



Modeling price clustering in high-frequency prices

Vladimír Holý & Petra Tomanová

To cite this article: Vladimír Holý & Petra Tomanová (2022) Modeling price clustering in high-frequency prices, Quantitative Finance, 22:9, 1649-1663, DOI: [10.1080/14697688.2022.2050285](https://doi.org/10.1080/14697688.2022.2050285)

To link to this article: <https://doi.org/10.1080/14697688.2022.2050285>



Published online: 28 Mar 2022.



Submit your article to this journal [↗](#)



Article views: 231



View related articles [↗](#)



View Crossmark data [↗](#)

Modeling price clustering in high-frequency prices

VLADIMÍR HOLÝ * and PETRA TOMANOVÁ 

Prague University of Economics and Business, Winston Churchill Square 4, 130 67 Prague 3, Czech Republic

(Received 22 April 2021; accepted 18 February 2022; published online 28 March 2022)

The price clustering phenomenon manifesting itself as an increased occurrence of specific prices is widely observed and well documented for various financial instruments and markets. In the literature, however, it is rarely incorporated into price models. We consider that there are several types of agents trading only in specific multiples of the tick size resulting in an increased occurrence of these multiples in prices. For example, stocks on the NYSE and NASDAQ exchanges are traded with precision to one cent but multiples of five cents and ten cents occur much more often in prices. To capture this behavior, we propose a discrete price model based on a mixture of double Poisson distributions with dynamic volatility and dynamic proportions of agent types. The model is estimated by the maximum likelihood method. In an empirical study of DJIA stocks, we find that higher instantaneous volatility leads to weaker price clustering at the ultra-high frequency. This is in sharp contrast with results at low frequencies which show that daily realized volatility has a positive impact on price clustering.

Keywords: High-frequency data; Price clustering; Generalized autoregressive score model; Double Poisson distribution

JEL Classifications: C22, C46, C58

1. Introduction

Over the last two decades, there has been a growing interest in modeling prices at the highest possible frequency which reaches fractions of a second for the most traded assets. The so-called ultra-high-frequency data possess many unique characteristics which need to be accounted for by econometricians. Notably, the prices are irregularly spaced with discrete values. Other empirical properties of high-frequency prices which can be incorporated into models include intra-day seasonality, jumps in prices, price reversal, and the market microstructure noise. For related models, see, e.g. Russell and Engle (2005), Robert and Rosenbaum (2011), Barndorff-Nielsen *et al.* (2012), Shephard and Yang (2017), Koopman *et al.* (2017), Koopman *et al.* (2018), and Buccheri *et al.* (2021).

We focus on one particular empirical phenomenon observed in high-frequency prices—price clustering. In general, price clustering refers to an increased occurrence of some values in prices. A notable type of price clustering is an increased occurrence of specific multiples of the tick size, i.e. the minimum price change. For example, on the NYSE and NASDAQ exchanges, stocks are traded with precision to one cent but multiples of five cents (nickels) and ten

cents (dimes) tend to occur much more often in prices. In other words, while one would expect the distribution of the second digit to be uniform, the probability of 0 and 5 is actually higher than 0.1 for each. This behavior can be captured by some agents trading only in multiples of five cents and some only in multiples of ten cents. It is well documented in the literature that this type of price clustering is present in stock markets (see, e.g. Lien *et al.* 2019), commodity markets (see, e.g. Bharati *et al.* 2012), foreign exchange markets (see, e.g. Sopranzetti and Datar 2002), and cryptocurrency markets (see, e.g. Urquhart 2017). Moreover, price clustering does not appear only in spot prices but in futures (see, e.g. Schwartz *et al.* 2004), options (see, e.g. ap Gwilym and Verousis 2013), and swaps (see, e.g. Liu and Witte 2013) as well. From a methodological point of view, almost all papers on price clustering deal only with basic methods and descriptive statistics of the phenomenon. The only paper, to our knowledge, that incorporates price clustering into a price model is the recent theoretical study of Song *et al.* (2020) which introduced the sticky double exponential jump-diffusion process to assess the impact of price clustering on the probability of default.

Our goal is to propose a discrete dynamic model relating price clustering to the distribution of prices and to study the high-frequency behavior of price clustering. We take a fundamentally very different approach than (Song *et al.* 2020) and incorporate the mechanism of an increased occurrence of specific multiples of the tick size directly into the model.

*Corresponding author. Email: vladimir.holy@vse.cz

This allows us to treat the price clustering phenomenon as dynamic and driven by specified factors rather than given. We also operate within the time series framework rather than the theory of continuous-time stochastic processes. In contrast to the existing literature on price modeling, we do not model log returns or price differences but rather prices themselves. Prices are naturally discrete and positive. When represented as integers, they also exhibit underdispersion, i.e. the variance lower than the mean. To accommodate for such features, we utilize the double Poisson distribution of Efron (1986). It is a less known distribution as noted by Sellers and Morris (2017) but was utilized in the context of time series by Heinen (2003), Xu *et al.* (2012) and Bourguignon *et al.* (2019). Modeling prices directly enables us to incorporate price clustering in the model. Specifically, we consider that prices follow a mixture of several double Poisson distributions with specific supports corresponding to agents trading in different multiples of the tick size. This mixture distribution has a location parameter, a dispersion parameter and parameters determining portions of trader types. In our model, we introduce time variation to all these parameters. We consider the location parameter to be equal to the last observed price resulting in zero expected returns. For the dispersion parameter, we employ dynamics in the fashion of the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986). Specifically, we utilize the class of generalized autoregressive score (GAS) models of Creal *et al.* (2013) and Harvey (2013) which allows to base dynamic models on any underlying distribution. In the high-frequency literature, the GAS framework was utilized by Koopman *et al.* (2018) for discrete price changes and (Buccheri *et al.* 2021) for log prices. To account for irregularly spaced observations, we include the last trade duration as an explanatory variable similarly to Engle (2000). Finally, we relate the trader portion parameters to the volatility process and other variables such as the price, the last trade duration and the volume. The resulting observation-driven model is estimated by the maximum likelihood method.

In the empirical study, we analyze 30 Dow Jones Industrial Average (DJIA) stocks in the first half of 2020. We first focus on price clustering from a daily perspective which is a common approach in the price clustering literature. Using a panel regression with fixed effects, we find a positive effect of daily volatility measured by realized kernels of Barndorff-Nielsen *et al.* (2008) on price clustering. This finding is in line with the results of ap Gwilym *et al.* (1998), Davis *et al.* (2014), Box and Griffith (2016), Hu *et al.* (2017), and Lien *et al.* (2019) among others. Next, we estimate the proposed high-frequency price model and arrive at a different conclusion—the instantaneous volatility obtained by the model has a negative effect on price clustering. The main message of the empirical study is therefore that the degree of aggregation plays a pivotal role in the relation between price clustering and volatility. While high daily realized volatility correlates with high price clustering, high instantaneous volatility has the opposite effect. The other explanatory variables have the expected effect in both the daily and high-frequency cases—the volume has a positive effect on price clustering while the price and the last trade duration are insignificant.

The rest of the paper is structured as follows. In Section 2, we review the literature dealing with high-frequency price models and price clustering. In Section 3, we propose the dynamic model accommodating for price clustering based on the double Poisson distribution. In Section 4, we use this model to study determinants of price clustering in high-frequency stock prices. We conclude the paper in Section 5.

2. Literature review

2.1. Some high-frequency price models

In the literature, several models addressing specifics of ultra-high-frequency data have been proposed. One of the key issues is irregularly spaced transactions and discreteness of prices. The seminal study of Engle and Russell (1998) proposed the autoregressive conditional duration (ACD) model to capture the autocorrelation structure of trade durations, i.e. times between consecutive trades. Engle (2000) combined the ACD model with the GARCH model and jointly modeled prices with trade durations. Russell and Engle (2005) again modeled prices jointly with trade durations but addressed discreteness of prices and utilized the multinomial distribution for price changes.

Another approach is to model the price process in continuous time. Robert and Rosenbaum (2011) considered that the latent efficient price is a continuous Itô semimartingale but is observed at the discrete grid through the mechanism of uncertainty zones. Barndorff-Nielsen *et al.* (2012) considered the price process to be discrete outright and developed a continuous-time integer-valued Lévy process suitable for ultra-high-frequency data. Shephard and Yang (2017) also utilized integer-valued Lévy processes and focused on frequent and quick reversal of prices.

Transaction data at a fixed frequency can also be analyzed as equally spaced time series with missing observations. In this setting, Koopman *et al.* (2017) proposed a state space model with dynamic volatility and captured discrete price changes by the Skellam distribution. Koopman *et al.* (2018) continued with this approach and modeled dependence between discrete stock price changes using a discrete copula. Buccheri *et al.* (2021) also dealt with multivariate analysis and proposed a model for log prices accommodating for asynchronous trading and the market microstructure noise. The latter two papers utilized the GAS framework.

2.2. Price clustering

The first academic paper on price clustering was written by Osborne (1962), where the author described the price clustering phenomenon as a pronounced tendency for prices to cluster on whole numbers, halves, quarters, and odd one-eighths in descending preference, like the markings on a ruler. Since then, there have been many studies focusing on this phenomenon—from Niederhoffer (1965) to very recent papers of Li *et al.* (2020), Song *et al.* (2020), and Das and Kadapakkam (2020)—showing that price clustering is remarkably persistent in time and across various markets.

Song *et al.* (2020) pointed out that, however, all studies are entirely focused on empirically examining price clustering in different financial markets. Except for the purely theoretical paper of Song *et al.* (2020) proposing the sticky double exponential jump-diffusion process to analyze the probability of default for financial variables, the studies related to price clustering are based on basic general methods and do not aim to incorporate the phenomenon into the dynamic price model.

The prevalent approach to price clustering examination is a linear regression model estimated by the ordinary least squares (OLS) method. Ball *et al.* (1985) and Kandel *et al.* (2001), and from the recent literature (Urquhart 2017, Hu *et al.* 2019, Li *et al.* 2020), used the classical regression with dummy variables to estimate the frequency of each level of rounding. The vast of the literature regressed price clustering on explanatory variables such as volatility and trade size, where price clustering is defined as the excess occurrence of multiples of nickles or dimes (see, e.g. Schwartz *et al.* 2004 and Ikenberry and Weston 2008 followed by Chung and Chiang 2006, Brooks *et al.* 2013, and Hu *et al.* 2017) or simply their frequency (see, e.g. Palao and Pardo 2012 and Davis *et al.* 2014). However, different definitions of the dependent variable representing price clustering can be found in the literature. Baig *et al.* (2019) defined the clustering as a sum of round clustering at prices ending by digit 0 and strategic clustering measured as a number of trades which decimals are equal to 01 or 99. ap Gwilym and Verousis (2013) defined the dependent variable as the percentage of price observations at integers, whereas (ap Gwilym *et al.* 1998) estimated the percentage of trades that occur at an odd tick. Ahn *et al.* (2005) regressed abnormal even price frequencies in transaction and quote prices on the firm and trading characteristics, and similarly, Chiao and Wang (2009) performed the analysis on the limit-order data. Cooney *et al.* (2003) estimated cross-sectional regressions of the difference in the percentage of even and odd limit orders on stock price and proxies for investor uncertainty.

