What Can We Learn From High Frequency Data in Finance - Thirty Years Later?

Michel M. Dacorogna

Partner at Prime Re Solutions, Zug, Switzerland

ETH Risk-Center Seminar Series, Zurich, March 20, 2018
The historical work presented here today is the result of a team work done at Olsen & Associates from 1986 to 2000.

I am indebted to all these very creative team members:

- Ulrich A. Müller
- Richard B. Olsen, who made it all possible
- Olivier V. Pictet
- Gilles Zumbach
- And all the visitors like Ramazan Gençay, Gennady Samorodnitsky or Gerhard Stahl and students like Fulvio Corsi or Philippe Hartman

I would like to thank F. Corsi and R. Gençay for providing me with updates on the research in this field.
In the late eighties, *high frequency* data meant *daily* data. The usual frequency used in models were monthly or quarterly data.

The approach for most of the work done in economics and finance was to *first develop models* and then look for *data to justify* them.

Most academics considered data beyond daily as *noise*, not very useful to understand financial markets.

Quite to the contrary, we set up our research program *first* to collect *data*, then to *find regularities* and only *afterwards to develop models* that would reproduce these regularities.
The company **Olsen & Associates** started in 1985 and I joined it on September 1986.
Selected O&A References


Data Frequency and Research in Finance

- Relating the *type of data available* for researchers, to *the effects and the models* that are discovered and developed with these different samples, provides insight into the development of research in finance.

![Graph showing available sample sizes and time scales](image)
The high-frequency data have opened great possibilities to test market micro-structure models, while traditionally low-frequency data are used for testing macroeconomic models. In between lies the whole area of financial and time series modeling, which is typically studied with daily or monthly data as, for instance, option pricing or GARCH models.
financial markets are the source of high-frequency data. The original form of market prices is tick-by-tick data: each “tick” is one logical unit of information arriving at a time $t_j$, like a quote or a transaction price.

<table>
<thead>
<tr>
<th>Time</th>
<th>CCY</th>
<th>PAGE NAME</th>
<th>* REUTER SPOT RATES</th>
<th>CCY</th>
<th>HI<em>EURO</em>LO FXFX</th>
</tr>
</thead>
<tbody>
<tr>
<td>0727</td>
<td>DEM</td>
<td>RABO</td>
<td>RABOBANK UTR</td>
<td>1.6290/00</td>
<td>* DEM</td>
</tr>
<tr>
<td>0727</td>
<td>GBP</td>
<td>MNBX</td>
<td>MOSCOW LDN</td>
<td>1.5237/42</td>
<td>* GBP</td>
</tr>
<tr>
<td>0727</td>
<td>CHF</td>
<td>UBZA U B S</td>
<td>ZUR</td>
<td>1.3655/65</td>
<td>* CHF</td>
</tr>
<tr>
<td>0727</td>
<td>JPY</td>
<td>IBJX</td>
<td>I.B.J LDN</td>
<td>102.78/83</td>
<td>* JPY</td>
</tr>
<tr>
<td>0727</td>
<td>FRF</td>
<td>BUEX</td>
<td>UE CIC PAR</td>
<td>5.5620/30</td>
<td>* FRF</td>
</tr>
<tr>
<td>0726</td>
<td>NLG</td>
<td>RABO</td>
<td>RABOBANK UTR</td>
<td>1.8233/38</td>
<td>* NLG</td>
</tr>
<tr>
<td>0727</td>
<td>ITL</td>
<td>BCIX B.C.I.</td>
<td>MIL 1592.00/3.00</td>
<td>* ITL</td>
<td>1596.00</td>
</tr>
<tr>
<td>0727</td>
<td>ECU</td>
<td>NWNT NATWEST</td>
<td>LDN 1.1807/12</td>
<td>* ECU</td>
<td>1.1820</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>XAU</th>
<th>SBZG 387.10/387.60</th>
<th>* ED3</th>
<th>4.43/ 4.56</th>
<th>* FED</th>
<th>PREB</th>
<th>* GOVA 30Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>XAG</td>
<td>SBCM 5.52/ 5.53</td>
<td>* US30Y YTM</td>
<td>7.39</td>
<td>4.31–4.31</td>
<td>86.14–15</td>
<td></td>
</tr>
</tbody>
</table>
In the late eighties, time series in economics and finance were *homogeneous (equally spaced in time)*. Tick-by-tick data are *inhomogeneous* and the time between two ticks contains in itself information.

