | \\ 0 & 0 & 0 & \dots & L_{T-1,T} \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \qquad \mathcal{T} = \begin{pmatrix} B^1_1 + S^0_1 & 0 & \dots & 0 & 0 \\ 0 & B^2_2 + S^1_2 & \dots & 0 & 0 \\ | & & & & | \\ 0 & 0 & \dots & B^{T-1}_{T-1} + S^{T-2}_{T-1} & 0 \\ 0 & 0 & \dots & 0 & S^{T-1}_T \end{pmatrix}$$

<u>Lemma 11</u> : $\mathcal{V}(\text{LInd}) = \sigma_G^2\, \mathcal{T}^{-1} \sum_0^{+\infty} [\, (\mathcal{L} + \mathcal{L}')\mathcal{T}^{-1}\,]^i$

We have: $\mathcal{V}(\text{LInd}) = \sigma_G^2\, [\mathcal{T} - (\mathcal{L}+\mathcal{L}')]^{-1} = \sigma_G^2\, [\, (Id - (\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1})\, \mathcal{T}\,]^{-1} = \sigma_G^2\, \mathcal{T}^{-1} [\, Id - (\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1}\,]^{-1}$

In order to calculate explicitly the inverse of the matrix $Id - (\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1}$ we are going to use the theory of the normed vectorial spaces. We choose for the space $\mathbb{R}^T$ the norm:

$$\| (x_1, x_2, \dots, x_T) \|_1 = \sum_{i=1,\dots,T} |\, x_i \,|$$

It induces on the space of the square matrixes of dimension T a matrix norm defined as:

$$\| M \|_1 = \| (m_{i,j})_{i,j=1,\dots,T} \|_1 = \text{Max}_{\, j=1,\dots,T} \sum_{i=1,\dots,T} |\, m_{i,j} \,|$$

$\| M \|_1$ corresponds to the maximum value that we get when we sum the coefficients $|\, m_{i,j} \,|$ on each column of M. We can prove that the norm of the matrix $(\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1}$ is strictly smaller than 1. Indeed, the right product $(\mathcal{L}+\mathcal{L}')$ time $\mathcal{T}^{-1}$ is equivalent to a division of the columns of $\mathcal{L}+\mathcal{L}'$ by the corresponding diagonal values of $\mathcal{T}$. More precisely, the p^{th} column of $\mathcal{L}+\mathcal{L}'$ is:

$$(L_{1,p} , L_{2,p} , \dots , L_{p-1,p} , 0, L_{p,p+1} , \dots , L_{p,T})'$$

and the p^{th} diagonal element of $\mathcal{T}$ is equal to:

$$b_{p,p} = (L_{p,p+1} + L_{p,p+2} + \dots + L_{p,T}) + (L_{0,p} + L_{1,p} + \dots + L_{p-1,p})$$

We have all the terms of the p^{th} column of $\mathcal{L} + \mathcal{L}'$ in the denominator, plus the quantity $L_{0,p}$. As the $L_{i,j}$ are all positive, the sum of the absolute values of the p^{th} column of $(\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1}$ is strictly smaller than 1 (if $L_{0,p} \neq 0$). This argument is valid for all the columns, even the last one, consequently the norm $\| (\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1} \|_1$ is also strictly smaller than 1 (if the first line of the table of the $\{L_{i,j}\}$ do not have any zero). The interest of this result lies in this proposition:

Proposition :

$\| \ \|$ is a matrix norm on the set of the square matrixes of dimension T.