Several extensions of the OLS method were employed to overcome certain issues. Verousis (2013) and Mishra and Tripathy (2018) argued that one encounters a simultaneity issue between trade size and price clustering when striving to examine a causal relationship between them. Hence, Verousis (2013) followed by Mishra and Tripathy (2018) used the two-stage least squares (2SLS) method. Moreover, to reflect the endogeneity of quote clustering in the spread model and the endogeneity of the spread in the quote clustering model, Chung *et al.* (2004) and Chung *et al.* (2005) estimated a structural model using three-stage least squares (3SLS) method. Meng and Verousis (2013) used the 3SLS method to formally examine the hypothesis of a substitution effect between price clustering and size clustering in the CDS market. Finally, Mbanga (2019) estimated robust regressions that eliminate gross outliers to examine the day-of-the-week effect in Bitcoin price clustering.

Another direction arises from the need to analyze panel data. The prevailing approach is a fixed-effects regression. Das and Kadapakkam (2020) included both firm and time fixed effects, whereas (Box and Griffith 2016) included fixed effects only for time and report that once they also included firm fixed effects, the results remained unchanged. Blau and

Griffith (2016) and Blau (2019) included month and year fixed effects, respectively, and used robust standard errors that account for clustering across both the cross-sectional observations and time series observations. On the other hand, Ohta (2006) picked random-effects model over the fixed effects model based on the results from the Hausman specification test.

A substantial part of the literature models price clustering as a binary variable. For that case, the straightforward approach is to use the logit or probit model. From one of the first papers using logistic regression to analyze price clustering, Ball *et al.* (1985) modeled three dependent variables taking value 1 if the price is rounded to the whole dollar, half-dollar, or quarter, respectively. Christie and Schultz (1994) estimated logistic regressions that predict the probability of a firm being quoted using odd eighths. Aitken *et al.* (1996) employed multivariate logistic regression to model three binary dependent variables that are equal to one if the final digit is 0; 0 or 5; and 0, 2, 4, 6, or 8 (even numbers), respectively. Brown and Mitchell (2008) examined the influence of Chinese culture on price clustering by logistic regressions where a binary dependent variable is equal to 1 if the last sale price ends in 4 and 0 if it ends in 8 since many Chinese consider the number 8 as *lucky* while 4 is considered as *unlucky*. From the literature employing probit models, Kahn *et al.* (1999) analyzed the propensity to set retail deposit interest rates at integer levels and (Sopranzetti and Datar 2002) analyzed the propensity for exchange rates to cluster on even digits. Moreover, Capelle-Blancard and Chaudhury (2007) modeled a binary dependent variable that is equal to one if the transaction price ends with 00, whereas (Liu 2011, Narayan and Smyth 2013) set the variable equal to one if the price ends at either 0 or 5, and 0 otherwise. Alexander and Peterson (2007) followed by Lien *et al.* (2019) used a bivariate probit model to take into account the dependence between price and trade size clustering.

Finally, Blau (2019) estimated a vector autoregressive process and examined the impulses of price clustering in response to an exogenous shock to investor sentiment. Besides the classical regression approaches, Harris (1991) and Hameed and Terry (1998) analyzed the cross-sectional data by static discrete price model.

To the best of our knowledge, the literature still lacks a discrete dynamic model to study the high-frequency behavior of the price clustering. Thus, in the next section, we propose a novel model which models high-frequency prices directly at the highest possible frequency and allows us to study the main drivers of price clustering such as price, volatility, volume, and trading frequency in the form of trade durations.

3. Dynamic price clustering model

3.1. Double Poisson distribution

Let us start with the static version of our model for prices. In the first step, we transform the observed prices to have integer values. For example, on the NYSE and NASDAQ exchanges, the prices are recorded with precision to two decimal places and we, therefore, multiply them by 100 to obtain

integer values. The minimum possible change in the transformed prices is 1. Empirically, the transformed prices exhibit strong *underdispersion*, i.e. the variance lower than the mean. In our application, the transformed prices are in the order of thousands and tens of thousands while the price changes are in the order of units and tens. We, therefore, need to base our model on a count distribution allowing for underdispersion. For a review of such distributions, we refer to Sellers and Morris (2017). Although not without its limitations, the double Poisson distribution is the best candidate for our case as the alternative distributions have too many shortcomings. For example, the condensed Poisson distribution is based on only one parameter, the generalized Poisson distribution can handle only limited underdispersion and the gamma count distribution as well as the Conway–Maxwell–Poisson distribution does not have the moments available in a closed form.

The double Poisson distribution was proposed in Efron (1986) and has a *location* parameter $\mu > 0$ and a *dispersion* parameter α . We adopt a slightly different parametrization than (Efron 1986) and use the logarithmic transformation for the dispersion parameter making α unrestricted. For $\alpha = 0$, the distribution reduces to the Poisson distribution. Values $\alpha > 0$ result in underdispersion while values $\alpha < 0$ result in overdispersion. Let Y be a random variable and y an observed value. The probability mass function is given by

$$P[Y = y | \mu, \alpha] = \frac{1}{C(\mu, \alpha)} \frac{y^y}{y!} \left(\frac{\mu}{y}\right)^{e^\alpha y} e^{\frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y}, \quad (1)$$

where $C(\mu, \alpha)$ is the *normalizing constant* given by

$$C(\mu, \alpha) = \sum_{y=0}^{\infty} \frac{y^y}{y!} \left(\frac{\mu}{y}\right)^{e^\alpha y} e^{\frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y}. \quad (2)$$

The log-likelihood for observation y is then given by

$$\begin{aligned} \ell(y; \mu, \alpha) &= -\ln(C(\mu, \alpha)) + y \ln(y) - \ln(y!) + e^\alpha y \ln\left(\frac{\mu}{y}\right) \\ &\quad + \frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y. \end{aligned} \quad (3)$$

Unfortunately, the normalizing constant is not available in a closed form. However, as Efron (1986) shows, it is very close to 1 (at least for some combinations of μ and α) and can be approximated by

$$C(\mu, \alpha) \simeq 1 + \frac{1 - e^\alpha}{12e^\alpha \mu} \left(1 + \frac{1}{e^\alpha \mu}\right). \quad (4)$$

Zou *et al.* (2013) notes that approximation (4) is not very accurate for low values of the mean and suggest approximating the normalizing constant alternatively by cutting off the infinite sum, i.e.

$$C(\mu, \alpha) \simeq \sum_{y=0}^m \frac{y^y}{y!} \left(\frac{\mu}{y}\right)^{e^\alpha y} e^{\frac{\alpha}{2} + e^\alpha y - e^\alpha \mu - y}, \quad (5)$$

where m should be at least twice as large as the sample mean. In our case of high mean, approximation (4) is sufficient while

approximation (5) would be computationally very demanding and we therefore resort to the former one. The expected value and variance can be approximated by

$$E[Y] \simeq \mu, \quad \text{var}[Y] \simeq \mu e^{-\alpha}. \quad (6)$$

The score can be approximated by

$$\begin{aligned} \nabla(y; \mu, \alpha) &\simeq \left(\frac{e^\alpha}{\mu} (y - \mu), e^\alpha (y \ln(\mu) - \mu \right. \\ &\quad \left. - y \ln(y) + y) + \frac{1}{2} \right)'. \end{aligned} \quad (7)$$

The Fisher information can be approximated by

$$\mathcal{I}(\mu, \alpha) \simeq \begin{pmatrix} \frac{e^\alpha}{\mu} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}. \quad (8)$$

3.2. Mixture distribution for price clustering

Next, we propose a mixture of several double Poisson distributions corresponding to trading in different multiples of tick sizes accommodating for price clustering. We consider that there are three types of traders—one who can trade in cents, one who can trade only in multiples of 5 cents, and one who can trade only in multiples of 10 cents. In Appendix 1, we treat a more general case with any number of trader types and tick size multiples. The distribution of prices corresponding to each trader type is based on the double Poisson distribution modified to have support consisting only of multiples of $k \in \{1, 5, 10\}$ while keeping the expected value $E[Y] \simeq \mu$ and the variance $\text{var}[Y] \simeq \mu e^{-\alpha}$ regardless of k . For a detailed derivation of the distribution, see Appendix 1. The distribution of prices for trader type $k \in \{1, 5, 10\}$ is given by

$$\begin{aligned} P[Y^{[k]} = y | \mu, \alpha] &= \mathbb{I}\{k | y\} P\left[Z^{[k]} = \frac{y}{k} \middle| \mu, \alpha\right], \\ Z^{[k]} &\sim \text{DP}\left(\frac{\mu}{k}, \alpha + \ln(k)\right), \end{aligned} \quad (9)$$

where DP denotes the double Poisson distribution and $\mathbb{I}\{k | y\}$ is equal to 1 if y is divisible by k and 0 otherwise. Note that for $k = 1$, it is the standard double Poisson distribution. Finally, the distribution of all prices is the mixture

$$\begin{aligned} P[Y = y | \mu, \alpha, \varphi_1, \varphi_5, \varphi_{10}] &= \sum_{k \in \{1, 5, 10\}} \varphi_k P[Y^{[k]} = y | \mu, \alpha], \end{aligned} \quad (10)$$

where the parameter space is restricted by $\mu > 0$, $\varphi_1 \geq 0$, $\varphi_5 \geq 0$, $\varphi_{10} \geq 0$ and $\varphi_1 + \varphi_5 + \varphi_{10} = 1$. Parameters φ_k , $k \in \{1, 5, 10\}$, are the portions of trader types and parameters μ with α have the same interpretation as in the double Poisson distribution. The log-likelihood for observation y is given by

$$\begin{aligned} \ell(y; \mu, \alpha, \varphi_1, \varphi_5, \varphi_{10}) &= e^\alpha y \ln\left(\frac{\mu}{y}\right) + \frac{\alpha}{2} + e^\alpha y - e^\alpha \mu \end{aligned}$$

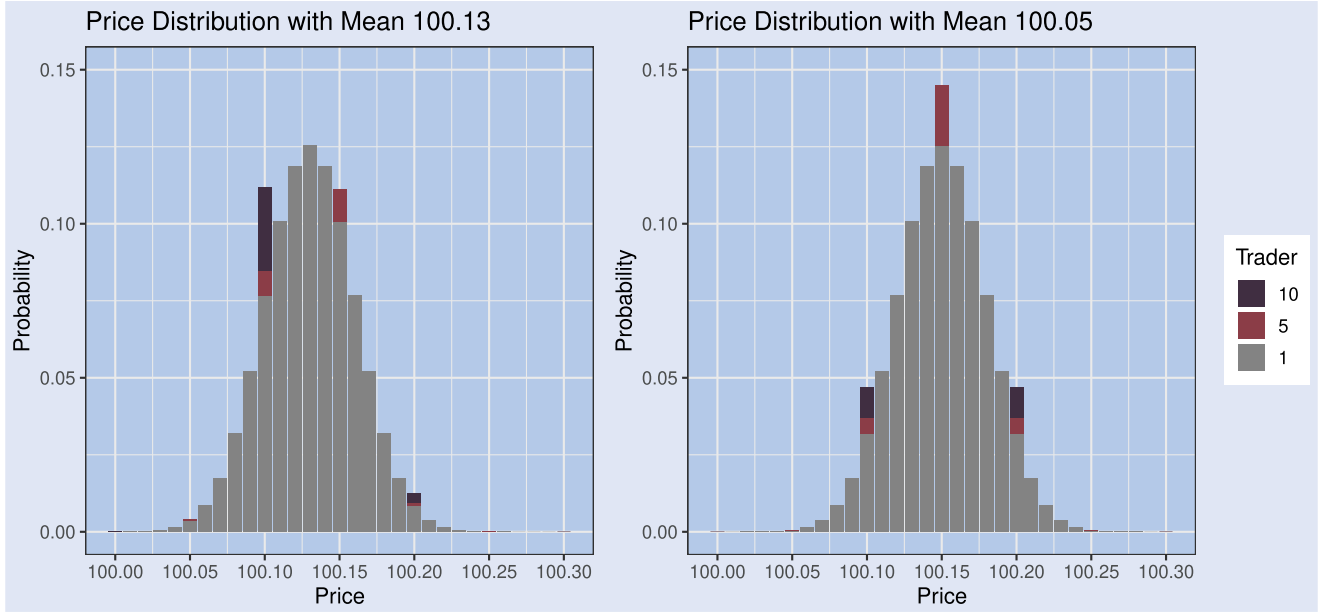


Figure 1. Illustration of the probability mass function for the mixture double Poisson distribution with parameters $\mu = 10013$ (left plot), $\mu = 10005$ (right plot), $\alpha = 7$, $\varphi_1 = 0.95$, $\varphi_5 = 0.02$, and $\varphi_{10} = 0.03$. The prices are reported in the original form with two decimal places.

