Affecting Market Efficiency by Increasing Speed of Order Matching Systems on Financial Exchanges – Investigation using Agent Based Model

Takanobu Mizuta
SPARX Asset Management Co. Ltd., Tokyo, Japan
Email: mizutata@gmail.com
http://www.geocities.jp/mizuta_ta/

Yoshito Noritake
Daiwa SB Investments Ltd., Tokyo, Japan

Satoshi Hayakawa
Osaka Exchange, Inc., Tokyo, Japan

Kiyoshi Izumi
School of Engineering, The University of Tokyo, Tokyo, Japan

Abstract—Recently, the speed of order matching systems on financial exchanges has been increasing due to competition between markets and due to large investor demands. There is an opinion that this increase is good for liquidity by increasing the number of traders providing liquidity. On the other hand, there is also the opposite opinion that this increase might destabilize financial markets and increase the cost of such systems and of investors’ order systems. We investigated price formations and market efficiency for various “latencies” (length of time required to transport data); while other settings remained the same, by using artificial market simulations which model is a kind of agent based models. The simulation results indicated that latency should be sufficiently smaller than the average order interval for a market to be efficient and clarified the mechanisms of the direct effects of latency on financial market efficiency. This implication is generally opposite to that in which the increase in the speed of matching systems might destabilize financial markets.

I. INTRODUCTION

Recently, the speed of order matching systems on financial exchanges has been increasing due to competition between markets and due to large investor demands. There is an opinion that this increase is good for liquidity by increasing the number of traders providing liquidity [1].

On the other hand, there is also the opposite opinion that this speed increase might destabilize financial markets by high-frequency traders (HFTs) and increase the cost of such systems and of investors’ order systems [2], [3]. In fact that IEX which is an alternative trading system started from 2013 in USA intentionally delays to provide price data to all traders for 350 microseconds [4].

The most important factor regarding this speed increase is “latency”; the length of time required to match orders and transport data. When latency is smaller, a matching system is faster. Such latency is very difficult to discuss only using the results of empirical studies. Because so many factors cause price formation in actual markets, an empirical study cannot be conducted to isolate the direct effect of latency to price formation. Furthermore, empirical studies cannot be conducted to investigate situations that have never occurred in actual financial markets.

We usually discuss whether changing market systems’ specification such as latency is good or bad for the market on the basis of their effects on price formation. An artificial market, which is a kind of agent based models, can isolate the direct effect of changes in matching systems, such as the speed increase to price formation, and can treat situations that have never occurred. These are strong advantages for an artificial market simulation, and the effects of several changing regulations have been investigated by using such simulations [5]–[8].

Not only academics but also financial regulators and stock exchanges are recently interested in agent based models such artificial market models to investigate regulations and rules of financial markets. Indeed, the Science article by Battiston et al. [9] described that ‘since the 2008 crisis, there has been increasing interest in using ideas from complexity theory (using network models, agent based models, and so on) to make sense of economic and financial markets’.

Many studies have investigated the effects of several changing regulations and rules by using artificial market simulations, for example, investigating effects of short selling regulations [10], [11], transaction taxes [12], financial leverages [11], [13], circuit breakers [14], price variation limits [15]–[17], cancel order tax [18], tick sizes [19], frequent batch auctions [20], and dark pools [21]–[23]. Of course, many artificial market simulation studies investigated the nature of financial markets, for examples, market impacts [24], [25], financial market crush [26]–[28], interaction between option markets and underlying markets [29], [30], effects of leveraged ETF [31] and effects of market makers and passive funds [32].

Indeed, effects of high-frequency traders (HFTs) [33]–[38] and of arbitrage trading between markets that have different latencies [39] to price formation were investigated using artificial market simulations. However, the direct effect of latency to price formation, its mechanisms, and how much of an increase in speed is needed for market efficiency have not been investigated.

Therefore, in this study we investigated the direct effects of latency on financial market efficiency by using an artificial
market model. We also implemented latency to the model by Mizuta et al. [23] and investigated price formations and market efficiency for various latencies; while other settings remained exactly the same.

