УДК 336.76(045)

A NEW WAY TO IDENTIFY HIGH-FREQUENCY TRADING

Perevalov D.V.,
student, Financial University, Moscow, Russia
dv.perevalov@yandex.ru

Abstract. Introduction. *Identification of high-frequency traders becomes an important task. Therefore, many financial market regulators and traders are engaged in this issue. This allows us to forecast the further development of the financial industry as a whole and helps to classify all market participants. In this paper, the main approaches to the definition of high-frequency trade are considered and a new approach to identification of high-frequency traders based on the clustering method is considered.*

Methods. *The study is based on the analysis of market data on USD/RUB instrument with a settlement for tomorrow. The data correspond to one trading day the trading on which took place at the currency section of the Moscow Exchange on April 13th, 2017. The most popular method of clustering — k-means — was used as a method of analysis.*

Results. *In this paper, a new approach to the identification of high-frequency traders based on the clustering method has been developed. This method provided a quantitative characteristic without elements of subjectivity to the issue of identifying high-frequency traders.*

Discussion. *The difficulty in confirming this technique lies in the fact that it is possible to verify the obtained results only on the marked data. That is data which have a markup of each exchange order in the form of the trader's name, who sent this order. Unfortunately, such information is closed within the trading platform. The use of other methods of machine learning can be a solution to this problem. For example, the fuzzy clustering to use.*

Keywords: *automated trading; direct market access; financial markets; financial regulators; high-frequency trading; identify HFT; k-means clustering; machine learning; market data*

НОВЫЙ СПОСОБ ИДЕНТИФИКАЦИИ ВЫСОКОЧАСТОТНЫХ ТРЕЙДЕРОВ

Перевалов Д.В.,
студент, Финансовый университет, Москва, Россия
dv.perevalov@yandex.ru

Аннотация. Актуальность. *Идентификация высокочастотных трейдеров становится важной задачей. Поэтому многие регуляторы рынка и трейдеры занимаются этим вопросом. Это позволяет нам прогнозировать дальнейшее развитие финансовой отрасли в целом и помогает классифицировать всех участников рынка. В этой статье рассматриваются основные подходы к определению высокочастотной торговли и рассматривается новый подход к идентификации высокочастотных трейдеров на основе метода кластеризации.*

Методы. *Исследование базируется на анализе рыночных данных по инструменту USD/RUB с расчетами на завтра. Данные соответствуют одному торговому дню, торговля по которому про-*

Advisor: ***Gisin V.B.,*** *Candidate of Physical and Mathematical Sciences, Professor, Department for Data Analysis, Decision Making and Financial Technologies, Financial University.*

Научный руководитель: ***Гисин В.Б.,*** *кандидат физико-математических наук, профессор Департамента анализа данных, принятия решений и финансовых технологий, Финансовый университет.*

ходила на валютной секции Московской биржи 13 апреля 2017 г. В качестве метода анализа был использован наиболее популярный метод кластеризации — k-means.

Результаты. *В представленной работе был разработан новый подход к идентификации высокочастотных трейдеров на основе метода кластеризации. Этот метод обеспечивал количественную характеристику без элементов субъективности в вопросе идентификации высокочастотных трейдеров.*

Перспективы. *Сложность подтверждения данной методики заключается в том, что проверить полученные результаты можно лишь на размеченных данных. То есть на таких данных, на которых присутствует разметка каждого приказа в виде имени трейдера, пославшего данный приказ. К сожалению, такая информация является закрытой в рамках торговой площадки. В качестве решения данной проблемы может послужить использование других методов машинного обучения. Например, использовать нечеткую кластеризацию.*

Ключевые слова: *автоматическая торговля; прямой доступ на рынки; финансовые рынки; финансовые регуляторы; высокочастотная торговля; определить HFT; k — означает группирование; машинное обучение; рыночные данные*

1. Introduction

High-frequency trading is classified as an algorithmic trade. The term “high-frequency trading” is relatively new and is not yet clearly defined. Therefore, financial regulators disagree on the definition of high-frequency trading and describe it in different ways. But these definitions have a number of common characteristics that they possess. This paper examines the main approaches to the definition of high-frequency trade and a new approach to identifying high-frequency traders based on the clustering method.

