Archives of Business Research – Vol.6, No.6
Publication Date: June. 25, 2018
DOI: 10.14738/abr.66.4216.
Hassane, M. (2018). Volatility and duration models for financial intaday data: formulation, estimation and evaluation. *Archives of Business Research,* 6*(6),* 66-84.

Volatility and duration models for financial intaday data: formulation, estimation and evaluation

Mamoudou Hassane
Lecturer, Money and International Finance,
Econometrics to the Faculty of Economics and Management of de
University Abdou Moumouni of Niamey. BP 12442 Niamey – Niger.

ABSTRACT

This paper develops and tests empirically counting models for high frequency data: BIN (1,1) model with Poisson process, to check if this model allows to capture the clusturing phenomenon in the case of high frequence data, concening stocks intaday data. The process of estimation of the model using data generating process (DGP), then using the acutal data coming from three stocks of NYSE place (BOEING, DISNEY, and AWK), involves good results that validate model for generalisation to BIN(n,n) and for works on density forcasting. In this paper we studie the issue of adequacy of BIN models to capture the activities of financial markets about stocks intraday data (volume, quote, prices), and help to forecast the evolution of financial markets activities.

Key words*:* ACD model, count data, BIN model, clustering, density forecast.
JEL classification: G14, G15.

I wish to thank Professor BAUWENS Luc, for the supervision of this work, Professor DEHEZ Pierre, Professor GIOT Pierre, all to Université Catholique de Louvain (Belgium). I am also grateful to Dr. VEREDAS David for his precious collaboration.

INTRODUCTION

It is usual to find time series consisting of *count data.* Such series record the number of events of a particular type occurring in a given interval. Since the data must consist of non-negative integers a model based on the normal distribution is not appropriate, although it might provide a reasonable approximation if the number of events observed in each time period is relatively large. Then for small numbers, the good distribution is a binomial process, but for a large number of observations the appropriate distribution is the Poisson. So, Poisson process should be used in the case of count data.

Considering an independent Poisson random variables, if $n_1,\ldots,n_q$ are independent with $n_i \sim Po(\ _i)$, then the total of all the counts is

$n_1 + n_2 + \ldots + n_q \sim Po(\ _1 + \ldots + \ _q)$ and the counts given the total are
$(n_1,\ldots,n_q)\big|N \sim Mult(N, p_1,\ldots, p_q)$ where $N = n_1 + \ldots + n_q$
and $p_i = \ _i/(\ _1 + \ldots + \ _q) \quad i = 1,\ldots,q.$

The conditional distribution is important for the analysis of log-linear models and it leads us to an analysis based on multinomial distribution.

Several authors as Engle and Russel (1998), Bauwens and Giot (1999) have previously worked on high frequency financial data through papers. These papers deal with the time between the

financials events as trades through autoregressive conditional duration (ACD) models, while the so-called BIN models used for count data deal with the number of events of the high frequency data (as trades) during fixed durations.

The autoregressive conditional duration (ACD) model of Engle and Russell (1998) is one of the most important models of the durations in econometric literature. It was formulated as follows:

$$x_t = \quad _t \quad _t, \qquad _t > 0, \;\; E(\ _t) = 1$$

where $x_t = \quad _t \qquad _{t\ 1}, \qquad t = 1,2,...$ is the length of time between financial events (trades), and the $_t's$ are independent identical distributed (i.i.d.), with

$$_t = \quad + \sum_{j=1}^{p} \quad _j x_{t\ j} + \sum_{j=1}^{q} \quad _j \quad _{t\ j}.$$

Here $_t = E\left(x_t \middle| F_{t\ 1}\right)$, the conditional expected waiting time. In practice Engle and Russell (1998) have used an exponential or Weibull distribution on the $\{\ _t\}$. Straightforward alternative structures would be to parameterize the $\log\ _t$ instead of the $_t$. This formulation called logACD proposed by Bauwens and Giot, allows to avoid to make constraints on parameters.

The two types of models are applied to high frequency data, particularly the financial data. The ACD models study the distribution of duration between the events (quote trades, volume or price duration), while the BIN models focus on the distribution of the number of events during a fixed length time. Then, the two types of models study the two faces of the same reality, but they could be considered as complementary than substitute.

The aim of this survey is to study the degree of relevance of BIN(1,1), the autoregressive form of the BIN models, in other words what is the degree of explanation of the financial market events, while in their paper, Rydberg and Shephard (2000) attempt to prove that, for modelling and forecasting the securities price changes on the stocks market, one could focus on N_i which are the count data.

COUNT DATA MODEL: BIN MODELS

In their paper Rydberg and Shephard (2000) proposed to model an asset price $p(r)$ at time r using a compound Poisson process

$$p(r) = p(o) + \sum_{r=1}^{N(r)} z_t, \qquad (1)$$

where $\{N(r)\}_{r\ 0}$ is a number of trades recorded up until r and z_t is the price movement or change associated with the t-th trade. Rydberg and Shephard (2000) specified $N(r)$ to be a counting process[1], modelled as Cox process – that is a Poisson process with a random intensity.

[1] The counting process, which is used in this context, states, that if $\{N(r)\}_{r\ 0}$ is a process with state space $U\{+\ \}$ and non-decreasing right continuous paths, then $\{N(r)\}_{r\ 0}$ is a counting process.