II. ARTIFICIAL MARKET MODEL

We constructed a simple artificial market model in which the latency was additionally implemented onto Mizuta et al.’s model [23], which was constructed on the basis of the model of Chiarella and Iori [40].

Mizuta et al.’s model [23] replicates high-frequency micro structures, such as execution rates, cancel rates, and one-tick volatility, which cannot be replicated with Chiarella and Iori’s model [40]. Chiarella and Iori’s model [40] is very simple but replicates long-term statistical characteristics observed in actual financial markets: fat-tail and volatility clustering. The simplicity of the model is very important for this study because an unnecessary replication of macro phenomena leads to models that are over-fitted and too complex and such models would prevent our understanding and discovering mechanisms affecting the price formation because of related factors increasing.

We explain the basic concept for constructing our artificial market model in the Appendix.

A. Latency Model

The model treats only one risk asset and non-risk asset (cash). The number of agents is n. First, agent 1 orders to buy or sell the risk asset; after that, agent 2 orders, then agents 3, 4, ..., n. After the last agent n orders, the cycle is repeated.

Time t increases by δt (the time interval of an order) every time an agent gives an order to the matching system. In this study, we assumed an order obeys the Poisson process. Therefore, δt is determined by random exponential variables distributed in an average agent order interval δω. Many previous empirical studies for various financial assets [41] showed that a δt roughly obeys an exponential distribution; however, they also showed that there is some separation from the distribution.

As we mentioned in Section I, latencies exist in actual financial markets when information of an order is transported to a matching system from investors, orders are matched in the system, order-book information is updated, and information is transported to the investors through the information delivering system. Shortening latencies of a matching system reduces time from an investor ordering to him/her receiving order-book information. In this study, for simplification, latencies are zero when information of an order is transported to a matching system from investors, orders are matched in the system, and order-book information is updated. Finite latency only exists in the transportation of traded price information from a matching system to an agent1 (Figure 1). In other words, agents observe the market price before δl from the true market price. Shortening δl reduces the time difference between true price and observed price. We define $P_t^e$ as an agent’s observed market price at time t.

B. Trading Market Model

Our market model adopts a continuous double auction to determine the market price of a risk asset. A continuous double auction is an auction mechanism in which multiple buyers and sellers compete to buy and sell some financial assets and where transactions can occur at any time whenever an offer to buy and an offer to sell match. The minimum unit of price change is $δP_t$. The buy order price is rounded off to the nearest fraction, and the sell order price is rounded up to the nearest fraction. When an agent orders to buy (sell), if there is a lower-priced sell order (a higher-priced buy order) than the agent’s order, dealing is immediately done. We call this a “market order”2. If there is not, the agent’s order remains in the order book. We call this a “limit order”.

C. Agent Model

When an agent j has a turn at ordering, the agent determines an order price $P_t^{e,j}$ and buys or sells as follows. Agents always order only one share. The quantity of holding positions is not limited, so agents can take any shares for long and short positions to infinity.

Agent j determines an order price and buys or sells as follows. Agents use a combination of a fundamental value and technical rules to form expectations on a risk asset’s returns. An expected return of agent j is

$$r_{i,j}^t = \frac{1}{w_{1,j} + w_{2,j} + u_j} \left( w_{1,j} \log \frac{P_t}{P_t^c} + w_{2,j} r_{h,j}^t + u_j r_{j}^t \right),$$

(1)

where $w_{i,j}$ is the weight of term $i$ of agent $j$ and is independently determined by random variables uniformly distributed in the interval $(0, w_{max})$ at the start of the simulation for each agent, $u_j$ is the weight of the third term of agent $j$ and is also independently determined by random variables uniformly distributed in the interval $(0, u_{max})$ at the start of the simulation for each agent, $P_t$ is a fundamental value that is constant3, $P_t^c$ is an observed market price of the risk asset at $t^c$, as we mentioned

1 Finite latency existing in other places is not considered to affect essential simulation results. Only the total latency for all places is important for the market price formation.