$$+ \ln \left(\sum_{k \in \{1,5,10\}} \varphi_k \mathbb{I}\{k | y\} \frac{\sqrt{k}}{C\left(\frac{\mu}{k}, k\alpha\right)} \left(\frac{y}{k}\right)^{\frac{y}{k}} e^{-\frac{y}{k}} \right). \quad (11)$$

Note that the last logarithm in (11) is not dependent on parameters μ and α besides the normalizing constant making the approximation of the score quite simple. Additionally, parameters φ_1 , φ_5 and φ_{10} appear only in the last logarithm in (11) making the approximation of the score for parameters μ and α independent of parameters φ_1 , φ_5 and φ_{10} . The approximations of the expected value and the variance as well as the score and the Fisher information for the parameters μ and α of the mixture distribution are therefore the same as for the regular double Poisson distribution presented in (6), (7), and (8), respectively when assuming $C(\mu/k, k\alpha) = 1$.

Figure 1 illustrates the probability mass function of the mixture distribution. As an example, we choose the portions of trader types as $\varphi_1 = 0.95$, $\varphi_5 = 0.02$, and $\varphi_{10} = 0.03$. The majority of traders, therefore, operate at one cent precision but some are restricted to trading in only five or ten cents, similarly to the observed behavior in the empirical study. The mixture distribution is based on the regular double Poisson distribution with all probabilities decreased by 5 %. Prices ending with 5 or 0 are then inflated according to distribution given by (9) for $k = 5$ with all probabilities decreased by 98 %. Finally, prices ending with 0 are further inflated according to the distribution given by (9) for $k = 10$ with all probabilities decreased by 97 %. Figure 1 shows how the probabilities can be decomposed for the three trader types.

3.3. Dynamics of time-varying parameters

Finally, we introduce time variation into parameters $\mu, \alpha, \varphi_1, \varphi_5, \varphi_{10}$. We denote the random prices as $Y_t \in \mathbb{N}_0$, $t = 1, \dots, n$ and the observed values as $y_t \in \mathbb{N}_0$, $t = 1, \dots, n$. We also utilize observed trade durations $z_t \in \mathbb{R}^+$, $t = 1, \dots, n$ and observed volumes $v_t \in \mathbb{R}^+$, $t = 1, \dots, n$. We assume that

Y_t follow the mixture double Poisson distribution proposed in Section 3.2 with time-varying parameters $\mu_t, \alpha_t, \varphi_{1,t}, \varphi_{5,t}$ and $\varphi_{10,t}$. The dynamics of the location parameter μ_t is given by

$$\mu_t = y_{t-1}. \quad (12)$$

This means that the expected value of the price is (approximately) equal to the last observed price, i.e. the expected value of the return is zero. This is a common assumption for high-frequency returns (see, e.g. Koopman *et al.* 2017).

For the dynamics of the dispersion parameter α_t , we utilize the generalized autoregressive score (GAS) model of Creal *et al.* (2013), also known as the dynamic conditional score (DCS) model by Harvey (2013). The GAS model is an observation-driven model providing a general framework for modeling time-varying parameters for any underlying probability distribution. It captures dynamics of time-varying parameters by the autoregressive term and the score of the conditional density function. Blasques *et al.* (2015) investigated information-theoretic optimality properties of the score function and showed that only parameter updates based on the score will always reduce the local Kullback–Leibler divergence between the true conditional density and the model-implied conditional density. Creal *et al.* (2013) suggested to scale the score based on the Fisher information. As the Fisher information for the parameter α_t is constant in our case, the score is already normalized and we therefore omit the scaling. Using (7) and (12), we let the dispersion parameter α_t follow the recursion

$$\alpha_t = c + b\alpha_{t-1} + a \left(e^{\alpha_{t-1}} \left(y_{t-1} \ln \left(\frac{y_{t-2}}{y_{t-1}} \right) - y_{t-2} + y_{t-1} \right) + \frac{1}{2} \right) + d \ln(z_t), \quad (13)$$

where c is the constant parameter, b is the autoregressive parameter, a is the score parameter and d is the

duration parameter. This volatility dynamics corresponds to the generalized autoregressive conditional heteroskedasticity (GARCH) model of Bollerslev (1986). Similarly to Engle (2000), we also include the preceding trade duration z_t as an explanatory variable to account for irregularly spaced observations. To prevent extreme values of durations, we use the logarithmic transformation.

The portions of trader types are driven by process

$$\begin{aligned} \eta_t = & f\eta_{t-1} + g_1 \ln(\mu_t) + g_2 (\ln(\mu_t) - \alpha_t) \\ & + g_3 \ln(z_t) + g_4 \ln(v_t), \end{aligned} \quad (14)$$

where f is the autoregressive parameter, g_1 is the parameter for the logarithm of the expected price, g_2 is the parameter for the logarithm of the variance of the price process $\mu_t e^{-\alpha_t}$, g_3 is the parameter for the logarithm of the preceding trade duration, and g_4 is the parameter for the logarithm of the volume v_t . The portions of trader types are then standardized as

$$\begin{aligned} \varphi_{1,t} &= \frac{e^{\eta_t}}{e^{\eta_t} + h_5 + h_{10}}, & \varphi_{5,t} &= \frac{h_5}{e^{\eta_t} + h_5 + h_{10}}, \\ \varphi_{10,t} &= \frac{h_{10}}{e^{\eta_t} + h_5 + h_{10}}, \end{aligned} \quad (15)$$

where $h_5 \geq 0$ and $h_{10} \geq 0$ are parameters capturing the representation of 5 and 10 trader types. The model can be straightforwardly extended to include additional explanatory variables in (13) and (14).

3.4. Maximum likelihood estimation

The proposed model based on the mixture distribution for price clustering (10) with dynamics given by (12), (13) and (15) can be straightforwardly estimated by the conditional maximum likelihood method. Let $\theta = (c, b, a, d, f, g_1, g_2, g_3, g_4, h_5, h_{10})'$ denote the static vector of all parameters. The parameter vector θ is then estimated by the conditional maximum likelihood

$$\hat{\theta} \in \arg \max_{\theta} \sum_{t=1}^n \ell(y_t; \mu_t, \alpha_t, \varphi_{1,t}, \varphi_{5,t}, \varphi_{10,t}), \quad (16)$$

where $\ell(y_t; \mu_t, \alpha_t, \varphi_{1,t}, \varphi_{5,t}, \varphi_{10,t})$ is given by (11).

Note that process (12) implies non-stationarity of our model, even though our main focus is on processes (13) and (14), in which we assume $b < 1$ and $f < 1$, respectively. This reflects a commonly used assumption of non-stationarity of prices. Non-stationary GAS models were discussed and applied e.g. by Gorgi *et al.* (2019), Harvey *et al.* (2019), and Holý and Zouhar (2021). In our case, the use of a non-stationary model is necessary as we require to model prices directly in order to capture price clustering.

For the numerical optimization in the empirical study, we utilize the PRincipal AXIS algorithm of Brent (1972). To improve numerical performance, we standardize the explanatory variables to unit mean. We also run the estimation procedure several times with different starting values to avoid local maxima.

4. Empirical results

4.1. Data sample

The empirical study is conducted on transaction data extracted from the NYSE TAQ database which contains intraday data for all securities listed on the New York Stock Exchange (NYSE), American Stock Exchange (AMEX), and Nasdaq Stock Market (NASDAQ). We analyze 30 stocks that form the Dow Jones Industrial Average (DJIA) index in June 2020. The extracted data span over six months from January 2 to June 30, 2020, except for Raytheon Technologies (RTX)[†] for which the data are available from April 3, 2020.

We follow the standard cleaning procedure for the NYSE TAQ dataset described in Barndorff-Nielsen *et al.* (2009) since data cleaning is an important step of high-frequency data analysis (Hansen and Lunde 2006). Before the standard data pre-processing is conducted, we delete entries that are identified as preferred or warrants (trades with the non-empty suffix indicator). Then we follow a common data cleaning steps and discard (i) entries outside the main opening hours (9:30–16:00), (ii) entries with the transaction price equal to zero, (iii) entries occurring on a different exchange than it is primarily listed, (iv) entries with corrected trades, (v) entries with abnormal sale condition, (vi) entries for which the price deviated by more than 10 mean absolute deviations from a rolling centered median of 50 observations and (vi) duplicate entries in terms of the time stamp. In the last step, we remain the entry with mode price instead of the originally suggested median price due to avoiding distortion of the last decimal digit of prices.

The first and last step has a negligible impact on our data and steps ii, iv, and vi have no impact at all. However, the third step causes a large deletion of the data which is, however, in line with (Barndorff-Nielsen *et al.* 2009). The basic descriptive statistics after data pre-processing are shown in Appendix 2. Number of observations ranges from 216 618 (TRV) to 3 099 279 (MSFT). Price clustering in terms of the excess occurrence of multiples of five cents and ten cents in prices ranges from 1.45 % (KO) to 11.52 % (BA). First, we analyze the price clustering using a common approach of fixed effects model on daily data in Section 4.2 to investigate whether the results for our dataset are in line with the existing literature. Then, we estimate the proposed dynamic price model in Section 4.3.

4.2. Analysis based on daily data

In this section, we investigate the main determinants of price clustering for which pervasive evidence is documented in the literature, namely price, volatility, trading frequency (which we measure in terms of trade durations), and volume. We use a panel regression with fixed effects to take into account the unobserved heterogeneity in both dimensions—stocks and days.

Let us define price clustering $p_{i,t}$ as the excess relative frequency of multiples of five cents and ten cents in prices of

[†] The RTX company results from the merge of the United Technologies Corporation and the Raytheon Company on April 3, 2020.

Table 1. Estimated coefficients and robust standard errors of models with fixed effects for stocks and days.

Variable	I	II	III
Price	− 1.7014*** (0.5469)	− 0.8067 (0.6615)	− 0.1245 (0.5323)
Volatility	0.5900*** (0.1901)		0.7008*** (0.1622)
Duration	− 0.0973 (0.1672)	− 0.4640*** (0.1347)	− 0.0069 (0.1739)
Volume		3.7544*** (0.4869)	3.9557*** (0.5025)

Notes: Label I refers to the model in (17) for which the volume parameter $\beta_4 = 0$; II to the model in (17) for which the volatility parameter $\beta_2 = 0$; III to the full model in (17). * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

stock i at day t . We model $p_{i,t}$ as

$$p_{i,t} = \gamma_i + \delta_t + \beta_1 \ln(\bar{\mu}_{i,t}) + \beta_2 \ln(\bar{\alpha}_{i,t}) + \beta_3 \ln(\bar{z}_{i,t}) + \beta_4 \ln(\bar{v}_{i,t}) + \varepsilon_{i,t}, \quad (17)$$

where γ_i is a stock-specific effect for the stock i , δ_t is a time effect for day t , and $\varepsilon_{i,t}$ is the error term. Parameters $\beta_1 - \beta_4$ corresponds to logarithmic explanatory variables, where $\bar{\mu}_{i,t}$ is an average price, $\bar{z}_{i,t}$ is an average duration and $\bar{v}_{i,t}$ is an average volume, where all averages are calculated for each stock i at each day t . Daily volatility $\bar{\alpha}_{i,t}$ is estimated by realized kernel estimator of Barndorff-Nielsen *et al.* (2008). We use Parzen kernel as suggested by Barndorff-Nielsen *et al.* (2009). See Holý and Tomanová (2021) for a comprehensive overview of quadratic covariation estimators.

