REM WORKING PAPER SERIES

Nonparametric determinants of market Liquidity

João A. Bastos, Fernando Cascão

REM Working Paper 0332-2024

July 2024

REM – Research in Economics and Mathematics
Rua Miguel Lúpi 20,
1249-078 Lisboa,
Portugal

ISSN 2184-108X

REM – Research in Economics and Mathematics

Rua Miguel Lupi, 20
1249-078 LISBOA
Portugal

Telephone: +351 - 213 925 912
E-mail: rem@iseg.ulisboa.pt

https://rem.rc.iseg.ulisboa.pt/

https://twitter.com/ResearchRem

https://www.linkedin.com/company/researchrem/

https://www.facebook.com/researchrem/

Nonparametric determinants of market liquidity

João A. Bastos Fernando Cascão

Lisbon School of Economics & Management (ISEG)
Universidade de Lisboa

Abstract

We examine the factors influencing equity market liquidity through explainable machine learning techniques. Unlike previous studies, our approach is entirely nonparametric. By studying daily placement orders for equity securities managed by a European asset management institution, we uncover multiple nonlinear relationships between market liquidity and placement characteristics typically not captured by a traditional parametric model. As expected, the results show that liquidity tends to increase in highly active markets. However, we also note that liquidity remains relatively stable within certain trading volume ranges. Price volatility, broker efficiency, and the market impact of the trade are important predictors of liquidity. Price volatility shows a linear relationship with bid-ask spreads, whereas broker efficiency and market impact have non-symmetric convex effects. Large bid-ask spreads are linked to increased uncertainty and weak economic activity.

Keywords: Market liquidity; Equity markets; Bid-ask spreads, Nonparametric models; Machine learning, Explainable AI.

1 Introduction

Equity markets provide a decentralized way of allocating resources. To operate efficiently, buyers and sellers must be able to transact easily, meaning that markets must be liquid. There is no single definition of 'liquid', as it depends on the context in which it is applied. Generally, it refers to the ability to convert an asset into a form of payment. In financial markets, liquidity can be understood as the capacity to trade a specific quantity of securities at the stated price without delay (Amihud et al., 2006). Market liquidity has implications in asset pricing, investment management, corporate finance, banking operations, and the development of monetary and fiscal policies, among other critical areas of

economic activity (Acharya and Pedersen, 2019). The significance of liquidity highlights the importance of understanding its determinants, which remains insufficiently explored in the existing literature (Schwartz and Peng, 2022).

This paper presents a novel approach to understanding the most important factors influencing liquidity in equity markets. Our liquidity measure is the quoted bid-ask spread divided by its midpoint. This metric serves as a natural proxy for liquidity in equity markets (Amihud and Mendelson, 1986; Stoll, 1989). The bid-ask spread is the difference between the highest price at which a prospective buyer is prepared to acquire an asset (known as the bid price) and the lowest price at which a seller is willing to divest that same asset (referred to as the asking price). This differential provides information on the ease of trading and the cost of immediate execution. Securities that are highly liquid and are traded frequently tend to exhibit narrower spreads, whereas those that lack liquidity or are thinly traded often display wider spreads. While the cost of transacting could seem small, the volume of transactions makes the overall economic effect far from negligible. High transaction costs increase the cost of capital for corporations and disrupt the efficiency of portfolio allocation for investors, thus diminishing overall welfare (Biais et al., 2005).

In contrast to prior literature, our approach is entirely nonparametric. This implies we avoid making assumptions regarding the relationship between bid-ask spreads and potential explanatory variables. Simply put, the market data determines the dependence of bid-ask spreads on explanatory variables. This is significant because our results indicate that the relationship between bid-ask spreads and their determinants is far from trivial and can be highly nonlinear. Moreover, any approach relying on linear or polynomial parametric regressions may fail to capture these dependencies. Using daily placement orders for equity securities managed by a European asset management institution, we constructed a novel dataset comprising many potential determinants of bid-ask spreads. These determinants include characteristics of the placement order, characteristics of the securities being traded, and systematic risk variables. To the best of our knowledge, this study presents the most comprehensive evaluation of potential determinants of bid-ask spread in the existing literature.

Our nonparametric model is based on ensembles of decision trees (Friedman, 2001). These predictive models consistently outperform other machine learning models on various problems involving tabular data (Grinsztajn et al., 2022; Curth et al., 2024), and this observation holds true in our dataset. However, these models are regarded as 'black boxes', making it difficult to understand the relationship between the input variables and the model's outputs. To understand the effects of the covariates on bid-ask spreads, we employed two recent frameworks from the realm of eXplainable Artificial Intelligence

(XAI): SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) and Accumulated Local Effects (ALE) (Apley and Zhu, 2020). SHAP values offer insights into the relative importance of the explanatory variables and the average sign of their partial effects. ALE plots further help us discern whether the relationships between bid-ask spreads and the covariates are positive or negative, linear or non-linear, convex or concave, and so on.

Our results show that the nonparametric model captures relationships between bid-ask spreads and the covariates that parametric linear models, commonly used in prior literature, fail to capture. Indeed, many of these relationships are highly nonlinear and non-monotonic. The findings indicate that trading features, including traded volume, broker efficiency, trade impact on the market, and price volatility, are among the most influential determinants of liquidity. However, the shape of the dependence varies for each variable. While volatility exhibits a linear and nearly monotonic impact on bid-ask spreads, broker efficiency, and market impact display a nonsymmetric convex effect. As anticipated, a higher traded volume corresponds to a lower bid-ask spread. This conclusion holds true whether we consider the average volumes of the day, the past 5 days, or the past 21 days before the respective transaction. We note, though, that within specific volume ranges, bid-ask spreads do not change with traded volume. Characteristics of the underlying entity, such as the country of the stock exchange where its equity is traded and its market capitalization, also play a crucial role. We observe significant variations in bid-ask spreads among different stock exchanges. Additionally, transactions involving entities with larger market capitalization tend to exhibit lower spreads.

To incorporate into the model common risk factors for the overall market, often referred to as the systematic risk factors (e.g. Tarashev et al., 2010; Gourieroux and Monfort, 2013), we have included a range of variables that evaluate risk across various regions, primarily focusing on the USA, Europe, and emerging markets. Our results show increased uncertainty and volatility in financial markets lead to higher bid-ask spreads. For instance, when the 10-year bond spread between Germany and Italy exceeds approximately 200 basis points, the bid-ask spread rises relative to the average spread. Moreover, an inverted yield curve of US Treasuries is linked to lower liquidity. We also highlight the significance of aggregated volatility measures, particularly the VIX index, which exhibits a positive and nearly monotonic relationship with bid-ask spreads.

A parametric linear regression model uncovers some effects of the explanatory variables on bid-ask spreads. Yet, it falls short in capturing several interesting relationships, such as the negative impact of daily volume on bid-ask spreads and the nonlinear mechanisms of the broker efficiency, market impact, and US yield curve. In summary, a nonparametric model, coupled with eXplainable Artificial Intelligence (XAI) techniques, has

a greater capacity to unveil patterns in equity market liquidity, with the added advantage of not requiring a predetermined model specification. To the best of our knowledge, this is the first study to apply XAI methodologies to understand bid-ask spreads. This approach can serve as a benchmark for evaluating market liquidity during stressful conditions, a matter of particular interest to regulatory authorities (e.g. ESMA, 2020).

The remainder of the article is structured as follows. Section 2 explores the concept of market liquidity, detailing its multifaceted dimensions and potential determinants. To contextualize our approach, we give some examples of the application of machine learning methodologies in other liquidity domains. Section 3 describes the data under analysis, outlines the determinants considered in our study, and presents the methodologies employed to predict and explain the dynamics of bid-ask spreads. Section 4 presents and discusses the empirical results, and Section 5 concludes.

2 Assessing market liquidity

In practice, there is no single measure of liquidity, even within the same asset class. This is attributed to its multifaceted nature, which can be categorized into five dimensions: tightness, immediacy, depth, breadth, and resilience (Sarr and Lybek, 2002; Bervas, 2006). First, 'tightness' addresses the transaction costs market participants incur when buying or selling a particular asset. A market is considered tight when these transaction costs are notably high. Second, 'immediacy' concerns the speed at which orders can be executed. In markets characterized by high immediacy, transactions among market participants occur rapidly, with minimal delays. The third dimension, 'depth', refers to the number of buying and selling orders clustered around equilibrium prices. A deep market has a robust pool of orders, making it less susceptible to rapid price fluctuations. The dimension 'breadth' assesses the diversity and volume of orders at different price levels. A market is broad when it has several buying and selling orders with significant trading volumes. Finally, 'resilience' characterizes the market's capacity to absorb and recover from unforeseen shocks or disruptions. Markets with resilience are equipped with a surplus of orders capable of rectifying imbalances swiftly. A thorough examination of metrics corresponding to each of these dimensions is provided by Díaz and Escribano (2020).

In equity markets, the bid-ask spread is a standard measure of liquidity (Amihud and Mendelson, 1986; Stoll, 1989). This spread represents the difference between the highest price a potential buyer is willing to pay for an asset and the lowest price at which a seller is prepared to divest it. It quantifies the level of illiquidity in a market concerning transaction and participation costs, representing the 'tightness' aspect of liquidity. A

noteworthy perspective on the spread is its portrayal as a round-trip cost, as half of its value is typically considered the cost incurred when executing the purchase or sale of securities immediately (Schwartz and Peng, 2022).