From an economic viewpoint these authors are typically interested in comparing the rate of return on holding the asset with that obtainable by other risky investments (opportunity cost) or riskless interest rate bearing accounts. In order to do this one has to compute the return over a fixed length of time 0. Then these returns will be based around the difference

$$\begin{aligned} p_i &= p\{(i+1) \} \quad p(i \ ) \\ &= \underset{t=1}{\overset{N\{(i+1) \}}{z_t}} \quad \underset{t=1}{\overset{N(i \ )}{z_t}} \\ &= \underset{t=N(i \ )+1}{\overset{N\{i+1)}{z_t}}. \end{aligned}$$

This shows that the number of trades in the interval $[i \ ,(i+1) \ ]$ plays a crucial role. To reflect this, Rydberg and Shephard specifies an expression as

$$N_i = N(i+1) \quad N(i \ ), \qquad (2)$$

the number of trades in that time interval[2]. This operation called "binning operation" consists to partitions time into sections and we count the number of trades in that interval. Notice that if $N_i = 0$, then $p_i = 0$, while $N_i > 0$ the prices can change. So, N_i is very important in determining the activity in the changes in the price level. For small values of there will be a negligible loss in information in doing this, compared to studying the complete record of the $\{N(r)\}$ process.

Let the $\{N_i\}$ and $\{z_t\}$ processes be stochastically independent and covariance stationary and assume that the $\{z_t\}$ are independent and identically distributed. Then, writing F_i as the information about the $\{N_i\}$ sequence available infinitesimally before time i by assuming the moments exist, will be

$$\begin{aligned} Var\left(p_i \middle| F_i \ \right) &= E\left\{Var\left(p_i \middle| N_i\right) \middle| F_i \ \right\} + Var\left\{E\left(p_i \middle| N_i\right) \middle| F_i \ \right\} \\ &= Var(z_t) E\left(N_i \middle| F_i \ \right) + E(z_t)^2 Var\left(N_i \middle| F_i \ \right). \end{aligned}$$

Thus predicting the variance of the price over the next period of length requires modeling the mean and variance of the future number of trades. In practice $E(z_t)$ will be too small and so what matters in the above setup is really only $E\left(N_i \middle| F_i \ \right)$. By setting $E(z_t) = 0$ then

$$\begin{aligned} Cov(p_i^2, p_{i+s}^2) &= E\left\{Cov\left(p_{i+s}^2, p_n^2 \middle| N_{i+s}, N_i\right)\right\} + Cov\left\{Var\left(p_{i+s} \middle| N_{i+s}\right), Var\left(p_i \middle| N_i\right)\right\} \\ &= Var(z_t)^2 Cov(N_i, N_{i+s}). \end{aligned}$$

Hence volatility clustering can be obtained with the autocorrelation of square price changes being proportional to the counts.

[2] Other derived financial activities could be used as object of count, as quote, and volume to obtain the count data.

A more specific result is obtained by assuming the $\{z_t\}$ have a first order moving average representation ((Rydberg and Shephard (2000)). Basically empirical modeling would require the assumptions that the $\{N_i\}$ and $\{z_t\}$ are stochastically independent.

Poisson process

Definition 1

The counting process $\{N(t), t \quad 0\}$, is a Poisson process of rate , $> 0,$ if

(i) $N(0) = 0$

(ii) The process has independent increments, in other words the distribution is memoryless;

(iii) The number of events in any interval of length t is Poisson process with mean t. That is for all $s, t \quad 0$

$$P\{N(t+s) \quad N(s) = n\} = e^{\ t(\ t)^n}, \ n = 0,1,...$$

Note that it follows from condition (iii) that a Poisson process has stationary increments and also that

$$E[N(t)] = \ t.$$

which explain why is called the rate of the Poisson process.

As a prelude to giving a second definition of a Poisson process we shall define the concept of function $f(\ )$ being $o(x).$

The function $f(\ )$ is said to be $o(x)$ if

$$\lim_{x \ 0} \frac{f(x)}{x} = 0.$$

Definition 2

The counting process $\{N(t), t \quad 0\}$, is said to be a Poisson process of rate , $> 0,$ if

(i) $N(0) = 0.$

(ii) The process has stationary and independent increment.

(iii) $P\{N(x) = 1\} = \ x + o(x)$

(iv) $P\{N(x) \quad 2\} = o(x)$.

N is a Poisson process with parameter .

$$P[N(t) = i] = \frac{(\ t)^i}{i!} e^{\ t}, \qquad t \quad 0.$$

Then, under these assumptions, $N(t) \sim P_O(\ )$; the duration between events follows exponential distribution of parameter : $d_i \sim \exp(\ )$, $i = 1,2,...$ and d_i are independent.

The hazard function is function of , and it *is constant*.

The particularity of BIN models is that *is random*. So, this last model, will be the topic of this survey.

The model and its properties

Structure of the model

In order to model the sequence $\{N_i\}$ Ridberg and Shephard (2000), suggested the BIN models

that specify the one-step ahead forecast distribution of $\{N_i\}$ series using a counting distribution. In particular they specify $N_i|F_i \sim P_O(\ _i)$, allowing $\ _i$ depending upon F_i, the information available infinitesimally before time i. Here $P_O(\ _i)$, denotes a Poisson distribution with mean $\ _i$, $\ _i$ is a linear function of past data as moving average models: BIN (1,1). Then the BIN(1,1) is given as follows

$$N_i|F_i \sim P_0(\ _i), \qquad _i = \ + N_{i\ 1} + \ _{i\ 1}, \qquad (3)$$

which is labeled a BIN(1,1) model. Sufficient conditions for $\ _i$ to be non-negative is that $\ , \ , \ \ 0$. This model is inspired to the GARCH model due to Bollerslev (1986) and Taylor (1986).