Table 1 reports estimated coefficients of three variants of the fixed effects model. The first variant models price clustering on price, volatility and duration, i.e. model in (17) where volume is skipped. The second model considers only the price, duration and volume as the explanatory variables and the third one is the full model in (17). We test the significance of the estimated coefficients using robust standard errors for which observations are clustered in both dimensions to account for serial as well as cross-sectional correlation. For illustration, Figure 2 shows fitted lines from univariate regressions with stock specific effects[†] for two stocks traded on NASDAQ—Apple Inc. (AAPL) and Microsoft (MSFT)—and two stocks traded on NYSE—Boeing (BA) and Visa Inc. (V). Figure 2 aims to depict the presence of the stock-specific effects: in descending order, BA, AAPL and MSFT have the highest level of price clustering while V represents a stock with an average level of price clustering from our Dow Jones dataset.

The results show that volatility is a highly significant driver of price clustering. Moreover, the effect is positive, which is in line with the overwhelming majority of literature, in particular, with the so-called *price resolution* hypothesis. We refer to Section 4.4 for a literature review on this topic and related implications of our models. Next, Table 1 shows that volume

is also highly significant and has a positive effect on price clustering according to expectations. For example, this is in line with results from the panel data analysis of Box and Griffith (2016) and Das and Kadapakkam (2020). Das and Kadapakkam (2020) suggested that the reason behind the significantly positive effect is algorithmic trading: as the trades become smaller in size due to increased algorithmic trading, lower levels of price clustering are observed.

The volatility is negatively correlated with duration; however, the duration does not bring additional information since it is no longer significant once the volatility is added in the model (see Model II vs. III). A similar result applies for volume and price (see Model I vs. III), where price turns to be insignificant once the volume is added to the model. We conclude that based on our model using daily data: (i) we do not find a significant effect of duration and price level on price clustering; (ii) the volatility and volume have highly significant positive effects on price clustering which are in line with expectations.

4.3. High-frequency analysis

Let us analyze the price clustering phenomenon at the highest possible frequency. First, we take a brief look at the relation between the individual explanatory variables and price clustering. We focus on the BA stock as its price clustering is the most pronounced. Figure 3 shows the average expected price, the average instantaneous variance obtained from the dynamic model, the average duration preceding the trade, and the average volume broken down by the second decimal of the price for the BA stock. We can clearly see that for prices ending with 0 and 5, the average variance and the average trade duration is much lower than for the other digits while the average volume is much higher. Note that succeeding durations show very similar behavior to preceding durations suggesting that price clustering tends to occur when trading is more intense.

Next, we estimate three versions of the proposed price clustering model for each of the 30 stocks. In the first version, we assume that there is no price clustering and set $f = g_1 = g_2 = g_3 = g_4 = h_5 = h_{10} = 0$. In the second version, we set only $f = g_1 = g_2 = g_3 = g_4 = 0$ and assume that there is price clustering present but is constant over time and does not depend on any variables. The third version is the dynamic model presented in Section 3.3 without any restrictions. We report the average log-likelihood and the Akaike information criterion (AIC) of the models in Table 2. We can see that adding price clustering to the model and subsequently adding dynamics to price clustering is very much worth of the extra few parameters as AIC is distinctly the lowest for the dynamic model for all stocks.

From now on we focus on the model with dynamic price clustering. Table 3 reports the estimated coefficients. We do not report standard deviations as they are very close to zero and all coefficients are significant at any reasonable level due to a huge number of observations. For all stocks, the coefficients in the volatility process c , b , a , and d have the same signs and fairly similar values demonstrating the robustness of the model. Parameter d is negative, which means that with longer durations, dispersion parameter α_t is lower

[†] Time effects are dropped for better visibility which does not alter the main result.

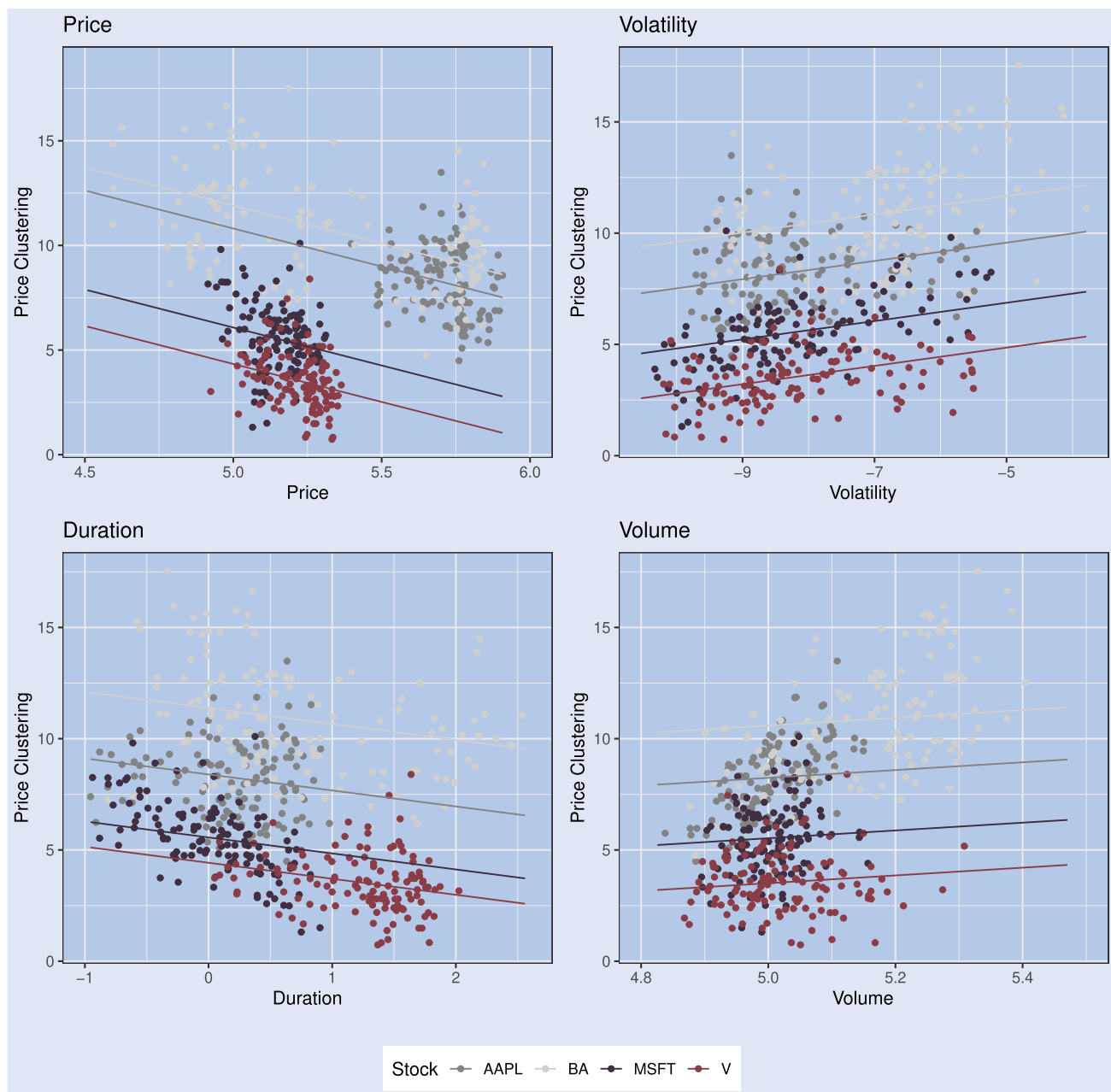


Figure 2. Daily data of four selected stocks and fitted lines from univariate panel regressions of price clustering on logarithmic price (top left), logarithmic volatility (top right), logarithmic duration (bottom left) and logarithmic volume (bottom right). All models are estimated with stock fixed effects.

and the instantaneous variance $\mu_t \exp(-\alpha_t)$ is higher. We attribute this behavior to the presence of a large amount of extremely short durations associated with small price changes. Note that for example, the BA stock has 50 % of durations shorter than 0.1 s and 19 % shorter than 0.0001 s. Engle (2000) observed the opposite relation between trade durations and volatility but based his results on data with a much lower frequency and without durations shorter than 1 s. This might indicate a change in the data structure over the years and a complex non-linear relation between trade durations and volatility. This topic is, however, beyond the scope of this paper.

Regarding the dynamics of price clustering, the autoregressive parameter f is stable across all stocks. The parameter for the expected price g_1 significantly varies for different stocks

suggesting its low informative value. This is in line with the daily analysis in which prices were found insignificant. The parameter for the instantaneous variance g_2 is positive for all stocks. The portion of one cent traders is, therefore, higher with higher variance and price clustering tends to occur when prices are less volatile. This is the most interesting result as it deviates from the behavior observed in the daily analysis. The parameter for the preceding trade duration g_3 significantly varies for different stocks, similarly to g_1 . When the preceding trade duration is the only explanatory variable included in the model; however, g_3 is positive for all stocks. Recall that durations have a positive effect on instantaneous variance as d is positive for all stocks. This implies that durations have an effect on instantaneous variance which in turn has an effect on price clustering. However, when controlling for instantaneous