2 Note that this definition is slightly different from a practical definition in actual trading markets.

3 We focused phenomena in time scale as short as the fundamental price remains static.

4 When a transaction does not occur at $t$, $P_t^e$ remains at the last market price, and at $t = 0$, $P_t^e = P_t$.
in Section II-A, $e^t_{c,j}$ is noise determined by random variables of normal distribution with an average 0 and a variance $\sigma_e$, $r^h_{b,j}$ is a historical price return inside an agent's time interval $\tau_j$, and $r^h_{o,j} = \log(P^t/P^{t-\tau_j})$, and $\tau_j$ is independently determined by random variables uniformly distributed in the interval $(1, \tau_{max})$ at the start of the simulation for each agent$^5$.

The first term of Eq. (1) represents a fundamental strategy: an agent expects a positive return when the market price is lower than the fundamental value, and vice versa. The second term of Eq. (1) represents a technical strategy: an agent expects a positive return when historical market return is positive, and vice versa.

After the expected return has been determined, the expected price is

$$P^t_{e,j} = P^t \exp(r^t_{e,j}).$$  

(2)

An order price $P^t_{o,j}$ is determined by random variables normally distributed in an average $P^t_{e,j}$ and a standard deviation $P_\sigma$, where $P_\sigma$ is a constant.

Buy or sell is determined by the magnitude relationship between $P^t_{e,j}$ and $P^t_{o,j}$, i.e.,

- when $P^t_{e,j} > P^t_{o,j}$, the agent orders to buy one share,
- when $P^t_{e,j} < P^t_{o,j}$, the agent orders to sell one share$^6$.

The remaining limit order, which an agent ordered $c$ times before, is canceled.

III. SIMULATION RESULTS

Mizuta et al. [23] searched for adequate model parameters verified by statistically existing stylized facts and market micro structures. They found parameters to replicate both long-term statistical characteristics and very short-term micro structures of actual financial markets. We explain how they verified their model in the Appendix. In this study, we set $\delta o = 1$ and had the other parameters remain the same as those of Mizuta et al.'s model [23]. When $\delta l = 0$, our model and that of Mizuta et al.'s model [23] are exactly the same. Specifically, we set

$^5$However, when $t < \tau_j$, $r^h_{b,j} = 0$.

$^6$However, during $t < c \times n = 20,000$, when $P^t_f > P^t_{o,j}$, the agent orders to buy one share, when $P^t_f < P^t_{o,j}$, the agent orders to sell one share.

n = 1,000, $w_{max,1} = 1$, $w_{max,2} = 10$, $u_{max} = 1$, $\tau_{max} = 10,000$, $\sigma_e = 0.06$, $P_\sigma = 30$, $c = 20$, $\delta P = 0.1$, and $P_f = 10,000$. We ran simulations until every agent ordered 10,000 times. We define $t_e$ as the end of the simulation.

We compared several statistical values of the simulation runs for various $\delta l/\delta o = 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2, 5$, and 10 not only under other parameters that were fixed but also the same random number table. We simulated these runs 100 times, changing the random number table each time, and used averaged statistical values of 100 runs.

A. Market Efficiency

We introduced the parameter “market inefficiency” $M_{ie}$ for directly measuring market efficiency,

$$M_{ie} = \frac{1}{t_e \sum_{t=1}^{t_e} \frac{|P^t_f - P^t_f|}{P^t_f}}.$$  

(3)
where $||$ is an absolute value. The $M_{ie}$ is always greater than zero, and $M_{ie} = 0$ means a market is perfectly efficient\(^7\). The larger the $M_{ie}$, the less efficient the market\(^8\).

Figure 2 shows $M_{ie}$ for various $\delta l/\delta o$. When $\delta l/\delta o < 0.5$, $M_{ie}$ is very stable; however, when $\delta l/\delta o > 0.5$, $M_{ie}$ increases, which means the market is less efficient. This indicates that $\delta l$ should be sufficiently smaller than $\delta o$ in order for the market to be efficient. Figure 3 shows differences in $M_{ie}$ when $\delta l/\delta o = 0.001$ and error bars (one standard deviation) for 100 runs for various $\delta l/\delta o$. When at least $\delta l/\delta o > 2$, the market is significantly less efficient than when $\delta l/\delta o = 0.001$.