2. Literature review

High-frequency trading is classified as an algorithmic trade. The term “high-frequency trading” is relatively new and is not yet clearly defined. Therefore, financial regulators disagree on the definition of high-frequency trading and describe it in different ways. But these definitions have a number of common characteristics that they possess. In this paper, an overview of ways to identify high-frequency traders as participants of the financial market is given.

In 2010, the United States Securities and Exchange Commission (SEC) gave one of the first definitions of the term “high-frequency trading”. It is typically used to refer to professional traders acting in a proprietary capacity that engages in strategies that generate a large number of trades on a daily basis. These traders could be organized in a variety of ways, including as a proprietary trading firm, as the proprietary trading desk of a multi-service broker-dealer, or as a hedge fund. As the regulator points out in his report, the proprietary firm dealing with HFT should have a number of characteristics: (1) the use of extraordinarily high-speed and sophisticated computer programs for generating, routing, and executing orders; (2) use of co-location services and individual data feeds offered by exchanges and others to minimize network and other types of latencies; (3) very short time-frames for establishing and liquidating positions; (4) the submission of numerous orders that are cancelled shortly after submission; and (5) ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions over-night) [8, p. 45].

In the definition given by the American regulator, there is an element of subjectivity. The lack of quantitative characteristics does not help to explicitly identify high-frequency traders [4, p. 14]. As a consequence, in 2011, the Investment Industry Regulatory Organization of Canada (IIROC) published a study in which market participants were compared according to a specific indicator. This indicator is the ratio of the number of applications for participation to the number of all market participants. The distribution of this characteristic was constructed. Those participants in the market whose indicator was more than 11.2:1 were attributed to the group of high-frequency traders [4, p. 17].

In March 2013, the Australian Securities and Investment Commission (ASIC) in its study determined the indicators used to define high-frequency traders. In addition to the indicator of the ratio of the number of applications to the number of transactions mentioned above, proposed by the Canadian regulator,

5 more features were considered: (a) order-to-trade ratios; (b) percentage of turnover traded within the day; (c) total turnover per day; (d) the number of fast messages; (e) holding times; (f) at-best ratios [1, p. 68]. Each participant was numerically evaluated according to each of the criteria. As a result, a rating table was received in which each participant had a rating from 6 to 24. To identify a group of high-frequency traders, market participants with the highest indicator were taken.

Some trading floors provide data that already contains a certain scheme of predetermination of bidders into classes. For example, data from the electronic platform (EBS) uses the division of all participants into two groups of agents: using (AT) and not using (HA) algorithms in their trade [3, p. 2046]. This approach is rough, but effective in terms of accuracy and simplicity.

Today, there is a number of empirical works devoted to the classification of financial market participants. One of them is the work by Kirilenko et al. The authors divided all the market participants into 6 categories: High Frequency Traders, Market Makers, Fundamental Buyers, Fundamental Sellers, Opportunistic Traders, and Small Traders [6, p. 14]. As the basis for the classification, the following characteristics were used: (i) daily trading volume; (ii) end-of-day position; (iii) intraday minute-by-minute inventory pattern [6, p. 12].

This approach has become the best starting point for further research in the field of classification of financial market participants. In another study, the authors used this approach and improved the quality of selection of high-frequency participants. In their approach, there were four criteria that a high-frequency trader should satisfy: (1) Trade more than a median of 5,000 contracts in all the days that this trader is active; (2) have a median (across days) end-of-day inventory position, scaled by total contracts the firm traded that day, of no more than 5%; (3) have a median (across days) maximum variation in inventory that day (maximum position minus minimum position that day), scaled by total contracts the firm traded that day, of less than 10% [2, p. 9].