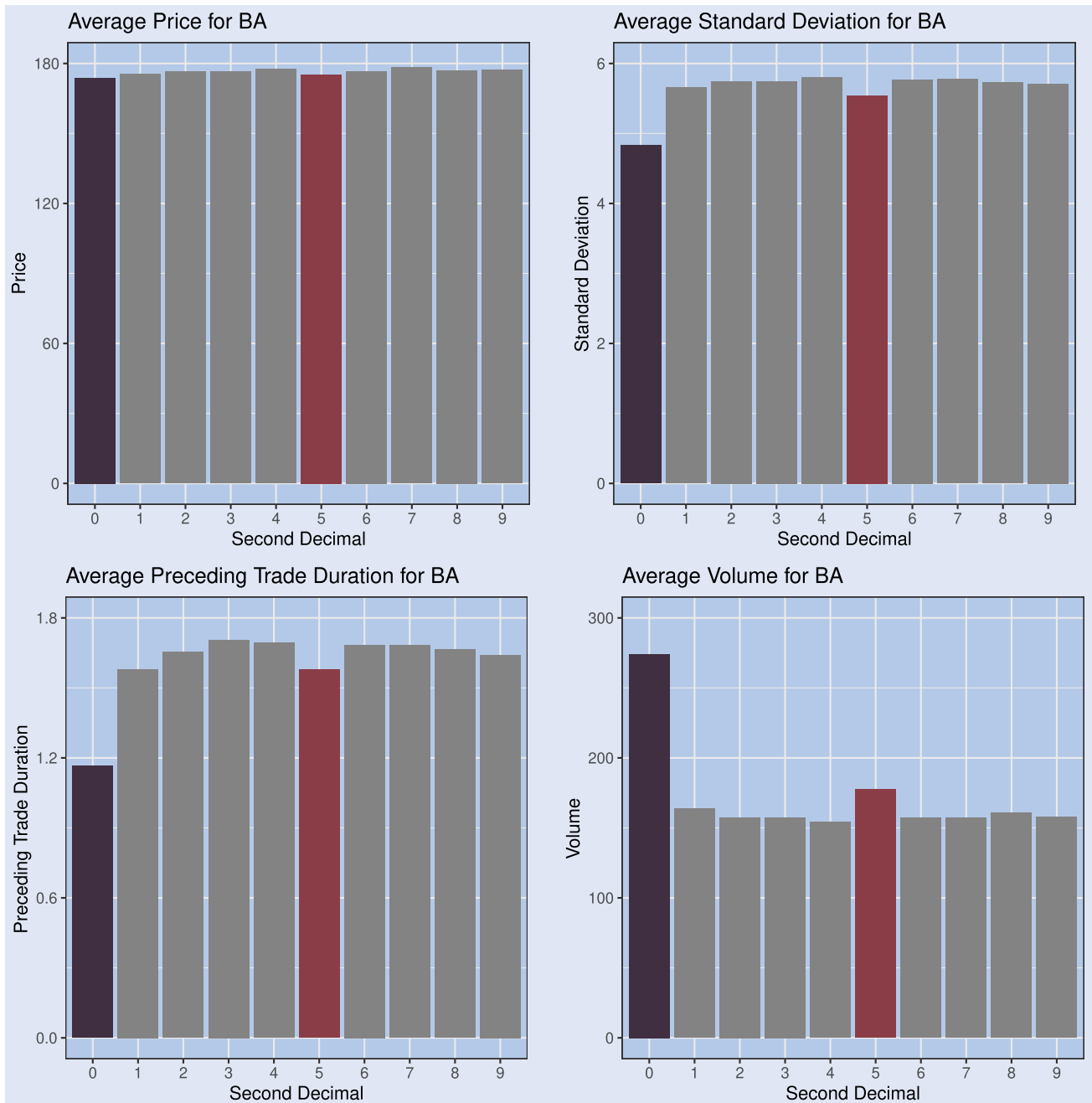


Figure 3. The average price (top left), the average standard deviation (top right), the average preceding trade duration (bottom left), and the average volume (bottom right) broken down by the second decimal digit of the BA stock prices.

variance, durations do not bring additional information to explain price clustering. These observations are in line with the daily analysis. Finally, the parameter for the volume g_4 is negative for all stocks. As in the daily analysis, higher volume is clearly associated with higher price clustering.

We omit parameters h_5 and h_{10} controlling strength of price clustering from Table 3 as they are not very informative for readers. It is far better to look at the average values of trader portions $\bar{\varphi}_1$, $\bar{\varphi}_5$ and $\bar{\varphi}_{10}$ reported in Table 4. The average portion of ten-cent traders ranges from 0.48 % for the TRV stock to 10.54 % for the BA stock. The average portion of five-cent traders ranges from 0.34 % for the TRV stock to 3.63 % for the BA stock. An example of the progression of trader type ratios is shown in Figure 4 for the BA stock on the first trading day of 2020.

As a benchmark, we compare the proposed model with the GARCH model based on the Student's t -distribution of Bollerslev (1987). For each stock, we estimate three specifications based on prices, price differences, and logarithmic returns. The direct use of prices results in a non-stationary model, similarly to the proposed model. As our goal is to compare models based on different data transformations, we report their log-likelihoods with respect to the original discrete prices. This approach is used, e.g. by Blasques *et al.* (2020). Table 5 reports the fit of the GARCH models. Based on these results, we cannot unambiguously determine which model is the best. The GARCH models perform better for 18 stocks while the proposed price clustering model performs better for 12 stocks. Among the GARCH models, the one based on price differences has the highest log-likelihood

Table 2. The average log-likelihood and AIC of the model without price clustering, the model with static price clustering, and the model with dynamic price clustering.

Stock	No PC		Static PC		Dynamic PC	
	Lik.	AIC	Lik.	AIC	Lik.	AIC
AAPL	-2.3548	12 080 434	-2.3292	11 949 022	-2.2953	11 775 068
AXP	-2.5117	2 284 414	-2.5054	2 278 659	-2.5008	2 274 477
BA	-2.8677	10 745 431	-2.8146	10 546 574	-2.7980	10 484 266
CAT	-2.6144	2 092 986	-2.6090	2 088 625	-2.6066	2 086 695
CSCO	-0.4650	1 525 535	-0.4636	1 520 852	-0.4606	1 510 984
CVX	-2.1513	4 094 545	-2.1476	4 087 384	-2.1444	4 081 369
DIS	-2.1382	5 620 020	-2.1295	5 597 162	-2.1231	5 580 359
DOW	-1.6701	1 981 601	-1.6685	1 979 699	-1.6668	1 977 626
GS	-3.1560	2 292 436	-3.1500	2 288 064	-3.1460	2 285 185
HD	-3.1018	3 043 787	-3.0969	3 039 026	-3.0935	3 035 655
IBM	-2.4432	2 689 846	-2.4393	2 685 501	-2.4360	2 681 876
INTC	-0.8455	3 165 039	-0.8442	3 160 026	-0.8407	3 147 091
JNJ	-2.3836	3 976 323	-2.3809	3 971 919	-2.3794	3 969 402
JPM	-2.0310	5 831 271	-2.0264	5 817 872	-2.0220	5 805 430
KO	-1.2238	2 758 822	-1.2230	2 757 081	-1.2212	2 753 113
MCD	-2.9611	2 787 186	-2.9563	2 782 675	-2.9526	2 779 159
MMM	-2.7229	2 356 746	-2.7180	2 352 455	-2.7149	2 349 844
MRK	-1.7024	3 763 397	-1.7011	3 760 622	-1.6998	3 757 820
MSFT	-1.7431	10 325 970	-1.7315	10 257 392	-1.7096	10 127 509
NKE	-2.1331	2 878 657	-2.1310	2 875 789	-2.1292	2 873 371
PFE	-0.8423	2 403 152	-0.8415	2 400 862	-0.8398	2 396 048
PG	-2.3049	3 882 847	-2.3024	3 878 712	-2.3007	3 875 756
RTX	-1.7279	1 602 691	-1.7238	1 598 954	-1.7207	1 596 096
TRV	-2.7871	1 135 714	-2.7861	1 135 292	-2.7848	1 134 797
UNH	-3.3882	3 716 005	-3.3834	3 710 792	-3.3816	3 708 786
V	-2.6542	5 058 066	-2.6485	5 047 243	-2.6434	5 037 606
VZ	-1.3165	3 025 851	-1.3156	3 023 684	-1.3141	3 020 201
WBA	-1.1807	1 507 681	-1.1787	1 505 148	-1.1748	1 500 227
WMT	-2.1291	3 718 970	-2.1262	3 714 030	-2.1233	3 708 967
XOM	-1.1969	4 893 716	-1.1951	4 886 388	-1.1895	4 863 486

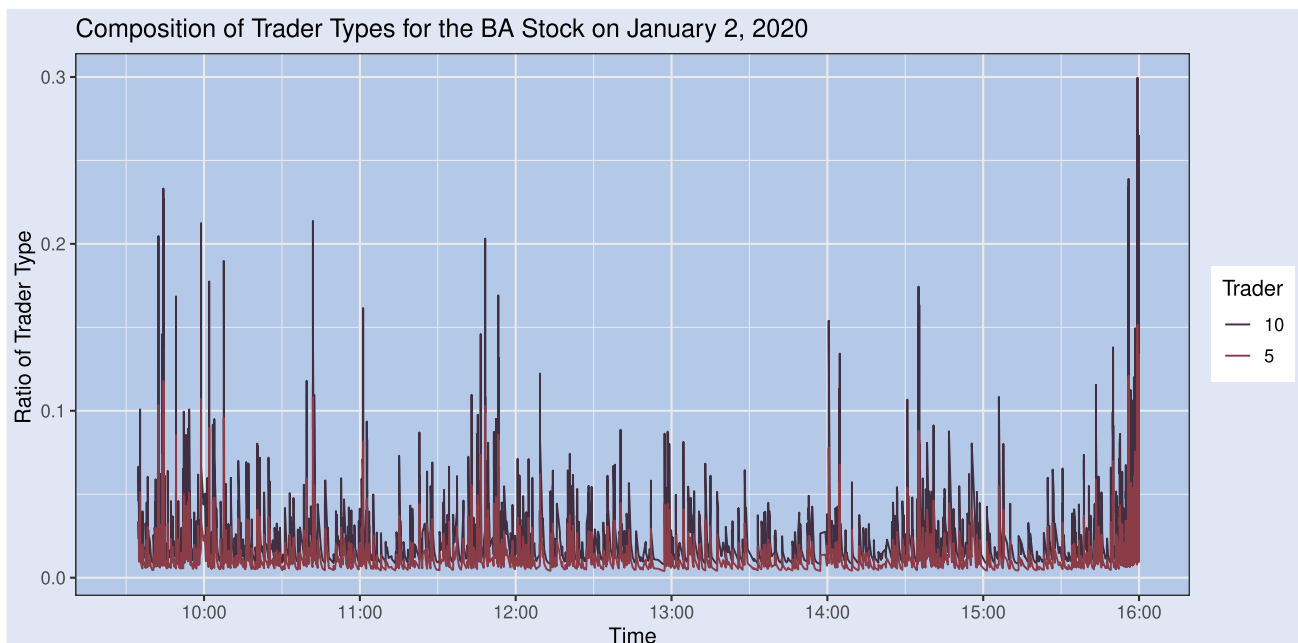


Figure 4. The time-varying portions of trader types obtained from the proposed price clustering model for the BA stock on January 2, 2020.

for 17 stocks. We can conclude that the performance of our proposed model is comparable to the GARCH model based on the Student's t -distribution, regardless of the price transformation.

4.4. Implications

Several hypotheses have been established to explain the price clustering phenomenon. The *attraction hypothesis* of Goodhart and Curcio (1991) essentially states that there exists a

Table 3. Estimated coefficients of the proposed dynamic price clustering model.