Some general characteristics of trading are recognized as relevant for assessing market liquidity. Schwartz and Peng (2022) underscore the importance of market depth, emphasizing that 'spreads are wider, market impact greater, and price discovery less accurate for thinner issues'. Furthermore, these authors consider that the structural mechanisms of the market, including the roles played by brokers and market makers, have a profound influence on the liquidity dynamics of individual securities. Also, upon this perspective, Schwartz et al. (2020) underscore that finding liquidity in the marketplace is a permanent challenge that 'depends not only on a trader's skill, but also on (1) the structure of the market within which participants are operating, (2) the regulatory environment, and (3) the traded stock itself'. Some authors have found a strong relationship between stock market liquidity and the business cycle (Næs et al., 2011; Lu-Andrews and Glascock, 2010). As Chordia et al. (2001) points out, if macroeconomic variables anticipate economic downturns, they might also anticipate liquidity levels and trading activity in equity markets.

When regressing market liquidity with explanatory variables, the methods considered have been linear regressions (Chordia et al., 2001; Brennan et al., 2012) and quantile regressions (Chuliá et al., 2023). Focusing on the tails of market liquidity, Chuliá et al. (2023) find nonlinear associations between liquidity measures and their determinants. These findings suggest using nonparametric techniques to uncover these nonlinear dynamics.

To the best of our knowledge, there is still no application of nonparametric methods in the context of market liquidity in the existing literature. However, there have been notable instances in other liquidity-related domains. Guerra et al. (2022) address liquidity from the standpoint of central bank supervision, developing a multiple-class liquidity scoring system based on banks' risk levels. Notably, they highlight the superior performance of machine learning techniques over logistic regression. Furthermore, Tavana et al. (2018) assess the liquidity risk of a US bank using an approach that integrates two distinct machine learning techniques: artificial neural networks and Bayesian networks. Neural networks are employed to approximate the bank's liquidity risk function, while Bayesian networks are used to identify the most influential factors in the model. These early studies underscore the potential of machine learning techniques for liquidity analysis and risk assessment.

3 Data and Methodology

3.1 Data description

We analyze daily placement orders for equity securities managed by a European asset management institution. The set of features of each placement order was extracted via Bloomberg and covers the period from 2017 through the first half of 2023. The asset management institution trades securities on several stock exchanges, primarily located in the USA and Europe. The raw data had 26,191 observations. To ensure data integrity and statistical robustness, we excluded anomalous observations. First, we observed that 12.8% of the placement orders had bid-ask spreads of zero, which were omitted from the analysis. We noticed many orders exhibiting unusually high bid-ask spreads, lacking any evident economic rationale. Following consultation with a director of the asset management institution, we opted to exclude the top 2.5% of observations featuring the highest bid-ask spreads. After these exclusions, we retained a total of 22,262 daily placement orders. We use the data from 2017 to 2022 to train and validate the statistical models. The first semester of 2023 data is withheld from the main analysis and exclusively reserved for an out-of-time testing exercise.

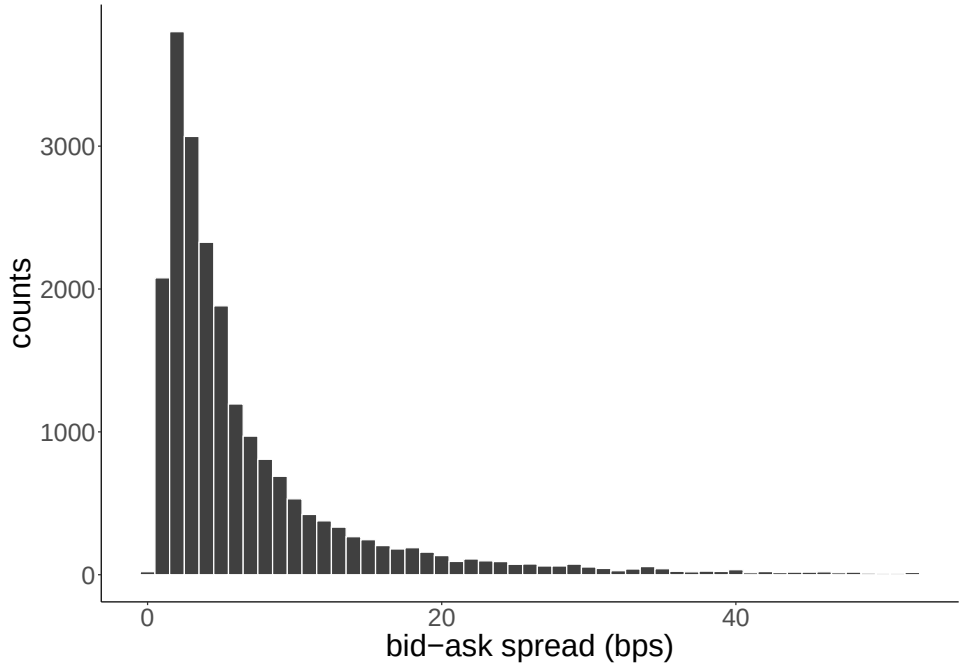

Figure 1: Distribution of bid-ask spread (in basis points) for the daily placement orders.

The target variable in our analysis is derived from the quoted bid-ask spread at the moment of the placement's creation. We model the difference between ask and bid prices

as a proportion of its midpoint,

$$\text{bid-ask spread} = \frac{\text{ask price} - \text{bid price}}{\frac{\text{ask price}+\text{bid price}}{2}} \times 10000. \tag{1}$$

This gives us a *relative* measure of bid-ask price differences expressed in basis points (bps). Henceforth, whenever we mention bid-ask spreads, we refer to the metric defined in Equation 1.

Year	**No. Obs.**	**Mean**	**Median**	$q_{0.05}$	$q_{0.95}$
2017	3330	5.3	3.5	1.1	16.8
2018	6430	6.2	4.1	1.3	18.7
2019	3602	6.5	4.0	1.2	19.4
2020	2437	9.0	5.1	1.3	29.6
2021	3441	7.3	4.0	1.1	25.1
2022	1952	9.1	6.1	1.2	27.6
2023	1070	7.9	5.8	1.3	22.7
All	22262	6.9	4.2	1.2	22.9

Table 1: Summary statistics of the bid-ask spreads by year: number of observations, the sample mean, sample median, and the 0.05 and 0.95 empirical quantiles. The 2023 figures pertain to the first semester.

Figure 1 shows the distribution of bid-ask spreads in basis points for our dataset of daily placement orders. This measure is characterized by positive skewness. Table 1 reports summary statistics for the bid-ask spreads. The year 2018 had the highest number of transactions. The mean bid-ask spreads increased during the 2020-2022 period. This can be attributed to the market turmoil during the COVID-19 pandemic. Indeed, the 95% quantile indicates a larger occurrence of extreme bid-ask spreads during this period.

Regarding potential determinants of bid-ask spreads, the dataset includes 27 variables specific to the order placements, 22 of which are numerical and 5 categorical. Additionally, we collected 11 variables that aim to measure systematic risk across the world's main markets, including the USA, Europe, and emerging markets. These variables enhance the model's ability to capture the impact of macroeconomic and systematic factors on market liquidity. We chose explanatory variables based on previous literature and intuitive reasoning concerning factors that could influence the bid-ask spreads.

Table 2 describes the numeric trading features present in the dataset. Several variables are associated with the traded volume at various lags. As highlighted in Biais et al. (2005), the trading volume makes the impact of transaction costs non-trivial. Furthermore, a positive relation between volume and liquidity has been suggested in Chordia et al. (2001). Consequently, we anticipate a negative relationship between traded volume and the bid-ask spread. Another crucial factor to consider is the price volatility of the

Variable	Description
aep_vwap_bp	Difference between average execution price and volume-weighted average price from order arrival until last fill, adjusted for limit price.
arrival_bp	Difference between the average execution price and the mid-price when the order was received.
avg_exe_px	Average execution price for the placement.
day_cl_px	Closing price on the day the order is completed.
day_op_px	Opening price on the day the order is received.
day_rev_bp	Difference between placement's last fill price and the closing price on the day the placement is completed.
day_volume	Volume of the security on the day of the placement.
leak_bp	Difference between the placement's first fill price and the placement's arrival price.
month_cl_px	Closing price one month before the order is received.
mkt_imp	Bloomberg's estimated market impact for the placement.
order_momentum	Percentage change in the price of the underlying during the order interval (order arrival to last fill) relative to the side.
percent_21adv	Placement size / 21-day average daily volume (up to, but not including the placement date).
post_momentum	Percentage change between open price and close price the day after the order is completed.
px_daily_range	Relative difference between high and low prices on the day the order is received.
value	Value of executed shares.
vwap	Volume-weighted average price from the placement arrival until that placement's last fill.
vwap_5min	Volume-weighted average price calculated over five minutes starting from that placement's last fill.
week_cl_px	Closing price one week before the order is received.
10day_abs_vol	Last 10 days of absolute security volatility.
10day_rel_vol	Last 10 days of the relative security volatility.
5day_volume	Average daily volume for the 5 days before placement arrival date.
21day_volume	Average daily volume for the 21 days before placement arrival date.

Table 2: Description of numeric trading features used in the analysis.

securities. Increased market volatility might adversely affect trading activity, potentially undermining liquidity (Chuliá et al., 2023). We also examine the 'market impact' of the trade (variable 'mkt_imp') estimated by Bloomberg. This is computed using a proprietary model that Bloomberg keeps confidential. A significant market impact can lead to limited market depth and breadth, resulting in wider bid-ask spreads (Schwartz and Peng, 2022). Several variables representing the prices of the securities at different moments in time were also included in the analysis. The stock price dynamics may affect the expectations of market participants, implying changes in demand for assets (Chuliá et al., 2023).

