Harnessing artificial intelligence for monitoring financial markets[1]

Matteo Aquilina[1,2], Douglas Araujo[1], Gaston Gelos[1,3], Taejin Park[1] Fernando Pérez-Cruz[1]

[1] Bank for International Settlements [2] Macquarie University [3] CEPR

Abstract

Signs of market dysfunction, such as open arbitrage positions in markets that are typically highly liquid, can be associated with periods of financial instability. Artificial intelligence and machine learning models may support high-frequency monitoring of financial markets, taking advantage of the natural wide availability of financial data to efficiently represent nonlinear dynamics. In this paper, we look at the dislocations in the Euro-Yen FX as a theory-based indicator of market dysfunction, and show that a properly set-up recurrent neural network perform well against the challenging autoregressive benchmark in forecasting these dislocations up to 60 business days ahead. And these networks appear to be quite stable: for example, models trained only on data up to end-2020 successfully identify higher probabilities of market dysfunctions in the lead up to the September 2022 UK LDI episode and the March 2023 banking stresses. Remarkably, these results obtain even with the known difficulties to forecast any currency movement at all, ample availability of intermediaries (and thus arbitrageurs) on both sides of this currency trade, and the relative facility with which these dislocations can be arbitraged away compared to other open arbitrage opportunities. Furthermore, our network assigns time-varying attention at the cross-section of input data, which regularises the network while explaining its predictions. Using these variable-specific weights for the periods that forecast market dislocations, large language models can support the monitoring process by, for example, helping to filter news about the highly weighted variables that might be associated with liquidity events. This model can shed light on drivers of market dysfunction and help central banks and financial supervisors anticipate market stresses.

Keywords: Market dysfunction, liquidity, arbitrage, artificial intelligence, financial stability

JEL classification: G14, G15, G17

[1] Work under active development and drafting. We appreciate the useful discussions with Tomaso Aste and team, with Alan Moreira, and comments received from seminar participants at the Journal of Banking and Finance paper development workshops on "Generative AI in Finance" in Dresden and in Montreal (2024). This paper does not necessarily represent the views of the Bank for International Settlement. All errors are our own. Corresponding author: gaston.gelos@bis.org.

Introduction

Markets often deviate from the law of one price or other no-arbitrage conditions. Stronger-than-usual deviations can be a symptom of lower market liquidity (Brandão-Marques et al (2015)) or looming materialisation of financial risk, especially if occurring in multiple markets (Pasquariello (2014)). In the extreme, such episodes can be the precursor of fully-fledged financial crises. While lower frequency monitoring of financial vulnerabilities is well-established conceptually and in practice (eg, Adrian et al (2015)), the real-time assessment of markets in search of impending financial stability risks is more challenging. Two obstacles loom large. First, observed episodes of severe financial stress are rare in the data, in part because policymakers work to prevent such occurrences. And second, risks can materialise in unique ways in each crisis through intricate and hidden channels, making it difficult to discern any common patterns through linear models.

These challenges can be addressed by combing economic theory with machine learning (ML) and artificial intelligence (AI) algorithms. Financial theory may guide the creation of indicators of market dysfunction that can gauge the "latent" chances of episodes of liquidity evaporation and financial instability, overcoming the lack of data on severe financial stresses. And ML/AI techniques can parse the large amounts of high-frequency data to diagnose the state of the markets in real time and help explain findings to supervisory teams or market participants. These techniques can process immense quantities of data in a non-linear way, helping to identify anomalies and unusual correlations across different markets, increasing the chance that major risks could be identified regardless of where they first surface. This paper introduces a combination of analytical tools that illustrates the description above.

Starting with theory-based indicators of financial market dysfunction, we focus our analyses on minute-by-minute price dislocations between the Euro-Yen pair and the same FX pair traded with the US dollar as a vehicle currency. This difference between the direct and intermediated exchange of a pair of currencies is known as the triangular arbitrage parity (TAP) deviation (Fenn et al (2009)). The TAP offers a clean proxy of market dysfunction because it is subject to less frictions: arbitraging discrepancies between currencies does not impose balance sheet constraints, use additional funding or require carrying positions given its immediate closing. Further, currency trading at scale is not usually subject to tax considerations or other meaningful costs. Normally, TAP discrepancies should not occur in such a pair of currencies from large economies stocked with multiple globally active banks with access to global markets, as they would be easily arbitraged away. Still, the Euro-Yen FX pair shows considerable TAP violations that are higher than other FX pairs, and dynamic over time. In fact, the daily average and standard deviations of the minute-by-minute TAP discrepancy are comparable to pairs that involve emerging market currencies, and much higher than other pairs of advanced economies. Another reason to explore the Euro-Yen TAP violation, measured by the daily standard deviation of its minute-by-minute difference, is that this series contains relevant information for financial stability beyond currently-used metrics. In particular, it is correlated with the lower frequency excess bond premia (Gilchrist and Zakrajsek (2012)) and global economic policy uncertainty (GEPU) index (Davis (2016)). But these metrics explain only half of the TAP deviation variation; its residue on these variables is associated with higher tail outcomes for cross-currency basis at a 3m tenor for Euro, Yen, but also Australian Dollar and British Pound up to multiple days ahead (although not for the Swiss Franc), signalling a potential association with forthcoming intermediary distress. For these reasons, the Euro-Yen TAP is taken as the closest to a model-free indicator of latent market dysfunction.

To test the ability of ML models to forecast Euro-Yen TAP deviations, we train recurrent neural networks (RNNs; Goodfellow et al. (2016)) on financial market variables. RNNs combine the empirically-verified power of neural networks to model non-linear and high-dimensional functional forms with the capability of keeping a stateful representation of time series; akin to non-linear, time-varying Kalmar filters.

This means that the model is capable of learning more complex time series-wise features of the data, such as changes in regimes, without being explicitly told about any structure of the Markov process, including the number of regimes or their transition probability. Since the daily data is quite noisy, and our primary focus is on forecasting the probability of a market illiquidity event around a range of dates rather than pinning down an exact date forecast, we focus on 20-day averages of this dislocation (roughly one month). Our goal then is to forecast this 20-day average in horizons that range from one up to 60 business days ahead. We evaluate these RNN-based forecasts by comparing them with a challenging autoregressive (AR) benchmark in two rigorous out-of-sample assessments. First, we use the Diebold-Mariano (1995) test to check whether the forecasts at least match the AR forecasts in terms of a symmetric loss function, namely the root mean squared error. This intuitive test probes how stable and small the forecasting error is over a long range of out-of-sample dates. However, given the RNN might learn slower-moving trends of the noisy data – which would still be desirable representations of the dislocation process for market monitoring goals – the AR test might not be fully informative as to the model performance. So we also use a second test that checks if the true data, unobserved by the econometrician, can be statistically explained by the forecast, by regressing the true data on the neural network and the AR forecasts for the same date. In all cases, the coefficient on the RNN forecast is statistically significant, even with the Newey-West correction.

Crucially, the set-up of the neural network is such that the model learns how to dynamically weight each input variable. Writing the time t vector of N covariates as $x_t = (x_{t,1}, x_{t,2}, \dots, x_{t,N})$, the model learns a flexible function that attributes time t-specific weights $\omega_t = (\omega_{t,1}, \omega_{t,2}, \dots, \omega_{t,N})$ to each variable, such that $\sum_{i=1}^{N} \omega_{t,i} = 1$. Since all variables are normalised to z-scale, these weights effective provide a competitive selection of the signal that the RNN effectively uses to forecast the target variable. A novel component in our network is that it learns from the data how "sharp" or "smooth" this competition for attention at each point in time must be. In any case, these weights act as a form of learned regularisation of the high-dimensional input space, similar in spirit to lasso (Tibshirani (1996)), ridge regression (Hoerl and Kennard (1970)) or elastic net regressions (Zou and Hastie (2005)). An important difference to these methods is that the regularisation is dynamic: unlike these models, who provide a single *global* variable selection valid for all periods, in our model, the contemporaneous and past information at each date influences the weights assigned to each variable, as in: $\omega_{t,i} = \Omega_i(x_t, x_{t-1}, \dots, x_1)$. As a by-product, an inspection of the weights in each period offers an explanation of the main drivers of that period's forecast.

Armed with these weights, we leverage the power of large language models (LLM) and availability of ample economically-rich data to better understand episodes of latent market dysfunction. In periods when the forecasts identify lower market liquidity ahead (ie, higher Euro-Yen TAP deviations), the variables with higher weights can be directly named in prompts to query LLMs about signals in texts that could be related to those variables, and their role in potential market dysfunction episodes. LLMs then inspect financial news and other texts to extract knowledge in a targeted way, helping to provide human-readable insights that could otherwise be difficult to find. We demonstrate episodes where timestamped queries are actually able to uncover key drivers of financial stability episodes a few days before they happen. More broadly, even when such a tool cannot offer forward-looking insights because some events are in practice very difficult to forecast, they can help analysts better understand factors driving recent movements in these input variables of relevance. Central banks and financial supervisors can consider adapting this monitoring system for use in their day-to-day financial stability monitoring activities.[2]

This work connects four literature streams. First, on arbitrage conditions and market dysfunctions. Second, on volatility prediction and financial stability. Third, on the use of AI/ML in finance, in particular

[2] The code and a proof of concept are planned to be made available through the BIS *gingado* open source library (Araujo (2023)).

to predict asset prices and returns (since arbitrage opportunities are differences between asset prices). And fourth, on the use of LLMs in finance.

The literature on understanding the potential drivers of market dysfunctions, financial stress and financial crises takes different approaches, but can be classified into two main ones. The first tries to predict different regimes in which the financial sector could be (an orderly and a dysfunctional one), while the second directly attempts to determine the variable of interest such as a measure of overall liquidity. Markov regime switching models have mostly been used for the former, while standard regression techniques, have been deployed to tackle the latter. We add to the literature by using more advanced models.

More recently, a burgeoning literature on AI/ML uses in finance has leveraged the numerical flexibility of these models to forecast or explain asset prices.[3] One theme is this literature is that these models enable analyses on high-dimensional settings, which in finance can be advantageous (Kelly et al (2023)). Our work draw from some of those insights in that the models trained have many more parameters than the number of data points – to the model's advantage. For example, Bali et al (2023) recently document the role of non-linearity and functional form flexibility in forecasting option returns out-of-sample. However, in contrast to a large share of the financial economics literature, who use feed-forward networks with no time structure, we deploy a model primarily based on recurrent neural networks (RNNs). Other works using RNNs include Sirignano and Cont (2019), who use RNNs to document universal features of asset prices, and Vuletić et al (2024), who combine RNNs with generative adversarial networks (GAN), model that learns to generate seemingly typical instances of its training data, to successfully forecast financial series.

Recurrent or not, the flexible way these models incorporate complex non-linearities, however, has the drawback that their predictions are in practice usually difficult to inspect (hence they are often called "black box" models). Attempts to address this drawback and make financial neural networks more interpretable while retaining considerable forecasting power include the Temporal Fusion Transformer (TFT) model of Lim et al (2021), who rely on a modified version of the transformer (Vaswani et al (2017)) model to learn how the different time periods of a series are weighted over time. This application builds on the concept of *attention* (Graves (2013), Bahdanau et al (2014)), or dynamic weighting of covariates' importance to explain a target variable. The attention mechanism is traditionally oriented at the time dimension, ie changing the importance of each lag in a sequence. Our work, in contrast, approaches explainability from a different angle, learning to weight different variables over time; ie $\omega_{t,i}$ can be interpreted as a time-varying *cross-sectional attention* score. This is similar in spirit to Chen et al (2018), Yoon et al (2019) and Arik and Pfister (2021), who apply variable-wise attention, but in our case we apply the concept over a time series through an RNN rather than instance-wise estimation. Closer to our work, Lim et al (2021) include variable selection networks in their neural network to learn when to let each variable pass their signal through; however, in our case the signals *compete* for the model's attention, ie high attention paid to one variable crowd out signals from other variables (the strength of this crowding out is learned by the network). This in turn allows an easy inspection of the main drivers of each forecast, a theme which we elaborate further down as it enables the use of LLMs to gather further human-readable insight on potential dysfunction movers.

Recent advances in large language models (LLMs) have further precipitated a transformation in the analytical capabilities available for financial market prediction and risk monitoring. These sophisticated architectures enable the systematic extraction and synthesis of latent signals from vast textual datasets, encompassing market commentary, corporate disclosures, regulatory filings, and social media postings, thereby expanding the analytical frontier beyond conventional structured data sources. Current implementations of LLM-based prediction systems predominantly concentrate on equity price forecasting,

[3] Kelly and Xiu (2023) offer a broad review of the literature.

leveraging two architectural paradigms: Retrieval-Augmented Generation (RAG), which enhance predictive accuracy through dynamic integration of relevant timely and contextual data, and LLM-based agents that are tasked to conduct specific analysis, strategy formulation, and decision-making processes within pre-defined operational parameters (Zhang et al (2024a); Zhang et al (2024b); Papasotiriou et al (2024); Iacovides et al (2024); Yu et al (2023); Elahi and Taghvaei (2024); Yu et al (2023)).