Figure 4 illustrates the relationship between time evolution of the observed prices and that of the true prices when $\delta l/\delta o > 1$ (top) and $\delta l/\delta o < 1$ (bottom). When $\delta l/\delta o > 1$ (top), sometimes $\delta l > \delta t$ (see the yellow middle arrows). In this case, an agent might decide a different order than when the observed price is the same as the true price. Therefore, when $\delta l/\delta o > 1$, it is expected that price formation might be different from the case without latency, $\delta l = 0$, because the observed price is sometimes different from the true price. On the other hand, when $\delta l/\delta o \ll 1$, it is almost $\delta l < \delta t$, and an agent usually knows the true price. Therefore, it is expected that price formation might be almost the same when $\delta l = 0$.

**B. Bid-ask Spread and Execution Rate**

In this section, we discuss mechanisms in which the latency is larger and a market is less efficient. Figure 5 shows bid-ask spreads for various $\delta l/\delta o$. We define a bid-ask spread $S$ as

\[
S = (P_{bb} - P_{ba})/P_f^2,
\]

where $P_{bb}$ is the highest limit buy order on an order book, and $P_{ba}$ is the lowest limit sell order on an order book. When $\delta l/\delta o > 0.5$, $S$ is wider.

Figure 6 shows the execution rates for various $\delta l/\delta o$. We define the execution rate as the number of market orders/number of all orders (market and limit orders). When $\delta l/\delta o > 0.5$, execution rates are also increasing. These suggest that increasing the execution rate reduces the limit orders to near the market price, the bid-ask spread becomes wider, and the market becomes less efficient. In the next section, we discuss a mechanism that increases the execution rate by latency.

**C. Mechanism that Increases Execution Rate**

In this section, we discuss a mechanism that increases the execution rate due to the latency using the results when $\delta l/\delta o = 0.001$ and 10. Figure 7 shows execution rates for various true prices. For large latency $\delta l/\delta o = 10$, the execution rates are larger than those for no latency $\delta l/\delta o = 0.001$, especially near $P_f = 10,000$. Table I lists the execution rates for the two cases and the averages of the estimated return $r^f_{e,j}$ of all agents. We broke down the execution rates into cases in which a market buy order matches a limit sell order and a market sell order matches a limit buy order. For $\delta l/\delta o = 10$, we broke down the execution rates into cases in which observed price $P_{ovs}$ is smaller than the true price $P_{tr}$ and $P_{ovs}$ is larger than $P_{tr}$. For $\delta l/\delta o = 10$, when $P_{ovs}$ is smaller than $P_{tr}$, there are more market buy orders than market sell orders and the average estimated return is positive. When $P_{ovs}$ is larger than $P_{tr}$, the opposite occurs.

Figure 8 shows that this mechanism is consistent with results in Table I. Eq. (1) shows that, when a market price is near the fundamental price ($P^f \sim P_f$), the second term (the technical strategy term) is more dominant than the first term (the fundamental strategy term) in the estimated return

\[\text{Fig. 5. Bid-ask spreads for various } \delta l/\delta o\]

\[\text{Fig. 6. Execution rates for various } \delta l/\delta o\]
Increasing Execution Rate especially near the fundamental price

Fig. 7. Execution rates for various true prices when $\delta l / \delta o = 0.001$ and $10$.

of agents ($r_{e,j}^t$). The technical strategy term indicates a positive (negative) estimated return when the historical return is positive (negative). On the left side of Figure 8 (the case in which an observed price is lower than the true price), the market price has upward momentum because the true current price is higher than past. Therefore, the technical strategy term is positive and the estimated returns of agents are frequently positive. Note that from Eq. (3), when $P_{e,j}^t$ is much higher than the true price, a buy order tends to become a market order. We discuss the case in which the upward trend has actually finished. If the agents knew the true price, their estimated returns would be almost zero because the technical strategy term would be almost zero, and they would not make market buy orders. However, if agents actually observe the past upward momentum because of the large latency, then they make market buy orders. On the right side of Figure 8 (the case in which an observed price is higher than the true price), the case in which an observed price is larger than the true price, the opposite occurs. These cases increase the execution rates, as shown in Table I and Figure 7.