Prices on the FX-market come *in pairs*: $p_{bid}$ and $p_{ask}$ as they are “quotes” from the market-makers.

The most important variable under study is the *logarithmic middle price* $x$. At time $t_j$ it is defined as:

$$x(t_j) = \frac{\ln(p_{bid}(t_j)) + \ln(p_{ask}(t_j))}{2} = \ln(\sqrt{p_{bid}(t_j) \cdot p_{ask}(t_j)})$$

The inhomogeneous time series $x(t_j)$ can be transformed to a homogeneous one by using an interpolation method, we use the index $i$ for the homogeneous time series:

$$x(t_i) = x(\Delta t; t_i) = \frac{\ln(p_{bid}(t_i)) + \ln(p_{ask}(t_i))}{2}$$
Interpolation Methods for Tick Data

- There are essentially two interpolation methods:
  1. linear interpolation
  2. previous tick

Both methods have their merits. Previous tick interpolation respects causality as it exclusively uses information already known at time $t_0 + i \Delta t$, whereas linear interpolation uses information from time $t_{j+1}$, which lies in the future of time $t_0 + i \Delta t$. While, linear interpolation is the appropriate method for a random process with identically and independently distributed (i.i.d.) increments.
The return at time $t_i$ is defined as

$$r(t_i) = r(\Delta t; t_i) = x(t_i) - x(t_i - \Delta t)$$

where $x(t_i)$ is a homogeneous sequence of logarithmic prices.

The realized volatility $v(t_i)$ at time $t_i$ is computed from historical data and it is also called historical volatility. It is defined as

$$v(t_i) = v(\Delta t, n, p; t_i) = \left\{ \frac{1}{n-1} \cdot \sum_{k=1}^{n} \left| r(\Delta t; t_i-n+k) - \mu \right|^p \right\}^{1/p}$$

where the regularly spaced returns $r$ are defined as above, and $n$ is the number of return observations. There are two time intervals, which are the return interval $\Delta t$, and the size of the total sample, $n\Delta t$. The exponent $p$ is often set to 2 so that $v^2$ is the realized variance of the returns about the mean $\mu = \frac{1}{n} \left[ \sum_{l=1}^{n} r(\Delta t; t_i-n+l) \right]$.
Variables of Interest (2/2)

Bid-Ask Spread and Tick Frequency

- In bid-ask price pairs, the ask price is higher than the bid price. The bid-ask spread is their difference. A suitable variable for research studies is the relative spread $s(t_j)$:

$$s(t_j) = \ln p_{\text{ask}}(t_j) - \ln p_{\text{bid}}(t_j)$$

where $j$ is still the index of the original inhomogeneous time series. The nominal spread, $p_{\text{ask}} - p_{\text{bid}}$, is in units of the underlying price, whereas the relative spread is dimensionless; relative spreads from different markets can directly be compared to each other.

- The tick frequency $f(t_i)$ at time $t_i$ is defined as

$$f(t_i) = f(\Delta t; t_i) = \frac{1}{\Delta t} N \{ x(t_j) \mid t_i - \Delta t < t_j \leq t_i \}$$

where $N \{ x(t_j) \}$ is the counting function and $\Delta t$ is the size of the time interval in which ticks are counted.
A Changing Distribution Shape

- We plot here the cumulative frequency of USD-JPY for returns measured at 10 min, 1 day, and 1 week on the scale of the cumulative Gaussian probability distribution (Q-Q plot).

- Normal distributions have the form of a straight line, which is approximately the case for the weekly returns with a moderate (excess) kurtosis of approximately 1.3.

- The distribution of 10-min returns, however, has a distinctly s-shape form, which is a sign of fat-tails.

- We are in presence of a non-stable distribution with fat-tails.
### Extreme risks over 6 hr for model distributions produced by Monte-Carlo simulations of synthetic time series fitted to USD-DEM, compared to empirical FX data studied through a tail estimation