This model is thus autoregressive moving average (ARMA) type for

$$N_i = \ _i + u_i \qquad (4)$$

$$= \ + N_{i\ 1} + \ _{i\ 1} + u_i$$

$$= \ + N_i + \ (N_{i\ 1} \ \ u_{i\ 1}) + u_i \qquad (5)$$

$$= \ +(\ \ +\ )N_{i\ 1} + u_i \ \ u_{i\ 1}, \qquad (6)$$

which is can be analyzed as standard multivariate ARMA models with white noise error term, where $u_i = N_i \ \ _i$ is such that $E(u_i|F_i\ ) = 0$.[3] Then u_i is conditional independent identically distributed ($u_i \sim c.i.i.d$). Many of the interesting features for the BIN model follow from this structure. $u_i = N_i \ \ _i$ is a martingale difference (MD) which appears as an innovation for N_i. This equation (as in the case of ACD(1,1)) shows that a BIN(1,1) process corresponds to a constrained ARMA(1,1) representation for N_i, with autoregressive coefficient $\ +\ $, and moving average coefficient $\ $, and with a MD error term if $\ +\ < 1$. The autocorrelation function (ACF) could be obtained by the standard formulae for the ARMA(1,1) model. Main features (mean, variance, autocorrelation function) of this model are described in following points.

Statistical properties of the BIN models

By definition, the conditional expectation of $\{N_i\}$ is equal to $\ _i$. Then, equation (3) allows us to forecast expected counts, based on the information set at the previous period.

If $\{N_i\}$ is generated by (3) for $n \ \ $ with $\ , \ , \ > 0$, then if $\ +\ < 1$, the unconditional expectation ($\ \ $) and variance ($\ ^2$) of $\{N_i\}$ are given by

$$= E(N_i) = \frac{\ }{1 \ \ (\ +\ )}, \qquad (7)$$

[3] u_i is consider as a Martingale since $\ _i$ is the compensatory of N_i.

$$^{2} = \frac{1 \quad ^{2} \quad 2}{1 \quad (\quad + \quad)^{2}} = \quad + \quad \frac{^{2}}{1 \quad (\quad + \quad)^{2}} \qquad (8)$$

And the autocorrelation function is derived as

$$_{1} = \frac{\{1 \quad (\quad + \quad)\}}{1 + \quad ^{2} \quad 2 \quad (\quad + \quad)}, \qquad _{s} = \quad _{1}(\quad + \quad)^{s \ 1}, \quad s = 2,3,...$$

Another way to compute $_{s}$ is $_{s} = \quad _{s \ 1}(\quad + \quad)$ (9)

From the expression of (^{2}), it is easy to check that (—) that measures the overdispersion, is greater than zero. N_i becomes more dispersed than u_i when increases. From equation (8), it is easy to check that is greater than , if is greater than zero, and implies that the series observations is overdispersed.

The properties of $_{i}$ are sometimes helpful, in particular $E(\ _{i}) =$ and

$$Var(\ _{i}) = \quad ^{2}$$

$$= \quad \frac{^{2}}{1 \quad (\quad + \quad)^{2}}$$

$$= \frac{^{2}}{\{1 \quad (\quad + \quad)\}\{1 \quad (\quad + \quad)^{2}\}}.$$

$$Cov(N_i, \ _{i}) = E(\ _{i}^{2}) \quad \{E(\ _{i})\}$$
$$= Var(\ _{i})$$

Generalization of BIN (1,1) model

BIN (1,1) model can be generalized as a BIN model of order p,q (BIN (p,q)) where

$$N_i | F_i \ \sim P_O(\ _{i}), \qquad _{i} = \quad + \sum_{j=1}^{p} \ _{j} N_{i \ j} + \sum_{j=1}^{q} \ _{j} \ _{i \ j}.$$

As in the standard ARMA case, p denotes the number of autoregressive terms in the model and q to denote the number of moving average ones.

We have: $> 0, \ _{j} > 0, \ _{j} > 0$ that leads to the fact that $_{i}$ is a non-negative sequence when $q \quad 1$ with probability one. The ARMA representation of this model can be written as follows:

$$N_i = \ _{i} + u_i$$

$$= \quad + \sum_{j=1}^{p} \ _{j} N_{i \ j} + \sum_{j=1}^{q} \ _{j} (N_{i \ j} \quad u_{i \ j}) + u_i$$

$$= \quad + \sum_{j=1}^{\max(p,q)} \ _{j} N_{i \ j} + u_i \quad \sum_{j=1}^{q} \ _{j} u_{i \ j},$$

where $u_i = N_i \quad _i$ is also a martingale difference sequence and $_j = \ _j + \ _j$. This model is covariance stationary when $\sum_{j=1}^{\max(p,q)} \ _j < 1.$

This last assumption allows us to write the following properties

$$= \frac{}{1 \quad \sum_{j=1}^{\max(p,q)} \ _j}, \quad Cov(N_i, \ _i) = Var(\ _i) = \ ^2 \quad .$$

$$\text{And} \qquad Var(u_i) = Var(N_i) + Var(\ _i) \quad 2Cov(N_i, \ _i)$$

Then, one can compute easily the autocorrelation function of $\{N_i\}$ by using results on variances and autocorrelations of ARMA $\{\max(p,q),q\}$ process as in the case of BIN(1,1) form. The relevant work is focus on BIN (1,1) model.

Numerical illustrations

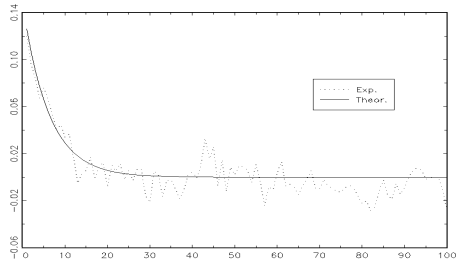
The first and second unconditional moments, and the autocovariances can be computed analytically as shown above. Then it is interesting to give numerical results about these moments and autocovariances for several sets of parameters.