Most studies on HFT use ready-made classifications provided by trade organizers and financial market regulators, while a small number of empirical works are devoted to identifying of high-frequency market participants. In the review, the main of them are briefly presented. Also in this paper, the main definitions of the concept of "high-frequency trading" are presented and approaches to the identification of high-frequency trading are highlighted both by regulators of financial markets and by the scientific community.

3. Data Description

Data in this article were obtained from the foreign exchange market by the Moscow Exchange (MOEX). The Moscow Exchange in multicast mode sends out market information to its customers. These clients are mainly brokerage companies. But as an exception some market players need data directly related to the trading platform.

In the table of this kind of data, each line gives information about a specific order sent to the foreign exchange market of the MOEX.

Example:

BUYSELL	TIME	ORDERNO	ACTION	PRICE	VOLUME
S	110137000624	1573	1	56.7325	200

Explanation: at 11:01:37 (624 μs), an order with the volume of 200 lots was sent to the foreign exchange market at a price of 56.7325 for sale with the ID1573 assigned.

We have access to information about each order, as in the example above. We also have information about the status of the order at the time of its execution or at the time of its cancellation.

Example:

BUYSELL	TIME	ORDERNO	ACTION	PRICE	VOLUME
S	134634862694	1573	0	56.7325	200

Explanation: at 11:46:34 (862 ms. 694 μs.) the order with identifiers 1573 was canceled.

The aim of the data handling from this table was to find the lifetime of a single order and its volume. Data relates to a financial instrument such as USD/RUB with settlement for tomorrow.

The result of data handling is finding two vectors (one-dimensional arrays). Each value of the first vector is the lifetime of the order expressed in microseconds. Each value of the second vector is the volume of the order expressed in lots.

Example:
Time vector … 235, 121998843 …

Volume vector ... 100, 5000 ...

Explanation: The lifetime and volume of the n order is 235 μs. and 100 lots respectively. Similarly, the lifetime and volume (*n* + 1) *of the order is 2 min. 1 sec. 998 ms. 843 μs. and in 5000 lots.*

4. Methodology and results

To present a new way of identifying high-frequency traders based on the clustering of two quantities: time and volume.

A. Data handling

According to the data described above, an algorithm is developed and a script written in the programming language R that finds the life time of each individual order and the volume corresponding to this order. Each vector contains 512,999 values.

B. Viewing data and deleting values as emissions

According to the presented figure the emissions along the volume axis are visible (there are five of them).

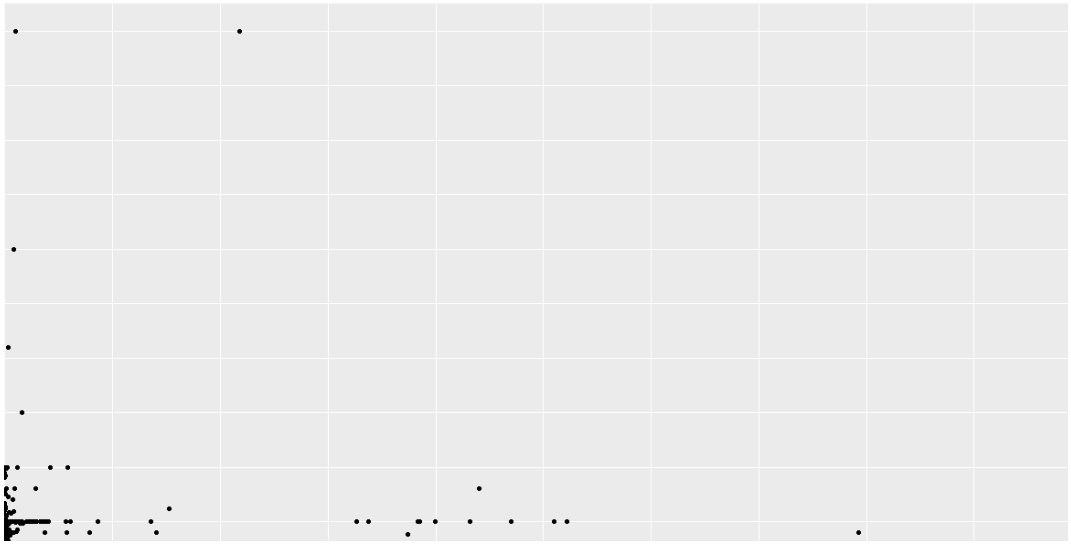

Figure 1. **One-point shows information about one order in the form of time (on the horizontal axis) and volume (on the vertical axis)**

After deleting these points, we get the following picture.