Stock	c	b	a	d	f	g_1	g_2	g_3	g_4
AAPL	6.09	0.08	0.29	-0.40	0.66	-0.33	0.61	-0.11	-0.69
AXP	5.32	0.10	0.35	-0.28	0.34	1.05	0.05	0.06	-0.74
BA	5.00	0.09	0.30	-0.29	0.39	-0.14	0.18	0.03	-0.71
CAT	5.60	0.04	0.26	-0.27	0.25	-0.28	0.29	-0.00	-0.71
CSCO	6.11	0.17	0.08	-0.35	0.81	0.28	0.40	-0.08	-0.37
CVX	5.79	0.11	0.35	-0.26	0.36	0.70	0.28	0.02	-0.64
DIS	6.05	0.12	0.33	-0.25	0.39	0.45	0.21	0.05	-0.64
DOW	5.71	0.15	0.33	-0.22	0.34	0.55	0.18	0.04	-0.70
GS	4.89	0.05	0.27	-0.31	0.29	1.06	0.12	0.02	-0.90
HD	5.01	0.08	0.29	-0.28	0.29	0.64	0.10	0.05	-0.83
IBM	5.79	0.08	0.32	-0.28	0.32	0.33	0.10	0.06	-0.80
INTC	6.29	0.14	0.13	-0.35	0.74	0.38	0.34	-0.06	-0.63
JNJ	5.76	0.13	0.40	-0.25	0.23	1.12	0.08	0.05	-0.74
JPM	5.76	0.17	0.44	-0.26	0.33	0.48	0.67	-0.07	-0.56
KO	5.97	0.25	0.33	-0.18	0.55	-0.05	0.62	0.01	-0.43
MCD	5.16	0.08	0.34	-0.27	0.31	1.69	0.02	0.06	-0.88
MMM	5.55	0.05	0.29	-0.27	0.25	0.93	0.06	0.06	-0.87
MRK	5.89	0.20	0.40	-0.23	0.45	0.20	0.71	-0.09	-0.53
MSFT	6.37	0.11	0.27	-0.38	0.72	0.17	0.39	-0.05	-0.72
NKE	6.01	0.10	0.35	-0.25	0.29	0.45	0.05	0.08	-0.70
PFE	5.19	0.38	0.28	-0.15	0.70	0.36	0.34	0.03	-0.33
PG	5.84	0.11	0.40	-0.23	0.41	-0.13	1.35	-0.25	-0.53
RTX	6.12	0.15	0.34	-0.24	0.32	-0.50	0.02	0.10	-0.65
TRV	5.07	0.05	0.30	-0.30	0.96	-0.66	0.57	-0.14	-0.02
UNH	4.58	0.07	0.28	-0.27	0.39	0.08	0.11	0.03	-0.73
V	5.91	0.06	0.32	-0.27	0.32	0.65	0.09	0.06	-0.91
VZ	5.79	0.27	0.38	-0.19	0.44	2.53	0.71	-0.05	-0.53
WBA	5.82	0.12	0.18	-0.34	0.68	0.38	0.26	-0.00	-0.62
WMT	6.19	0.12	0.38	-0.24	0.34	-2.33	0.15	0.08	-0.71
XOM	6.11	0.24	0.31	-0.18	0.57	-0.51	0.98	-0.06	-0.52

Specifically, c is the constant parameter, b is the autoregressive parameter, a is the score parameter, and d is the duration parameter, which enters the dynamic equation for the dispersion parameter α_t . Further, f is the autoregressive parameter, g_1 is the parameter for the logarithm of the expected price, g_2 is the parameter for the logarithm of the variance of the price process $\mu_t e^{-\alpha_t}$, g_3 is the parameter for the logarithm of the preceding trade duration, and g_4 is the parameter for the logarithm of the volume v_t , which enter the process for the portions of trader types.

particular preference (basic attraction) for certain numbers, especially for the rounded ones. The *negotiation hypothesis* of Harris (1991) assumes that traders use discrete price sets to lower the costs of negotiating. Once the set of prices is reduced, the traders reach agreements more easily since the amount of information that must be exchanged between negotiating traders decreases. Christie and Schultz (1994) argued in the *collusion hypothesis* that the lack of odd-eighth quotes on NASDAQ cannot be explained by the negotiation hypothesis, trading activity, or other variables thought to impact spreads, which suggests that NASDAQ dealers might implicitly collude to maintain wide spreads. However, assessing these hypotheses is out of the scope of our paper. In this section, we focus only on the most studied hypothesis in the literature – the price resolution hypothesis.

The *price resolution hypothesis* of Ball *et al.* (1985) considers the source of price clustering to be the uncertainty. It states that when the amount of information in the market is low and the volatility becomes higher, the market participants incline to round their prices, and consequently, the price clustering increases. This hypothesis was confirmed by many

Table 4. The average values of the time-varying parameters of the proposed dynamic price clustering model, where μ_t is the location parameter for the price process and h_1 , h_5 and h_{10} are parameters capturing representation of 1, 5, and 10 trader types.

Stock	$\bar{\mu}$	$\bar{\alpha}$	$\bar{\varphi}_1$	$\bar{\varphi}_5$	$\bar{\varphi}_{10}$
AAPL	291.80	8.43	92.37	0.72	6.91
AXP	98.43	7.01	95.05	1.81	3.15
BA	175.80	6.85	85.84	3.63	10.54
CAT	117.91	6.99	95.40	1.85	2.75
CSCO	42.22	10.26	98.07	0.54	1.38
CVX	88.26	7.62	96.26	1.32	2.42
DIS	112.50	7.90	94.18	2.25	3.56
DOW	36.46	7.69	97.36	1.36	1.29
GS	193.14	6.40	95.49	1.06	3.45
HD	215.44	6.63	95.72	1.44	2.84
IBM	124.05	7.38	96.01	1.64	2.34
INTC	57.72	9.82	97.83	0.65	1.52
JNJ	140.37	7.63	96.78	1.36	1.85
JPM	103.65	8.03	95.47	2.54	1.98
KO	48.33	8.88	98.29	0.67	1.04
MCD	183.86	6.74	96.12	0.72	3.16
MMM	149.92	7.02	95.87	1.17	2.96
MRK	79.28	8.43	97.69	1.23	1.08
MSFT	169.29	9.11	94.57	0.84	4.59
NKE	88.58	7.67	96.89	1.60	1.51
PFE	34.46	9.30	98.07	1.01	0.91
PG	116.41	7.61	96.82	1.87	1.31
RTX	62.28	8.13	95.83	2.06	2.11
TRV	111.57	6.59	99.18	0.34	0.48
UNH	271.73	6.29	96.01	1.08	2.91
V	180.13	7.34	95.59	1.03	3.37
VZ	55.60	8.85	97.90	1.13	0.97
WBA	45.73	8.91	97.49	0.72	1.79
WMT	118.65	7.98	96.57	1.56	1.87
XOM	45.82	8.86	97.02	1.54	1.44

Values of $\bar{\mu}$ are in dollars and values of $\bar{\varphi}_1$, $\bar{\varphi}_5$, and $\bar{\varphi}_{10}$ are in percent.

studies. The studies found that price clustering increases with volatility using different data and measures. For example, Ahn *et al.* (2005) computed the volatility as the inverse of the daily return standard deviation, while (Ikenberry and Weston 2008) used the standard deviation of returns over the sample period, Box and Griffith (2016) used the standard deviation of 15-minute continuously compounded midpoint returns over the trading day, Schwartz *et al.* (2004) used the difference in the high and low prices for the day, and (Lien *et al.* 2019) utilized the transitory volatility defined as the coefficient of variation of intraday trade prices. Davis *et al.* (2014) found that price clustering is positively related to volatility, however, only when a non-high-frequency trading firm provides liquidity. On the contrary to the vast majority of the literature, Blau (2019) reported based on panel regressions that the volatility is negatively related to price clustering, where the volatility is measured as the standard deviation of residual returns obtained from estimating a Fama and French 3-factor model.

Our results from the daily analysis show that the realized volatility is highly significant and positively related to the price clustering. This finding is in line with the price resolution hypothesis. Interestingly, instantaneous volatility obtained from the proposed dynamic price model has

Table 5. The average log-likelihood and AIC of the GARCH model based on prices, price differences, and logarithmic returns.

Stock	GARCH Prices		GARCH Diff.		GARCH Returns	
	Lik.	AIC	Lik.	AIC	Lik.	AIC
AAPL	-2.5746	13 208 099	-5.2727	27 049 463	-4.3884	22 512 972
AXP	-2.4059	2 188 119	-2.3377	2 126 170	-2.3763	2 161 238
BA	-2.7394	10 264 832	-2.6292	9 851 685	-2.6574	9 957 573
CAT	-2.5208	2 018 017	-2.4113	1 930 365	-2.4510	1 962 110
CSCO	-1.5249	5 002 357	-2.0551	6 741 593	-2.1396	7 018 554
CVX	-2.0584	3 917 643	-2.0497	3 901 118	-2.0949	3 987 076
DIS	-2.0628	5 421 767	-2.0354	5 349 755	-2.0673	5 433 680
DOW	-1.6886	2 003 496	-4.5152	5 357 203	-2.3970	2 844 000
GS	-3.0694	2 229 509	-2.9351	2 131 952	-2.9922	2 173 438
HD	-2.9571	2 901 792	-2.8380	2 784 909	-2.8896	2 835 555
IBM	-2.3741	2 613 784	-2.2938	2 525 361	-2.3335	2 569 071
INTC	-1.9481	7 292 699	-2.6590	9 953 934	-2.6006	9 735 148
JNJ	-2.2352	3 728 754	-2.1734	3 625 695	-2.2070	3 681 711
JPM	-1.9525	5 605 765	-1.9862	5 702 612	-1.9881	5 707 947
KO	-1.9314	4 354 157	-3.8296	8 633 353	-3.9745	8 959 837
MCD	-2.8049	2 640 119	-2.6921	2 534 034	-2.7435	2 582 349
MMM	-2.6323	2 278 283	-2.5214	2 182 343	-2.5606	2 216 259
MRK	-1.7283	3 820 785	-4.2919	9 488 111	-2.4060	5 318 993
MSFT	-2.5715	15 233 489	-4.0619	24 062 626	-3.6716	21 750 487
NKE	-2.0372	2 749 121	-2.0211	2 727 453	-2.0467	2 761 959
PFE	-2.1853	6 235 010	-3.4096	9 728 002	-4.0341	11 509 851
PG	-2.1326	3 592 595	-2.0727	3 491 644	-2.1039	3 544 311
RTX	-1.7441	1 617 736	-2.7315	2 533 649	-2.0640	1 914 459
TRV	-2.6650	1 085 969	-2.5548	1 041 069	-2.6048	1 061 427
UNH	-3.2196	3 531 152	-3.0794	3 377 429	-3.1228	3 425 016
V	-2.5259	4 813 622	-2.4277	4 626 415	-2.4777	4 721 700
VZ	-1.9834	4 558 623	-3.8440	8 834 862	-6.5427	15 037 233
WBA	-2.3269	2 971 343	-3.0885	3 943 860	-3.2108	4 100 059
WMT	-2.0344	3 553 640	-2.0067	3 505 265	-2.0152	3 520 046
XOM	-1.9437	7 947 244	-4.0050	16 375 114	-4.0749	16 661 060

a negative effect on price clustering. The results do not contradict since they explore price clustering from different perspectives. The result based on daily data holds for low-frequency traders whose price resolution is influenced by the uncertainty in a negative way, i.e. the higher daily volatility, the higher price clustering. On the other hand, the presence of high-frequency traders is typically associated with increased volatility (see, e.g. Roşu 2019, Shkilko and Sokolov 2020, and Boehmer *et al.* 2020). Moreover, high-frequency traders generally do not incline to price rounding (see Davis *et al.* (2014)). Consequently, the higher the instantaneous volatility is, the higher portion of high-frequency traders is, which lowers the price clustering.

5. Conclusion

We have proposed a dynamic price model to capture agents trading in different multiples of the tick size. In the literature, this empirical phenomenon known as price clustering was mostly approached only by basic descriptive statistics rather than a proper price model. By analyzing 30 DJIA stocks from both daily and high-frequency perspectives, we have revealed dissension between the two time scales. While daily realized volatility has a positive effect on price clustering, instantaneous volatility obtained by the proposed model has a negative effect. We argue that volatility on lower frequency affects low-frequency traders through the resolution hypothesis while

volatility on higher frequency affects only high-frequency traders who do not tend to price clustering.

Our price model brings several advantages over existing approaches to price clustering since it operates at the highest possible frequency. Consequently, (i) it allows practitioners and academics to study price clustering from a new perspective, and (ii) it can be seen as an extension or a new component for high-frequency price modeling. For example, our price model can be used for simulations of intra-day price processes, which typically serve for optimization of trading strategies, assessment and calibration of models, asset pricing, a decision-making process of investments and hedging strategies, pricing derivatives and other financial instruments. We show that price clustering is not negligible since the excess relative frequency of multiples of five and ten cents can make up to 11.52 % in our dataset. Consequently, neglecting the price clustering phenomenon in the simulations of the price process can have serious consequences for the subsequent usage resulting in unexpected losses caused by, e.g. biased estimates for prices of derivatives, a sub-optimal choice of trading or hedging strategies and models. Our high-frequency price model can also help academics and practitioners to investigate and understand the price clustering phenomenon as such. Finally, high-frequency traders and market makers can directly incorporate our approach into their price models to take into account this phenomenon.