Numerical simulations allow by using (7), (8), and (9) to compute the degree of overdispersion and to get a figure (Figure 1) that plots the autocorrelation function for four set parameters[4].

Similarly as in the case for the ARCH, GARCH, and ACD class of models, close to one implies a slowly decreasing autocorrelation function, and a large value of implies a large degree of overdispersion. Figure 1 gives the graphs of the theoretical and empirical autocorrelation function.

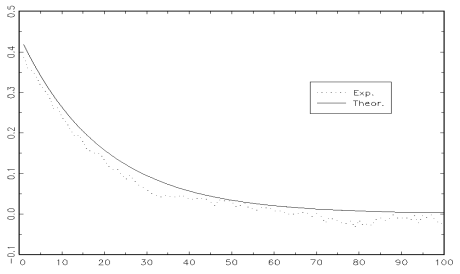
Figure 1: Theoretical and Empirical autocorrelation functions of the BIN (1,1) model

Figure 1a: gamma = 0.10, delta = 0.75



Lags

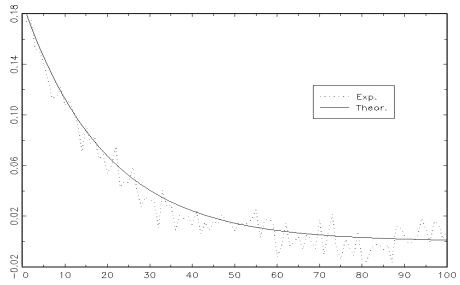
Figure 1b: gamma = 0.20, delta = 0.75



Lags

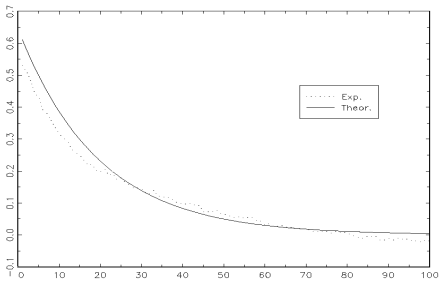
[4] The model is BIN(1,1) with the unconditional mean set equal to one, i.e. $= 1$

Figure 1c: gamma = 0.10, delta = 0.85



Lags

Figure 1d: gamma = 0.30, delta = 0.65



Lags

The figure 1c seems to exhibit the best-fitted representation of the model according to the residuals. Thus, the real values of the parameters γ and δ should be closed to 0.10 and 0.85 respectively. The missing parameter here is the time interval of the count data denoted Δ. It will be taken into account in the case of the empirical work using the actual data.

In other hand, we can examine the experimental overdispersion implied by changes in values of gamma and delta in the following table. The results are derived from a simulated data that allows a Data Generating Process.

Table 1: Overdispersion of BIN model

Coefficients	Overdispersion Ratio
$\gamma = 0.10, \ \delta = .75$	1.014 (1.017)
$\gamma = 0.20, \ \delta = .75$	1.184 (1.187)
$\gamma = 0.10, \ \delta = .85$	1.031 (1.050)
$\gamma = 0.30, \ \delta = .65$	1.339 (1.386)

The overdispersion ratio is defined as the ratio of standard deviation / mean, computed according to the formula (7) and (8). We parameterize λ to one, thus $\alpha = 1- \gamma - \delta$. In brackets we have the theoretical overdispersion ratio.

Results of Table 1 exhibit that the overdispersion ratio is an increasing function of γ and decreasing function of δ in BIN (1,1) model.

Estimation by Likelihood method

Consider $N_1,\ldots,N_T$ be the T non-negative integers events count observations for the dependent variable that is a random dependent variable which represents the number of events (here financial: quote, price or volume) that have occurred during the observation period i. Let that the events, which occur within each period, are independent and have constant rate of occurrence, then, N_i can by this fact follow a Poisson distribution with conditional probability

density function:

$$f_p(N_i, \ _i) = \begin{array}{l} \dfrac{e^{\ i}(\ _i)^{N_i}}{N_i!} \qquad \text{for} \quad _i > 0 \ \text{ and } \ N_i = 0,1,\ldots \\ 0 otherwise \end{array}$$

with expected value and variance $_i$ (the rate of event occurrence, that must be greater than zero) is assumed to be an exponential linear function of a vector of explanatory variables, x_i:

$$E(N_i) \qquad _i = \exp(x_i \ \ )$$

A constant term as the first element of x_i is included in the program, and one can include any number of explanatory variables.

The Poisson regression model that is the standard model for count data, is a non linear regression. This regression model is hence based upon the Poisson distribution with intensity parameter that depends on covariates regressors. In the case of missing of stochastic variation, and with exact parametric dependence, with exogenous covariates, then we have the standard Poisson regression. The mixed Poisson regression is obtained if the function relating and the covariates is stochastic, likely because it involves unobserved random variables, then assumptions must be done to take into account the random term for obtaining the precise form or to come back to the standard Poisson model.

The appropriate data are cross-sectional for applied work, which consist of T independent observations, indexed by i (N_i, x_i)[5]. N_i is the number of occurrence of the event object of study, and x_i is the vector of linearly independent regressors that are thought to determine N_i . A regression model based on this conditional distribution with a k-dimensional vector of covariates, $x_i = (x_{1i},\ldots,x_{ki})$, and parameters , through a continuous function $(x_i, \ )$, such that $E[N_i|x_i] = \ (x_i, \ )$. That is to say N_i given x_i is Poisson-distributed with density

$$f(N_i|x_i) = \frac{e^{\ i} \ _i^{N_i}}{N_i!}, \quad N_i = 0,1,2,\ldots \quad (1)$$

The log-linear form is the parameterization of the such that

$$_i = \exp(x_i \ \ ) \qquad (2)$$

to keep $> 0.$

The Poisson distribution property allows us to write $V(n_i|x_i) = E(n_i|x_i)$, with n_i considered as the realization of random variable N_i; then,

$$E(n_i|x_i) = \exp(x_i \ \ )$$
$$= \exp(x_{1i} \ _1)\exp(x_{2i} \ _2).....\exp(x_{ki} \ _k).$$

[5] Here x_i contains autoregressive components $(\ _{i\ 1}, N_{i\ 1})$ in the case of $BIN(1,1)$ model.