Variable	Description
country	Country of the exchange where the security is traded.
market_cap	Market capitalization of the underlying company (large, mid, small, or micro).
news_heat	Average value of news publication heat for the underlying company within the order interval (0 to 4).
sector	Sector of the security being traded.
side	Side of the order (buy or sell).

Table 3: Description of categorical variables used in the analysis.

Schwartz and Peng (2022) emphasize that finding liquidity in the marketplace remains an ongoing challenge, influenced by the trader's skill, among other factors. Hence, we also considered variables measuring the efficiency of the broker/trade, such as the variable 'leak bp' – the difference between the placement's first fill price and the placement's arrival price.

Table 3 shows the categorical trading features used in the analysis. Five primary attributes were collected: (i) order side (buy or sell), (ii) country of the stock exchange where the security is traded, (iii) entity's sector of the underlying security, (iv) entity's market capitalization, and (v) a variable indicating the extent of the entity's news publication activity. These variables capture the side of the order and important elements about the traded entity and the exchange where the security is listed, providing valuable context for assessing bid-ask spreads. Brennan et al. (2012) find that buy and sell orders impact illiquidity differently – market-wide sell illiquidity is generally greater than buy illiquidity. The same authors suggest that liquidity problems are more prominent for smaller stocks.

Schwartz and Peng (2022) also suggests that markets can be thinner for small-cap and mid-cap stocks because, at any given moment, only a few individuals (if any) may actively seek to buy or sell shares. The variable representing the average heat level of the entity's news publication within the order interval refers to the amount of unexpected publication activity compared to the previous 45 days. Aman and Moriyasu (2022) evaluate the impact of two major information sources – corporate disclosure and press media – on market liquidity spreads. Their analysis suggests that bid-ask spreads tend to widen upon corporate news disclosure. Conversely, greater press media coverage is linked to a narrower spread. This variable consolidates all publications, serving as a proxy for the market's interest in the underlying entity.

Table 4 shows the systematic risk variables used in the analysis. We examined three market volatility metrics: the CBOE options implied volatility index (VIX index), commonly employed in this context (Brennan et al., 2012; Chuliá et al., 2023), along with its counterparts for the Eurozone (VSTOXX index) and emerging markets (VXEEM index).

Variable	Description
cds_brazil	Credit Default Swap for Brazil.
cds_china	Credit Default Swap for China.
cds_turkey	Credit Default Swap for Turkey.
cds_uk	Credit Default Swap for the United Kingdom.
spread_ge_fr_10yr	10-year sovereign bonds spread between Germany and France.
spread_ge_it_10yr	10-year sovereign bonds spread between Germany and Italy.
spread_ge_pt_10yr	10-year sovereign bonds spread between Germany and Portugal.
spread_2_10yr_us	US Treasury yield curve (10-year yield minus 2-year yield).
vol_eu	Euro STOXX 50 volatility index VSTOXX.
vol_us	CBOE volatility index (VIX).
vol_emerging	CBOE emerging markets ETF volatility index.

Table 4: Description of systematic risk variables used in the analysis.

Higher aggregated market volatility has been linked to greater liquidity demand (Kim et al., 2023).

If macroeconomic variables anticipate economic downturns, they may also predict reduced trading activity and liquidity in equity markets (Chordia et al., 2001). Some authors find a strong relation between stock market liquidity and the business cycle (Lu-Andrews and Glascock, 2010; Næs et al., 2011). During economic downturns, investors typically demand a higher price for liquidity. Consequently, the illiquidity premium tends to increase when economic activity is low. We excluded macroeconomic variables, such as GDP growth, unemployment rate, consumption, or investment, from the analysis since they are only available at low frequencies. Instead, we used 10-year European sovereign spreads and the US Treasury yield curve as proxies for the macroeconomic and financial landscape during each placement. In the absence of a common European debt market, we selected the 10-year yield spreads between Italy's, France's, and Portugal's sovereign debt and Germany's debt. Portugal was included due to the significant number of securities from the Portuguese stock exchange in the dataset. Germany was the benchmark for comparing European sovereign spreads due to its consistent financial stability and creditworthiness.

As mentioned in Chordia et al. (2001), an increase in default risk might increase the perceived risk of holding inventory, subsequently reducing liquidity. Consequently, our analysis incorporates the credit default swap (CDS) prices of four countries in different regions: Turkey (MENA), China (East Asia), Brazil (South America), and the United Kingdom (Europe).

The relationship between bid-ask spreads and the systematic risk variables is expected to be nonlinear due to the asymmetric response of liquidity to market movements. Indeed, lack of liquidity tends to increase more severely during market declines than it decreases during market upswings (Chordia et al., 2001; Chuliá et al., 2023).

3.2 Predictive models

Our predictive models are well-known in the finance literature and include a linear model with and without regularization penalty, ensembles of decision trees, and feedforward neural networks. Below is a concise summary of these models; please refer to James et al. (2023) for a more comprehensive coverage.

3.2.1 Linear models

Let Y denote the observed bid-ask spread and $\boldsymbol{X}$ denote the characteristics of the placement order, which are listed in Tables 2, 3 and 4. Let n denote the number of observations in the training data. The simplest approach for modeling the bid-ask spreads is to assume a linear relationship with the placement order characteristics,

$$f(\boldsymbol{X}; w_0, \boldsymbol{w}) = w_0 + \boldsymbol{w} \cdot \boldsymbol{X}, \tag{2}$$

where $(w_0, \boldsymbol{w})$ is a vector of coefficients. These coefficients are obtained by minimization of the loss function

$$J(w_0, \boldsymbol{w}) = \sum_{i=1}^{n} (Y - w_0 - \boldsymbol{w} \cdot \boldsymbol{X})^2 + \lambda ||\boldsymbol{w}||_1, \tag{3}$$

where $\lambda \geq 0$ is a constant and $|| \cdot ||_1$ denotes the L1 norm. When $\lambda = 0$ we have the workhorse of applied econometrics: the linear regression model estimated via ordinary least squares. When $\lambda > 0$ we obtain the LASSO (*Least Absolute Shrinkage and Selection Operator*) regression. LASSO introduces regularization by adding an L1 penalty on the coefficients' magnitude to the cost function. The penalty parameter λ determines the extent of regularization. For sufficiently large λ values, LASSO regression eliminates less influential explanatory variables by setting their coefficients to zero.

3.2.2 Tree-based ensembles

A decision tree is a nonparametric model that divides the regressor space into mutually exclusive regions $\{R_m\}_{m=1}^{M}$. The region R_m where an observation falls depends on a sequence of if-then-else tests performed on its regressor values, $\boldsymbol{X}$. For example, a sequence of tests can be something like $X_3 < c$ and $X_1 > c'$ and $X_4 > c''$. Consequently, region R_m is defined as:

$$R_m = \{\boldsymbol{X} : (X_3 < c) \wedge (X_1 > c') \wedge (X_4 > c'')\}. \tag{4}$$

Formally, a decision tree model is represented by the equation:

$$f(\boldsymbol{X}; \boldsymbol{w}) = \sum_{m=1}^{M} w_m \cdot \mathrm{I}(\boldsymbol{X} \in R_m), \tag{5}$$

where $\mathrm{I}(\boldsymbol{X} \in R_m)$ is an indicator function that evaluates to 1 if its argument is true and 0 otherwise. The weight w_m represents the output of the model when an observation falls into region R_m. In its simplest form, w_m is calculated as the average Y-value of all observations within the training data that belong to that specific region:

$$w_m = \frac{\sum_{i=1}^{n} Y_i \cdot \mathrm{I}(\boldsymbol{X}_i \in R_m)}{\sum_{i=1}^{n} \mathrm{I}(\boldsymbol{X}_i \in R_m)} \tag{6}$$

Decision trees are not particularly accurate models on their own, yet they serve as the building blocks of powerful models based on ensembles of decision trees. These ensembles leverage the collective predictions of multiple trees to enhance predictive power.

Random forests (Breiman, 2001) is a simple yet powerful method for combining multiple individual decision trees. The process starts by creating a specified number of bootstrap samples from the data, each containing the same number of observations as the original dataset. Then, a decision tree is constructed for each of these bootstrap samples. Only a random subset of explanatory variables is considered when dividing the data into subsets for increased diversity among decision trees. Let's suppose that we have generated B bootstrap samples and let $f_b(\boldsymbol{X}; \boldsymbol{w})$ denote a decision tree trained on a specific bootstrap sample. The prediction provided by a random forest is simply the average of the individual predictions given by the B trees:

$$f(\boldsymbol{X}; \boldsymbol{w}, B) = \frac{1}{B} \sum_{b=1}^{B} f_b(\boldsymbol{X}; \boldsymbol{w}) \tag{7}$$

An alternative strategy for building ensembles is the 'gradient boosting machine' (Friedman, 2001). The predictions given by the gradient boosting machine involve summing the predictions of a collection of B decision trees $\{f_b(\boldsymbol{X}; \boldsymbol{w})\}_{b=1}^{B}$:

$$f(\boldsymbol{X}; \boldsymbol{w}, B) = \sum_{b=1}^{B} f_b(\boldsymbol{X}; \boldsymbol{w}). \tag{8}$$

The initial tree, $f_1(\boldsymbol{X}; \boldsymbol{w})$, is a standard decision tree trained on the original data. The subsequent decision trees, $\{f_b(\boldsymbol{X}; \boldsymbol{w})\}_{b=2}^{B}$, are incrementally added to the committee. However, each new tree is trained on the errors produced by the trees already present in the committee. This process aims to rectify the errors made by the existing set of trees. During each iteration, the tree to be added is the one that minimizes the regularized loss function:

$$\sum_{i=1}^{n} L\left(Y_i, \hat{Y}_i^{(b-1)} + f_b(\boldsymbol{X}_i; \boldsymbol{w})\right) + \gamma T + \frac{1}{2}\lambda ||\boldsymbol{w}_b||^2, \tag{9}$$

where $L(\cdot)$ is the squared-error loss:

$$L\left(Y_i, \hat{Y}_i^{(b-1)} + f_b(\boldsymbol{X}_i; \boldsymbol{w})\right) = \left(Y_i - \hat{Y}_i^{(b-1)} - f_b(\boldsymbol{X}_i)\right)^2. \tag{10}$$

The last two terms in Equation (9) are regularization terms that penalize complex trees, thereby preventing the committee from overfitting the training data. The parameter γ is a penalty on the number of terminal nodes in a tree, denoted by T, while λ is a penalty on the magnitude of the tree weights $\boldsymbol{w}_k$. A gradient descent algorithm minimizes the loss function when adding new trees. There are several efficient implementations of gradient boosting. This study uses the most popular of these implementations: the 'Extreme Gradient Boosting' algorithm (XGBoost) (Chen and Guestrin, 2016).