Notably, an emerging strand of research has extended these methodological frameworks toward financial market risk monitoring applications, which align closely with our research objectives. These monitoring implementations demonstrate particular efficacy in synthesizing multi-modal data streams, incorporating traditional numerical indicators alongside diverse textual sources such as earnings call transcripts and contextual news feeds to construct more comprehensive risk assessment frameworks (Cao et al (2024)). However, a critical limitation of using LLMs for market predictions stems from the inherent probabilistic nature of LLMs, which generate outputs based on statistical patterns in training data rather than through causal reasoning or accurate insight. This fundamental characteristic introduces significant risk when these models are deployed for direct market forecasting and integrated with real-time trading algorithm. Therefore, applications of LLM-based financial market analysis are most appropriately positioned as complementary analytical tools that generate supplementary insights for human evaluation and interpretation, rather than as autonomous predictive systems, thereby preserving critical human judgment in the complex process of financial decision-making and risk assessment.

The next section provides context for our work. Section 3 describes the methodology we use, and Section 4 describes the datasets for the empirical estimation. Section 5 presents the results and the last Section concludes.

2. Background and contribution

2.1 Policy-relevant market dysfunctions

Public authorities responsible for supervising and regulating financial markets typically have among their objectives that of promoting fair, orderly, and efficient markets.[4] Although the specific definitions may vary slightly, they generally stipulate that the trading process should be efficient and orderly. This ensures that prices reflect available information and correspond to the fundamental value of the underlying assets.

The ideal benchmark for these authorities is an infinitely liquid market devoid of frictions or impediments: participants can trade any quantity at the current price, and the price of any asset reflects its fundamental value. However, in reality, this benchmark is unattainable. Intermediaries are needed to connect buyers and sellers, market makers require funding to execute trades, and the trading process demands time and effort, among other factors. Many of these frictions take a central role in some asset pricing paradigms, eg intermediary asset pricing (He and Krishnamurthy (2013, 2018)).

These frictions can persist and eventually lead to market distortions, which can manifest in two non-mutually exclusive ways. Firstly, prices can deviate from an asset's fundamental value, and secondly, the trading process can become particularly costly, or in other words, the market may become illiquid. Both are relevant as they interfere with an appropriate pricing and allocation of risk. We describe the former case in section 2.2 below, and the latter in section 2.3.

[4] For instance, a review of regulators' website reveals that this is the case in central banks in Australia, Brazil, Canada, the European Union, Indonesia, Singapore the UK and the US, and possibly others.

2.2 Price deviations

The financial literature has a long history of highlighting persistent price differences among essentially equivalent asset pairs, ie situations when the law of one price does not hold in practice. This is prevalent across asset classes. Looking at corporate bonds, Dammon et al (1993) noted considerable price disparities between essentially equivalent bonds issued by the same company. Securities that provide access to shares of the same company also violate the law of one price (Lamont and Thaler (2003)); since a depositary receipt is a certificate representing shares in a particular stock, its price should align with that of the underlying stock. This phenomena takes place even in markets with a trade turnover measured in hundreds of billions or trillions of US dollars. For example, in the US bond market, Fleckenstein et al. (2014) highlighted price differentials between bonds with and without inflation protection and Foucault et al (2014) underscored significant mispricing in triplets of currencies, which usually rely on the US dollar as a *vehicle* currency; this market is also studied by Ranaldo and Santucci de Magistris (2022). Other potential benchmarks include covered interest rate parity - the relationship between the spot and forward price of two currencies and their interest rates, ETF discounts compared to the net asset value of the assets they invest in, and depositary receipts parity.

Pasquariello (2014) conducts a comprehensive study of such dislocations across foreign exchange, interest rate, and equity markets. He shows that these dislocations significantly increase in proximity of financial stress episodes, such as the Asian crisis in 1997, the LTCM episode in 1998, and the global financial crisis in 2008. This indicates that such dislocations are indeed connected to the orderly and efficient functioning of markets. However, there's no need to limit the focus to episodes where overall financial stability was at risk, like the 2008 crisis; examining such dislocations can help assess market functionality even when deviations are not as substantial.

2.3 Lack of liquidity

Orderly functioning of financial markets involves examining the trading process itself. Specifically, it is useful to consider instances where, irrespectively of whether prices reflect fundamentals or not, the trading process breaks down and transactions among market participants become challenging or particularly expensive. In other words, one can assess the overall level of market liquidity.

A liquid market is one where participants can transact quickly, at low cost and at a price that is close to the fundamental value of an asset. For various reasons, including the characteristics of the assets, the type of participants that tend to trade it as well as the way in which the trading process is organised.[5] The liquidity of a single asset is not static and can fluctuate over time, particularly during periods of high uncertainty regarding its value. However, significant reductions in liquidity are, in themselves, indicative of markets not functioning efficiently and orderly. In some instances, the trading process can completely break down, leading to what are termed "liquidity black holes" (Persaud (2003)): episodes where the trading process becomes destabilising rather than stabilising. Brunnermeier and Pedersen (2009), distinguish between the interrelated concepts of *funding* liquidity and *market* liquidity, which interact with each other and can be the source of such black holes. Similarly, Tirole (2011) argues that these episodes can occur due to either adverse selection or to shortages of financial capability in financial intermediaries.[6]

[5] Foucault et al (2013) provide an excellent introduction to the concept of market liquidity and discuss the reasons why it varies among different securities and across time.

[6] Kyle (1985) and a larger market microstructure literature point out the relevance of informed vs uninformed traders in driving trade flow and asset volatility.

Hence, in addition to violations of the law of one price, episodes of substantial illiquidity are of interest for regulators and supervisors because they signal a reduction in the efficiency of market functioning.

2.4 Combining the different dislocations

Several additional considerations are relevant for the measurement of financial stability risk associated with the dislocations discussed above. First, it is crucial to establish a benchmark on the basis of which dislocations are measured. The existence of frictions implies that prices will diverge and liquidity may not always be available in large quantities. Dislocations may vary over time and by asset: equities and currencies are much more liquid than real estate, and there could be relationships between the level and volatility of dislocations. For instance, the IMF (2015) highlights how bonds that tend to be less liquid also experience less resilience in liquidity.

Second, it's important to note that liquidity shortages and price dislocations can occur concurrently or independently within a market. All else being equal, supervisors are likely to be more concerned if both types of dislocations occur simultaneously, as this would indicate multiple aspects of the market are malfunctioning. A possibility is that a lack of liquidity could cause price dislocations: as mispricing arises between two different assets that are theoretically linked, market participants may be unable to trade against such mispricing and return to an equilibrium position due to low liquidity.

A third relevant issue is whether supervisors should be more concerned about a large dislocation in a single market or by many relatively small ones in multiple markets. The answer is not straightforward and likely to depend and is likely dependent on several potential factors. If the market subject to a large disruption is a very important one such as – say – the government bond market, which is used as a reference in many other markets, it may well be right for supervisors to pay particular attention to this single market. However, deviations across multiple markets may suggest broader issues prevalent across the entire financial system, indicating an overall higher risk.

A fourth feature is related to how dislocations develop over time. In this case too there is no single way of ranking different developments in terms of their seriousness. On one hand, gradual changes may lead to dislocations that persist for a prolonged period. These episodes could merely reflect the market transitioning to a new equilibrium or might signal enduring problems in the markets that warrant further examination. On the other hand, some dislocations could arise suddenly but then quickly dissipate. While supervisors might not be overly concerned about these short-lived spikes, if such episodes occur frequently, they could indicate the presence of market frictions that could lead to larger dislocations in the future.

For all the issues above, forecasting indicators of market dysfunction can provide useful and actionable signal that might, if persistent, indicate looming illiquidity events.

2.4 Forecasting market dysfunctions

There is a substantial body of literature focused on understanding the potential drivers of market dysfunctions, financial stress and financial crises. The approach taken varies depending on the specific questions being asked. One class of questions deals with assessing and predicting different regimes in which the financial sector may operate. Although theoretically there could be numerous regimes, in practice, the literature has focused on estimating two: a tranquil (or normal) regime and a stressed one. Another type of question involves directly assessing which explanatory factors are more important to forecast a variable of interest, which is often a measure of market liquidity. Typically, the focus had been on stock markets where data has been most readily available. But over the past two decades, this research

has expanded to include other markets—especially government bonds and foreign exchange—and to highlight the commonality of liquidity across them.

Markov (regime) switching models are the most widely used tool to answer the first type of questions. Originally developed by Hamilton (1989), they have been used mainly to predict changes in the business cycle and forecast the likelihood of a recession. Regime switching models of this type, estimate two types of parameters: the variables of interest (say GDP growth) in each of the regimes and the probability of switching from one regime to the other. While in the early incarnations of these models such switching probabilities were fixed, more recent developments have incorporated time-varying ones. The assumption underpinning these models is that an economic variable would behave differently in the two regimes: for instance countries may experience growth periods in which output is steadily increasing and not volatile intermitted by recessions with declining output.

While mostly used in macroeconomics, recent attempts have been made to deploy these models to analyse financial stress. However, the focus has been mostly on low-frequency data, typically quarterly or monthly. Hauptmann et al (2014) and Duprey and Klaus (2017) use regime switching models to forecast turbulence in the S&P500 and in a European-wide measure of stress respectively. Hauptmann et al (2014) aim to assess whether Markov models can help construct better-performing portfolios and find that incorporating signals from the switching model can enhance performance and reduce risk compared to standard buy-and-hold strategies. The latter study looks at whether leading indicators affect the probability of entering a financial stress regime across European countries. It finds that system-wide stress predictors like the debt service ratio and property prices are useful. However, most predictors only became significant during the 2008 financial crisis, suggesting heavy reliance on that episode for future stress forecasts.

Among the approaches that attempt to forecast dysfunctions or lack of liquidity directly, Karolyi et al. (2012) examine liquidity commonality in stock markets across countries and over time, ie whether liquidity moves together across space and time. Their empirical tests assess whether such commonality is driven predominantly by supply or demand side factors. The former relate to the drivers of liquidity supply such as the funding needs of intermediaries, while the latter are mostly related to correlated trading by investors and investors' sentiment. By exploiting both cross-sectional and time-series data, they conclude that demand factors, and in particular correlated investor flows (mainly mutual funds), are more important for stock market liquidity. They note that liquidity commonality is especially pronounced during periods of high volatility and market downturns, likely driven by agents liquidating positions across multiple assets.

But the determinants of liquidity are not related simply to a single asset class. Chordia et al. (2005) analyse the interaction between sovereign bond and stock liquidity, focusing on the United States. They identify several interesting patterns: for instance, Friday is the least liquid day in both markets, while July to September are the most liquid months. These findings suggest that any liquidity prediction model should incorporate such seasonal patterns. Additionally, they find that changes in volatility and liquidity within and across asset classes are associated with future liquidity levels, indicating the relevance of both market interconnections and of an autoregressive component in liquidity determination. In other words, any shock to liquidity and volatility can be used to predict the future level of liquidity. They also highlight that changes in bond fund flows can predict liquidity in both stock and bond markets, extending the findings of Karolyi et al. (2012).

Foreign exchange markets have been studied significantly less than other markets. Karnaukh et al. (2015) and Ranaldo and Santucci de Magistris (2022) are the most comprehensive studies of liquidity determinants. Both papers emphasize that liquidity deteriorates when intermediaries are less able to respond to shocks, whether due to operational reasons (e.g., reduced liquidity outside London hours and during holidays) or risk management issues (e.g., higher funding costs or volatility). They find economically significant effects: a one standard deviation increase in the VIX is associated with a 17% increase in liquidity

costs. Additionally, currencies more exposed to systemic risk experience higher liquidity commonality. A key finding is the high degree of liquidity commonality across foreign exchange, bond, and stock markets, with liquidity in one market deteriorating when it is lower in another. Gradojević et al (2020) looks more specifically at currency TAP deviations at an ultra-high-frequency definition, and finds that a denoised version can be forecasted at that frequency. Further, this forecastability depends on volatility risk and correlation between currencies, which can generally speaking that a low or a high value (ie, can be interpreted as weakly following a Markov process).

Recently, Aliyev et al. (2024) examined the resilience of liquidity in foreign exchange, sovereign bond, and stock markets over the past 25 years across various geographical areas. They found that while the mean and standard deviation of bid-ask spreads have declined across all markets studied, higher moments of the distribution, especially skewness, show different patterns. Skewness has increased significantly in equity and sovereign bond markets, while it exhibits an inverted-U shape in currency markets. This suggests that episodes of illiquidity are now more common in sovereign bonds and stock markets. The authors also find that structural breaks in mean spreads correlate with macroeconomic shocks, while breaks in skewness are more closely associated with changes in regulation and market microstructure. Overall, modern markets may have traded off smaller average spreads for higher liquidity skewness.

A study that combines the Markov switching approach with the direct one is Grillini et al (2019) who focus on different liquidity regimes for a number of European stock markets. They find strong evidence for the existence of two regimes in stock liquidity and indeed that these regimes are very persistent: the probability of remanning in each regime is above 90% in all the specifications.