In the likelihood-based models, the joint density of the dependent variables is specified.

By assuming that the scalar random variable N_i, given the vector of regressors x_i and parameter vector , is distributed with density $f(N_i|x_i,)$[6]. The likelihood principle performs as estimator of the value that maximizes the joint probability of observing the sample values $N_1,...,N_T$. This probability is called the likelihood function, and it appears as a function of parameters conditional on the data. It is formulated as:

$$L( ) = \prod_{i=1}^{T} f(n_i|x_i,), \qquad (3)$$

This formulation allows suppressing the dependence of $L( )$ on the data and has assumed independence over i. This definition could be extended to time series data by allowing x_i to include lagged dependence and independent variables, even if it implicitly assumes cross-section data.

So, maximizing the likelihood function is equivalent to maximizing the log-likelihood function

$$l( ) = \ln L( ) = \sum_{i=1}^{T} \ln f(n_i|x_i,) \qquad (4)$$

Under the so-called regularity conditions that are conditions on continuity and differentiation, the Maximum Likelihood Estimator (MLE) $\hat{}_{ML}$ is the solution to the first order conditions.

$$\frac{l}{} = \sum_{i=1}^{T} \frac{\ln f_i}{} = 0 \qquad (5)$$

where $f_i = f(n_i|x_i,)$ and $\frac{l}{}$ is a $q \cdot 1$ vector.

The data generating process for n_i has density $f(n_i|x_i, _0)$ where $_0$ is the true parameter value. That is to say the asymptotic distribution of the MLE is usually obtained under the assumption that the density is correctly specified. Then, under the regularity conditions, $\hat{} \xrightarrow{P} _0$, so the MLE is consistent for $_0$.

Then,

$$\sqrt{n}(\hat{}_{ML} - _0) \xrightarrow{d} N[0, A^{-1}], \qquad (6)$$

where the $q \cdot q$ matrix A defined as

$$A = \lim_{T} \frac{1}{T} E \sum_{i=1}^{T} \frac{^2 \ln f_i}{} \Big|_{_0} \qquad (7).$$

[6] = in the BIN(1,1) case.

Simulation

In this section, we perform an algorithm for reference sample generation then, we compute the characteristics of theoretical model (after a parameterization), and the parameters from the generate sample. We did comparison between the two generated models. We also compute the autocorrelation function (see Figure1).

By Monte Carlo simulation of discrete distribution we can perform the estimation of Poisson distribution, but there exists in Gauss the command that allows to perform the estimation of a Poisson process.

Comparison analysis

By Data Generating Process (DGP) we realize a 1,000 observations sample.

The simulation allows us, by a play on parameters to obtain the following experimental results in Table 2 below, in brackets we have the theoretical results.

Table 2: mean s.d. skewness and kurtosis of simulated data

Coefficients	mean	s.d.	skewness	kurtosis
$\gamma = 0.10, \delta = .75$	1.021 (1.000)	1.035 (1.018)	1.050	4.153
$\gamma = 0.20, \delta = .75$	0.976 (1.000)	1.156 (1.187)	1.494	5.882
$\gamma = 0.10, \delta = .85$	1.009 (1.000)	1.041 (1.050)	1.101	4.384
$\gamma = 0.30, \delta = .65$	0.920 (1.000)	1.232 (1.387)	1.813	7.317

We parameterize λ to one, thus $\alpha = 1- \gamma - \delta$ in all computations.

The results allow us to draw the following conclusion.

$= 1.00$, $= 0.05$, $= 0.10$, $= 0.85$, we have the best fitted model.

Then, the means of respectively theoretical and empirical results are 1.000 and 1.009. The standard deviations of the theoretical and empirical results are respectively 1.050 and 1.041. The theoretical and empirical overdispersions are respectively 1.050 and 1.032, and are the lower overdispersion values for theoretical and empirical results. The graph of the autocorrelation function in this case indicates that it is the best-fitted model.

It is easy to check that the overdispersion coefficient, which is equal to, $\frac{N}{N}$ increases with .

Another tools to check the adequacy of the model are the skewness and the kurtosis.

The skewness measures the locations indicate the number around which the sample data are centered. It indicates the direction in which a frequency distribution (or frequency curve or

frequency polygon) leans. Then, the skewness equal to zero implies a symmetric distribution. We can also have the case of negative or positive skewness.
The kurtosis measures a distribution's peakedness, the degree to which one narrow range of values contains a large fraction of sample data. So, skewness indicates whether the histogram "leans to the left (negative value)" or "leans to the right (positive value)", and kurtosis indicates how peaked it is. Their formulas are

$$sk = \frac{M_3^2}{M_2^3}, \qquad kur = \frac{M_4}{M_2^2},$$

where M_k is the kth moment of a sample of ungrouped data.

The simulation results give $sk = 1.10$, which that the frequency curve (or density curve) leans to the right, then there are more values to the right of the model than to the left; $kur = 4.38 > 3$, that implies a peak curve.

These results are conforming to Poisson density function.