<table>
<thead>
<tr>
<th>Probabilities $(p)$</th>
<th>1/1 year</th>
<th>1/5 year</th>
<th>1/10 year</th>
<th>1/15 year</th>
<th>1/20 year</th>
<th>1/25 year</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Models:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0.4%</td>
<td>0.5%</td>
<td>0.6%</td>
<td>0.6%</td>
<td>0.7%</td>
<td>0.7%</td>
</tr>
<tr>
<td>Student 3</td>
<td>0.5%</td>
<td>0.8%</td>
<td>1.0%</td>
<td>1.1%</td>
<td>1.2%</td>
<td>1.2%</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>1.5%</td>
<td>2.1%</td>
<td>2.4%</td>
<td>2.6%</td>
<td>2.7%</td>
<td>2.9%</td>
</tr>
<tr>
<td>HARCH</td>
<td>1.8%</td>
<td>2.9%</td>
<td>3.5%</td>
<td>4.0%</td>
<td>4.3%</td>
<td>4.6%</td>
</tr>
<tr>
<td><strong>USD rates:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USD-DEM</td>
<td>1.7%</td>
<td>2.5%</td>
<td>3.0%</td>
<td>3.3%</td>
<td>3.5%</td>
<td>3.7%</td>
</tr>
<tr>
<td>USD-JPY</td>
<td>1.7%</td>
<td>2.4%</td>
<td>2.9%</td>
<td>3.2%</td>
<td>3.4%</td>
<td>3.6%</td>
</tr>
<tr>
<td>GBP-USD</td>
<td>1.6%</td>
<td>2.3%</td>
<td>2.6%</td>
<td>2.9%</td>
<td>3.1%</td>
<td>3.2%</td>
</tr>
<tr>
<td>USD-CHF</td>
<td>1.8%</td>
<td>2.7%</td>
<td>3.1%</td>
<td>3.5%</td>
<td>3.7%</td>
<td>4.0%</td>
</tr>
<tr>
<td>USD-FRF</td>
<td>1.6%</td>
<td>2.3%</td>
<td>2.8%</td>
<td>3.0%</td>
<td>3.3%</td>
<td>3.4%</td>
</tr>
<tr>
<td>USD-ITL</td>
<td>1.8%</td>
<td>2.8%</td>
<td>3.4%</td>
<td>3.8%</td>
<td>4.1%</td>
<td>4.4%</td>
</tr>
<tr>
<td>USD-NLG</td>
<td>1.7%</td>
<td>2.5%</td>
<td>2.9%</td>
<td>3.2%</td>
<td>3.4%</td>
<td>3.6%</td>
</tr>
</tbody>
</table>
There is no privileged time interval at which the data and the generating process should be investigated.

We analyze the dependence of mean volatility on the time interval on which the returns are measured.

The scaling law is empirically found for a wide range of financial data and time intervals in good approximation:

$$\left\{ \mathbb{E}[|r|^p] \right\}^{1/p} = c(p) \Delta t^D(p)$$

where $\mathbb{E}$ is the expectation operator, and $c(p)$ and $D(p)$ are deterministic functions of $p$.

This form for the left part of the equation is so that, for a Gaussian random walk, a constant drift exponent of 0.5 whatever the choice of $p$. 

M. Dacorogna Prime Re Solutions
High Frequency Data in Finance
Empirical Scaling Laws. On the y-axis, the natural logarithm of the mean absolute return ($p = 1$) is reported. The sample period is January 1, 1987, to December 31, 1995.

All results indicate a very general scaling law that relates time intervals and realized volatility and applies to different currencies as well as to commodities such as gold and silver or stock indices.
Intraday volatility in terms of mean absolute returns is plotted in the two top histograms of the figure for USD-DEM.

Both histograms indicate distinctly uneven intraday-intraweek volatility patterns. The daily maximum of average volatility is roughly four times higher than the minimum.

The pattern can be explained by considering the structure of the world market, which consists of three main parts with different time zones: America, Europe, and East Asia.

Similar patterns are found on the stock markets (U-shape) that have opening and closing hours.
The intradaily and intraweekly \textit{seasonality of volatility} is a dominant effect that \textit{overshadows} many further stylized facts of high-frequency data. In order to continue the research for stylized facts, we need a powerful treatment of this seasonality.

Our model for the seasonal volatility fluctuations introduces a \textit{new time scale} such that the transformed data in this new time scale do not possess intraday seasonalities.

The construction of this time scale utilizes two components: the \textit{directing process}, $\theta(t)$, and a \textit{subordinated price process} generated from the directing process $x(t) = x^*[\theta(t)]$. The process $x^*$ does not have intraday seasonality.
The autocorrelation function of volatility decays at a *hyperbolic rate* rather than an exponential rate.

To illustrate the presence of the long memory, *two curves*, one *hyperbolic* and one *exponential*, are drawn in the Figure together with the empirical autocorrelation functions.

The hyperbolic curve approximates the autocorrelation function much more closely than the exponential curve.