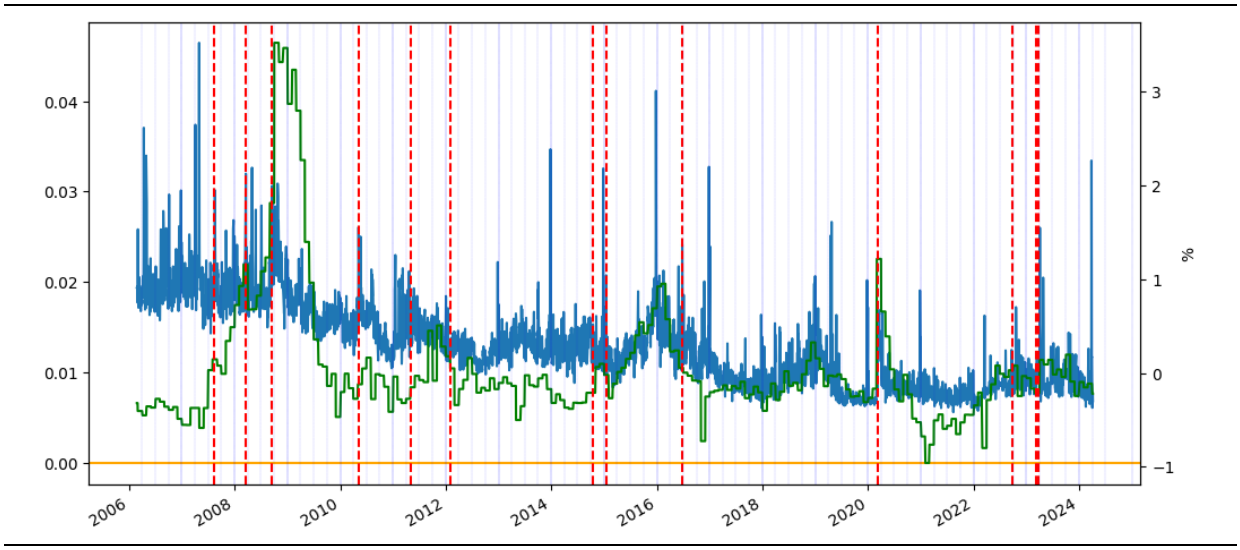
2.5 Euro-Yen TAP deviation as a measure of market dysfunction

Leveraging all these aforementioned insights, we choose as testing ground for our model the challenging but relevant task of forecasting TAP deviations for the Euro-Yen pair. Further, we claim that this metric is a model-free indicator of relevance for market stability.

In principle, no FX should experience TAP violations: for sufficiently large intermediaries, taking advantage of these opportunities should be straightforward: if the EUR/JPY exchange rate (direct) is not exactly the same as EUR/USD followed by USD/JPY (intermediated), an arbitrageur could buy the more affordable one while selling the other (time subscripts are ignored as the transaction is instantaneous). Also, this type of transaction does not usually fall into tax rules, even for larger banks. But it is impressive that this FX pair in particular seems to concentrate particular market illiquidity metrics, considering the sizeable population of large, global banks available in each country. Lastly, this TAP deviation is ultimately calculated from underlying exchange rates, which have limited forecastability (Rossi (2013)).

In addition to its practical forecasting challenges, the Euro-Yen TAP deviation is relevant. Figure 1 plots the TAP deviation with the excess bond premium (EBP; Gilchrist and Zakrajšek (2012), Favara et al (2016)). The latter explains 14.6% of the variation of contemporaneous TAP deviations – even though the EBP has a lower frequency. Another lower frequency variable of policy and market importance, global policy uncertainty (Davis (2016)), has R2 of 41.4% (Figure 2).

Daily Euro-Yen TAP violations and the excess bond premium — Figure 1



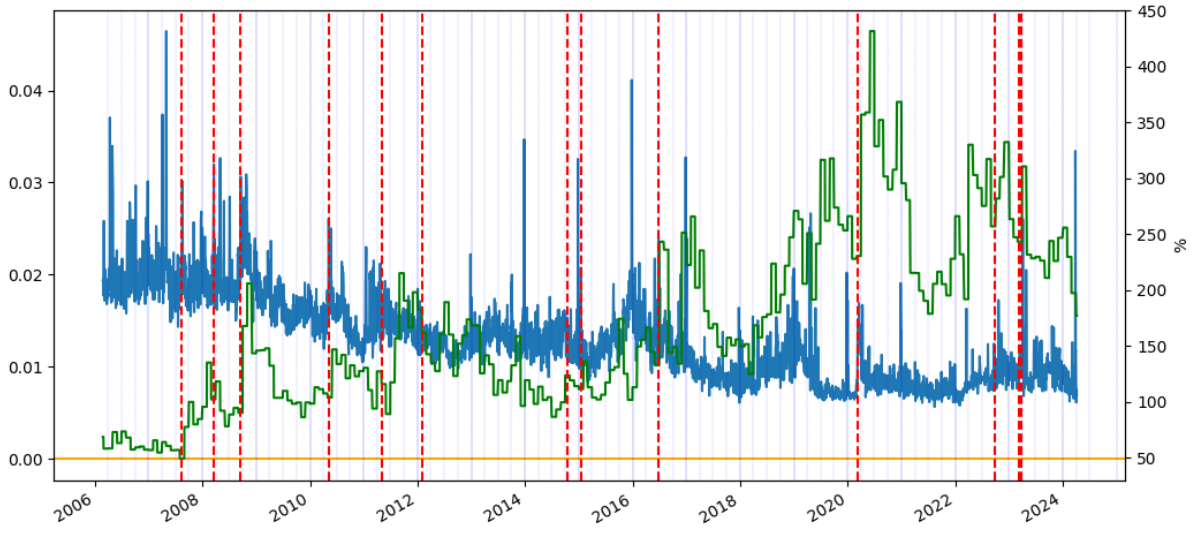
Blue: daily standard deviation of the minute-by-minute Euro-Yen TAP violation; green: excess bond premium.

Source: LSEG, Favara et al (2016), authors' calculations.

In spite of these interesting correlations with variables shown to have practical and policy importance, our measure of market dysfunction carries additional economic signal. Regressing the Euro-Yen TAP deviation on these two variables and another metric that is related to financial market distress, the bond volatility index MOVE, explains only slightly more than half of the variation of TAP deviation (R2 of 53.7%), with all three variables statistically significant (Table 1).

Daily Euro-Yen TAP violations and global economic policy uncertainty

Figure 2



Blue: daily standard deviation of the minute-by-minute Euro-Yen TAP violation; green: global economic policy uncertainty.

Source: LSEG, Davis (2016), authors' calculations.

Euro-Yen TAP deviations retain significant signal after controlling for key variables

Table 1

OLS Regression Results

Dep. Variable:	USDEURJPY_Std	R-squared:	0.537
Model:	OLS	Adj. R-squared:	0.537
Method:	Least Squares	F-statistic:	187.4
Date:	Wed, 11 Dec 2024	Prob (F-statistic):	1.35e-114
Time:	02:19:38	Log-Likelihood:	19856.
No. Observations:	4553	AIC:	-3.970e+04
Df Residuals:	4549	BIC:	-3.968e+04
Df Model:	3		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0174	0.001	31.820	0.000	0.016	0.018
ebp	0.0017	0.000	7.770	0.000	0.001	0.002
MOVE	1.943e-05	4.5e-06	4.320	0.000	1.06e-05	2.83e-05
GEPU_current	-3.677e-05	2.17e-06	-16.943	0.000	-4.1e-05	-3.25e-05

Omnibus:	1700.665	Durbin-Watson:	0.561
Prob(Omnibus):	0.000	Jarque-Bera (JB):	12555.774
Skew:	1.593	Prob(JB):	0.00
Kurtosis:	10.486	Cond. No.	860.

Notes:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 12 lags

Source: LSEG, Favara et al (2016), Davis (2016), Bloomberg, authors' calculations.

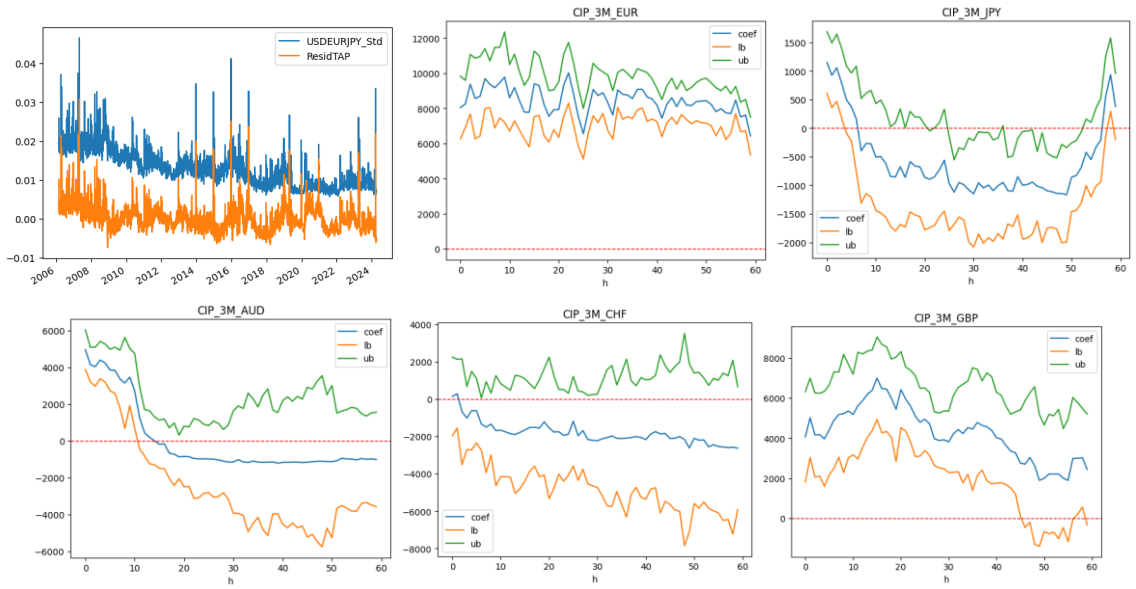
In a more direct illustration of the economic significance of this measure, the residual from this regression (ie, orthogonal to EBP, policy uncertainty and the MOVE) is associated with higher violations of covered interest parity such as widening of cross-currency basis spreads at the 3 month tenor (Figure 3). These regressions are similar to local projections (Jordà (2005)), albeit one for each currency's cross-currency swap spreads against the US dollar. These spreads are analysed in absolute numbers, since any deviation from zero essentially represents a less-than-smooth local market for US dollars. We project the impact on the 99th quantile, that is, positive values mean that market dysfunctions as measured by our TAP deviations deteriorate global local US dollar funding conditions; conversely negative values mean that the tail becomes less extreme, in effect normalising markets. Given the nature of the TAP deviations and the fact that we are using residuals, these analyses illustrate the impact of impulses or jumps in market dysfunction, although here we abstain from making any causal claim. Instead, we point at how these shocks anticipate meaningful movements at the tail of market variables that speak to CIP deviations.

Two things are worth noting: first, the residual (top left panel) retains much of the same higher-frequency dynamics of the TAP deviation, and some movements that develop over multiple days but then converge to the mean eventually. A second observation is that the effect of this residual is different across currencies: tail outcomes in the Euro increase persistently across the lifetime of the swap contract (top centre panel); interestingly, the same increase in the Japanese currency side (top right panel) is more short-lived and in fact, there is zero or retraction effect on the tail for the Yen-US dollar swap spread tail. Perhaps this could be related to this currency's safe haven characteristics. Other currencies that are not directly

involved also respond differently: the Australian dollar (bottom left panel) sees an intense effect that lasts for a reasonable time although definitely not as long as for the Euro, while the same occurs for the British Pound (bottom right panel), with the effect lasting longer. Another safe haven currency, the Swiss Fran (bottom centre panel) does not record any meaningful effect on its tail outcomes.

TAP violation residual is meaningfully associated with higher tail of CIP deviations in some currencies against the US dollar

Figure 3



Source: LSEG, Bloomberg, authors' calculations.

In short, the metric we use to illustrate market dysfunctions – the daily standard deviation of the Euro-Yen minute-by-minute TAP violations – can both anticipate at a higher frequency signals of economic importance that are only available at a lower frequency, such as the EBP and policy uncertainty index, as well as carry significant new economic content on its own.

3. Methodology

Now that we established the importance of the metric we use for a globally-relevant measure of market dysfunction, this section presents the neural network machinery that will extract insights from market variables (further described in Section 4) to forecast the TAP deviations.

AI and ML techniques can be used to measure and predict the market dysfunctions outlined in the previous section. We focus on types of neural networks in our analysis that can efficiently represent time series, described in subsections 3.1 and 3.2 below.[7] The design of these algorithms affords the possibility to tailor the inflow of signal from the covariates all the way to the prediction made by the model at each time step. We take advantage of this flexibility and design a neural network whose prediction architecture can at the same time extract meaningful signal and inform the econometrician as to which

[7] Other widely used ML methods based on completely different algorithms, such as elastic net (Zou and Hastie (2005)) and XGBoost (Chen and Guestrin (2016)), are not used as they do not hold a stateful representation of the time series as RNNs do.

variables these signals came from, at each point in time. This and other design decisions are discussed in subsection 3.3.