If the latency is large, agents cannot quickly change their estimated prices when the market trend has finished. Agents then make unnecessary market orders, and such market orders increase the execution rate. As we mentioned in Section III-B, increasing the execution rate reduces limit orders to near the market price, widens the bid-ask spread, and makes the market becomes less efficient. This mechanism causes a market to be inefficient when $\delta l / \delta o > 1$.

IV. CONCLUSION AND FUTURE WORK
We constructed a simple artificial market model in which the latency was also implemented onto Mizuta et al.’s model [23] and investigated price formations and market efficiency for various latencies; while other settings were maintained exactly the same.

We discussed how much of an increase in speed is needed for market efficiency. The simulation results indicated that latency should be sufficiently smaller than the average order interval for a market to be efficient. This implication is generally opposite to that in which the increase in the speed of matching systems might destabilize financial markets.

We clarified the mechanisms of the direct effects of latency on financial market efficiency. If the latency is large, agents cannot quickly change their estimated prices when the market trend has finished. Agents then make unnecessary market orders, and such market orders increase the execution rate. We argued that increasing the execution rate reduces limit orders to near the market price, widens the bid-ask spread, and makes the market becomes less efficient. This indicates that latency should be sufficiently smaller than the average order interval for a market to be efficient.

For future work, we will investigate the case of a large amount of orders for less than one minute after very important news. We did not consider this case for specific and very short spans in the simulations of this study. We implemented only normal agents replicating general investors; however, latency was more important, especially for HFTs whose investment strategies are market maker, arbitrage, and so on. We should discuss the latencies for different types of agents for future work.

For more detailed discussions, we should compare the simulation results to those from studies using other methods, e.g., empirical studies and theoretical studies. An artificial market can isolate the direct effect of changes in market systems to price formation, and can treat situations that have never occurred. However, outputs of artificial market simulations may not be accurate or credible forecasts in actual markets. It is an important for artificial market simulations to show possible mechanisms affecting price formation through many runs and gain new insight; conversely, a limitation of artificial market simulations is that their outputs may, but not certainly, occur in actual financial markets.
A. Basic Concept for Constructing Model

An artificial market, which is a kind of agent-based models, can isolate the pure contribution of these system changes to the price formation and can treat the changes that have never been employed [5]–[8]. These are the strong points of the artificial market simulation study.

However, outputs of the artificial market simulation study would not be accurate or credible forecasts of the actual future. The artificial market simulation needs to show possible mechanisms affecting the price formation by many simulation runs, e.g. searching for parameters, purely comparing before/after the changing, and so on. The possible mechanisms shown by these simulation runs will give us new intelligence and insight about effects of the changes to price formation in actual financial markets. Other study methods, e.g. empirical studies, would not show such possible mechanisms.

Indeed, artificial markets should replicate macro phenomena existing generally for any asset and any time. Price variation, which is a kind of macro phenomena, is not explicitly modeled in artificial markets. Only micro processes, agents (general investors), and price determination mechanisms (financial exchanges) are explicitly modeled in artificial markets. Macro phenomena are emerging as the outcome interactions of micro processes. Therefore, the simulation outputs should replicate general macro phenomena at least to show that simulation models are probable in actual markets.

However, it is not a primary purpose for the artificial market to replicate specific macro phenomena only for a specific asset or a specific period. An unnecessary replication of macro phenomena leads to models that are over-fitted and too complex. Such models would prevent our understanding and discovering mechanisms affecting the price formation because of related factors increasing.

Indeed, artificial market models that are too complex are often criticized because they are very difficult to evaluate [6]. A too complex model not only would prevent our understanding mechanisms but also could output arbitrary results by over-fitting too many parameters. Simpler models harder obtain arbitrary results, and are easier evaluated.