We believe the model to be sufficient for its purpose – capturing price clustering and allowing to explain it. For the

model to be able to compete with other high-frequency price models, however, it would have to be improved. The main limitation lies in the underlying distribution. We have to study how well the double Poisson distribution, which we have used, captures the observed prices in more detail. However, due to our specific problem, we require the distribution to be defined on positive integers and allow for underdispersion. The range of possible alternatives is therefore severely limited as it is not a typical situation in count data analysis. Furthermore, the specification of the dynamics could be enhanced. We could include a separate model for durations and we could add a seasonality component to the volatility. One possible direction for the future research is therefore to comprehensively assess the suitability of the double Poisson distribution for prices, extend the specification of the proposed dynamic model, and compare it with various models for price differences.

Concerning the empirical study, our focus has been on the price variance, whether it is daily realized volatility or instantaneous variance. Nevertheless, we have also included the expected price, the preceding trade duration, and the volume as control variables. These are the most common variables in the price clustering literature. However, other factors such as the spread and the investor sentiment could also be considered. In the context of the proposed high-frequency price model, any variable could be straightforwardly included in the price clustering dynamics. Price clustering at levels other than 5 and 10 cents can be added to the model as well. Analyzing the effects of these factors and price clustering levels is the second possible direction of the future research.

Acknowledgments

Computational resources were supplied by the project ‘e-Infrastruktura CZ’ (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by the Czech Science Foundation under project 19-02773S, the Internal Grant Agency of the Prague University of Economics and Business under project F4/53/2019, and the Institutional Support Funds for the long-term conceptual development of the Faculty of Informatics, Prague University of Economics and Business.

ORCID

Vladimír Holý  <http://orcid.org/0000-0003-0416-0434>
 Petra Tomanová  <http://orcid.org/0000-0001-8072-3193>

References

- Ahn, H.J., Cai, J. and Cheung, Y.L., Price clustering on the limit-order book: Evidence from the stock exchange of Hong Kong. *J. Financ. Markets*, 2005, **8**, 421–451.
- Aitken, M., Brown, P., Buckland, C., Izan, H.Y. and Walter, T., Price clustering on the Australian stock exchange. *Pacific-Basin Finance J.*, 1996, **4**, 297–314.
- Alexander, G.J. and Peterson, M.A., An analysis of trade-size clustering and its relation to stealth trading. *J. Financ. Econ.*, 2007, **84**, 435–471.
- ap Gwilym, O., Clare, A. and Thomas, S., Extreme price clustering in the London equity index futures and options markets. *J. Bank. Finance*, 1998, **22**, 1193–1206.
- ap Gwilym, O. and Verousis, T., Price clustering in individual equity options: Moneyness, maturity, and price level. *J. Futures Markets*, 2013, **33**, 55–76.
- Baig, A., Blau, B.M. and Sabah, N., Price clustering and sentiment in bitcoin. *Finance Res. Lett.*, 2019, **29**, 111–116.
- Ball, C.A., Torous, W.N. and Tschoegl, A.E., The degree of price resolution: The case of the gold market. *J. Futures Markets*, 1985, **5**, 29–43.
- Barndorff-Nielsen, O.E., Pollard, D.G. and Shephard, N., Integer-valued Lévy processes and low latency financial econometrics. *Quant. Finance*, 2012, **12**, 587–605.
- Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N., Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. *Econometrica*, 2008, **76**, 1481–1536.
- Barndorff-Nielsen, O.E., Hansen, P.R., Lunde, A. and Shephard, N., Realized kernels in practice: Trades and quotes. *Econom. J.*, 2009, **119**, 1–32.
- Bharati, R., Crain, S.J. and Kaminski, V., Clustering in crude oil prices and the target pricing zone hypothesis. *Energy Econ.*, 2012, **34**, 1115–1123.
- Blasques, F., Holý, V. and Tomanová, P., Zero-inflated autoregressive conditional duration model for discrete trade durations with excessive zeros, 2020. <https://arxiv.org/abs/1812.07318>.
- Blasques, F., Koopman, S.J. and Lucas, A., Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, 2015, **102**, 325–343.
- Blau, B.M., Price clustering and investor sentiment. *J. Behavioral Finance*, 2019, **20**, 19–30.
- Blau, B.M. and Griffith, T.G., Price clustering and the stability of stock prices. *J. Bus. Res.*, 2016, **69**, 3933–3942.
- Boehmer, E., Fong, K. and Wu, J., Algorithmic trading and market quality: International evidence. *J. Financ. Quant. Anal.*, 2020, **56**, 2659–2688.
- Bollerslev, T., Generalized autoregressive conditional heteroskedasticity. *J. Econom.*, 1986, **31**, 307–327.
- Bollerslev, T., A conditionally heteroskedastic time series model for speculative prices and rates of return. *Rev. Econ. Statist.*, 1987, **69**, 542–547.
- Bourguignon, M., Rodrigues, J. and Santos-Neto, M., Extended Poisson INAR(1) processes with equidispersion, underdispersion and overdispersion. *J. Appl. Stat.*, 2019, **46**, 101–118.
- Box, T. and Griffith, T., Price clustering asymmetries in limit order flows. *Financ. Manage.*, 2016, **45**, 1041–1066.
- Brent, R.P., *Algorithms for Minimization Without Derivatives*, 1972 (Prentice-Hall: Englewood Cliffs).
- Brooks, R., Harris, E. and Joymungul, Y., Price clustering in Australian water markets. *Appl. Econ.*, 2013, **45**, 677–685.
- Brown, P. and Mitchell, J., Culture and stock price clustering: Evidence from The Peoples’ Republic of China. *Pacific-Basin Finance J.*, 2008, **16**, 95–120.
- Buccheri, G., Bormetti, G., Corsi, F. and Lillo, F., A score-driven conditional correlation model for noisy and asynchronous data: An application to high-frequency covariance dynamics. *J. Bus. Econ. Stat.*, 2021, **39**, 920–936.
- Capelle-Blancard, G. and Chaudhury, M., Price clustering in the CAC 40 index options market. *Appl. Financ. Econ.*, 2007, **17**, 1201–1210.

- Chiao, C. and Wang, Z.M., Price clustering: Evidence using comprehensive limit-order data. *Financ. Rev.*, 2009, **44**, 1–29.
- Christie, W.G. and Schultz, P.H., Why do NASDAQ market makers avoid odd-eighth quotes? *J. Finance*, 1994, **49**, 1813–1840.
- Chung, H. and Chiang, S., Price clustering in e-mini and floor-traded index futures. *J. Futures Markets*, 2006, **26**, 269–295.
- Chung, K.H., Kim, K.A. and Kitsabunnarat, P., Liquidity and quote clustering in a market with multiple tick sizes. *J. Financ. Res.*, 2005, **28**, 177–195.
- Chung, K.H., Van Ness, B.F. and Van Ness, R.A., Trading costs and quote clustering on the NYSE and NASDAQ after decimalization. *J. Financ. Res.*, 2004, **27**, 309–328.
- Cooney, J.W., Van Ness, B. and Van Ness, R., Do investors prefer even-eighth prices? Evidence from NYSE limit orders. *J. Bank. Finance*, 2003, **27**, 719–748.
- Creal, D., Koopman, S.J. and Lucas, A., Generalized autoregressive score models with applications. *J. Appl. Econom.*, 2013, **28**, 777–795.
- Das, S. and Kadapakkam, P.R., Machine over mind? Stock price clustering in the era of algorithmic trading. *North Amer. J. Econ. Finance*, 2020, **51**, 100831.
- Davis, R.L., Van Ness, B.F. and Van Ness, R.A., Clustering of trade prices by high-frequency and non-high-frequency trading firms. *Financ. Rev.*, 2014, **49**, 421–433.
- Efron, B., Double exponential families and their use in generalized linear regression. *J. Am. Stat. Assoc.*, 1986, **81**, 709–721.
- Engle, R.F., The econometrics of ultra-high-frequency data. *Econometrica*, 2000, **68**, 1–22.
- Engle, R.F. and Russell, J.R., Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, 1998, **66**, 1127–1162.
- Goodhart, C. and Curcio, R., The clustering of Bid/Ask prices and the spread in the foreign exchange market, 1991. <https://www.fmg.ac.uk/publications/discussion-papers/clustering-bidask-prices-and-spread-foreign-exchange-market>.
- Gorgi, P., Koopman, S.J. and Lit, R., The analysis and forecasting of tennis matches by using a high dimensional dynamic model. *J. R. Statist. Soc.: Series A (Statistics in Society)*, 2019, **182**, 1393–1409.
- Hameed, A. and Terry, E., The effect of tick size on price clustering and trading volume. *J. Bus. Finance. Account.*, 1998, **25**, 849–867.
- Hansen, P.R. and Lunde, A., Realized variance and market microstructure noise. *J. Bus. Econ. Stat.*, 2006, **24**, 127–161.
- Harris, L., Stock price clustering and discreteness. *Rev. Financ. Stud.*, 1991, **4**, 389–415.
- Harvey, A., Hurn, S. and Thiele, S., Modeling directional (circular) time series, 2019. <https://doi.org/10.17863/cam.43915>
- Harvey, A.C., *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*, 1st ed. 2013 (Cambridge University Press: New York).
- Heinen, A., Modelling time series count data: An autoregressive conditional Poisson model, 2003. <https://ssrn.com/abstract=1117187>.
- Holý, V. and Tomanová, P., Streaming approach to quadratic covariation estimation using financial ultra-high-frequency data. *Comput. Econ.*, 2021. <https://doi.org/10.1007/s10614-021-10210-w>.
- Holý, V. and Zouhar, J., Modelling time-varying rankings with autoregressive and score-driven dynamics, 2021. <https://arxiv.org/abs/2101.04040>.
- Hu, B., Jiang, C., McInish, T. and Zhou, H., Price clustering on the shanghai stock exchange. *Appl. Econ.*, 2017, **49**, 2766–2778.
- Hu, B., McInish, T., Miller, J. and Zeng, L., Intraday price behavior of cryptocurrencies. *Finance Res. Lett.*, 2019, **28**, 337–342.
- Ikenberry, D.L. and Weston, J.P., Clustering in US stock prices after decimalisation. *Eur. Financ. Manag.*, 2008, **14**, 30–54.
- Kahn, C., Pennacchi, G. and Sopranzetti, B., Bank deposit rate clustering: Theory and empirical evidence. *J. Finance*, 1999, **54**, 2185–2214.
- Kandel, S., Sarig, O. and Wohl, A., Do investors prefer round stock prices? Evidence from Israeli IPO auctions. *J. Bank. Finance*, 2001, **25**, 1543–1551.
- Koopman, S.J., Lit, R. and Lucas, A., Intraday stochastic volatility in discrete price changes: The dynamic skellam model. *J. Am. Stat. Assoc.*, 2017, **112**, 1490–1503.
- Koopman, S.J., Lit, R., Lucas, A. and Opschoor, A., Dynamic discrete copula models for high-frequency stock price changes. *J. Appl. Econom.*, 2018, **33**, 966–985.
- Li, X., Li, S. and Xu, C., Price clustering in bitcoin market – An extension. *Finance Res. Lett.*, 2020, **32**, 101072.
- Lien, D., Hung, P.H. and Hung, I.C., Order price clustering, size clustering, and stock price movements: Evidence from the Taiwan stock exchange. *J. Empir. Financ.*, 2019, **52**, 149–177.
- Liu, H.C., Timing of price clustering and trader behavior in the foreign exchange market: Evidence from Taiwan. *J. Econ. Finance*, 2011, **35**, 198–210.
- Liu, H.C. and Witte, M.D., Price clustering in the U.S. dollar/Taiwan dollar swap market. *Financ. Rev.*, 2013, **48**, 77–96.
- Mbanga, C.L., The day-of-the-week pattern of price clustering in bitcoin. *Appl. Econ. Lett.*, 2019, **26**, 807–811.
- Meng, L. and Verousis, T., A substitution effect between price clustering and size clustering in credit default swaps. *J. Int. Financ. Markets, Inst. Money*, 2013, **24**, 139–152.
- Mishra, A.K. and Tripathy, T., Price and trade size clustering: Evidence from the national stock exchange of India. *Q. Rev. Econ. Finance*, 2018, **68**, 63–72.
- Narayan, P.K. and Smyth, R., Has political instability contributed to price clustering on Fiji's stock market? *J. Asian. Econ.*, 2013, **28**, 125–130.
- Niederhoffer, V., Clustering of stock prices. *Oper. Res.*, 1965, **13**, 258–265.
- Ohta, W., An analysis of intraday patterns in price clustering on the Tokyo stock exchange. *J. Bank. Finance*, 2006, **30**, 1023–1039.
- Osborne, M.F.M., Periodic structure in the Brownian motion of stock prices. *Oper. Res.*, 1962, **10**, 345–379.
- Palao, F. and Pardo, A., Assessing price clustering in European carbon markets. *Appl. Energy*, 2012, **92**, 51–56.
- Robert, C.Y. and Rosenbaum, M., A new approach for the dynamics of ultra-high-frequency data: The model with uncertainty zones. *J. Financ. Econom.*, 2011, **9**, 344–366.
- Roşu, I., Fast and slow informed trading. *J. Financ. Markets*, 2019, **43**, 1–30.
- Russell, J.R. and Engle, R.F., A discrete-state continuous-time model of financial transactions prices and times: The autoregressive conditional multinomial-autoregressive conditional duration model. *J. Bus. Econ. Stat.*, 2005, **23**, 166–180.
- Schwartz, A.L., Van Ness, B.F. and Van Ness, R.A., Clustering in the futures market: Evidence from S&P 500 futures contracts. *J. Futures Markets*, 2004, **24**, 413–428.
- Sellers, K.F. and Morris, D.S., Underdispersion models: Models that are 'Under the Radar'. *Commun. Statist. – Theory Methods*, 2017, **46**, 12075–12086.
- Shephard, N. and Yang, J.J., Continuous time analysis of fleeting discrete price moves. *J. Am. Stat. Assoc.*, 2017, **112**, 1090–1106.
- Shkilko, A. and Sokolov, K., Every cloud has a silver lining: Fast trading, microwave connectivity, and trading costs. *J. Finance*, 2020, **75**, 2899–2927.
- Song, S., Wang, Y. and Xu, G., On the probability of default in a market with price clustering and jump risk. *Math. Financ. Econ.*, 2020, **14**, 225–247.
- Sopranzetti, B.J. and Datar, V., Price clustering in foreign exchange spot markets. *J. Financ. Markets*, 2002, **5**, 411–417.
- Urquhart, A., Price clustering in bitcoin. *Econ. Lett.*, 2017, **159**, 145–148.
- Verousis, T., Trade size clustering and the cost of trading at the London stock exchange. *Int. Rev. Financ. Anal.*, 2013, **27**, 91–102.
- Xu, H.Y., Xie, M., Goh, T.N. and Fu, X., A model for integer-valued time series with conditional overdispersion. *Comput. Stat. Data. Anal.*, 2012, **56**, 4229–4242.
- Zou, Y., Geedipally, S.R. and Lord, D., Evaluating the double Poisson generalized linear model. *Accident Anal. Prevention*, 2013, **59**, 497–505.