Validation of the model: Monte Carlo method
For purpose we used the data obtained by DGP for Maximum Likelihood estimation, and we get the following results in Table 3, 4, 5, 6. In brackets the fixed value of the parameter. The figures represent the graphs of counts and forecast counts.

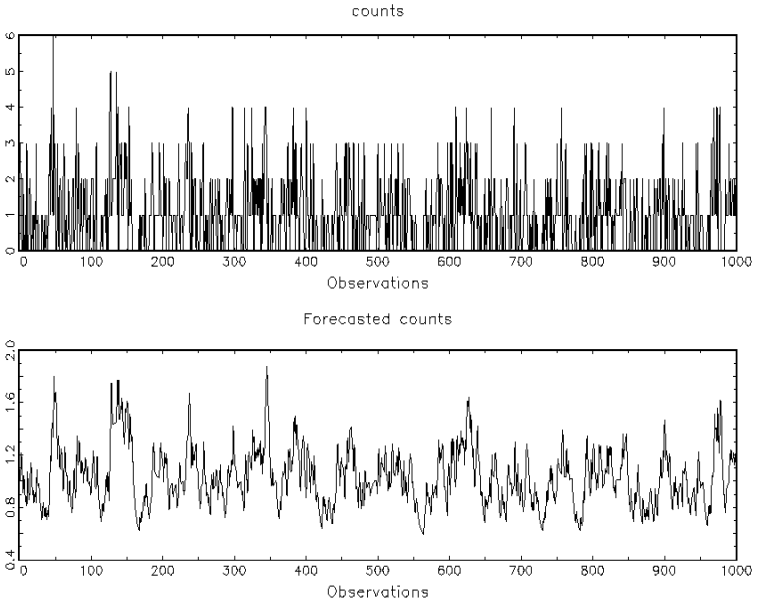
Table 3: Estimation results by MLE using DGP count data

Parameter	**parameter value**	**t-stat**
alpha	0.143 (0.150)	3.071
gamma	0.109 (0.100)	4.743
delta	0.752 (0.750)	12.981
Q(10)	79.87	
Q(10)*	9.77	

Q(10) and Q(10)* correspond respectively to the Ljung-Box Q-statistic of order 10 on counts (N_i) and Q-statistic on the residual u_i defined in the BIN(1,1) model. If Q is more than 18.307, then there is autocorrelation of order 10 for a threshold of 5 %.

The t-statistic must be compare to 1.96. If the t-stat is greater than 1.96 then the parameter is significant for a threshold of 5 %.

Figure 2: graphs of counts and forecast counts for estimation of Table 3



The results in Table 3 and the graphs of figure 2 show that the BIN(1,1) model behaves well, then the model may candidate for estimation and tests.

Table 4: Estimation results by MLE using DGP count data

Parameter	**parameter value**	**t-stat**
alpha	0.068 (0.050)	4.546
gamma	0.244 (0.200)	9.963
delta	0.694 (0.750)	22.397
Q(10)	1416.44	
Q(10)*	13.97	

Figure 3: graphs of counts and forecast counts for estimation of Table 4

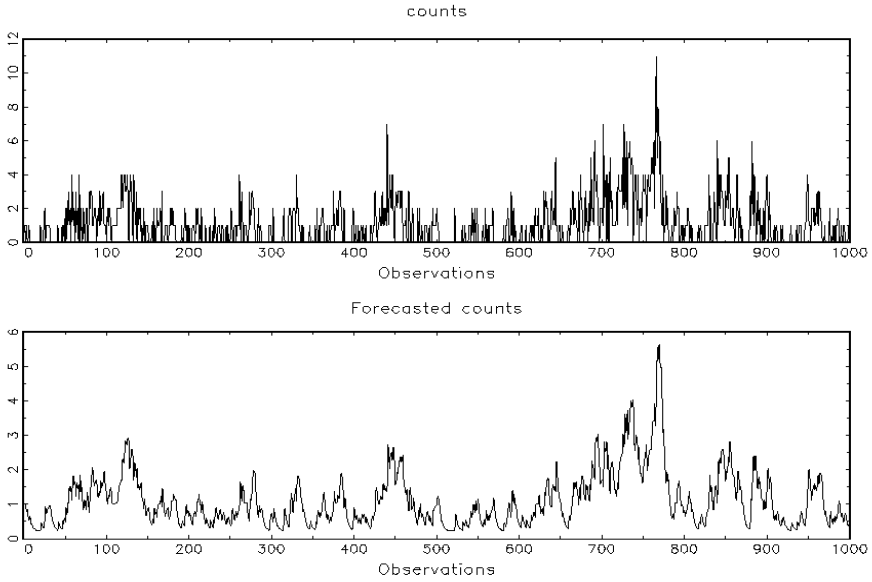
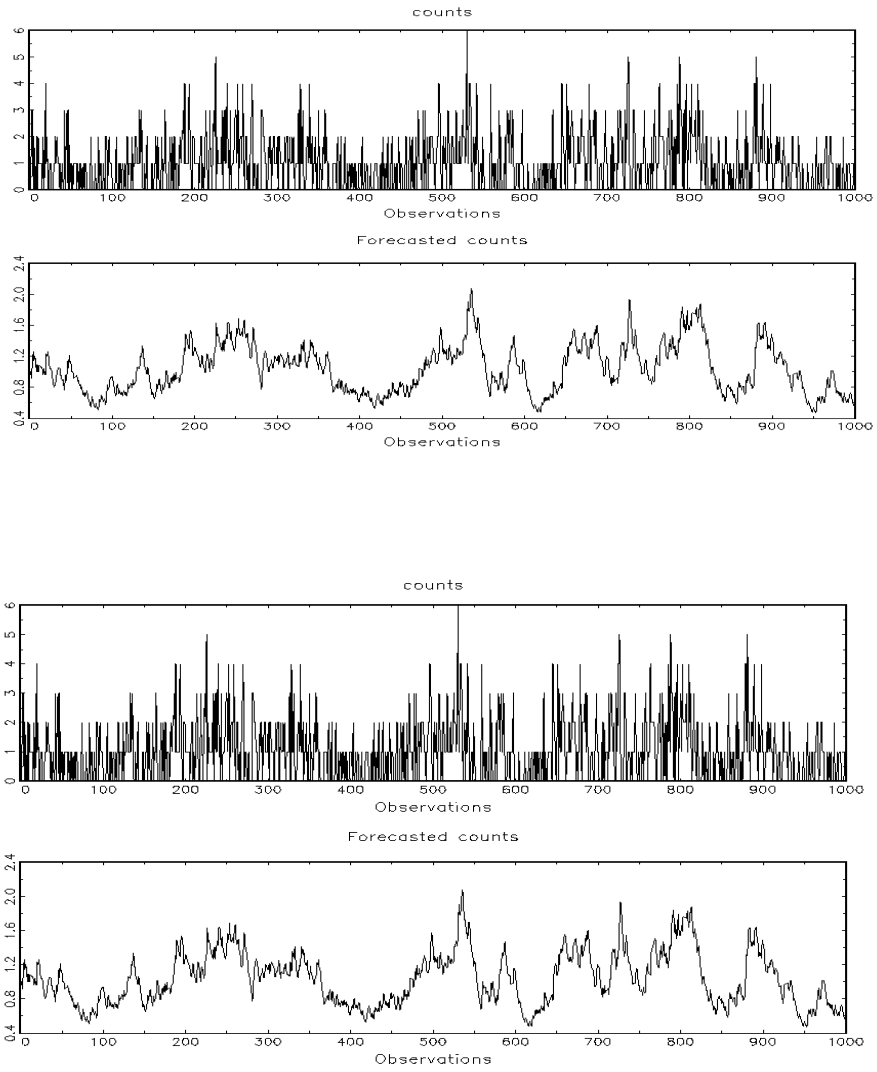