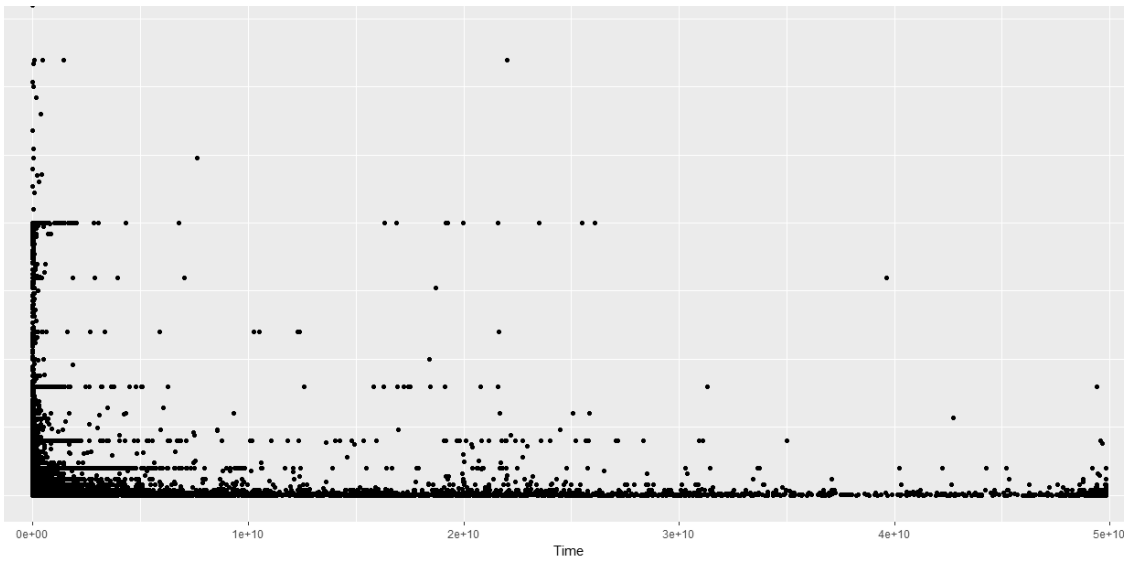

Figure 2. **Emission-free schedule**

C. Normalization of vectors' values

Since the sampling range of the time vector is much higher than that of the volume vector, it is necessary to normalize the values for each vector. This can be achieved by dividing each vector by its standard deviation.