Appendices

Appendix 1. Derivation of distribution for specific trader types

Let there be m types of traders that can trade only in k_1, \dots, k_m multiples of the tick size, respectively. For trader type $k \in \{k_1, \dots, k_m\}$, we derive the distribution of prices $P[Y^{[k]} = y | \mu, \alpha]$. We require the distribution to be based on the double Poisson distribution, to have the support consisting of multiples of k , to have the expected value $E[Y^{[k]}] \simeq \mu$ and to have the variance $\text{var}[Y^{[k]}] \simeq \mu e^{-\alpha}$. We can modify any integer distribution $P[Z^{[k]} = y | \mu, \alpha]$ to have support consisting only of multiples of k as

$$P[Y^{[k]} = y | \mu, \alpha] = \mathbb{I}\{k | y\} P[Z^{[k]} = \frac{y}{k} | \mu, \alpha], \quad (\text{A1})$$

where $\mathbb{I}\{k | y\}$ is equal to 1 if y is divisible by k and 0 otherwise. We assume that $Z^{[k]}$ follows the double Poisson distribution with parameters $\mu^{[k]}$ and $\alpha^{[k]}$, i.e. $Z^{[k]} \sim \text{DP}(\mu^{[k]}, \alpha^{[k]})$. The expected value of $Y^{[k]}$ is

$$\begin{aligned} E[Y^{[k]}] &= \sum_{y=0}^{\infty} y P[Y^{[k]} = y | \mu, \alpha] \\ &= \sum_{y=0}^{\infty} y \mathbb{I}\{k | y\} P[Z^{[k]} = \frac{y}{k} | \mu, \alpha] \\ &= \sum_{y=0}^{\infty} ky P[Z^{[k]} = y | \mu, \alpha] \\ &= k E[Z^{[k]}] \\ &\simeq k \mu^{[k]}. \end{aligned} \quad (\text{A2})$$

The variance of $Y^{[k]}$ is

$$\begin{aligned} \text{var}[Y^{[k]}] &= \sum_{y=0}^{\infty} (y - E[Y^{[k]}])^2 P[Y^{[k]} = y | \mu, \alpha] \\ &= \sum_{y=0}^{\infty} (y - E[Y^{[k]}])^2 \mathbb{I}\{k | y\} P[Z^{[k]} = \frac{y}{k} | \mu, \alpha] \\ &= \sum_{y=0}^{\infty} (ky - kE[Z^{[k]}])^2 P[Z^{[k]} = y | \mu, \alpha] \\ &= k^2 \text{var}[Z^{[k]}] \\ &\simeq k^2 \mu^{[k]} e^{-\alpha^{[k]}}. \end{aligned} \quad (\text{A3})$$

Our last requirements $E[Y^{[k]}] \simeq \mu$ with $\text{var}[Y^{[k]}] \simeq \mu e^{-\alpha}$ lead to the system of equations

$$\begin{aligned} \mu &= k \mu^{[k]} \\ \mu e^{-\alpha} &= k^2 \mu^{[k]} e^{-\alpha^{[k]}} \end{aligned} \quad (\text{A4})$$

with the solution

$$\mu^{[k]} = \frac{\mu}{k}, \quad \alpha^{[k]} = \alpha + \ln(k). \quad (\text{A5})$$

Everything together gives us the distribution

$$\begin{aligned} P[Y^{[k]} = y | \mu, \alpha] &= \mathbb{I}\{k | y\} P[Z^{[k]} = \frac{y}{k} | \mu, \alpha], \\ Z^{[k]} &\sim \text{DP}\left(\frac{\mu}{k}, \alpha + \ln(k)\right). \end{aligned} \quad (\text{A6})$$

Note that the mixture distribution of all prices

$$P[Y = y | \mu, \alpha, \varphi_{k_1}, \dots, \varphi_{k_m}]$$

$$= \sum_{k \in \{k_1, \dots, k_m\}} \varphi_k P[Y^{[k]} = y | \mu, \alpha] \quad (\text{A7})$$

has approximately the same expected value and variance as the distribution of $Y^{[k]}$. This is based on the identity

$$\begin{aligned} E[g(Y)] &= \sum_{y=0}^{\infty} g(y) P[Y = y | \mu, \alpha, \varphi_{k_1}, \dots, \varphi_{k_m}] \\ &= \sum_{y=0}^{\infty} g(y) \sum_{k \in \{k_1, \dots, k_m\}} \varphi_k P[Y^{[k]} = y | \mu, \alpha] \\ &= \sum_{k \in \{k_1, \dots, k_m\}} \varphi_k \sum_{y=0}^{\infty} g(y) P[Y^{[k]} = y | \mu, \alpha] \\ &= \sum_{k \in \{k_1, \dots, k_m\}} \varphi_k E[g(Y^{[k]})] \\ &= E[g(Y^{[k]})], \end{aligned} \quad (\text{A8})$$

where $g(\cdot)$ is any function satisfying that $E[g(Y^{[k]})]$ are the same for all k .

Appendix 2. Descriptive statistics of cleaned data

Descriptive statistics are given in Table A1.

Table A1. The table reports a number of observations (#Trades), sample mean (Mean P) and standard deviation (SD P) of prices, sample mean (Mean D) and standard deviation (SD D) of durations, and price clustering (PC) calculated as the excess relative frequency of multiples of five cents and ten cents in prices.

Stock	#Trades	Mean P	SD P	Mean D	SD D	PC [%]
AAPL	2 671 590	291.71	34.67	1.10	2.71	8.28
AXP	467 623	98.59	17.84	6.24	13.30	4.02
BA	1 886 402	176.06	60.40	1.55	5.01	11.52
CAT	413 148	118.07	14.32	7.08	15.25	3.79
CSCO	1 712 341	42.26	4.14	1.71	5.54	2.08
CVX	964 508	88.34	15.87	3.03	6.42	3.03
DIS	1 327 041	112.56	16.38	2.20	4.67	4.80
DOW	606 116	36.53	8.33	4.82	10.44	2.24
GS	376 058	193.37	31.28	7.78	16.89	3.53
HD	503 522	215.65	28.69	5.80	12.12	3.57
IBM	563 345	124.12	15.30	5.19	10.60	3.34
INTC	2 065 813	57.84	5.53	1.42	4.22	2.13
JNJ	846 988	140.42	9.87	3.45	7.14	2.68
JPM	1 448 425	103.69	17.38	2.02	4.30	3.71
KO	1 140 050	48.35	5.78	2.56	6.63	1.45
MCD	483 508	184.02	20.76	6.05	12.52	3.16
MMM	445 633	150.04	14.01	6.56	13.99	3.51
MRK	1 118 215	79.30	5.40	2.61	5.55	1.94
MSFT	3 099 279	169.25	16.27	0.94	2.40	5.96
NKE	687 619	88.66	11.31	4.25	8.78	2.50
PFE	1 439 433	34.48	3.06	2.03	5.73	1.56
PG	855 186	116.44	6.85	3.41	7.14	2.79
RTX	470 061	62.28	4.87	3.04	5.49	3.37
TRV	216 618	111.75	17.08	13.49	29.25	1.84
UNH	561 256	271.93	26.72	5.21	12.16	3.39
V	965 718	180.20	18.36	3.03	6.07	3.68
VZ	1 162 040	55.61	2.60	2.52	6.15	1.62
WBA	668 123	45.78	4.87	4.38	10.54	2.22
WMT	886 255	118.66	6.14	3.30	6.82	2.85
XOM	2 057 230	45.84	9.29	1.42	3.43	2.25