A small local maximum at a lag of around 1 average business day (one-fifth of a week in $\theta$-time); a small local maximum at a lag of 2 business days and maxima at 3 and 4 business days also exist: a *market-dependent persistence* of absolute returns.
The Heterogeneous Market Hypothesis

- Financial markets are made of traders with *different trading horizons*. In the heart of the trading mechanisms are the *market makers* (now also High Frequency Traders HFT)

- A next level up are the *intraday traders* who carry out trades only within a given trading day

- Then there are day traders who may carry positions overnight, *short-term* traders and *long-term* traders

- Each of these classes of traders may have their own trading tool sets and may possess a *homogeneous appearance* within their *own classes*

- What we see is in the price time series is the *sum of these activities*
Modelling the Data

A Signature of Market Heterogeneity: Fine and Coarse Volatility

For exploring the behavior of various traders, we analyze volatility measured at different frequencies. We define a fine and a coarse volatility as:

\[ v_f(t_i) = \sum_{k=1}^{n} |r(\Delta t'; t_i - 1 + k\Delta t')| \quad \text{and} \quad v_c(t_i) = |\sum_{k=1}^{n} r(\Delta t'; t_i - 1 + k\Delta t')| \]

where \(\Delta t' \equiv \Delta t/n\)

In the figure below, we illustrate this definition where at every time point, \(t_i = t_{i-1} + 6\Delta t'\), both quantities are simultaneously defined.
Lead-Lag Correlation

- Lagged correlation reveals *causal relations* and information flow structures in the sense of *Granger causality*.

- If two time series were generated on the basis of a *synchronous information flow*, they would have a *symmetric lagged correlation* function $\rho_{-\tau} = \rho_\tau$.

- Here we see that the deviations between $\rho_{-\tau}$ and $\rho_\tau$ are significant, there is asymmetry in the information flow and a causal relation: coarse volatility predicts better fine volatility than the other way around.
The idea is to define a set of *partial volatilities*, $\sigma_j$, measured at different frequencies:

$$\sigma_{j,t}^2 = \mu_j \sigma_{j,t-1}^2 + (1 - \mu_j) \left( \sum_{i=1}^{k_j} r_{t-i} \right)^2$$

where $k_j = p^{j-2} + 1$ for $j > 1$ with $k_1 \equiv 1$, $\mu_j = e^{-2/(k_j+1-k_j)}$, and $p$ can be chosen arbitrary but here is 4.

Then, we define the HARCH process as a *linear combination* of these partial volatilities:

$$\sigma_t^2 = C_0 + \sum_{j=1}^{n} C_j \sigma_{j,t}^2 \text{ with } r_t = \sigma_t \varepsilon_t$$

and $\varepsilon_t$ is iid $\mathcal{N}(0,1)$.

It is easy to prove that a necessary stationarity condition is $\sum_{j=1}^{n} k_j C_j < 1$.

We define the *impact of the component j* as $I_j = k_j C_j$. 
The result of the optimization procedure is a set of $C_j$ coefficients from which the component impacts $I_j$ are calculated. The sum of impacts $I_j$ must be below one for stationarity of the process.

The impact of the fourth component is the weakest among all impacts. The fourth component has a typical time horizon of around 12 hr - too long for intraday dealers and too short for other traders. This naturally explains the weakness of that component.

The short-term components have, in all cases, the largest impacts. These short-term components model essentially the intraday dealers and the market makers who are known to dominate the markets.

The similarity in the impacts of the USD-CHF and USD-DEM are plausible as it is well known that the Swiss National Bank policy was tightly tied to the USD-CHF to the USD-DEM rates.

The relative weakness of the longer-term components for the GBP-USD is another relevant piece of information. In the late years of the $X X^{th}$, the long-term investors were reluctant to invest in this market since 1992 and were more concentrated on the cross rate GBP-DEM.
We construct a time series of realized hourly volatility, $v_{h,t}$, from our time series of returns as follows:

$$v_{h,t} = \sum_{i=1}^{a_h} r_{t-i}^2$$

where $a_h$ is the aggregation factor. In this case, we use data points every 10 min in $\theta$-time, so the aggregation factor is $a_h = 6$

Forecasts of four different models are compared to the realized volatility. The one-step ahead forecasts are based on hourly returns in $\theta$-time.