3.2.3 Feedforward neural network

The feedforward neural network (FNN) is a classical neural network architecture. This model is characterized by the unidirectional flow of information from the inputs through intermediate hidden layers ending at the output layer. Each layer is composed of a set of computational units. Each unit is a function of a linear combination of the outputs from the units preceding it. In particular, the units in the first hidden layer compute a function of a linear combination of the input data. For example, the output of the j-th unit in the first hidden layer is given by:

$$h_j^{(1)} = g^{(1)}\left(w_{0j}^{(1)} + \boldsymbol{w}_j^{(1)} \cdot \boldsymbol{X}\right), \tag{11}$$

where $g^{(1)}$ is the *activation function* for the first hidden layer. Likewise, the output of the j-th unit in the k-th hidden layer is:

$$h_j^{(k)} = g^{(k)}\left(w_{0j}^{(k)} + \boldsymbol{w}_j^{(k)} \cdot \boldsymbol{h}^{(k-1)}\right), \tag{12}$$

Finally, the output of the model computes a function of a linear combination of the values provided by the last hidden layer,

$$Y = g^{(o)}\left(w_0^{(o)} + \boldsymbol{w}^{(o)} \cdot \boldsymbol{h}^{(K)}\right). \tag{13}$$

The activation function for the hidden layers $\{g^{(k)}\}_{k=1}^{K}$ is the *rectified linear function* (or ReLU):

$$\text{ReLU}(z) = \max(0, z) \tag{14}$$

In regression problems where the dependent variable is numeric, the output unit typically computes a linear combination of the values provided by the last hidden layer; that is, $g^{(o)}$ is the identity function. However, bid-ask spreads are strictly positive. Therefore, we also used a ReLU function for $g^{(o)}$, ensuring that the model output matches the range of bid-ask spreads.

The model weights, $\boldsymbol{w}$, are estimated by the backpropagation algorithm (Rumelhart et al., 1986). This process involves multiple steps, known as epochs, aiming to minimize a squared loss function. The minimization occurs through the gradient descent algorithm known as Adam (Kingma and Ba, 2014).

3.3 Model optimization and validation

The models described above include parameters known as 'hyperparameters' that are not learned during training. Therefore, we must identify the hyperparameters that optimize the performance of each method. Specifically, the LASSO regression has a single hyperparameter: the penalty parameter λ. To optimize the random forest models, we must determine the ideal number of trees within the ensemble and the parameters that govern the complexity of the individual trees. To optimize the gradient boosting machine, we further identify an appropriate learning rate for the gradient descent algorithm that minimizes the loss function. For the neural network, we must choose the number of hidden layers (typically 1 or 2) and the number of units within each layer. We also need to choose the number of training epochs and the learning rate used by the gradient descent algorithm.

We conduct a grid search over all possible hyperparameter combinations to obtain the set with the lowest expected out-of-sample error. We implemented a 5-fold cross-validation on the data from 2017 to 2022 to estimate the out-of-sample error. We start by randomly dividing the data into five distinct folds. Then, we iterate through the five subsets, treating one as the validation set while using the other four for estimation. This is repeated until each subset has served as the validation set. The best hyperparameters are the one that gives the smallest average root mean square error (RMSE) across the five folds:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}, \tag{15}$$

where $\hat{Y}$ denotes the model predictions.

3.4 Explanation methodologies: SHAP and ALE

Simpler models, such as linear models, offer the advantage of being straightforward to interpret and comprehend. They provide well-defined statistical tools to extract conclusions from the model. However, the simplicity of these models can be a limitation when the true relationship between the explanatory variables and the target variable is complex or highly nonlinear. In such cases, simpler models may yield inaccurate results as they struggle to capture complex patterns in the data. On the other hand, complex models – such as ensemble methods or neural networks – are 'black boxes' that do not provide a straightforward path to understanding how they arrive at their predictions. Several methodologies have recently been proposed to address this issue. These techniques were designed to extract and quantify the influence of the model's inputs on its outputs, providing insights into the factors driving the predictions.

One of these methodologies is SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017). SHAP is based on the concept of Shapley values proposed in the context of cooperative game theory (Shapley, 1953). Shapley values define a way to distribute the final payoff among individuals based on their contributions to the collective effort. This player-payoff assignment has four essential properties: (i) efficiency, (ii) symmetry, (iii) dummy player, and (iv) additivity (see, e.g. Roth, 1988; Aas et al., 2021). Let $\boldsymbol{X}_{\setminus j}$ denote the subset of $\boldsymbol{X}$ that excludes regressor X_j, that is $\boldsymbol{X}_{\setminus j} \equiv \boldsymbol{X} \setminus X_j$, and let S denote all possible subsets of $\boldsymbol{X}_{\setminus j}$. For instance, if we have $p = 3$ regressors, $\boldsymbol{X} = \{X_1, X_2, X_3\}$, and we exclude regressor X_1 from $\boldsymbol{X}$, then $S = \{\emptyset, X_2, X_3, \{X_2, X_3\}\}$. The SHAP value for X_j is a weighted sum of its marginal contribution to a prediction over all possible coalitions that exclude it:

$$\phi(X_j) = \sum_{S \subseteq \boldsymbol{X}_{\setminus j}} \frac{|S|!\,(p-1-|S|)!}{p!} \left[f_{S \cup X_j}(S \cup X_j) - f_S(S) \right]. \tag{16}$$

Each observation and explanatory variable is assigned a SHAP value. However, calculating these values can be computationally expensive, as it requires obtaining the model's predictions for every possible combination of variables. To address this issue, Lundberg et al. (2018) introduced an algorithm for obtaining SHAP values when the model is based on decision trees. This algorithm reduces the computational complexity from exponential to polynomial time, ensuring the efficient computation of SHAP values, even for complex ensembles with many decision trees and explanatory variables.

While SHAP values provide information on the relative importance of the explanatory variables and the sign of the partial effects, accumulated local effects (ALE) (Apley and Zhu, 2020) allow us to understand whether the relationship between bid-ask spreads and the covariates are positive or negative, linear or non-linear, convex or concave, and so on. For example, ALE plots show linear dependence if the true relationship is linear. ALE plots are the state-of-the-art approach for visualizing relationships between a target variable and the regressors when the regressors are not independent, which is the case here.

Suppose we want to understand how regressor X_j affects the model's output. First, we divide its range using a grid with K bins. Let $\{Z_k\}_{k=0}^{K}$ denote the set of X_j values that define the boundaries of these bins. For instance, the first bin encompasses all X_j values between Z_0 and Z_1, while the last bin comprises all X_j values between Z_{K-1} and Z_K. Typically, the Z_k values are selected as the (k/K)-quantiles of the empirical distribution of X_j, where Z_0 is chosen slightly below the smallest observation, and Z_K equals the largest observation.

Let $\mathcal{I}_k$ denote the set of indices corresponding to the observations where $Z_{k-1} < X_j \leq Z_k$, with $k > 0$. Each bin's observation count is n_k. The average *local* effect of X_j within

a specific bin is given by:

$$\frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \left[f(Z_{k,j}, \boldsymbol{X}_{i,\setminus j}) - f(Z_{k-1,j}, \boldsymbol{X}_{i,\setminus j}) \right]. \tag{17}$$

The sum in Equation 17 loops over all observations in a given bin. For each of these observations, we obtain the difference between the model predictions with X_j equal to the upper limit of the bin, Z_k, and X_j equal to the lower limit of the bin, Z_{k-1}. We divide this sum by the number of observations in that bin, n_k, to obtain the average local effect of X_j on the model's output.