If the matrix M satisfies to: $\|M\| < 1$

Then the matrix series $S = \sum_0^{+\infty} M^i$ has a limit and $S (Id - M) = (Id - M) S = Id$

We get this way the formula: $\mathcal{V}(\text{LInd}) = \sigma_G^2 \, \mathcal{T}^{-1} [Id - (\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1}]^{-1} = \sigma_G^2 \, \mathcal{T}^{-1} \sum_0^{+\infty} [(\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1}]^i$

Appendix F: $H_p(t)$, $H_f(t)$ and the price index M

The expression of $H_p(t)$ is:

$$[\, H_p(t)\,]^{I^t} = \Pi_{i \le t < j} \left[\left(\Pi_{k'} p_{k',i} \right)^{1/(\Theta + (j-i))} \right] = \Pi_{i=0,\dots,t} \left[\Pi_{j>t} \left(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})} \right) \right]$$

For each i between 0 and t, the prices in the square brackets are a sub-sample of F_i, as exemplified in **Figure 5b** (for i = 2). This expression is close to a geometric average of the prices; the total mass of the weights is: $\sum_{j>t} \left(\sum_{k'} \inf(p_{k',i}) \right) = \sum_{j>t} \sum_{k'} (\Theta + (j-i))^{-1} = \sum_{j>t} L_{i,j} = L_{i,t+1} + L_{i,t+2} + \dots + L_{i,T} = B^t_i$. The two averages $\left[\Pi_{j>t} \left(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})} \right) \right]^{1/B^t_i}$ and M_i are not necessarily equal because a partial mean can vary around the global one. However, one can reasonably assume that: $\left[\Pi_{j>t} \left(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})} \right) \right]^{1/B^t_i} = M_i \exp(\upsilon(i,t))$, with $\upsilon(i,t) \approx 0$ and $E[\upsilon(i,t)] = 0$. The quantity $\upsilon(i,t)$ capture the variability of the average when it is calculated on a sub-sample of F_i. For $H_p(t)$, it gives :

$$[\, H_p(t)\,]^{I^t} = \Pi_{i=0,\dots,t} \left[\Pi_{j>t} \left(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})} \right) \right] = \Pi_{i=0,\dots,t} \left[M_i \exp(\upsilon(i,t)) \right]^{B^t_i}$$

$$= \left[\Pi_{i=0,\dots,t} M_i^{B^t_i} \right] \exp\left(B^t_0 \upsilon(0,t) + B^t_1 \upsilon(1,t) + \dots + B^t_t \upsilon(t,t) \right)$$

$$[\, H_p(t)\,] = \left[\Pi_{i=0,\dots,t} M_i^{B^t_i} \right]^{1/I^t} \exp\left((B^t_0 / I^t) \upsilon(0,t) + \dots + (B^t_t / I^t) \upsilon(t,t) \right)$$

If we denote $\upsilon^t = (B^t_0 / I^t) \upsilon(0,t) + \dots + (B^t_t / I^t) \upsilon(t,t)$, we get:

$$[\, H_p(t)\,] = \left[\Pi_{i=0,\dots,t} h_i^{B^t_i} \right]^{1/I^t} \exp(\upsilon^t) \qquad \text{with } \upsilon^t \approx 0 \text{ and } E[\upsilon^t] = 0.$$

Similarly for the future, that is for the sell-side, we establish that:

$$[\, H_f(t)\,] = \left[\Pi_{j=t+1,\dots,T} h_j^{S^t_j} \right]^{1/I^t} \exp(\nu^t) \qquad \text{with } \nu^t \approx 0 \text{ and } E[\nu^t] = 0$$

Table 1 : Times of noise equality in Case, Shiller (1987)

City	Atlanta	Chicago	Dallas	San Francisco
Θ	12.89	9.11	6.77	4.20

Table 2: Relevant repeat-sales for the time interval [t',t]