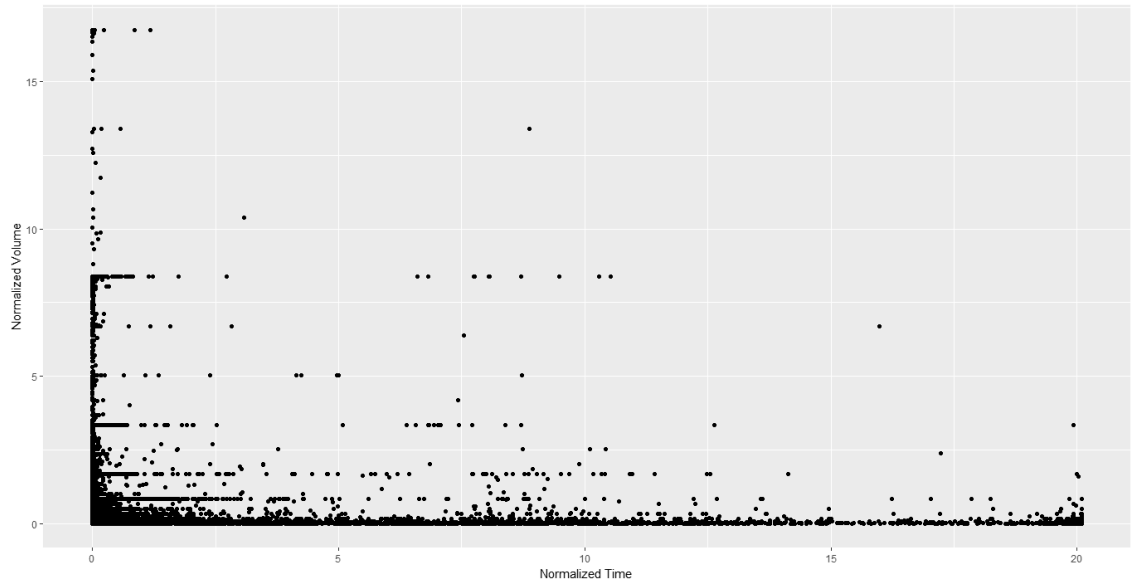

Figure 3. **Normalization of the values of the time and volume vectors by dividing by the standard deviation**

D. Finding the optimal number of clusters

To use the *k-means* clustering method, it is necessary to find out the number of clusters to which our data can be divided. To do this, we will sequentially set the number of clusters from 1 to 10 for splitting our data and for each splitting we shall look at such an indicator as the *total within-cluster sum of squares* [5, p. 300].

Within-cluster sum of squares is the sum of the squares of the deviations of each observation from the centroid of the cluster.

Total within-cluster sum of squares will be obtained if we calculate for each cluster its within-cluster sum of squares, and then sum them.

Next, we need to look at the dependence of the number of clusters and the indicator of the total within-cluster sum of squares.

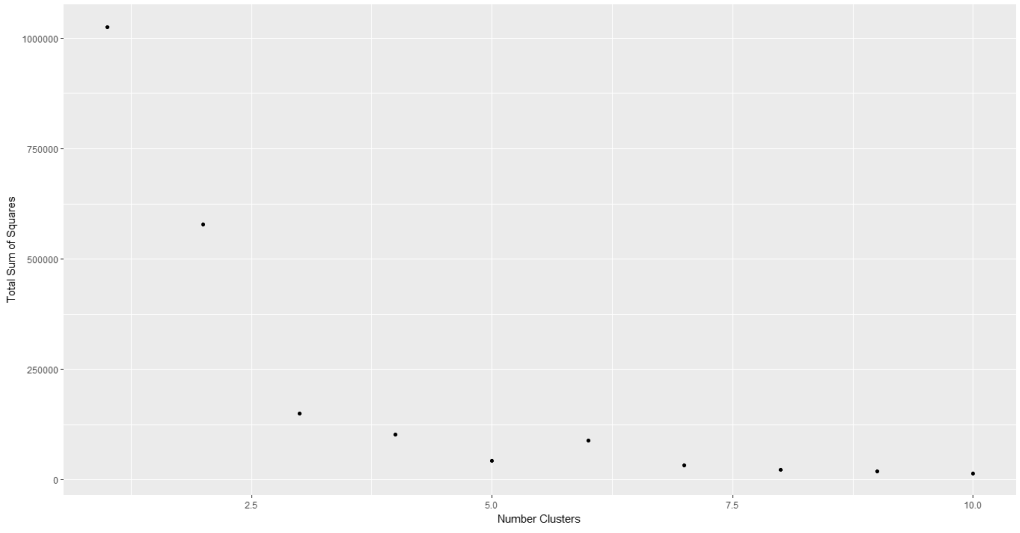

Figure 4. **The relationship between the number of clusters of a partition and the total within-cluster sum of squares corresponding to this number**

It can be seen from the graph that the indicator of the total within-cluster sum of squares practically does not change from the moment when the data is divided into three clusters. Therefore, we will break our data into three clusters.

E. Clustering data using k-means.

After finding the optimum number of clusters, we need to break our data into three clusters [5, p. 278].