Now, let $k(X_j)$ represent the index of the bin where a specific value of X_j is located, with $k(X_j) = 1$ if X_j falls within the first bin, $k(X_j) = 2$ if it lies within the second bin, and so on. The 'accumulated local effect' at value X_j is simply the sum of the local effects from the first bin up to the bin where X_j is located:

$$f_A(X_j) = \sum_{k=1}^{k(X_j)} \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \left[f(Z_{k,j}, \boldsymbol{X}_{i,\setminus j}) - f(Z_{k-1,j}, \boldsymbol{X}_{i,\setminus j}) \right]. \tag{18}$$

This sum accumulates the local average effects up to a given value of X_j. The plot of $f_A(X_j)$ as a function of X_j provides a visualization of the dependence of the X_j across its range. It should be noted that the values $f_A(X_j)$ initiate at zero and subsequently ascend or descend based on the sign of the average local effects. In other words, $f_A(X_j)$ estimates the dependency of the model's output on X_j apart from a constant factor. Therefore, Apley and Zhu (2020) suggest subtracting the average ALE in all bins to center the ALE plot at zero,

$$f_A(X_j) \leftarrow f_A(X_j) - \frac{1}{n} \sum_{k=1}^{K} n_k \times f_A(Z_{k,j}). \tag{19}$$

4 Empirical results

4.1 Prediction accuracy

The five models under consideration vary in degrees of freedom and capacity to overfit the training data. Therefore, evaluating the model's performance requires using data not seen during their training. Our initial focus is on the data between 2017 and 2022, aiming to determine the out-of-sample accuracy of the observations during this period. This is accomplished through a 5-fold cross-validation, as detailed in Section 3.3. Furthermore, we use the first semester of 2023 to obtain *out-of-time* accuracy measures. To achieve this, we trained the models using data from 2017 to 2022 and assessed their performance using data from 2023.

Model	Out-of-sample			Out-of-time		
	RMSE	R^2	ρ	RMSE	R^2	ρ
OLS regression	6.47	27.88	52.79	6.65	21.68	46.56
LASSO regression	6.47	27.95	52.86	6.66	21.62	46.50
Random forest	4.24	69.51	83.37	5.47	49.58	70.41
Gradient boosting machine	4.06	71.59	84.60	5.65	50.00	70.71
Feedforward neural network	4.82	60.13	77.53	7.15	21.85	46.74

Table 5: Out-of-sample accuracy in 2017-2022, and out-of-time accuracy in 2023. Out-of-sample accuracy is obtained from a 5-fold cross-validation procedure. The accuracy metrics are the root mean square error (RMSE), the R^2, and Pearson correlation coefficient (ρ).

In addition to the RMSE used for optimizing the hyperparameters of the models, we have also calculated an R^2 metric that compares the sum of squared residuals of the trained models with that of a naïve model that always predicts the average outcome $\bar{Y}$:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{20}$$

Furthermore, we computed the Pearson correlation between the predicted and actual values.

$$\rho = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}} \tag{21}$$

Table 5 reports the out-of-sample accuracy measures in the period 2017–2022 and the out-of-time accuracy in 2023. The two tree-based models better predict bid-ask spreads in the cross-validation setting. In particular, these models exhibit a notable prediction accuracy when compared to both the linear model estimated through ordinary least squares and the linear model estimated using the LASSO penalized loss function[1]. This highlights the importance of modeling bid-ask spreads with non-linear models, in contrast with the practice in the literature. Additionally, the tree-based models outperformed the neural networks. This is also unsurprising since tree-based models outperform neural networks on many problems with tabular data (Grinsztajn et al., 2022; Curth et al., 2024). In the out-of-time validation exercise, the relative rank of the models remains. We also note that the neural network does not generalize well beyond the training set, as its out-of-time performance in the 2023 test sample drops significantly, reaching levels similar to the linear models. Most metrics indicate that the gradient boosting machine outperforms the random forest in terms of predictive accuracy, so the former was selected to derive the nonparametric determinants of bid-ask spreads.

[1]The performance of the OLS regression did not change substantially when we considered the logarithm of the bid-ask spread as the dependent variable, giving an R^2 of 28.51 and 23.78 in the cross-validation and out-of-time settings, respectively.

4.2 Parametric determinants

We start by analyzing the determinants of bid-ask spreads given by a linear regression model estimated by least squares. Table 6 presents the estimated model.

Variable	**Coefficient**	**std. err.**	**t-stat**	**p-value**
Intercept	16.484	1.012	16.283	0.000
10day_abs_vol	0.218	0.044	5.001	0.000
10day_rel_vol	0.097	0.031	3.093	0.002
5day_volume	-1.4×10^{-9}	6.8×10^{-9}	-0.201	0.841
21day_volume	-3.9×10^{-8}	6.9×10^{-9}	-5.559	0.000
aep_vwap_bp	-0.007	0.002	-4.508	0.000
arrival_bp	0.005	0.002	2.238	0.025
avg_exe_px	-0.001	0.010	-0.144	0.885
cds_brazil	0.006	0.002	3.773	0.000
cds_china	-0.007	0.006	-1.266	0.206
cds_turkey	4.4×10^{-4}	7.7×10^{-4}	0.567	0.571
cds_uk	0.047	0.007	6.442	0.000
country_belgium	-3.276	0.390	-8.394	0.000
country_denmark	-4.773	0.391	-12.201	0.000
country_france	-4.061	0.144	-28.129	0.000
country_germany	-3.892	0.179	-21.704	0.000
country_great_britain	-3.691	0.175	-21.143	0.000
country_italy	-1.006	0.285	-3.527	0.000
country_netherlands	-3.813	0.206	-18.519	0.000
country_portugal	7.083	0.281	25.191	0.000
country_spain	0.380	0.267	1.425	0.154
country_sweden	-4.158	0.306	-13.610	0.000
country_switzerland	-4.038	0.249	-16.184	0.000
country_other	3.808	0.367	10.366	0.000
day_cl_px	9.1×10^{-5}	5.5×10^{-4}	0.168	0.867
day_op_px	6.6×10^{-5}	6.5×10^{-4}	0.102	0.919
day_rev_bp	1.5×10^{-3}	3.5×10^{-4}	4.278	0.000
day_volume	5.6×10^{-9}	1.4×10^{-9}	1.892	0.058
leak_bp	-0.002	0.001	-1.629	0.103
market_cap_large	-12.245	0.963	-12.718	0.000
market_cap_mid	-9.413	0.956	-9.850	0.000
market_cap_small	-7.295	1.106	-6.599	0.000
mkt_imp	-1.169	0.331	-3.531	0.000
month_cl_px	3.3×10^{-5}	3.3×10^{-4}	0.237	0.813
news_heat	-0.508	0.041	-12.252	0.000
order_momentum	-0.299	0.186	-1.610	0.107
percent_21adv	0.491	0.054	9.068	0.000
post_momentum	0.042	0.026	1.621	0.105
px_daily_range	48.017	3.364	14.275	0.000
sector_consumer_discretionary	0.905	0.217	4.175	0.000
sector_consumer_staples	0.489	0.229	2.134	0.033
sector_energy	0.652	0.256	2.549	0.011
sector_financials	2.109	0.217	9.716	0.000

Variable	Coefficient	SE	t-stat	p-value
sector_health_care	1.907	0.224	8.497	0.000
sector_industrials	2.787	0.222	12.575	0.000
sector_information_technology	0.138	0.214	0.646	0.518
sector_materials	3.362	0.251	13.415	0.000
sector_real_estate	1.915	1.256	1.525	0.127
sector_utilities	0.621	0.330	1.883	0.060
side_sell	0.093	0.091	1.025	0.306
spread_ge_fr_10yr	-0.007	0.009	-0.777	0.437
spread_ge_it_10yr	0.004	0.001	2.732	0.006
spread_ge_pt_10yr	-0.007	0.002	-3.966	0.000
spread_2_10yr_us	0.003	0.002	1.657	0.098
value	-1.8×10^{-7}	8.5×10^{-8}	-2.091	0.037
vol_emerging	-0.052	0.022	-2.407	0.016
vol_eu	0.023	0.020	1.108	0.268
vol_us	0.066	0.018	3.691	0.000
vwap	0.009	0.012	0.741	0.459
vwap_5min	-0.007	0.003	-2.445	0.014
week_cl_px	5.6×10^{-5}	2.6×10^{-4}	0.212	0.832

Table 6: Linear regression model for bid-ask-spreads.

Concerning the variables representing the average traded volume in different time periods ('day_volume', '5day_volume' and '21day_volume'), only the variable with the longest lag – 21 days before the order placement – is significant at the 5% level. The negative sign of the coefficient indicates that a higher '21day_volume' is related to a narrower spread. However, contrary to expectations, the sign of the coefficient for the daily volume ('day_volume') is positive. Furthermore, we find a positive and significant association between price volatility (measured by '10day_abs_vol', '10day_rel_vol', and 'px_daily_range') and bid-ask spreads. The impact of the trade on the market (measured by the variable 'mkt_impact' estimated by Bloomberg) is also significantly related to bid-ask spreads. This variable has a negative effect on the observed bid-ask spread. The variables related to the absolute security price at various lags ('avg_exe_pc', 'day_cl_px', 'week_cl_px' and 'month_cl_px') are not statistically significant at the 5% significance level. This is also the case for the broker efficiency proxy, measured by the difference between the arrival and fill prices ('leak_bp').

Portugal has the stock exchange with the highest mean bid-ask spreads, followed by countries representing less than 1% of the dataset (grouped in the dummy variable 'country_other'), and then Spain. It is estimated that the stock exchanges in Portugal and this group of other countries have bid-ask spreads on average 7.1 bps and 3.8 bps higher than those of the US stock exchanges (the base category), respectively. A narrower spread was estimated for other European markets compared to the US, possibly because

the dataset is overrepresented by their most liquid securities.

Moreover, a negative and significant relationship was found between the size of the underlying entity and the estimated spread, suggesting that entities with larger market capitalization tend to have greater liquidity. A similar conclusion is drawn regarding the heat of news publication for the underlying company, with a higher index corresponding to increased liquidity. Regarding the activity sector, we observed wider spreads for all sectors compared to the reference sector, communication services. Materials and industrials were identified as the activity sectors with the widest spreads.