We use an out-of-sample period of 5 years of hourly data, which represents more than 43,000 observations to compare the accuracy of four forecasting models to the realized hourly volatility with three quality measures ($Q_d$, direction quality, $Q_r$, the realized potential, $Q_f$, the improvement of the absolute forecasting errors).
The forecast accuracy is remarkable for all ARCH-type models. In more than two-thirds of the cases, the forecast direction is correctly predicted and the mean absolute errors are smaller than the benchmark errors for all models.

For all measures, three parameter models perform better than the benchmark and the HARCH performs the best.

The realized potential $Q_r$ is the only measure that consistently improves with dynamic optimization.

<table>
<thead>
<tr>
<th>USD-DEM</th>
<th>$Q_d$</th>
<th>$Q_r$</th>
<th>$Q_f$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Static Optimization</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benchmark</td>
<td>67.7% (67.6%)</td>
<td>54.2% (54.3%)</td>
<td>0.000</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>67.8% (67.3%)</td>
<td>58.5% (59.7%)</td>
<td>0.085 (0.072)</td>
</tr>
<tr>
<td>HARCH(7c)</td>
<td>69.2% (68.7%)</td>
<td>58.3% (59.2%)</td>
<td>0.134 (0.129)</td>
</tr>
<tr>
<td>EMA-HARCH(7)</td>
<td>69.4% (68.8%)</td>
<td>60.7% (62.5%)</td>
<td>0.140 (0.128)</td>
</tr>
<tr>
<td><strong>Dynamic Optimization</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benchmark</td>
<td>67.7% (67.4%)</td>
<td>54.2% (54.6%)</td>
<td>0.000</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>67.0% (66.0%)</td>
<td>59.5% (59.8%)</td>
<td>0.074 (0.057)</td>
</tr>
<tr>
<td>HARCH(7c)</td>
<td>67.7% (66.8%)</td>
<td>60.1% (60.8%)</td>
<td>0.113 (0.102)</td>
</tr>
<tr>
<td>EMA-HARCH(7)</td>
<td>68.8% (67.7%)</td>
<td>62.4% (62.9%)</td>
<td>0.133 (0.117)</td>
</tr>
</tbody>
</table>
Since Thirty Year, What Has Changed?

- Nowadays High Frequency Trading (HFT) represents 80% of the activity of the main exchanges and probably also of the FX market.

- The markets have experienced few flash-crashes with quick recoveries, but an episode like 1987 (sort of precursor) is not excluded.

- All trading is now fast, with technological improvements originally attached to HFTs permeating throughout the market place.

- One might have expected that when things are fast the market structure becomes irrelevant - the opposite is actually the case. At very fast speeds, only the microstructure matters.

- All HFTs are strategic because their goal are generally to be the “first in line” to trade.
Since Thirty Year, What Does Remain?

- **Financial markets** remain the privileged place of the *price discovery* process
- Volatility clustering and fat tails are still present
- The *market makers* dictate the conditions in the market
- *Short-term volatility* continues to *increase* favored by the increase of HFT (J. Hasbrouck, 2016)
- *Volume* continues to *increase* reaching levels unheard of
- *Gaussian assumptions* are still *prevailing* among practitioners and academics
Since Thirty Year, New Research Developments

- Development of models along the *HARCH approach* (MC-HARCH Zumbach and Lynch 2001, HAR-RV Corsi 2009)

- Research concentrates on *market micro-structure* and the effects of HFT (O’Hara 2015, Hasbrouck 2016, Mahmoodzadeh and Gençay, 2017)

- The *flash crashes* have also attracted a lot of attention among researchers on HFD (Gençay *et al.* 2016)


- Studies on the effects of *regulation on HFT trading* (in Canada, Malinova *et al.* 2018)
Conclusions

- Collecting and analyzing HFD in the late eighties was a real innovation, big-data 25 years before big-data!

- Olsen & Associates was an exceptional place for developing creative research recognized by both academics and professionals.

- Access to HFD was at the origin of a new dynamics in research and in the markets that gave birth to the HFT and refined mathematical methods.

- Our understanding of markets has evolved: heterogeneous market agents.

- HFT represents a challenge to research but also generally to society. The role of financial markets as the price discovery process is put in danger.

- The goal of HFT is to buy and sell at the same time and thus pocket the spread, rather than intelligently discover the information.

- Academic analyses show that the nominal spread has decreased thanks to HFT. They neglect the fact that the short-term volatility has increased and thus the effective spread for long-term traders has actually increased. Researchers should concentrate on studying this effective spread rather than looking at nominal spread.
Some Recent References*


*) Subjective and by no means exhaustive