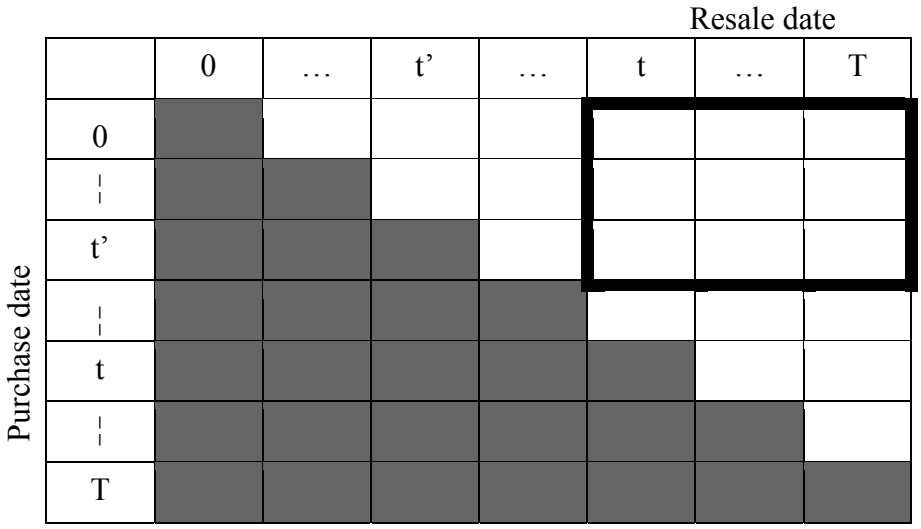


Table 3a: Real distribution for the repeat-sales sample

	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 3b: 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 4: Relevant repeat-sales for [t',t+1] and quantity of information associated

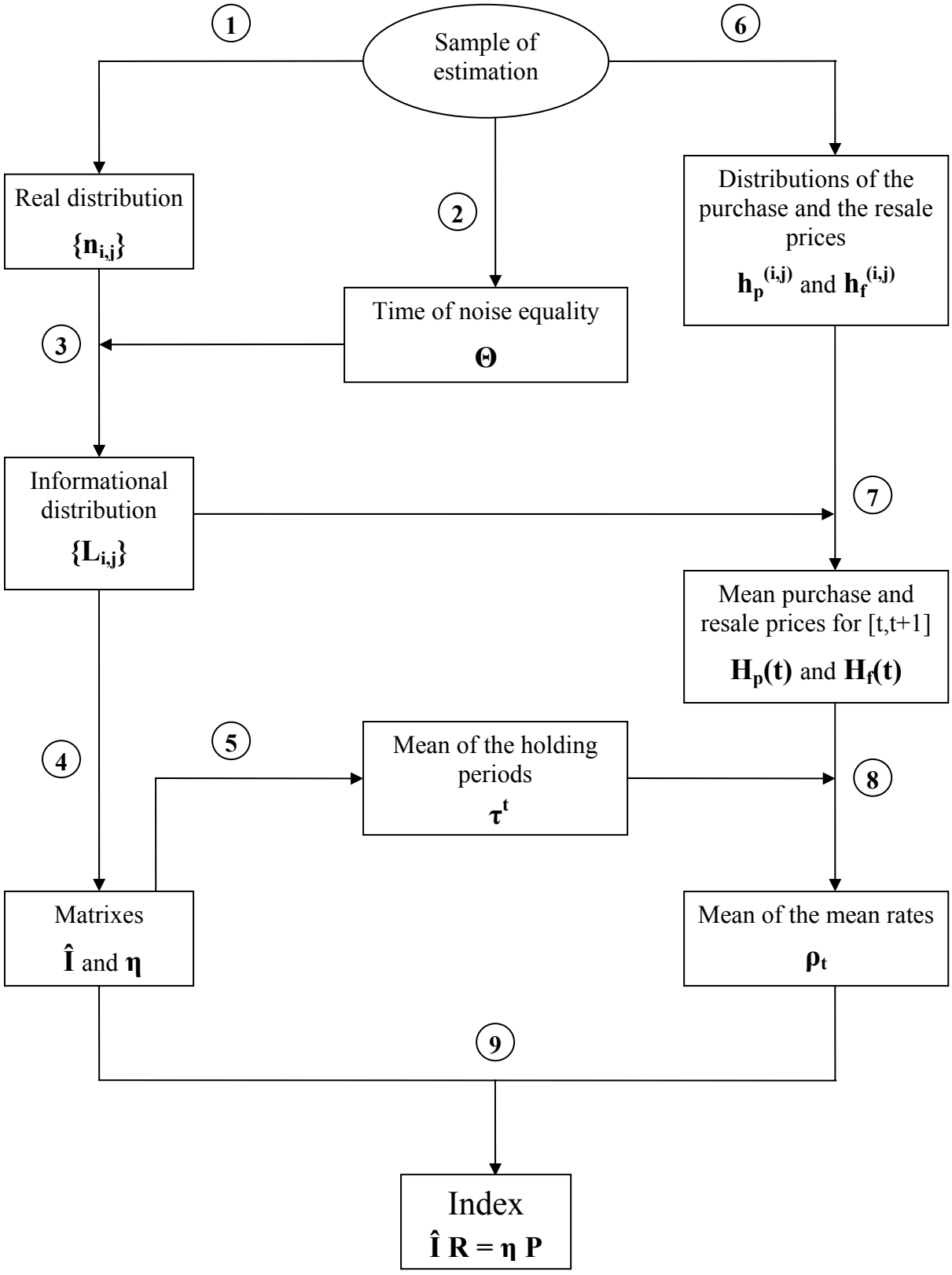
	0	…	t'	…	t	t+1		T	**Sum**
0			$L_{0,t'}$		$L_{0,t}$	$L_{0,t+1}$		$L_{0,T}$	B_0^t
⋮									⋮
t'					$L_{t',t}$	$L_{t',t+1}$		$L_{t',T}$	$B_{t'}^t$
⋮									⋮
t						$L_{t,t+1}$		$L_{t,T}$	⋮
⋮									⋮
T									⋮
					Sum	$S^{t'}_{t+1}$	…	$S^{t'}_T$	$I^{[t',t+1]}$