Some of the systematic risk variables are significant at a 5% significance level. Specifically, a higher VIX index and a broader 10-year bond spread between Germany and Italy are associated with lower market liquidity. Conversely, the 10-year bond spread between Germany and Portugal and the emerging markets volatility index exhibits an opposite effect. Credit default swap curves with positive and significant associations to bid-ask spreads are observed for the UK and Brazil.

4.3 Nonparametric determinants

The global importance of a variable to a model can be determined by calculating the mean absolute SHAP values for the individual observations:

$$\sum_{i=1}^{n} |\phi_i(X_j)| \,. \tag{22}$$

The SHAP values for individual observations are usually represented in a beeswarm plot. Figure 2 shows the beeswarm plot for our dataset. First, variables are ordered by their global importance, that is, by their mean absolute SHAP value. To simplify the exposition, we report the results for the covariates with a mean absolute SHAP value greater than 0.2.

In Figure 2, we can observe that the dummy variable indicating the US as the country where the security was traded is the most important variable on average. Each dot represents an observation. The dot's x-position represents its SHAP value. Shades of grey display the variable's value. Naturally, the dummy variable 'country_usa' only has two values (1 and 0) corresponding to the extreme shades of the palette. The value 1 is associated with positive SHAP values, whereas 0 is associated with negative SHAP values. So, this dummy has a positive effect on the bid-ask spreads. The larger spread for securities traded in the US may be attributed to the overrepresentation of securities with greater liquidity when traded in other European markets. The opposite effect is observed for the dummy that indicates whether or not the security traded belongs to an entity with a large capitalization level ('market_cap_large'). Considering the continuous

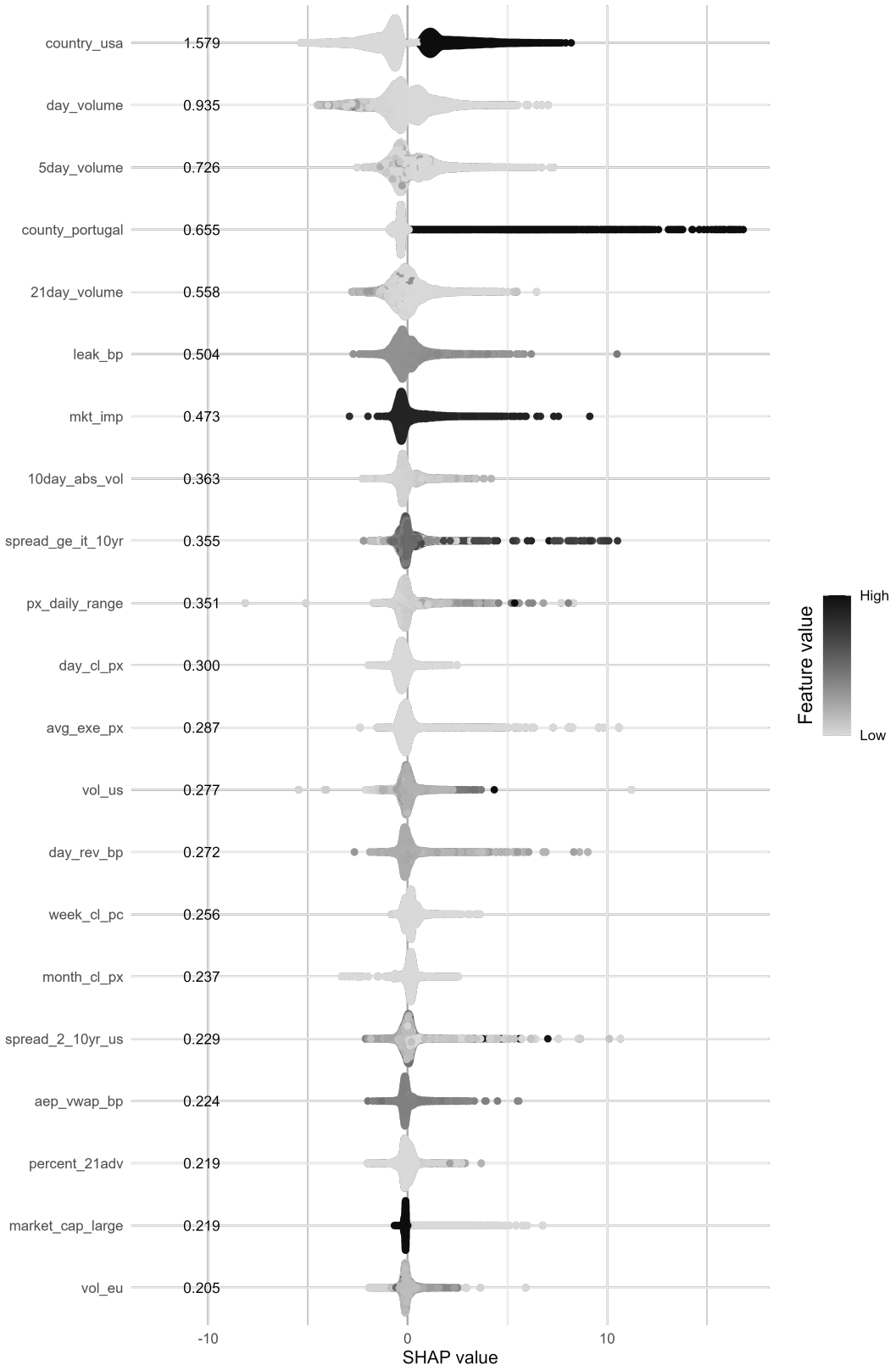

Figure 2: Beeswarm plot for the most influential features according to the SHAP methodology.

variable that measures the spread between the 10-year bonds of Germany and Italy ('spread_ge_it_10y'), we can see a positive association between this variable and bid-ask spreads. For some variables, say the volume of the traded security ('day_volume'), the partial effect's direction is unclear. In those cases, we can inspect the ALE plot. Notably, the volume of the traded security is the second most important variable for the nonparametric model, while not statistically significant at the 5% level in the parametric model.

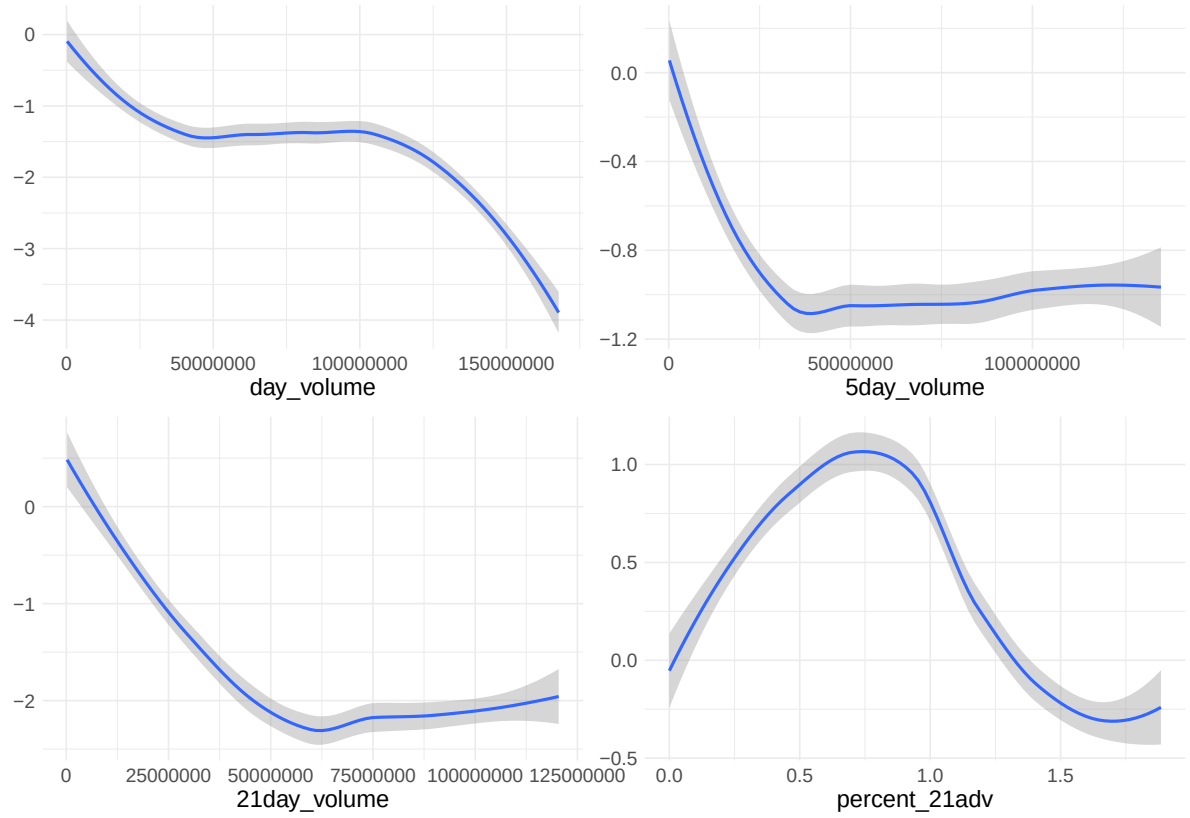

Figure 3: ALE plots per variable.