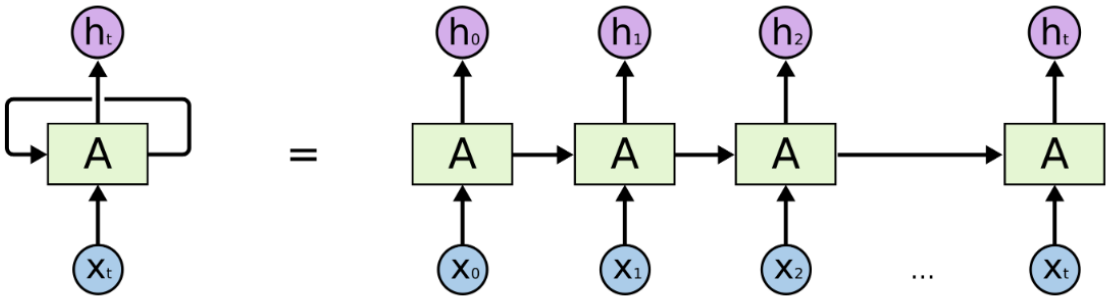
The ability of these models to predict the TAP violations for the Euro-Yen pair is evaluated in two complementary forms. The first, and most stringent evaluation, is a formal Diebold-Mariano test that compares the neural network to an equivalent AR model, using the root mean squared error as a loss function. The second evaluation directly checks whether the predictions are contemporaneously related to the target variable: if yes, the former can help predict the latter even if it does not match its scale or the higher frequency movements. In fact, this second evaluation addresses the cases where the neural network produces sufficiently smooth series (similar to trends in time series decompositions) that is still, nonetheless, informative of the overall movement of the variable even if it abstracts from its higher frequency noise. These tests are described in subsection 3.4.

3.1 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are an architecture for deep neural networks that operate with sequential data, such as time series, speech, or text (Goodfellow et al. (2016)). RNNs take a new input, update their state (also called "memory"), and produce an output at each time step (illustrated in Figure 4). The output of an RNN – the target variable - depends solely on the current input data and on the stored memory. As a result, the memory efficiently stores all relevant information from previous inputs. Economic uses of RNNs include inflation forecasting (Almosova and Andresen (2023)) and well-performing submissions to "M" competitions. In fact, the winner of the M4 competition (Petropoulos and Makridakis (2020)) combines RNNs with exponential smoothing (Smyl (2020)).

RNNs function similarly to Kalman filters (Gelb (1976)) on state-space models, with the main difference being that the relationships between inputs, outputs, and the memory update functions are nonlinear. Further, the nonlinearities are flexibly learned from the data, as opposed to modelled explicitly by the analyst. For example, in traditional econometrics a prototypical nonlinear model is the inclusion of a squared regressor. This needs to be explicitly included by the econometrician. In contrast, all a neural network requires is the set of regressors and the corresponding dependent variables to model, with sufficient examples, the (potentially much more complex) nonlinearity form that best fits the data.

Scheme of a typical recurrent neural network — Figure 4



X_t represents the input to the RNN, *A* the hidden layer and memory, and h_t the output. The right-side plot unfolds the RNN structure to illustrate left-to-right how each input only affects the current state of the network.

Source: (Olah, 2015)

However, early types of RNNs are unable to learn long-term relationships because the memory is constantly updated with new information. Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber (1997)) address this vanishing memory problem by adding a set of "gates" to the memories of the RNNs.[8] These gates are themselves neuronal layers whose output is a number between 0 and 1. This allows the gates to automatically control, based on data, the addition of new relevant information to the memory, the removal of information that is no longer relevant, and the retention of long-term concepts that are useful to predict the output. LSTMs also include a gate to decide which memories are active for each output.[9]

Exploring the structure of an LSTM networks provides insight about how and why they function well; see Figure 5. The "long-term" information (also known as "carry state") flows from left to right, represented by the vector c_t. At each period, new observed inputs x_t enter from the bottom, while a vector of outputs (also known as "hidden states") h_t are produced at the top and transmitted to the subsequent stage too, where another version of this LSTM with the same parameters will repeat this process. The top line in the Fig. 2 scheme represents the memory of the LSTM, c_t. The values of this vector can be modified by a nonlinear combination of the new input x_t and the previous state h_{t-1}. The LSTM has three gate functions:

1. Forget Gate (Labelled 1): This is a single layer fully connected neural network with a sigmoid activation function σ that constraints the output to be between 0 and 1. If the sigmoid output is zero, the network will erase the memory. Conversely, an output of one will allow the information to pass through fully. Formally, the forget gate is:

$$f_t = \sigma(b^f + W^f x_t + U^f h_{t-1}).$$

2. Input Gate (Labelled 2): This consists of two networks. The first is a single layer fully connected neural network with a sigmoid activation function that decides whether to add information to the memory (the opposing operation to the forget gate). The second is another single layer fully connected neural network with a hyperbolic tangential activation function ($tanh$) that limits the output to a range between -1 and +1. This second layer determines the specific information to be added, while the first network determines *how much* of this new information will effectively be added to the long-term carry state c_t. Formally, the inputs are:

$$i_t = \sigma(b^i + W^i x_t + U^i h_{t-1})$$
$$\tilde{c}_t = \tanh(b^c + W^c x_t + U^c h_{t-1}),$$

with $\odot$ representing the Hadamard point-wise multiplication. After an LSTM cell receives the carry state from the previous time step, c_{t-1} and it goes through the forget gate and the input gate in sequence, it becomes the cell's current carry state, as such:

$$c_t = c_{t-1} \odot f_t + \tilde{c}_t \odot i_t$$

3. Output Gate (Labelled 3): This sigmoid network determines which part of the (now updated) memory is exposed at time t to be the LSTM cell's output. If the value is zero, no memory is exposed. If the value is one, the updated memory c_t is presented as the output h_t. Formally:

$$o_t = \sigma(\mathrm{b}^o + \mathrm{W}^o \mathrm{x}_t + \mathrm{U}^o \mathrm{h}_{t-1})$$

[8] LSTMs, combined with connectionist temporal classification (CTC) of Graves et al (2006), solved the speech-to-text and OCR problems in the early 2010s. A comprehensive introduction to RNN and LSTMs can be found in Olah (2015).

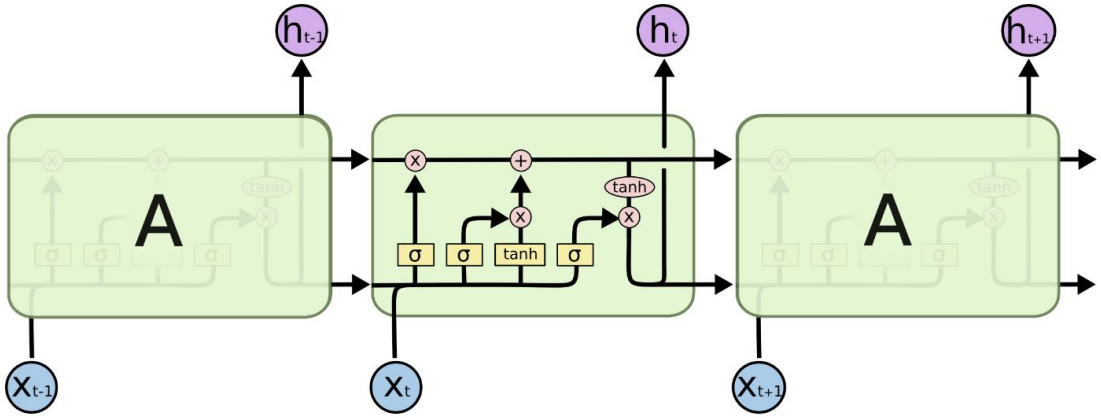
[9] While recent advances in AI are underpinned by another neural network, the Transformers (Vaswani et al (2017)), LSTMs remain a competitive technique for data of the scale typically used in economics and financial settings. RNNs (including LSTM networks) process data sequentially and are not well-suited for parallel processing and thus do not take full advantage of parallel computing power. Transformers resolve this issue by eliminating recurrency through positional encoding. This allows these networks to train on much larger amounts of data, underpinning their outperformance with respect to LSTMs.

$$h_t = o_t \odot \tanh c_t.$$

In all above regressions, $x_t \in \mathbb{R}^N$ is the column vector of N covariates, c_t, $h_t \in \mathbb{R}^{d_{LSTM}}$ are vectors of the same dimension $d_{LSTM} > 0$. Each element of $\{W^f, W^i, W^c, W^o\} \in \mathbb{R}^{d_{LSTM} \times N}$ are the parameter matrices related to contemporaneous inputs and each element of $\{U^f, U^i, U^c, U^o\} \in \mathbb{R}^{d_{LSTM} \times d_{LSTM}}$ are the parameter matrices related to the previous state. The elements of $\{b^f, b^i, b^c, b^o\} \in \mathbb{R}^{d_{LSTM}}$ are the column vectors of constants (known in the computer science literature as "biases"). And the operations described above by the Hadamard multiplication are independently applied to each vector component. The specific parameters used in our setting are described in the Annex [to be developed].

Scheme of an LSTM network

Figure 5



X_t represents the input to the RNN, C_t the hidden layer and memory, and h_t the output.

Source: (Olah; 2015)

3.2 Forecasts with RNNs

We directly forecast the 20-day average of the our TAP deviation measure (daily standard deviation of the minute-by-minute Euro-Yen TAP dislocation, or the difference between the direct and intermediated spot exchange rate) at each horizon $h \in \{20, 40, 60\}$ days.

This direct forecasting approach (as opposed to iterated forecasts) builds on a considerable literature on RNN-based forecasting. Many of these papers use two LSTM layer groups: one or more LSTM layers act as *encoders*, summarising known input into a summary vector that encodes the state at time *t*, and correspondingly *decoders* are LSTMs that estimate a future time series given this vector. For example, Wen et al (2017) encode the input data $x_t, x_{t-1}, \ldots, x_1$ into a vector μ_t, which is then used by densely connected neural layers to produce (quantile) forecasts at different horizons. Lim et al (2021) also encode the input data known into a state vector, which is then used to initialise the state vector of a decoder LSTM to produce forecasts. While not new in this regard, our approach in contrast explores more directly the recurrent nature of the LSTM by using a shifted time series, whereby data x_t is used to directly forecast y_{t+h} given a state h_{t-1}. This simplifies training and avoids having to cram all relevant information from complete history into a single vector, making for more efficient direct forecasts.

Our models are multi-variate, similar to most financial economics forecasting but in contrast to a large part of the ML literature, which focuses on univariate analyses (including text, which is an univariate

sequence of words or sentences). This implies that in high-dimensional settings, where the number of different variables is sufficiently large, there might be a need for the model to regularise in order to maintain good predictive ability. One way to regularise is akin to L1- or L2-norm regularisation (eg, lasso Tibshirani (1996) or ridge regression (Hoerl and Kennard (1970)): to constraint the signals from variables, letting through only those that are particularly salient. In other words, regularisation is done through a form of variable selection. However, these methods conduct *global variable selection*, ie the same variables are valid for all time periods. In contrast, our models perform *time-varying variable selection*: each period in time attributes a given set of weights to (potentially) different variables. This is close to the variable selection network proposed by Lim et al (2021), with the difference each variable's weights compete with each other as they must sum to one; the strength with which these weights respect to salient features is given by a "temperature" parameter, which is learned by the network. Chen et al (2018), Yoon et al (2019) and Arik and Pfister (2021) also use a model of instance-wise attention, albeit for each unit instead of only along a time series as in our model.

3.3 Prediction architecture

The purpose of our model is to map an extensive panel of daily financial market variables into h-days ahead predictions of the market anomalies for a window of w days. In what follows, we illustrate the technique with an example using $h = 60$ and $w = 20$, that is, a model that forecasts the average of the 20-day window that starts 60 days into the future. We use a novel architecture inspired by insights that financial market variables can interact nonlinearly, and that these interactions (and the importance of each variable) can change over time. The model also flexibly deals with seasonality. Importantly, in addition to the predictions, the network offers interesting by-products such as the time series of variable importance.

The architecture of the network is shown in Figure 6 below. Each rectangle is a different layer of the network. Starting from the top towards the bottom, the network takes in two types of data: a time series of two categorical variables (quarter and a dummy for days around quarter-ends) and a time series of 770 continuous variables representing financial market series and their variations (eg, long differences). The former data enters the network through the *InputsCat* layer and the latter through the *InputsCont* layer. The data is then sequentially processed by the network, as shown in the directed graph, eventually producing the value to be predicted by the layer named *Output* at the bottom. Rather than going through each layer, it can be instructive to analyse its building blocks separately. These are as follows:

Temporal features. Seasonality can play a key role in financial market series. The variables of interest in this particular case, for example, display considerable changes at end-of-quarter periods. In addition, there might be other, less obvious ways in which seasonality can play a role. For this reason, we use embed each quarter $Q_t \in \{1, 2, 3, 4\}$ and a quarter-end dummy $QEnd_t \in \{0, 1\}$, defined as 1 for the five days before and five days after quarter-ends. These embeddings map each possible value of Q_t and of $QEnd_t$ to a real-valued vector of dimension $d_{TempFeat}$, as such: $M_Q : Q_t \to \mathbb{R}^{d_{TempFeat}}$. In this case, we use $d_{TempFeat} = 4$. This is done in the layers *Embedding_quarter* and *Embedding_quarter_end*, respectively. They result in two time series of embeddings, one of each time variable. They are then concatenated (*ConcatTempFeat*) to result in a single vector time series of dimension $2 \times d_{TempFeat}$. An LSTM layer (*TemporalFeatures*) encodes this time series into a more flexible, recursive representation that summarises all temporal features, now with dimension d_{LSTM}. Following the downstream arrows, these temporal features inform the decision on the variable specific-weights, ω_t, as well as the final prediction.