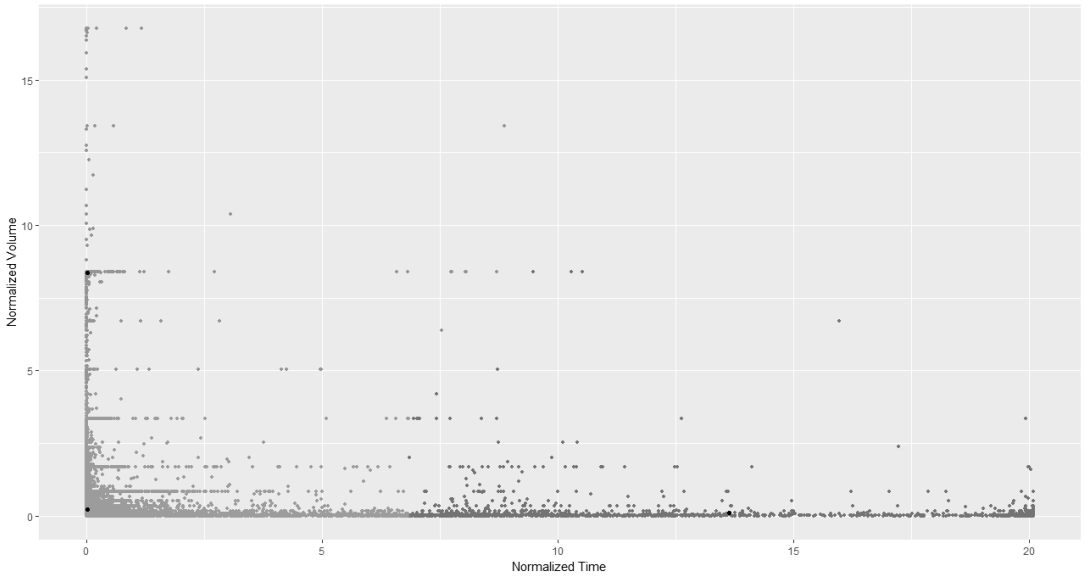

Figure 5. **The first level of clustering into three clusters**

Cluster 1 contains 503,856 values, cluster 2–2,333, cluster 3–6,805. Black dots indicate centroids of clusters.

F. Hierarchical clustering

We do the operations with (ii) to (v) over the data belonging to the 1 cluster. By such actions, we will perform the semblance of a method of hierarchical clustering [7, p. 1120]. As a result, we get the following picture.