ALE plots show the marginal effect of each explanatory variable on bid-ask spreads. We applied a LOESS smoother to the ALE values to eliminate potential statistical artifacts. In addition, a confidence interval is plotted, quantifying the uncertainty associated with the estimated smoothing curve. Figure 3 shows the ALE plots for the average traded volume for the day, 5 days, and 21 days before the order placement. Given that the curves decrease with volume, we can infer that matching bid and ask prices in highly active markets is easier. For daily volume ('day_volume'), on the most active days, the bid-ask spread can be 4 bps lower than on the days with the lowest activity. This contrasts with the positive coefficient for 'day_volume' in the OLS regression. Furthermore, the relationship between market activity and liquidity is far from linear. For certain ranges of traded volume, the bid-ask spreads do not change with market activity. For example,

bid-ask spreads remain relatively constant for 5-day volumes above approximately 40 million euros. We also find a strong non-linearity when analyzing the size of the placement relative to the average daily volume of the previous 21 days ('percent_21adv'). In this case, the relationship is concave, reaching a maximum when the size of the placement is approximately 75% of the average volume of the past 21 days.

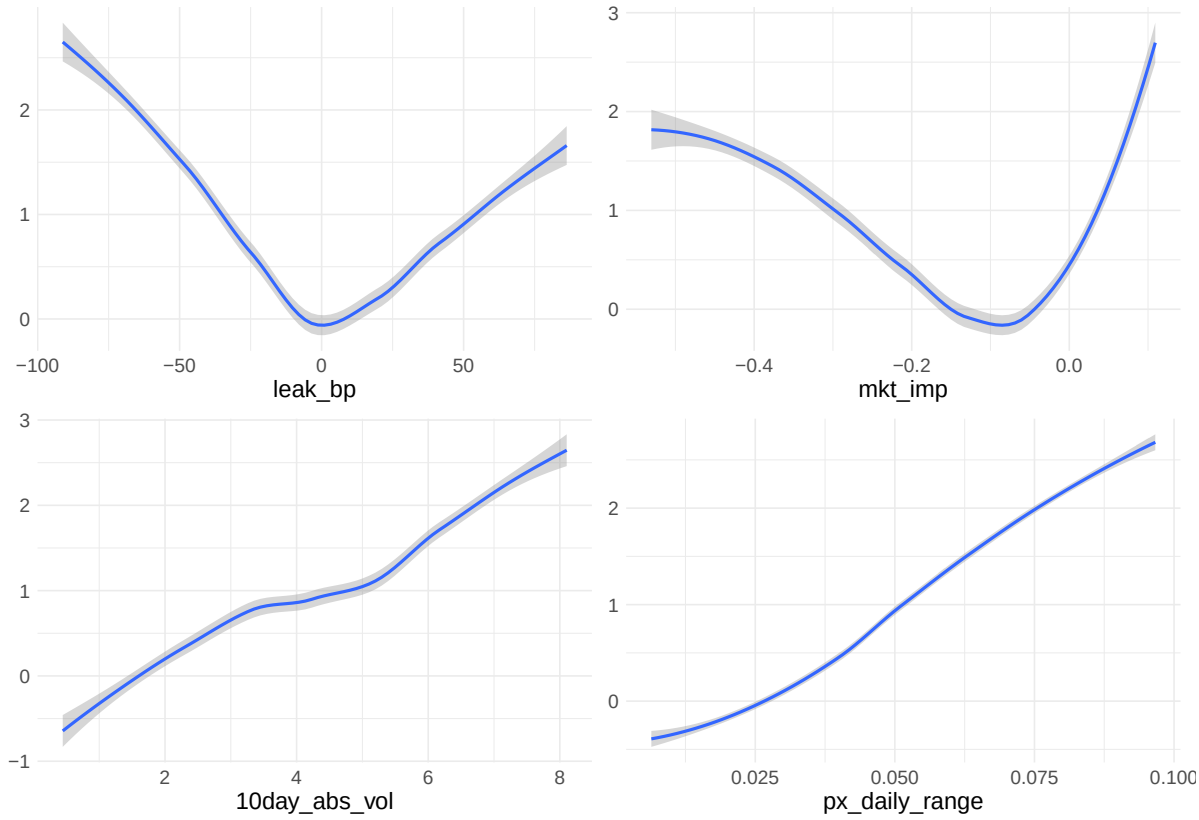

Figure 4: ALE plots per variable.

The first row in Figure 4 shows the ALE plots for the broker efficiency proxy ('leak_bp') and the market impact estimated by Bloomberg ('mkt_imp'). Bid-ask spreads exhibit a dependence with a convex shape on these variables. Variations of up to 3 bps are observed in the bid-ask spreads. It is worth noting that the broker efficiency proxy is not significant in the OLS model (t-stat = -1.63). In contrast, the second row in Figure 4 shows a positive and approximately linear relationship between bid-ask spreads and the price volatility, measured through the last 10 days ('10day_abs_vol') and the intraday relative range ('px_daily_range'). In the OLS model, 'px_daily_range' and '10day_abs_vol' are highly significant, with t-stats of 14.28 and 5.00, respectively.

Figure 5 shows ALE plots for variables related to the underlying security price. In particular, it reports ALE plots for the difference between the average execution price and the volume-weighted average price from order arrival until last fill, adjusted for limit price ('aep_vwap_bp'), the average execution price for the placement ('avg_exe_px'), the

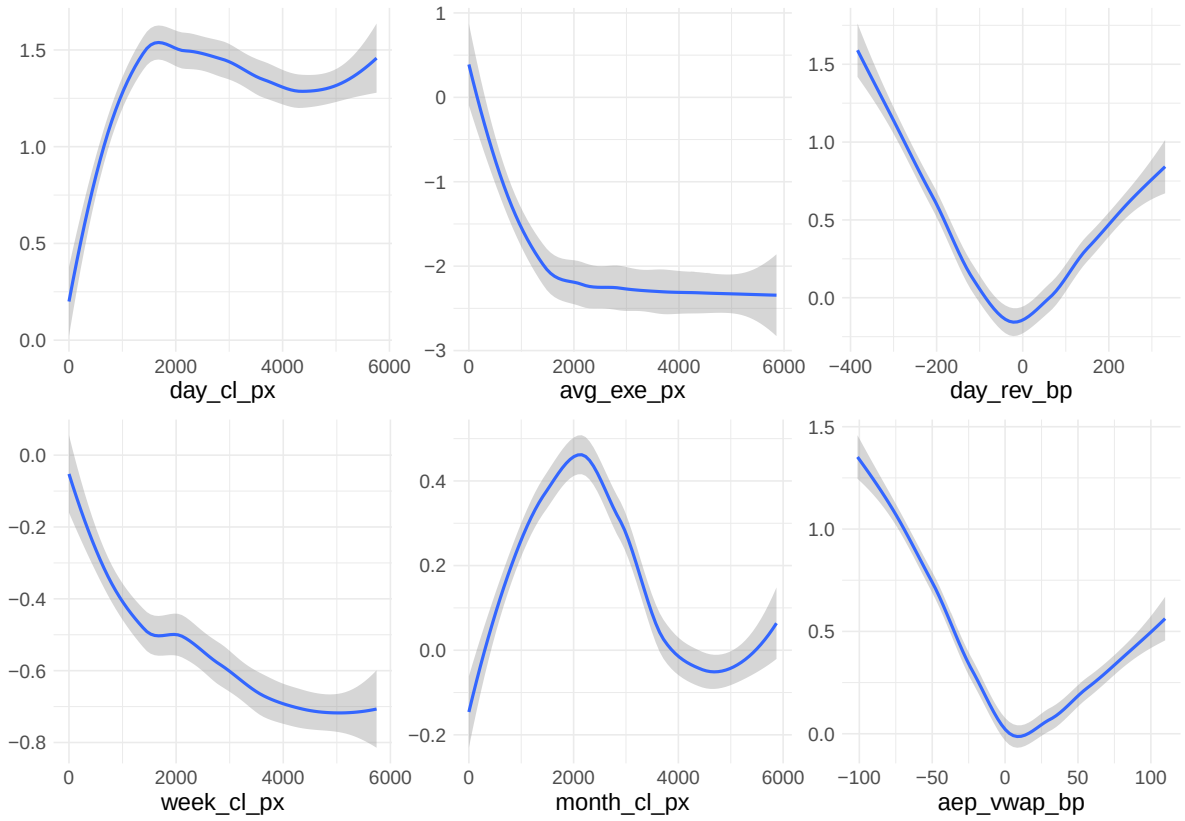

Figure 5: ALE plots per variable.

closing price on the day the order is completed ('day_cl_px'), the difference between the placement's last fill price and the closing price on the day the placement is completed ('day_rev_bp'), the closing price one month before the order is received ('month_cl_px'), and the closing price one week before the order is received ('week_cl_px'). Once more, we observe a wide range of dependencies of bid-ask spreads on the covariates. The variables 'day_cl_px' and 'avg_exe_px' exhibit plateaus when the execution prices cross certain thresholds, while the variables 'day_rev_bp' and 'aep_vwap_bp' show convex shapes. Overall, varying values of these variables may impact the bid-ask spread by approximately 2 bps.