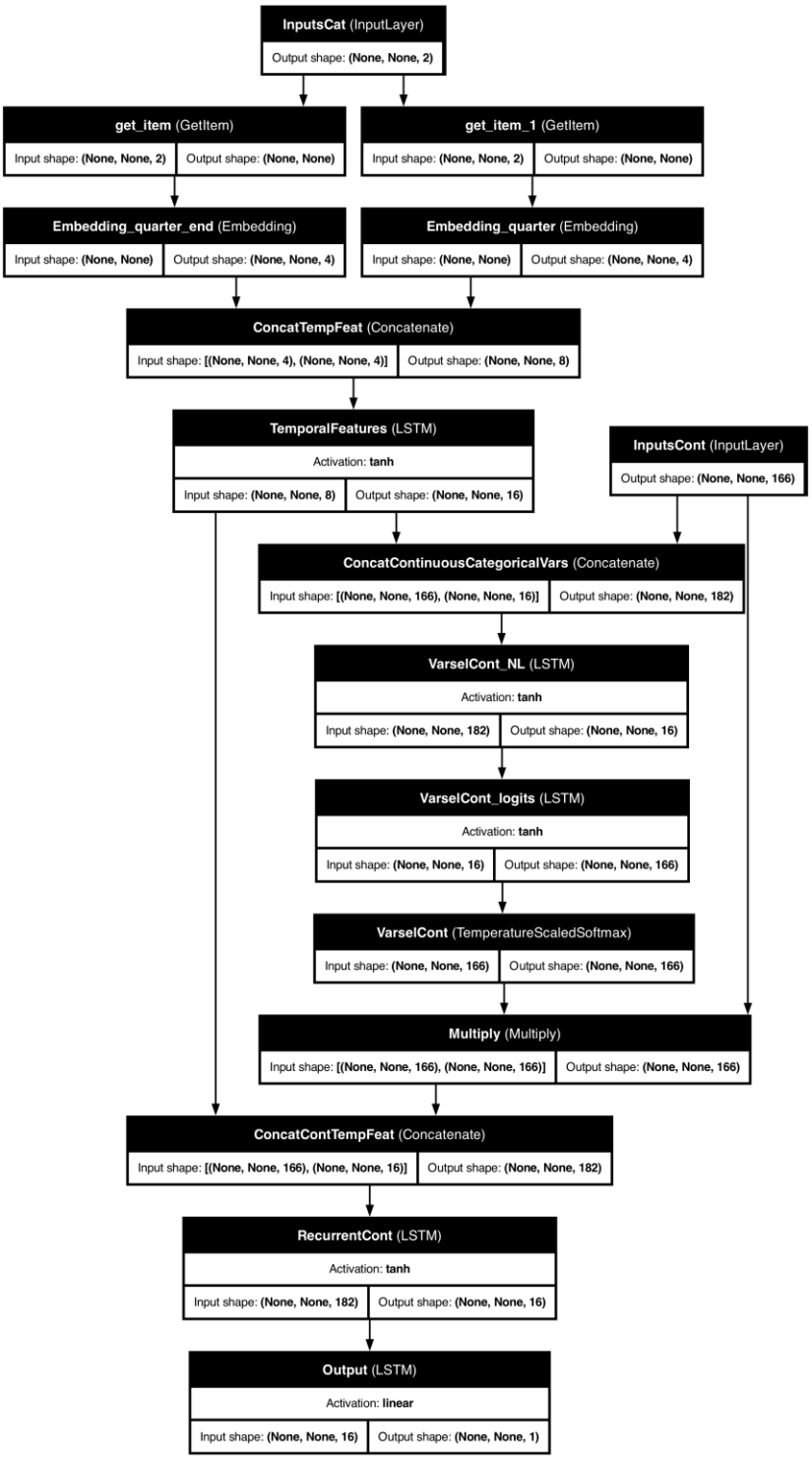
Competitive variable selection. The vector time series of temporal features is concatenated with the time series of the N different financial market variables into a single, large time series vector with dimension $d_{TempFeat} + N$ (*ConcatContinuousCategoricalVars*). This large panel time series is then fed to an LSTM layer (*VarselCont_NL*) that combines nonlinearly all these signals. At the end of this process, the

network combines the information into an *N*-dimensional vector (*VaselCont_logits*). Then at each point in time, these N values undergo a softmax transformation, which maps each result into a positive number, with all summing to one. The "temperature" or smoothness of this transformation is a learned parameter (*VarselCont*). The result is a time series of equal dimension to the financial market inputs, but with weights between 0 and 1 for each variable. These weights change over time, depending on the data up until that time (due to recurrent nature of the *VarselCont_NL* layer). Further, the weights are "competitive" in the sense that a higher weight for one variable necessarily implies lower weights for the others, as they all must sum to one in each period. The temperature parameter determines just how competitive this process is (lower values make it more competitive). All variables are normalised, thus the weight is independent of each variable's scale. These time-varying, variable specific weights are then multiplied by each variable's series (*Multiply*), resulting in competitively weighted variables. For some variables, the weight might be very close to zero.

Final transformation. Once weighted, the variables pass through another LSTM layer with nonlinear activation to extract increasingly refined features to predict the anomaly in question, and another LSTM with a linear result, which serves to compose the final result without having its range constrained by the nonlinearity.

Layers of one version of the prediction network

Figure 6



The order of the dimensions in the shapes of each layer is: sample, time dimension, feature dimension. "None" means that it is unspecified.

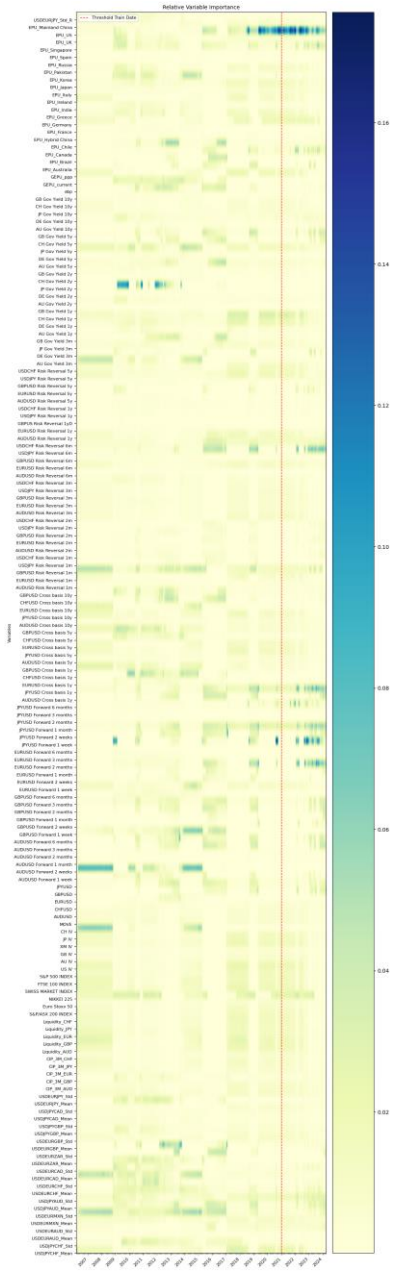
Source: Authors.

The competitive variable selection component offers an interesting by-product: the possibility of inspecting how the network weights the variables at each point in time. As described above, the network learns how to look at variables up until period t to select variables that best help it predict anomalies at $t + h$. In other words, due to the recurrent nature of the network, each weight only depends on the financial market variables data up until period. By passing the data through the network *up until the VarselCont layer only*, the result is a time series of variable-level weights. Inspecting these weights allows an inspection of which signals the network think are more informative for prediction at each point in time. Further, the competitive selection instead of a variable-level gating helps the model strongly regularise, which is important in a high-dimensional context, especially due to the high level of correlation between variables (Giannone et al (2022)).

The result of this process is seen on Figure 7: every row is a different input variable, and the horizontal axis represents the time dimension; the red dashed line shows when the network training ended. Darker colours correspond to variables with higher weight in each particular date. One interesting observation is that the network strongly prefers to regularise: the temperature with which the network settled is 0.27 (higher numbers smoothen the distribution, lower numbers sharpen it). Another advantage is the possibility to directly inspect the variables that matter at each time, for each prediction horizon.

In fact, inspecting the different weights assigned to each variable over time might provide the econometrician with some clues as to the underlying process, or at least point to other data whose movements might be correlated with the market dysfunctions of interest. For example, after the Global Financial Crisis (GFC), Japanese government yields were a key covariate for this model, while more recently this variable has been downplayed in favour of economic policy uncertainty in the United States, a strong driver of the forecasts since around the Covid-19 pandemic.

Variable selection weights to predict TAP deviations in the 4th month ahead — Figure 7



4th month ahead: 20-day average (approximately one month) that begins after a 60-day forecasting horizon (approximately 3 months).

Source: Authors.

3.4 Forecast evaluation

The models' out-of-sample forecasts are compared to a corresponding auto-regressive (AR) model using the Diebold-Mariano (DM) test (Diebold and Mariano (1995), Diebold (2015)), using the root mean squared error as a loss metric. Formally, let l represent the size of the context used to train LSTM models, h the direct forecasting horizon and w the number of days over which the dependent variable is averaged (all three measured in days); then the forecasts using our model (generically named G) and the corresponding AR model are:

$$\hat{y}^{G}_{t+h,w} = G^{l}_{h,w}(x_t) + \epsilon^{G}_{t};$$

$$\hat{y}^{AR}_{t+h,w} = \alpha + \sum_{i=0}^{l} \beta_i \, y_{t,w} + \epsilon^{AR}_{t},$$

with the subscript w in $y_{t,w}$ a short-hand notation for an average over w consecutive days that begin in t.

We evaluate these forecasts with a symmetric loss function, namely the root mean squared error, to assess the null hypothesis that these two forecasts are equal. The key idea is that forecasting models should be able to forecast the underlying metric well enough to at least statistically match an AR model's forecast at a symmetric, balanced metric such as RMSE. Once this is achieved, any ability to identify in advance periods of market dysfunction can be assigned to a reliable, unbiased forecasting of the data generating process (or at least, as biased as an AR would be). In other words, if a model was optimised to blindly estimate periods where market values signal dysfunction, the underlying prediction model would be dependent on forecasting errors and on the econometrician model error (which itself could be biased if the rule is heuristically simple, for example).

While this test provides assurance on the forecasts, especially considering the high bar that is set in terms of the variable itself and the benchmark, a question arises about the ability of the forecasts to adequately capture the *dynamic* of the target value. Of particular concern is to fairly measure the scenario where the neural network predictions result in smooth forecasts – by virtue of the model's design, which tends to abstract away from higher-frequency noise. In fact, such smooth forecasts can even be considered better from a user perspective, avoiding see-saw movements that might overburden monitoring processes. One way to directly measure the ability to forecast the TAP violations even in a scenario where the predictions end up estimating the target variable's "trend"-like series, is to regress the target on the contemporaneous prediction. A statistically significant loading on the forecast would represent an additional confirmation of the relevance of the estimates.

4. Data

For our analysis we focus on one of the most important markets in the world, namely the foreign exchange market. Beyond the deep liquidity with a daily turnover of trillions of US dollars (Drehmann and Sushko (2022)), foreign exchange-based metrics offer a clean way to assess market dislocations, since these trades are typically not subjected to taxes, are not hampered by restrictions to short-selling or other intraday regulatory limits, and do not rely on a convergence to parity (Pasquariello (2014)).

Our main source of data is LSEG Tick History. Tick History provides historical information at different levels of aggregation (from tick by tick to daily) for a number of asset classes. Its coverage goes as far back as 1996 for many times series and is widely used by industry practitioners. The dataset consists of recorded trades and quotes from a number of real-time feeds across more than 500 trading venues including all types of participants (i.e., not just dealers).

We obtain 1-minute snapshots of bid and ask quotes and calculate the mid point, as well as the bid-ask spread and its higher moments. We then average these high frequency observations to obtain daily observations. We clean the data by only using local trading hours (9:30 to 16:00) and removing weekends.[10] We also eliminate data when the order book is crossed (i.e., ask price is lower than or equal to the bid price).

[10] When training the models and predicting we also remove end of quarter data, as these exhibit a seasonal pattern of lower liquidity including eg due to window dressing (Garcia et al (2021)), which is not relevant for our purposes.

Our measure of financial market dislocation is the extent to which triangular arbitrage parity (TAP) fails to hold (Lyons and Moore (2009), Pasquariello (2014)). In a perfectly functioning market, the exchange rate between – say – the Euro and the Japanese yen should be equal to the exchange rate between the Euro and the US dollar multiplied by the exchange rate between the US Dollar and the Japanese Yen. If this is not the case, then there are frictions in the market that prevent arbitrageurs from realigning the prices of the various currencies. The larger the deviation between the *theoretical* exchange rate and the observed exchange, the bigger the dysfunction.