Therefore, we constructed an artificial market model that is as simple as possible and do not intentionally implement agents to cover all the investors who would exist in actual financial markets.

B. Verification of the Model

In many previous artificial market studies, the models were verified to see whether they could explain stylized facts such as a fat-tail, volatility-clustering, and so on [5]–[8]. A fat-tail means that the kurtosis of price returns is positive. Volatility-clustering means that the square returns have positive autocorrelation, and the autocorrelation slowly decays as its lag becomes longer. Many empirical studies, e.g. that of Sewell [44], have shown that both stylized facts (the fat-tail and volatility-clustering) exist statistically in almost all financial markets. Conversely, they also have shown that only the fat-tail and volatility-clustering are stably observed for any asset and in any period because financial markets are generally unstable.

Indeed, the kurtosis of price returns and the autocorrelation of the square returns are stably and significantly positive, but the magnitudes of these values are unstable and very different depending on asset and/or period. The kurtosis of price returns and the autocorrelation of the square returns were observed to have very broad magnitudes of about 1 ~ 100 and about 0.01 ~ 0.2, respectively [44].

For the above reasons, an artificial market model should replicate these values as significantly positive and within a reasonable range as we mentioned. It is not essential for the models to replicate specific values of stylized facts because these stylized facts’ values are unstable in actual financial markets.

Table II lists statistics in which there is no latency. All statistics are averages of 100 simulation runs, and all the following figures use the average of 100 simulation runs. We define 20,000 time steps as 1 day because the number of trades within 20,000 time steps is almost the same as that in actual markets per day. All statistics; execution rates,

### Appendix

#### Table I

<table>
<thead>
<tr>
<th>δt/δo</th>
<th>P_{ovs} &lt; P_{tr}</th>
<th>P_{ovs} &gt; P_{tr}</th>
<th>Avg. Estimated Return</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>32.5%</td>
<td>3.5%</td>
<td>0.28%</td>
</tr>
<tr>
<td>0.001</td>
<td>32.5%</td>
<td>3.5%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

#### Table II

<table>
<thead>
<tr>
<th>Statistics without Latency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Execution rate</td>
</tr>
<tr>
<td>Cancel rate</td>
</tr>
<tr>
<td>Number of trades / 1 day</td>
</tr>
<tr>
<td>Standard deviations</td>
</tr>
<tr>
<td>Autocorrelation</td>
</tr>
<tr>
<td>Coefficient for lag</td>
</tr>
<tr>
<td>Square return</td>
</tr>
<tr>
<td>Kurtosis</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

The artificial market simulation needs to show possible mechanisms affecting the price formation by many simulation runs, and all the following figures use the average of 100 simulation runs. We define 20,000 time steps as 1 day because the number of trades within 20,000 time steps is almost the same as that in actual markets per day. All statistics; execution rates.
cancel rates\textsuperscript{10}, standard deviations of returns for one tick and one day\textsuperscript{11}, kurtosis of price returns, and the autocorrelation coefficient for square returns\textsuperscript{12} are of course almost the same as the results of Mizuta et al.\textsuperscript{23} because the models do not differ except for the latency. Mizuta et al.\textsuperscript{23} showed that this mode replicated very short term micro structure, execution rates, cancel rates, and standard deviations of returns for one tick and replicated long-term statistical characteristics, fat tail, and volatility clustering, observed in real financial markets. Therefore, the model was verified to investigate the effect of the latency on market stability.

\textbf{Disclaimer & Acknowledgments}

Note that the opinions contained herein are solely those of the authors and do not necessarily reflect those of SPARX Asset Management Co., Ltd., Daiwa SB Investments Ltd. and Osaka Exchange, Inc. This research was partially supported by CREST, JST and JSPS KAKENHI Grant Number 15H02745.

\textbf{REFERENCES}

\begin{itemize}
\item [27] M. Paddock, R. Hayes, A. Todd, S. Yang, P. Beling, and W. Scherer, “An agent based model of the e-minis s amp p 500 applied to flash crash analysis,” in Computational Intelligence for Financial Engineering Economics
\end{itemize}