Figure 6 presents the ALE plots for the most influential systematic risk variables. Overall, we conclude that poor macroeconomic conditions and market illiquidity are closely related. The bid-ask spread tends to be wider when the 10-year bond spread between Germany and Italy ('spread_ge_it_10y') is large, particularly when exceeding the 225 bps threshold. Analyzing the US debt market through the yield curve ('spread_2_10yr_us'), we also observe an asymmetric response of bid-ask spreads to market dynamics. Indeed, this variable provides information about market expectations for future economic conditions, including economic growth, inflation, and monetary policy. An inverted yield curve – where longer-term yields are lower than shorter-term yields, as represented in the first

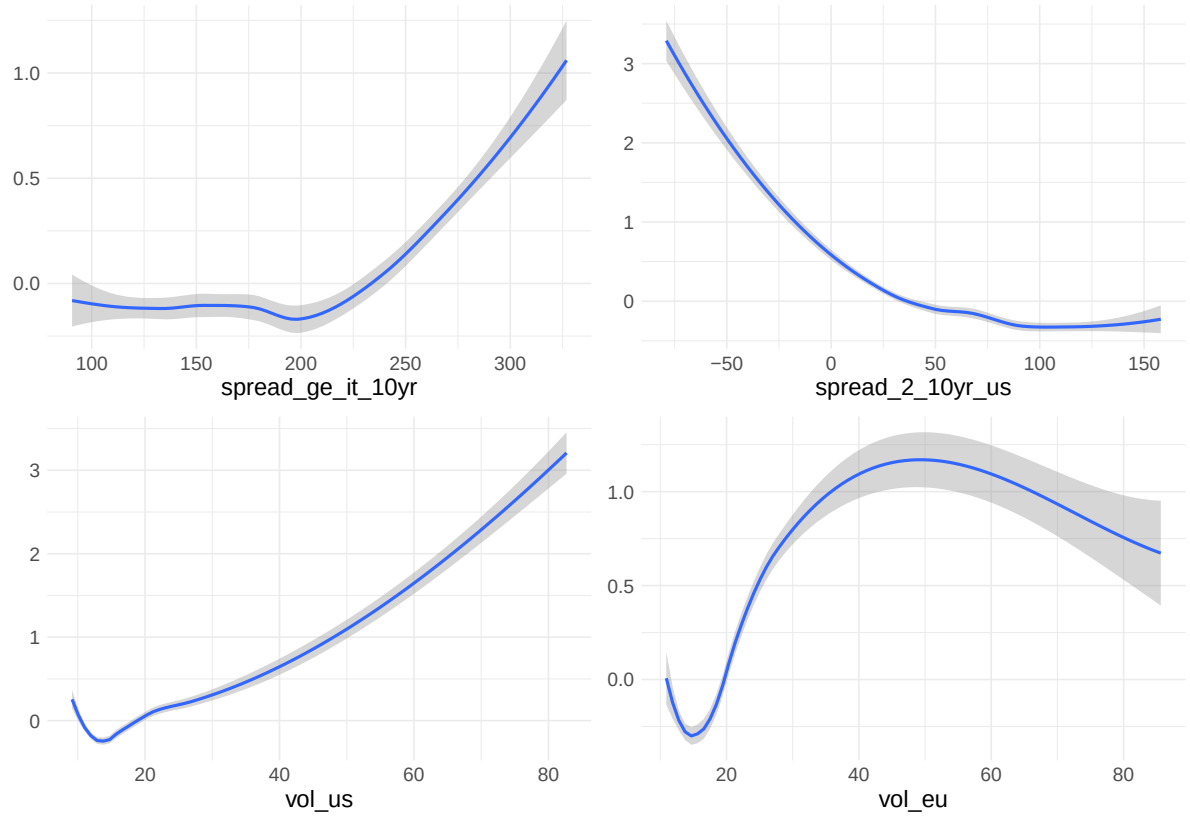

Figure 6: ALE plots per variable.

part of the curve – signals investors' pessimism about the economic outlook, which leads to lower market liquidity.

The plot in the bottom left corner of Figure 6 presents the ALE for the VIX index ('vol_us'). As mentioned in the literature (e.g. Chuliá et al., 2023; Kim et al., 2023), higher forward-looking implied volatility positively affects bid-ask spreads. Our results demonstrate that this link has a close-to-monotonic trend compared to the average predicted spread. The equivalent index for the Eurozone ('vol_us') also shows a positive trend, though much less pronounced, possibly due to the higher proportion of securities traded on US stock exchanges in the data under analysis.

5 Conclusion

Using a nonparametric model, this paper analyzes the factors influencing market liquidity, quantified as the relative difference between quoted ask and bid prices. With this approach, we refrained from making assumptions about the relationship between bid-ask spreads and potential explanatory variables. That is, this relationship was determined by the data itself. The nonparametric model also captures non-linear relationships that a parametric linear model fails to detect. Using a novel dataset of placement orders for

equities managed by a European asset management institution, we explored the most influential determinants and their relationship with bid-ask spreads. The tools to achieve this goal were SHAP values and accumulated local effects (ALE) plots.

Our results suggest that the country of the stock exchange and the size of the underlying entity are relevant in explaining the spreads on traded securities. Moreover, market activity strongly defines spreads, with higher volumes associated with lower spreads. In contrast, there is a positive effect of high price volatility, lower broker efficiency, and a higher estimated impact on the market. Along the same lines, the variables related to the increase in price differences within the trading interval are positively related to spreads. Through systematic risk variables, we also demonstrate that periods marked by heightened uncertainty and weak economic activity are associated with lower market liquidity. More precisely, in periods of higher sovereign bond spreads between Germany and Italy, an inverted US Treasury yield curve or higher aggregate volatility can adversely affect liquidity.

Overall, our results contribute to the existing literature by showing how trading activity, security characteristics, or the macroeconomic environment influence bid-ask spreads. This work opens up room for further research. The analysis aimed to assess market liquidity in terms of transaction and participation costs. However, liquidity has other dimensions that are also important to consider for a comprehensive understanding of liquidity in equity markets. Furthermore, there is interest in exploring how this approach applies to other asset classes, such as fixed-income.

It is also important to note that the results are influenced by the securities traded by the asset management institution that provided the data. Applying this approach to all traded securities on one or several stock exchanges over a specific time period would offer a more comprehensive view of market liquidity.

Acknowledgements

This work was supported by Fundação para a Ciência e a Tecnologia [grant number UIDB/05069/2020].

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502.

Acharya, V. V. and Pedersen, L. H. (2019). Economics with market liquidity risk. *Critical Finance Review*, 8(1-2):111–125.

Aman, H. and Moriyasu, H. (2022). Effect of corporate disclosure and press media on market liquidity: Evidence from Japan. *International Review of Financial Analysis*, 82:102167.

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, 17(2):223–249.

Amihud, Y., Mendelson, H., Pedersen, L. H., et al. (2006). Liquidity and asset prices. *Foundations and Trends® in Finance*, 1(4):269–364.

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086.

Bervas, A. (2006). Market liquidity and its incorporation into risk management. *Financial Stability Review*, (8):63–79.

Biais, B., Glosten, L., and Spatt, C. (2005). Market microstructure: A survey of micro-foundations, empirical results, and policy implications. *Journal of Financial Markets*, 8(2):217–264.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brennan, M. J., Chordia, T., Subrahmanyam, A., and Tong, Q. (2012). Sell-order liquidity and the cross-section of expected stock returns. *Journal of Financial Economics*, 105(3):523–541.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chordia, T., Roll, R., and Subrahmanyam, A. (2001). Market liquidity and trading activity. *The Journal of Finance*, 56(2):501–530.

Chuliá, H., Mosquera-López, S., and Uribe, J. M. (2023). Nonlinear market liquidity: An empirical examination. *International Review of Financial Analysis*, page 102532.

Curth, A., Jeffares, A., and van der Schaar, M. (2024). Why do random forests work? understanding tree ensembles as self-regularizing adaptive smoothers.

Díaz, A. and Escribano, A. (2020). Measuring the multi-faceted dimension of liquidity in financial markets: A literature review. *Research in International Business and Finance*, 51:101079.

ESMA (2020). Guidelines on liquidity stress testing in UCITS and AIFs. Guidelines & Recommendations ESMA34-39-897, European Securities and Markets Authority.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232.

Gourieroux, C. and Monfort, A. (2013). Allocating systemic risk in a regulatory perspective. *International Journal of Theoretical and Applied Finance*, 16(07):1350041.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 507–520. Curran Associates, Inc.

Guerra, P., Castelli, M., and Côrte-Real, N. (2022). Machine learning for liquidity risk modelling: A supervisory perspective. *Economic Analysis and Policy*, 74:175–187.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2023). *An introduction to statistical learning*. Springer.

Kim, D., Li, C., and Wang, X. (2023). Liquidity dry-ups in equity markets. *International Review of Financial Analysis*, 86:102536.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lu-Andrews, R. and Glascock, J. L. (2010). Macroeconomic effects on stock liquidity. *Available at SSRN 1662751*.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *ArXiv*, abs/1802.03888.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Næs, R., Skjeltorp, J. A., and Ødegaard, B. A. (2011). Stock market liquidity and the business cycle. *The Journal of Finance*, 66(1):139–176.

Roth, A. E. (1988). *The Shapley value: essays in honor of Lloyd S. Shapley*. Cambridge University Press.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Sarr, A. and Lybek, T. (2002). Measuring liquidity in financial markets. IMF Working Paper 2002/232, International Monetary Fund.

Schwartz, R. A., Francioni, R., and Weber, P. (2020). Market liquidity: An elusive variable. *The Journal of Portfolio Management*, 46(8):7–26.

Schwartz, R. A. and Peng, L. (2022). *Market Liquidity*, pages 1001–1005. Springer International Publishing, Cham.

Shapley, L. S. (1953). A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games II*, pages 307–317. Princeton University Press, Princeton.

Stoll, H. R. (1989). Inferring the components of the bid-ask spread: Theory and empirical tests. *The Journal of Finance*, 44(1):115–134.

Tarashev, N. A., Borio, C. E., and Tsatsaronis, K. (2010). Attributing systemic risk to individual institutions. BIS Working Paper 308, Bank for International Settlements.

Tavana, M., Abtahi, A.-R., Di Caprio, D., and Poortarigh, M. (2018). An artificial neural network and bayesian network model for liquidity risk assessment in banking. *Neurocomputing*, 275:2525–2554.