We study TAP violations in a representative currency pair, although our approach can easily be extended to other pairs. Specifically, we study the EUR-JPY with the USD as the vehicle currency. As seen on Table 2, this pair is markedly more volatile than others, even if it relates to two widely-traded currencies from advanced economies and a number of globally active banks. Other pairs of relevance, and lagged values of the pair of choice, are used as explanatory variables, namely: EUR-AUD, EUR-CAD, EUR-CHF, EUR-GBP, EUR-JPY, EUR-MXN, EUR-ZAR, JPY-AUD, JPY-CHF and JPY-GBP with the USD as the vehicle currency, and JPY-AUD, and JPY-GBP using the EUR as the vehicle currency. Note that these pairs include both advanced and emerging market economies, two currencies that are usually considered safe havens (CHF and JPY) and also advanced economy currencies that are usually associated with commodities (AUD and CAD).

For each of these currency pairs and vehicle currency, we calculate the daily standard deviation of the minute-by-minute difference between the actual currency pair exchange rate and the theoretical rate using the vehicle currency. These data are described in Table 2 below.

Triangular arbitrage parity violations

Daily standard deviation of minute-by-minute arbitrage deviations — Table 2

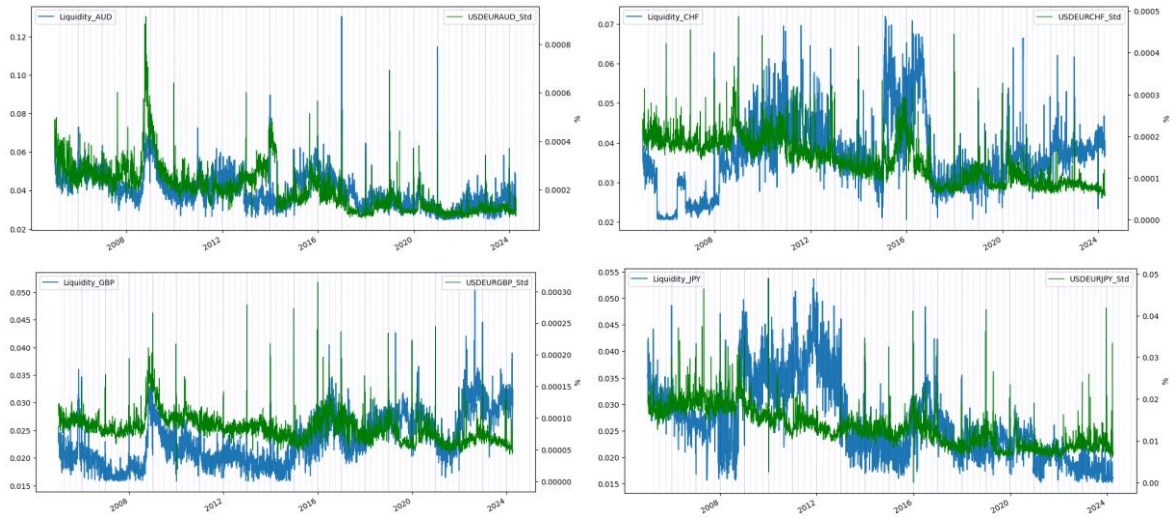
Currency pair (vehicle currency)	Count	Mean	Std. dev.	Min	25%	50%	75%	Max
EUR-AUD (USD)	4513	0.0002	0.0001	0.0001	0.0002	0.0002	0.0002	0.0009
EUR-CAD (USD)	4513	0.0002	0.0001	0.0001	0.0001	0.0002	0.0002	0.0005
EUR-CHF (USD)	4513	0.0001	0.0001	0.0001	0.0001	0.0001	0.0002	0.0004
EUR-GBP (USD)	4513	0.0001	0.0000	0.0000	0.0001	0.0001	0.0001	0.0003
EUR-JPY (USD)	4513	0.0128	0.0045	0.0056	0.0091	0.0122	0.0155	0.0465
EUR-MXN (USD)	4513	0.0036	0.0016	0.0000	0.0027	0.0033	0.0042	0.0135
EUR-ZAR (USD)	4513	0.0038	0.0020	0.0000	0.0026	0.0033	0.0045	0.0274
JPY-AUD (USD)	4513	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JPY-CHF (USD)	4513	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JPY-GBP (USD)	4513	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JPY-AUD (EUR)	4513	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
JPY-GBP (EUR)	4513	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Source: LSEG Tick History, authors' calculations.

We include as predictor variables a series of bid-ask spreads to represent a currency pair's liquidity, as well as a market-wide measure of dislocation based on an unweighted average of all the daily TAP deviation standard deviations and the bid-ask spreads (all are described in Table 3). Our measure of TAP deviations tracks well also the evolution of liquidity, as seen on Figure 8.

Liquidity and TAP deviations by currency

Figure 8



Top left panel: USD-AUD bid-ask spread (blue; left axis) and daily std. dev. of minute TAP deviations for EUR-AUD with USD as vehicle currency (green; right axis). Top right panel: USD-CHF bid-ask spread (blue; left axis) and daily std. dev. of minute TAP deviations for EUR-CHF with USD as vehicle currency (green; right axis). Bottom left panel: USD-GBP bid-ask spread (blue; left axis) and daily std. dev. of minute TAP deviations for EUR-GBP with USD as vehicle currency (green; right axis). Bottom right panel: USD-JPY bid-ask spread (blue; left axis) and daily std. dev. of minute TAP deviations for EUR-JPY with USD as vehicle currency (green; right axis).

Source: LSEG Tick History; author's calculations.

Our selection of potential predictors also includes adapted variables that have proven to be useful when used in machine learning models for prediction of returns (Gu et al (2020)). After all, a price dislocation is simply a difference between two prices and therefore the factors determining prices are likely to be important for dislocations too. Hence in each model the explanatory variables include lags of all the TAP deviations in Table 2 and the liquidity variables in Table 3. Conversely, while the volatility of bid-ask spreads over a rolling window of time, and momentum variables for both the TAP deviations and liquidity are known to be able to forecast financial variables, we did not include them as the recurrent nature of the model should have the ability to calculate these time-wise moments from the data if they are indeed relevant.

Foreign exchange bid-ask spreads

As a percentage of the bid-ask midpoint

Table 3

Currency pair	Count	Mean	Std. dev.	Min	25%	50%	75%	Max
USD-AUD	4513	0.0388	0.0083	0.0247	0.0329	0.0374	0.0442	0.1306
USD-CHF	4513	0.0375	0.0089	0.0207	0.0316	0.0365	0.0414	0.0719
USD-EUR	4513	0.0223	0.0054	0.0134	0.0181	0.0218	0.0260	0.0819
USD-GBP	4513	0.0232	0.0046	0.0157	0.0195	0.0224	0.0267	0.0518
USD-JPY	4513	0.0262	0.0073	0.0152	0.0206	0.0243	0.0307	0.0537

Source: LSEG Tick History, authors' calculations.

We include the excess bond premium (Gilchrist and Zakrajšek (2012)) as an explanatory variable given its information-rich content on credit original sentiment, which also correlates with business cycle movement in the quarters ahead. In fact, the co-movement between the excess bond premium and our preferred measure of market dysfunction from the TAP deviations is striking, even though the excess bond premium is calculated from a completely different dataset and relates to the US corporate bond market (see Figure *9* below). Since this series has a monthly frequency, its values are repeated daily within the reference month.

In addition, each model included as explanatory variables temporal features to better fit any dynamics that are related to specific dates of a week, month, quarter or year (described in Table 4). These variables capture variations in liquidity conditions that are seasonal or periodic, as well as any factors around quarter-ends that could create or amplify liquidity shortages (eg, simple seasonal festivities or more serious window-dressing concerns as in Garcia et al (2021)).

Temporal features

Table 4

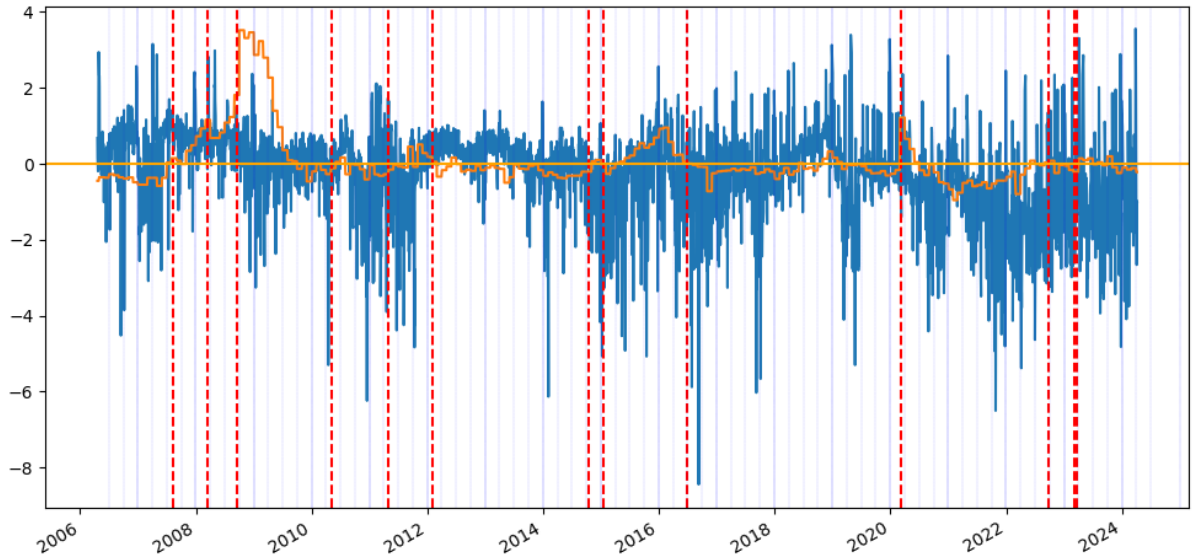
Variable	Count	Mean	Std. dev.	Min	25%	50%	75%	Max
Day of week	4513	2.005	1.413	0	1	2	3	4
Month of year	4513	6.529	3.446	1	4	7	10	12
Period around quarter-ends	4513	0.120	0.325	0	0	0	0	1

Source: Authors' calculations.

Our dependent variable is a transformation of the daily mean of the 5-minute deviations of TAP. Namely, we use the natural logarithm of its absolute value. This transformation ensures the model is capturing deviations in either direction of the trade, since both equally offer arbitrage opportunities. This transformed value serves as the dependent variable for the ML prediction models. Figure *9* presents this series, along with the excess bond premium and dates of market events.

TAP deviation of Euro-Yen pair and excess bond premium

Figure 9



Blue: log(abs(Euro-Yen TAP)); orange: excess bond premium. Red lines, in sequence: 9pct fall in Nasdaq (14 Apr 2000), BNP Paribas freezes US funds (9 Aug 2007), demise of Bear Sterns (16 Mar 2008), Lehman Brothers bankruptcy (15 Sept 2008), stock market flash crash (6 May 2010), first "local" peak of Greek sovereign yield (May 2011), "global" peak of Greek sovereign yield (Feb 2012), US Treasury flash event (15 Oct 2014), Swiss Franc-Euro FX rate cap (15 Jan 2015), Brexit (23 Jun 2016), Covid-19 declared pandemic (11 Mar 2020), LDI crisis (28 Sept 2022), SVB (10 Mar 2023), UBS buys CS (24 Mar 2023).

Source: LSEG Tick History; Favara et al (2016); author's calculations.