$B_0^t = L_{0,t+1} + \ldots + L_{0,T}$ $\quad B_{t'}^t = L_{t',t+1} + \ldots + L_{t',T}$ 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)

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

Figure 1: Algorithmic decomposition of the repeat-sales index



Legend of the **Figure 1**

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 realised between t_i and t_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 the holding period is including [t,t']. The diagonal elements of the diagonal matrix $\boldsymbol{\eta}$ are equal to the sums (on the rows or on the columns) of the components of the matrix $\hat{I}$.

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

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

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

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 calculated as 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:

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

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 and $H_f(t)$ for the resale, and the fictitious holding period τ^t

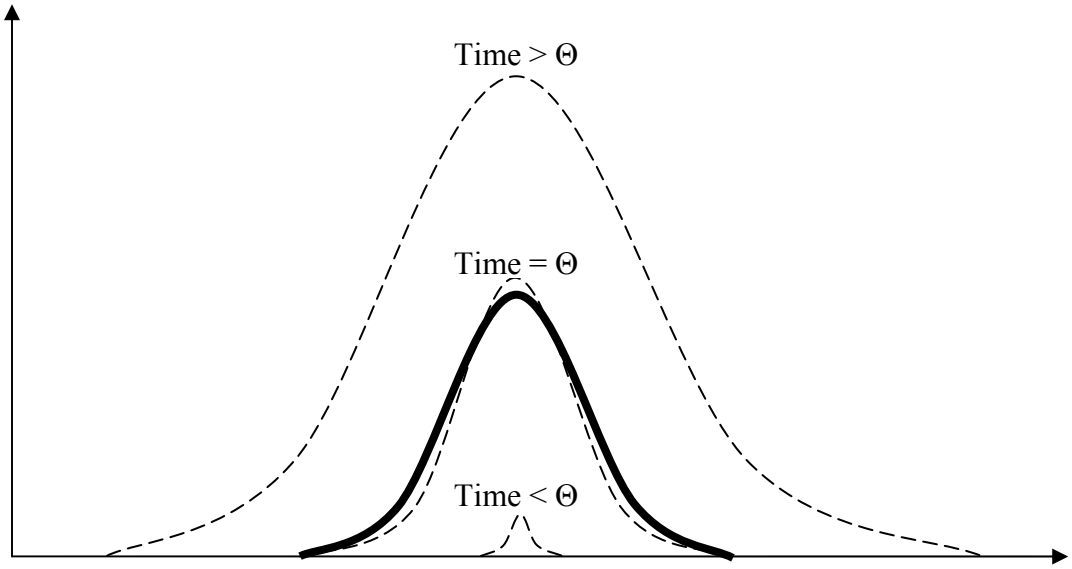
$$\mathbf{\rho_t = \left(1 / \tau^t \right) * \ln \left[H_f(t) / H_p(t) \right]}$$