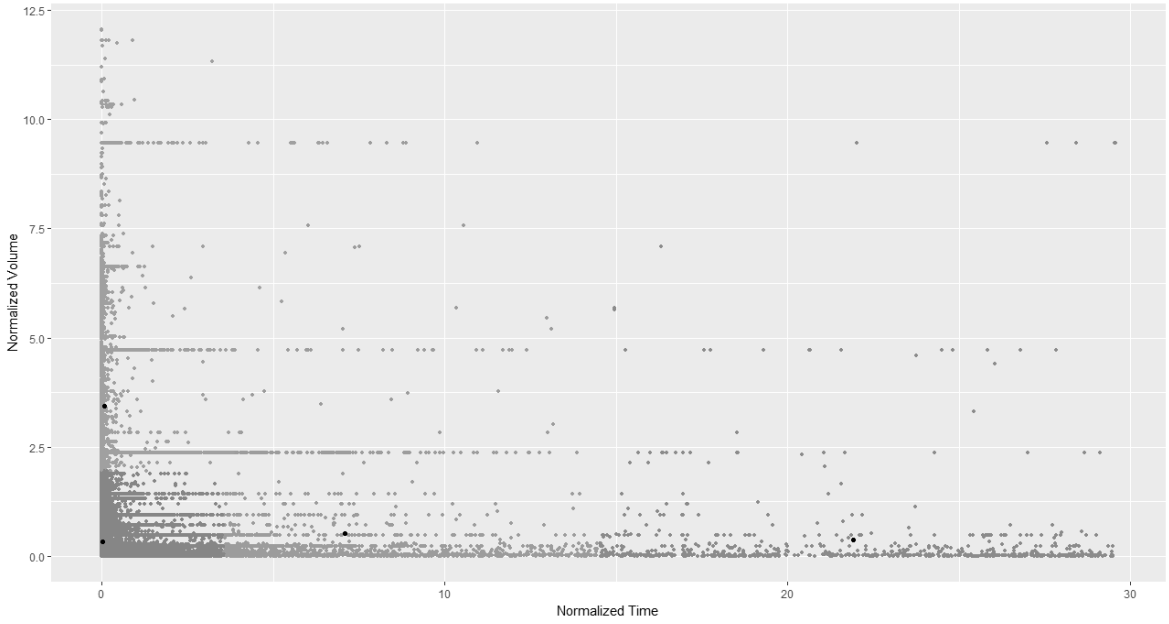

Figure 6. **The second level of clustering into four clusters**

Cluster 1 contains 1,758 values, cluster 2–763, cluster 3–459,290, cluster 4–42,045.

Based on the results of the work, it can be said that orders belonging to the third cluster (in the second stage of clustering) are orders generated by high-frequency traders: these orders have a short life time and relatively small volume, and also because this cluster contains a little less than 90% of all orders.

5. Conclusion

Most studies on HFT use ready-made classifications provided by trade organizers and financial market regulators, while a small number of empirical works are devoted to identifying of high-frequency market participants. In the review, the main of them are briefly presented. Also in this work, a new approach to the identification of high-frequency traders based on the clustering method has been developed. This method provided a quantitative characteristic without elements of subjectivity to the issue of identifying high-frequency traders.

References

1. Australian Securities and Investments Commission (ASIC), Report 331: Dark liquidity and high-frequency trading. 2013.
2. Baron M., Brogaard J., Kirilenko A. The trading profits of high frequency traders. 2012.
3. Chaboud A.P. et al. Rise of the machines: Algorithmic trading in the foreign exchange market // *The Journal of Finance*, 2014, vol. 69, no. 5, pp. 2045–2084.
4. Investment Industry Regulatory Organization of Canada (IIROC), The HOT Study Phases I and II of IIROC's Study of High Frequency Trading Activity on Canadian Equity Marketplaces. 2012.
5. Jain A.K., Murty M.N., Flynn P.J. Data clustering: a review // *ACM computing surveys (CSUR)*. 1999, vol. 31, no. 3, pp. 264–323.
6. Kirilenko A.A. et al. The flash crash: High frequency trading in an electronic market. 2016.
7. Rani Y., Rohil H. A study of hierarchical clustering algorithm // ter S & on Te SIT-2. 2013. 113 p.
8. U.S. Securities and Exchange Commission (SEC), Concept Release on Equity Market Structure. Concept Release, 17 CFR PERT 242, Concept Release, no. 34–61358; File no. S7–02–10. 2010.

ИНФОРМБЮРО

АВТОМАТИЧЕСКАЯ ТОРГОВЛЯ НА РЫНКЕ ФОРЕКС

В настоящее время существует много торговых роботов, каждый из которых обещает трейдерам получение огромных и постоянных доходов без особых усилий. Однако найти по-настоящему эффективного советника для ведения автоматической торговли на рынке Форекс достаточно непросто. Уровень доходности того или же иного Форекс советника будет определяться главным образом количественно. Если фактор прибыльности меньше чем единица, то сразу же исключайте его из списка, потому как такой уровень доходности абсолютно не оправдывает возможные риски. Самым главным индикатором риска считается так называемая просадка, представляющая собой процент, который может быть потерян трейдером от самого последнего максимума активов до их следующего минимума. Благодаря данному индикатору можно получить краткое описание изменений, которые следует ожидать от советников, а также изображения возможного снижения цены.

Другим не менее значимым показателем является так называемая средняя просадка. Как правило, определяется такая величина посредством суммирования всех процентов и последующего разделения получившейся суммы на общее число просадок, с которыми сталкивался советник.