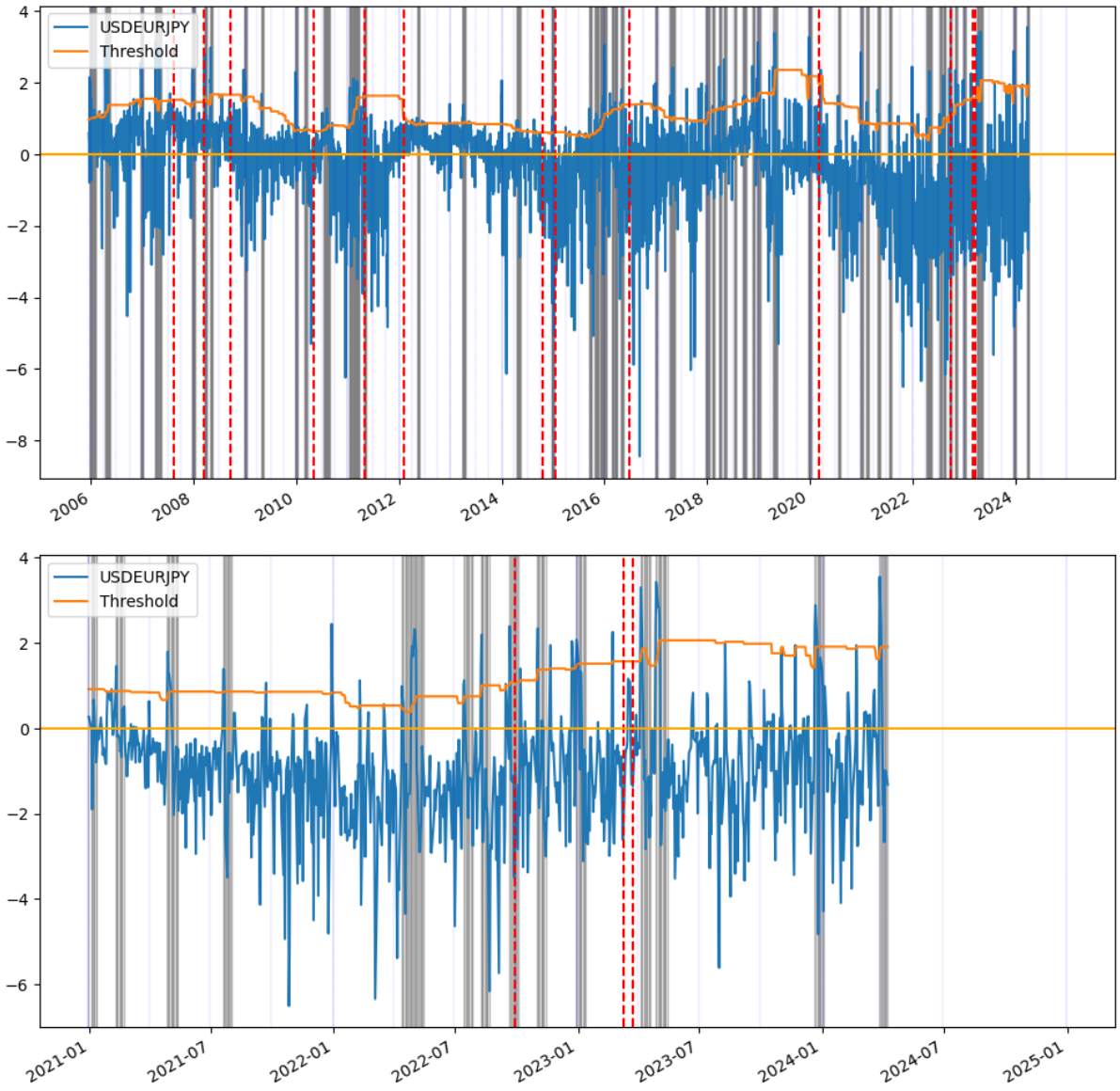
The dependent variables and all continuous variables (ie, except for the temporal features) are normalised to a zero-mean, unit-standard deviation variable, following common practice in neural network training.

We also use the (transformed) dependent variable to create a series of "yellow flags" or signs that a potential liquidity shortage may occur in specific markets. The reason to calculate such a variable and not directly use known market events is the latter are rare. Another reason is to test how well the model identifies latent or simmering liquidity shortage periods. This definition is based on a simple heuristics to establish the periods of potential alert for deteriorating market liquidity conditions. Using a rolling window of 250 days up to a date *t*, we find the 95% percentile of that window and compare the value at *t* with that threshold. If the value is higher than the threshold for 2 or more consecutive days, we consider the whole window of time between the first *t* above threshold until at least 10 days later to be a single event. The result is shown in Figure 10, where the grey bars represent such periods.

Note from Figure *10* that while many such adverse liquidity periods match well with known market events, they do not necessarily overlap. Also, by virtue of the definition we use, the identified periods tend to be more frequent – again, by design, as they are supposed to proxy for periods of latent dysfunctional liquidity. Another interesting observation is that these periods tend to cluster over time.

Identified periods of adverse market liquidity

Figure 10



Grey bars represent the periods identified as potentially dysfunctional. Both panels have the same data; the bottom panel focuses on the out-of-sample period (from Jan 2021 onwards).

Source: LSEG Tick History; author's calculations.

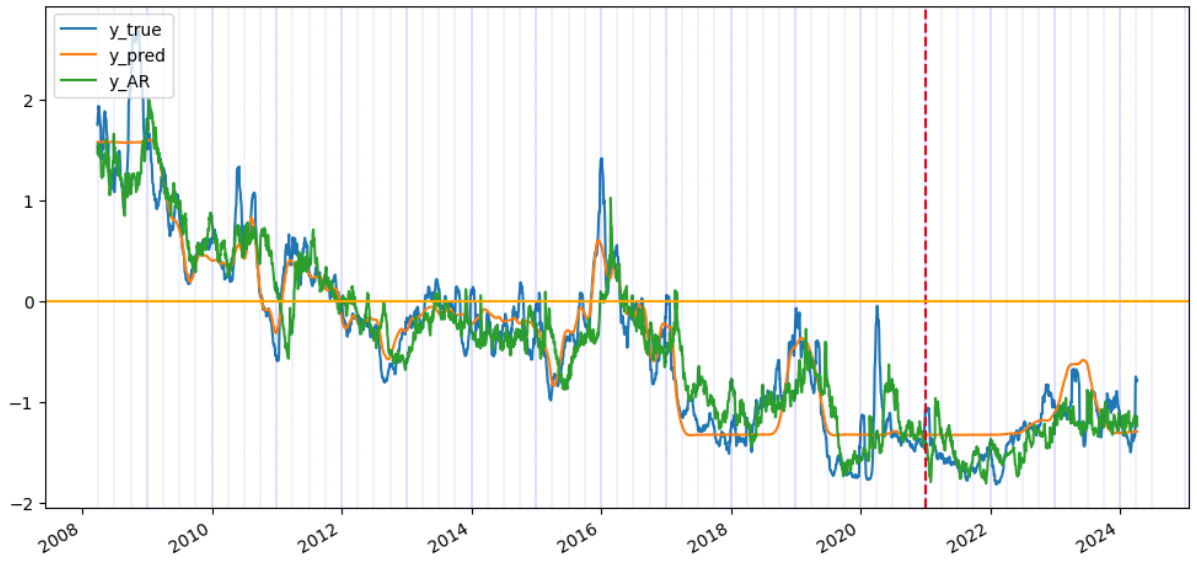
5. Preliminary findings[11]

This section introduces the preliminary findings from the forecasting exercise, the identification of liquidity periods of interest, and the use of LLMs for summarisation of text content that could be related to events of interest.

5.1 Market dysfunction prediction

Relatively simple LSTM models, regularised by the dynamic variable selection feature, perform competitively with respect to an AR benchmark, although slightly different configurations along the dimension size perform better at different horizons. Figure 11 presents the results for the 20-day average of Euro-Yen TAP violations, forecasted with 20, 40 and 60 days in between, respectively.

Performance of LSTM model across different forecasting horizons — Figure 11



[11] The current version of this work only studies the Euro-Yen TAP; other currency pairs and potentially other market dysfunction metrics are the topic of current development.



Top panel: h=20, w=20. Centre panel: h=40, w=20. Bottom panel: h=60, w=20.

Source: author's calculations.

These findings imply that pricing deviations are indeed forecastable to some extent, and that exploring the non-linearities in the cross-sectional and along the time dimension might be key to extract this signal. It is plausible that a wider search along the parameter space of these models might yield even better performances, although that remains an empirical question.

The question then arises about the reason why such price deviations can be forecasted. While exploring this question more fully is outside the scope of this work, the results in Figure 11 suggest some informal exploration of potential channels. One potential reason for the predictability of these arbitrage violations is that it could depend on the state of large financial intermediaries. Lending credence to this channel are results such as Haddad and Muir's (2021), demonstrating that the returns from assets that are more intermediated tend to be more predictable.

Along these lines, it is worth focusing on the completely out-of-sample performance in key recent periods. There are signs that these models were picking up higher probability of market dysfunction in the

lead up to September 2022 LDI event in the UK and especially in the lead up to the March 2023 banking stresses. This even after the network has not been trained in *years* worth of data. In contrast, periods such as Covid-19 were essentially not forecasted by the model, consistent with this being an event that started from completely outside the financial system.

5.2 Estimating periods of lower liquidity

While the forecasting model is competitive with respect to the AR benchmark on average, as measured by the RMSE, it is the forecasts' tail outcomes what counts for its use in foreseeing adverse liquidity events. In this section, we report the performance of the models discussed in Section 5.1 to identify potential liquidity events, ie the greyed periods in the bottom panel of Figure 10.

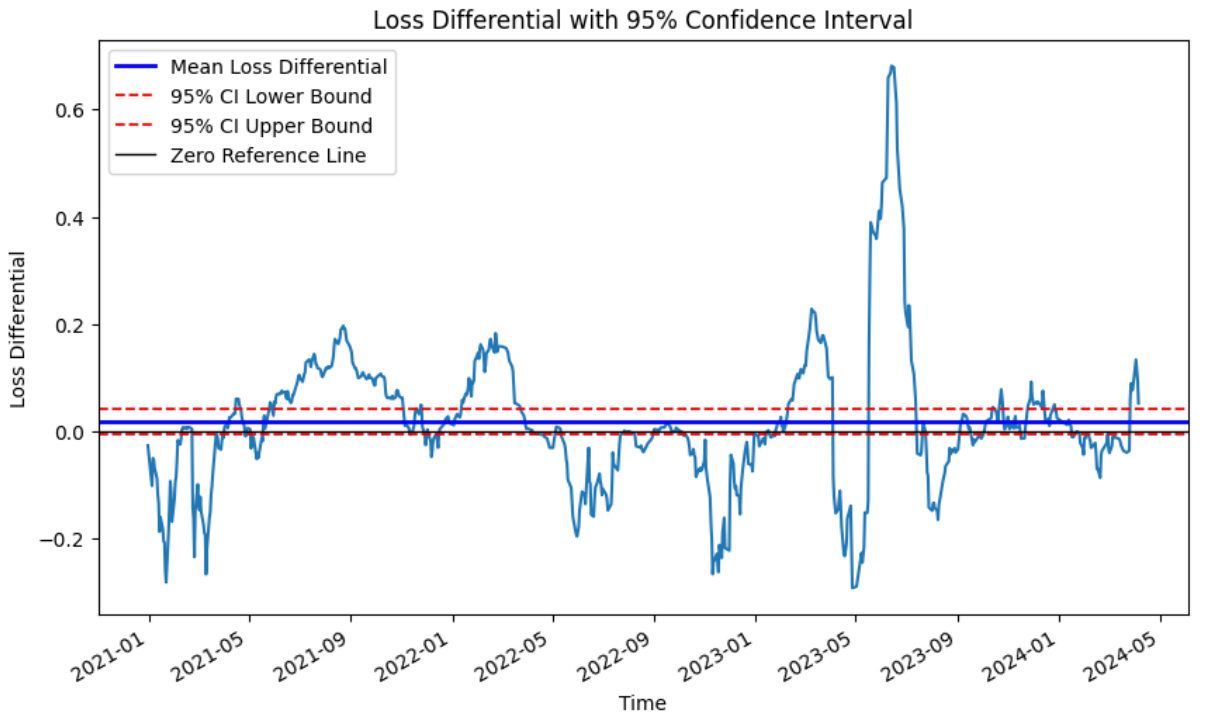
[Work in progress]

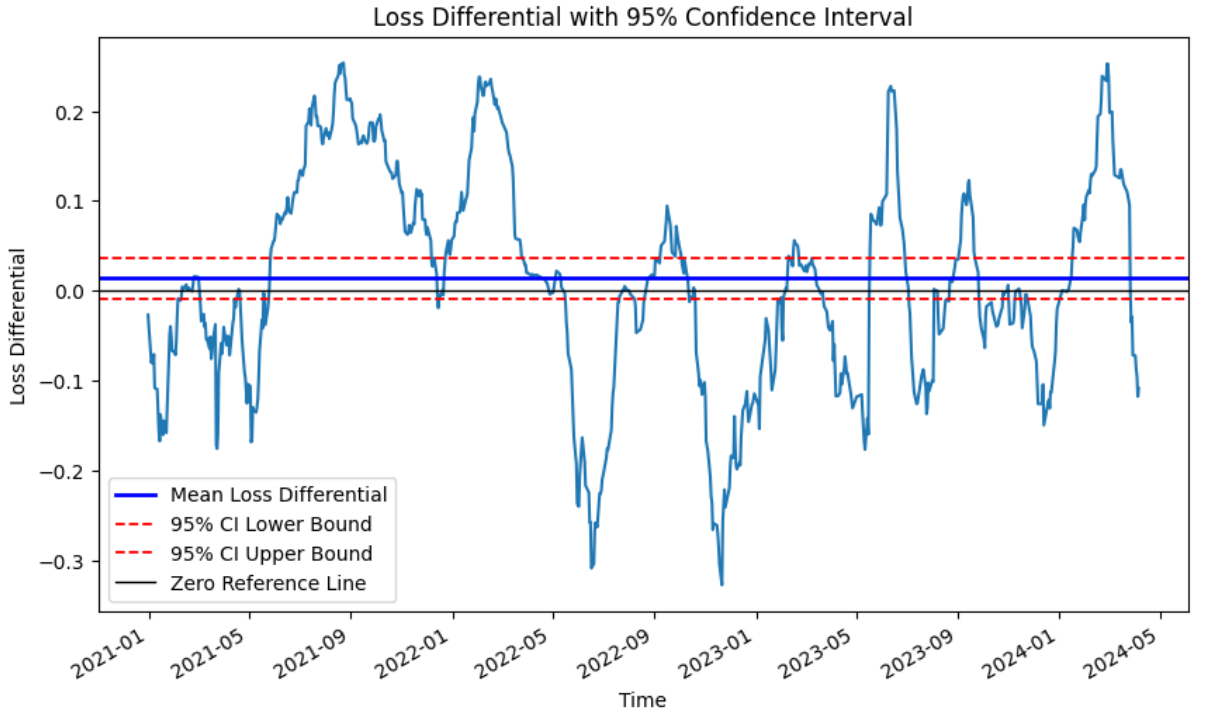
5.3 Formal evaluation of predictive power