9. The vector of the monoperiodic growth rates of the index, R, is the solution of the equation:

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

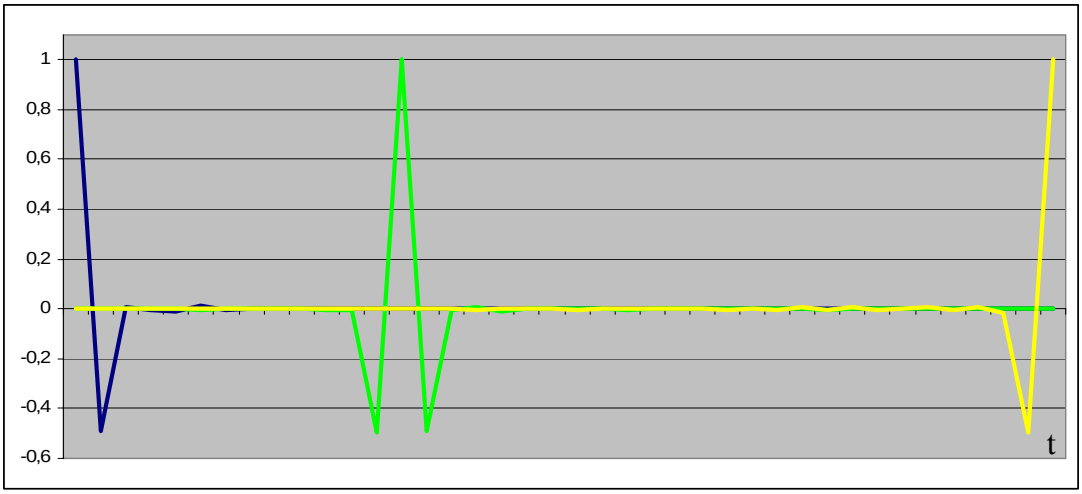
where P represents the vector the $(\rho_0, \rho_1, \ldots, \rho_{T-1})$

Figure 2 : time of noise equality



Continuous line : constant noise $2\sigma_N^2$ dotted lines : Gaussian noise σ_G^2 * time

Figure 3a: Correlograms of the r_t, for the dataset ω_0



Blue : t = 0 Green : t = 13 Yellow : t = 39

Figure 3b : Theoretical bias for ω_0 (%)

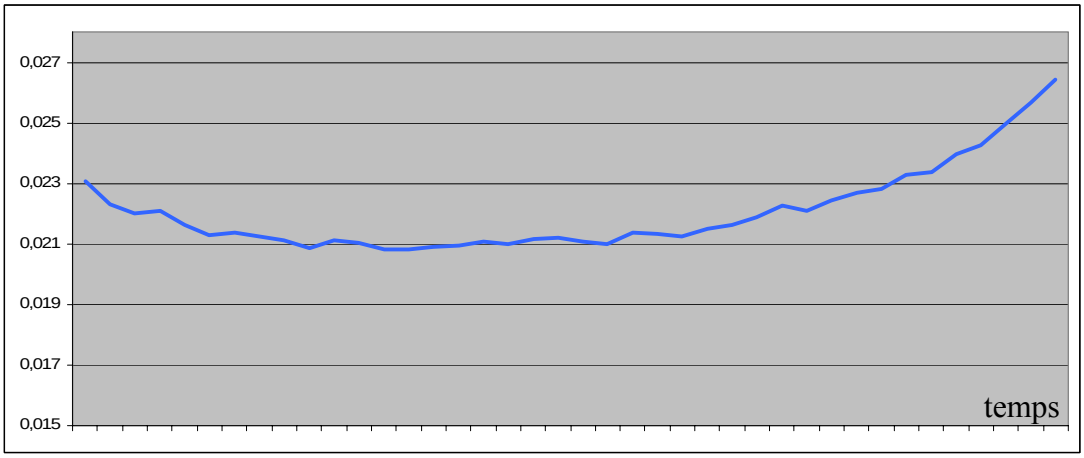


Figure 4: Is there a functional relation between the RSI and a price index?