Table 5: Estimation results by MLE using DGP count data

Parameter	parameter value	t-stat
alpha	0.038 (0.050)	2.464
gamma	0.085 (0.100)	5.312
delta	0.88 (0.850)	34.996
Q(10)	159.88	
Q(10)*	15.24	

Figure 4: graphs of counts and forecast counts for estimation of Table 5



It easy to check through the table 3, 4, 5 and their corresponding graphs, that the model has a good behaviour with a MLE thus, it can be used for estimation and for forecasting. There is an absence of residual autocorrelation in the three cases above. Then we can apply it for actual data.

APPLICATION TO NYSE DATA

The data of three stocks

For empirical analysis, we choose a financial activity on two stocks traded on the NYSE: BOEING, DISNEY and AWK (American Water Work). The data that have been previously used for ACD by Bauwens, Giot and Veredas, were extracted from the Trade and Quote (TAQ) the database of the NYSE (for more detail, see CORE discussion paper of Bauwens and Giot (1999)). For the three stocks, we choose quote data for BOEING, quotes volume for DISNEY, and trades for AWK.

Before using these data, we must transform durations to counts. We get the data under the count data form according to Veredas method, that is program, which transforms durations to

counts. The count is made for a fixed length of time. In our case we use the interval of 0.25 second, 1 second and 4 seconds, and estimations are performed using each form of data. The estimation method used here is the Maximum Likelihood method. The results are analyzed through the section 3.2 below.

Estimation results

Table 6: Estimation results by MLE using quote count data for BOEING stock

Parameter	**parameter value**		
Δ	0.25	1	4
α	0.009 (0.002)	0.094 (0.021)	0.743 (0.153)
γ	0.054 (0.006)	0.146 (0.017)	0.257 (0.028)
δ	0.909 (0.014)	0.760 (0.034)	0.557 (0.057)
Q(10)	829.49	711.44	304.05
Q(10)*	41.08	28.10	5.60

In brackets we have the standard deviation. Q(10) and Q(10)* are the Ljung-Box Q-statistic of order 10 on counts (N_i) and Q-statistic on the residual u_i defined in the BIN(1,1) model. If Q is more than 18.307, then there is autocorrelation of order 10 for a threshold of 5 %.

Δ is the fixed length of time. (The number of cases is 10,491 for Δ = 0.25; 2,623 for Δ = 1; and 656 for Δ = 4).

Table 7: Estimation results by MLE using quote volume count data for DISNEY stock

Parameter	parameter value		
Δ	0.25	1	4
α	0.002 (0.000)	0.020 (0.007)	0.548 (0.169)
γ	0.009 (0.001)	0.049 (0.008)	0.255 (0.037)
δ	0.984 (0.002)	0.931 (0.013)	0.608 (0.064)
Q(10)	375.60	332.57	523.48
Q(10)*	410.29	44.52	11.46

In brackets we have the standard deviation. Q(10) and Q(10)* are the Ljung-Box Q-statistic of order 10 on counts (N_i) and Q-statistic on the residual u_i defined in the BIN(1,1) model. Δ is the fixed length of time. (The number of cases is 13,795 for Δ = 0.25; 3,449 for Δ = 1; and 863 for Δ = 4).