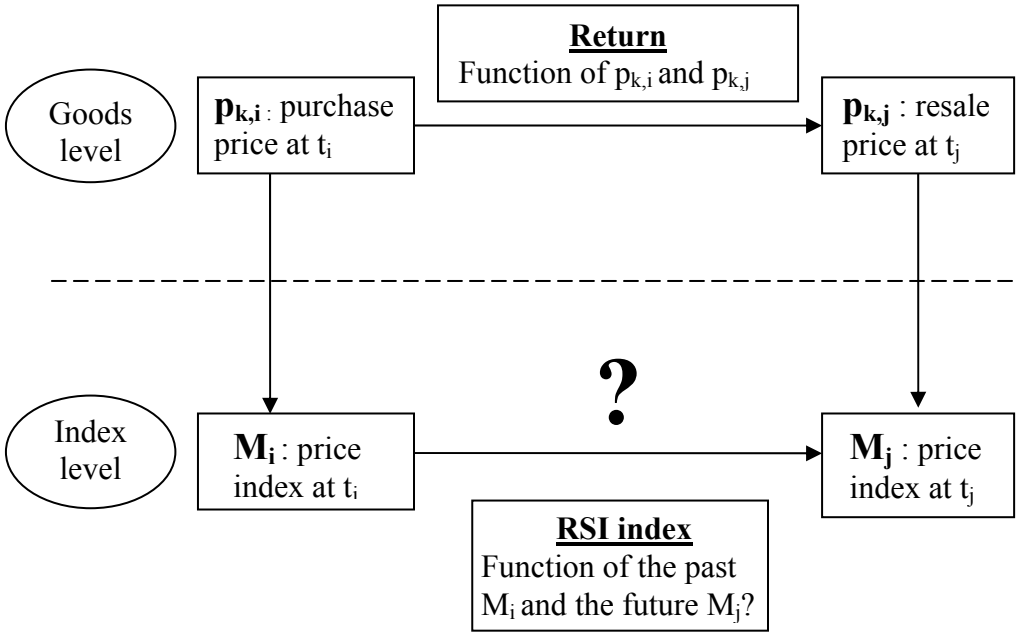


Figure 5a: Subset E_t for the price index at t

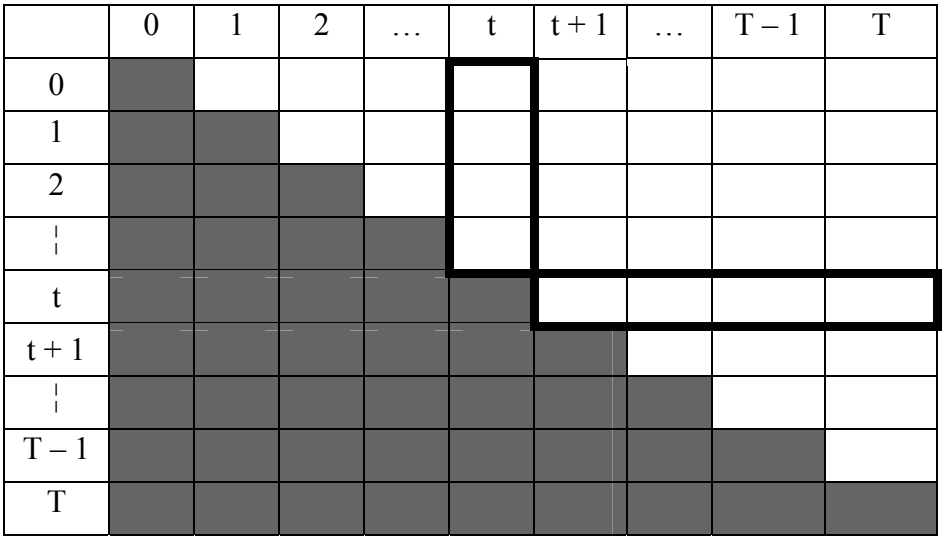


Figure 5b: Subset of F_2 for $H_p(t)$