Table 8: Estimation results by MLE using trades count data for AWK stock

Parameter	parameter value		
Δ	0.25	1	4
α	0.004 (0.001)	0.013 (0.006)	0.468 (0.339)
γ	0.012 (0.002)	0.023 (0.005)	0.112 (0.034)
δ	0.973 (0.006)	0.963 (0.010)	0.771 (0.117)
Q(10)	147.74	187.35	189.66
Q(10)*	26.41	21.78	12.06

In brackets we have the standard deviation. Q(10) and Q(10)* are the Ljung-Box Q-statistic of order 10 on counts (N_i) and Q-statistic on the residual u_i defined in the BIN(1,1) model. Δ is the fixed length of time. (The number of cases is 26,225 for $\Delta = 0.25$; 6,557 for $\Delta = 1$; and 1,640 for $\Delta = 4$).

We can see that when the length of fixed time Δ increases, the values of α and γ increase too. Inversely, the increases of Δ make δ decreasing. That implies the volatility of the clustering phenomenon, and it is the basic motivation of the ARMA representation, to capture the law of the process of the high frequency counts data in financial market microstructure system.

The good results are given by the fixed length $\Delta = 4$, where there is no residual autocorrelation, in the estimation of the count data of the three stocks. Then, when Δ increases the estimation and tests results may be better, and the volatility of the clustering is more perceptible.

CONCLUSION

The aim of this survey is to built and test BIN(1,1) model for the counts data. Before generating the data by parametrization, we used these data for estimation by the ML method for the BIN(1,1) validation. The results exhibit a good behaviour of this model, so it could be applied to the actual data. It is what we did for empirical analysis. For purpose, we use the transformed data (durations to counts), for different fixed length of time. The results of estimation of three stocks that are object of trade on the NYSE (BOEING, DISNEY, and AWK) by the Maximum Likelihood method allow to draw the following remarks. There is less dependence between the high frequency data when we consider a large value of the fixed length of time, the value of gamma becomes more and more large which increases the dispersion. The model could be generalized to take into account other variables and it could be used for density forecasting.

References

Aman, U. and David, E. A. G. (1998) "Handbook of applied economic statistics." Marcel Dekker, inc. New York.

Bauwens, L. (1999) "Recent developments in the econometrics of financial markets Using intra-day data." CORE discussion paper 1403, Université Catholique de Louvain (UCL).

Bauwens, L. and Giot, P. (1999) "The logarithmic ACD model: an application to the bid-ask quote process of three NYSE stocks." CORE discussion paper 9789, (UCL).

Bauwens, L. and Veredas, D. (1999) "The stochastic conditional duration model: a latent factor model for the analysis of financial duration." CORE discussion paper 9958, UCL

Bera, A. K., and Higgins, M. L. "ARCH models: properties, estimation and testing." Journal of Economic Survey, vol. 7, No. 4.

Colin Cameron and Pravin K. Trivedi (1998) "Regression analysis of count data." Cambridge University Press.

Dawid, P. A. (1979) "Some misleading arguments involving conditional independence." JRSS B 41, No. 2, pp. 249-252.

Devolder, P. (1993) "Finance stochastique." Edition de l'Université de Bruxelles.

Droesbeke, J. Fichet B. and Tassi P. (1994) "Modelisation ARCH: Theorie statistique et application dans le domaine de la Finance." Editions de l'Université de Bruxelles, Bruxelles.

Dyckman T. and Thomas L. (1977) "Fundamental Statistics for Business and Economics." Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Engle, R. F. (1995), "ARCH selected readings". Oxford University Press.

Engle, R. F., and Granger, C. W. J. (1991) "Long run economic relationships: Readings in cointegration". Oxford University Press.

Engle, R. and Russell, J.(1998) "Autoregressive conditional duration, a new model for for irregularly spaced transaction data." Econometrica 66, 1127-1162.

GAUSS Applications (1995) "Constrained Maximum Likelihood." Aptech Systems, Inc. Maple Valey, USA.

Giot, P. (1999) "Time transformations, intra-day data and volatility models." CORE discussion paper 9944, UCL.

Greene, W. H. (1997), "Econometrics Analysis." Prentice-Hall Inc. New York.

Harvey, A. C. (1990) "The Econometric Analysis of Time Series." Second edition, Harvester Wheatsheaf.

----------(1993) "Time Series Models." Second Edition, Harvester Wheatsheaf.

Hendry, D. F. and ali (1993) "Cointegration, error correction, and the econometric analysis of non-stationary data." Oxford University Press.

Laure, B. and Gervais, J. P. (1998) "L'Analyse des séries chronologiques: spécification et estimation des modèles univariés et multivariés." CODESRIA, document spécial n° 9.

Lewis P. and Orav E. (1989) "Simulation methodology for Statisticians operations Analysts and Engineers." Vol. 1 Wadsworth, Inc., Belmont, California.

Mills, T. C. (1999) "The Econometric Modeling of financial Time Series." Second edition, Cambridge University Press.

Nerlove M., Grether D. M. and Carvalho (1979) "Analysis of Economic Time Series: A Synthesis." Academic Press, inc. New York.

Randolph, N. (1995) "Probability, Stochastic Processes, and Queuing Theory: The Mathematics of computer performance Modeling." Springer-Verlag New York, Inc.

Ridberg, T. and Shephard, N. (1999) "BIN models for trade-by-trade data. Modelling the number of trades in a fixed interval." Nuffield College, Oxford, working paper series 1999-w14.

Rubinstein R. Y. (1981) "Simulation and the Monte Carlo method." John Wiley & Sons, Inc., USA.

Sheldon M. Ross (2000) "Introduction to Probability Models." Seventh Edition, Academic Press, San Diego, USA.

Sigles

ACD: Autoregressive Conditional Durations
ARCH: Autoregressive Conditional Heteroscedasticity
ARMA: Autoregressive Moving Average
DGP: Data Generating Process
GARCH: General ARCH
MLE: Maximum Likelihood Estimation