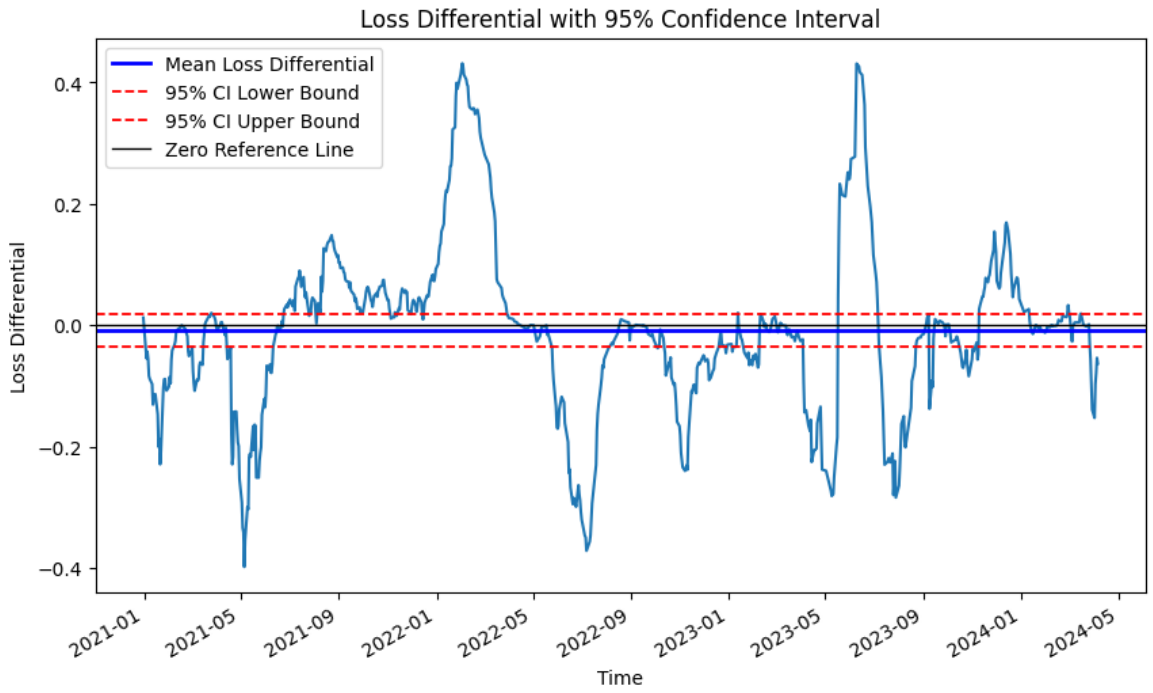
We compare the forecasting ability of our neural network using the RMSE formula as a loss function. Comparing the results against an autogressive model (AR) that uses the same offset h as the network, and whose lags are the same as the training context window,[12] We employ the Diebold-Marino test due to Diebold and Mariano (1995) and further explained in Diebold (2015) (henceforth DM) on out-of-sample predictions to avoid comparing potentially overfitted results. This test intuitively compares the difference between two prediction loss functions to a normal distribution under the null hypothesis that they are statistically the same. Due to the random walk nature of financial market variables, we consider results that either match or beat the AR benchmark to be satisfactory.

Diebold-Mariano test comparing predictions with RNNs and AR — Figure 12

[12] In practice, this results in an almost similar result as using only one lag, but is used for fairness of comparison since longer-lagged coefficients might capture richer seasonality structures that would elude a single-lag AR model.


Loss Differential with 95% Confidence Interval
Mean Loss Differential
95% CI Lower Bound
95% CI Upper Bound
Zero Reference Line
Loss Differential
0.6
0.4
0.2
0.0
−0.2
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
2023-09
2024-01
2024-05
Time


Loss Differential with 95% Confidence Interval
Mean Loss Differential
95% CI Lower Bound
95% CI Upper Bound
Zero Reference Line
Loss Differential
0.2
0.1
0.0
−0.1
−0.2
−0.3
2021-01
2021-05
2021-09
2022-01
2022-05
2022-09
2023-01
2023-05
2023-09
2024-01
2024-05
Time



Mean loss and confidence interval bounds calculated with the Diebold-Mariano test using Newey-West standard errors and root mean squared error loss function. Top panel: h=20, w=20. Centre panel: h=40, w=20. Bottom panel: h=60, w=20.

Source: Authors.

With respect to the regressions, the forecasts from the models for each of the three forecasting periods are shown to serve as statistically significant approximators for the target variable. The results in Table 5 demonstrate that not only the loading of the future TAP deviation on the prediction is statistically significant, it explains a meaningful portion of the former – substantially more so for shorter-term forecasts but a still meaningful 12.0% R2 for the 60-day ahead forecast. These analyses confirm the relevant of the predictions obtained with RNNs.

Forecast loading test comparing predictions with RNNs and AR Table 5

OLS Regression Results

Dep. Variable:	y_true	R-squared:	0.482
Model:	OLS	Adj. R-squared:	0.482
Method:	Least Squares	F-statistic:	667.2
Date:	Tue, 10 Dec 2024	Prob (F-statistic):	1.81e-104
Time:	22:22:00	Log-Likelihood:	126.33
No. Observations:	718	AIC:	-248.7
Df Residuals:	716	BIC:	-239.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2342	0.061	3.845	0.000	0.115	0.354
y_pred	1.3583	0.053	25.829	0.000	1.255	1.462

Omnibus:	27.960	Durbin-Watson:	0.023
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17.793
Skew:	0.250	Prob(JB):	0.000137
Kurtosis:	2.412	Cond. No.	16.2

OLS Regression Results

Dep. Variable:	y_true	R-squared:	0.120
Model:	OLS	Adj. R-squared:	0.119
Method:	Least Squares	F-statistic:	14.24
Date:	Wed, 11 Dec 2024	Prob (F-statistic):	0.000174
Time:	00:29:21	Log-Likelihood:	-63.988
No. Observations:	718	AIC:	132.0
Df Residuals:	716	BIC:	141.1
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.5226	0.219	-2.388	0.017	-0.952	-0.093
y_pred	0.6332	0.168	3.773	0.000	0.304	0.963

Omnibus:	137.521	Durbin-Watson:	0.013
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.004
Skew:	-0.022	Prob(JB):	5.03e-07
Kurtosis:	2.016	Cond. No.	17.0

Notes:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using

Top panel: h=40, w=20. Bottom panel: h=60, w=20.

Source: Authors.

5.4 Generative AI as an interpretative and 'early warning' tool

Generative AI and LLMs can be used to assist supervisors and regulators in validating and interpreting the anomalies predicted using AI and ML models. This process also aids in assessing the potential causes of future dislocations, based on a real-time search of financial news and other relevant information.

Identifying probable predictors of market dislocations is beneficial for supervisors, even without an assessment of what the potential causes of such future dislocations. Supervisors could use this information to prepare for ex post intervention, preventing more severe consequences. However, an evaluation of the likely underlying causes of such dislocations would enhance the tool's usefulness even further. Supervisors would then be able to address the causes of such dislocations, preventing situations from deteriorating before they occur. By leveraging the capabilities of LLMs to sift news in real time and aggregate information, supervisors could have access to potential explanations and decide on an appropriate course of action.

To illustrate this potential, we tested the capacity of generative LLMs using the recent global financial market disruption on 5 August 2024 (Aquilina et al (2024)). The goal was to determine whether generative LLMs can identify critical news concisely, ensuring that financial supervisors are aware of emerging risk signals. For this purpose, we compiled 200-300 financial market news articles per day from a globally renowned market news source for the two business days leading up to the disruption. The generative LLM was then tasked with summarising the most critical news items that a financial supervisory institute should monitor to detect potential upcoming market dysfunctions, such as liquidity shortages.[13] The LLM was instructed to condense this information into three bullet points, including the reasons and relevance for each. The table below summarises the results.

1 August 2024	2 August 2024
1. **US Treasury Yields and Basis Trade Concerns:** U.S. Treasury yields experienced a significant drop, with two-year yields falling 17.8 basis points—the largest daily decline since December 2023. This movement is linked to the Fed's potential interest rate cuts and a record volume of basis trades involving Treasury futures and overnight repo markets. The extreme positions in these markets, valued at over $1 trillion each, pose a substantial risk if unwound suddenly, potentially leading to severe liquidity shortages and market disruptions. The highly leveraged nature of these trades could trigger instability in the broader financial system if market conditions shift abruptly, necessitating close monitoring by supervisory authorities. 2. **Bank of Japan's Hawkish Stance and Yen Volatility:** The Bank of Japan's (BOJ) shift towards a more hawkish policy stance, including potential further rate hikes, has already led to significant volatility in the yen, impacting global carry trades. The BOJ's actions could lead to further strengthening of the yen, disrupting global markets that have heavily relied on the yen as a funding currency. The yen's volatility could trigger a rapid unwinding of carry trades, leading to substantial market fluctuations and liquidity challenges, particularly in emerging markets and other highly leveraged financial sectors. 3. **Bank of England's Rate Cut and Potential Policy Reversal:**	1. **Sharp Decline in U.S. Treasury Yields and Aggressive Rate Cut Expectations:** U.S. Treasury yields, particularly the two-year and ten-year, experienced significant drops, with the two-year yield falling below 4% for the first time since May 2023. This decline is driven by weaker-than-expected employment data and increasing market expectations of aggressive interest rate cuts by the Federal Reserve. The market is now pricing in a 120 basis point cut by the end of the year, which could lead to increased volatility and liquidity concerns in the bond market. Sudden shifts in yield curves and rate expectations can cause liquidity strains and lead to disorderly market conditions, particularly if leveraged positions in Treasury futures and overnight repo markets begin to unwind. 2. **Emerging Markets and Asian Equity Market Volatility:** Emerging Asian markets, particularly South Korea and Taiwan, saw significant declines, with indices dropping 4.4% and 3.7% respectively. This sell-off was triggered by fears of a U.S. economic slowdown and was exacerbated by the Fed's potential end to its restrictive policy stance. The broader impact included a significant decline in the MSCI International All Asia Pacific Index. Market volatility in these regions can lead to capital outflows, currency instability, and broader financial instability, which could spill over into other emerging markets and potentially affect global financial stability. 3. **Increased Bearish Sentiment Among Hedge Funds:**

[13] The prompt is: "Attached is a collection of financial market news for a specific day. Summarise the most critical news for a financial supervisory institute to monitor any potential market dysfunction (eg shortage of liquidity) in the coming days. Summarise in three bullets according to the importance and relevance. Briefly explain why they are important."

The Bank of England (BoE) recently cut interest rates from a 16-year high, signalling a cautious approach to future cuts. This decision reflects growing concerns about the UK economy's resilience and inflation dynamics. The BoE's actions could influence market expectations and create volatility in UK financial assets, with potential spillover effects on global markets. The BoE's policy decisions may impact investor confidence and market stability, especially if future rate cuts or reversals are perceived as insufficient to address underlying economic weaknesses. Monitoring the BoE's communications and market reactions will be crucial.	Global hedge funds have been increasingly adding bearish bets, particularly in equity markets, driven by fears of a faster-than-anticipated U.S. economic slowdown. This sentiment was reinforced by recent weak economic data, leading to a correction in the Nasdaq Composite and broader equity markets. Hedge funds reduced exposure across multiple sectors, signaling a shift towards a more defensive posture. A continued shift towards bearish sentiment among major financial players could lead to increased market stress and a tightening of liquidity, particularly in riskier asset classes. Supervisory bodies need to monitor these developments to mitigate potential systemic risks.

The results show that the generative LLM successfully identified critical and relevant news that, based on information available before the event, are now known to have triggered the disruption. For instance, the LLM pinpointed an aggressive shift in policy stance by major central banks in response to indications of a U.S. economic slowdown, rising market volatilities and bearish market sentiments. Moreover, it accurately explained that the Bank of Japan's hawkish stance could lead to a further strengthening of the yen, potentially disrupting global markets that heavily rely on the yen as a funding currency. This, in turn, could trigger a rapid unwinding of carry trades—a central issue widely discussed in the aftermath of the disruption. Notably, this task is accomplished without any prior training (ie zero-shot learning). By utilising an information retrieval system, the LLM is able to combine its extensive knowledge base with the external data source of news articles, a process known as retrieval-augmented generation (RAG).

The important component is that this neural network architecture enables a targeted inspection, based on the variables identified in Figure 7.

In addition to using LLMs to sift news sources to highlight potential causes of future dislocations, one can also use a multi-agent framework in the spirit of Park (2024) to harness a collaborative network of AI agents and coordinate their answers in the pursuit of the common objective to explain the potential cause of future dislocations.

6. Preliminary conclusions

In this paper, we explore how directly forecasting indicators of market dysfunction can support financial stability monitoring. Given the multidimensionality of financial data and the nonlinearities in their relationship, both in the cross-section and over time, we find that neural networks with recurrent layers are able to successfully withstand the challenging AR benchmark in out-of-sample tests, and in fact whose predictions covary with the target variable even while being smoother. This result suggests policymakers could have *some* insight into potential looming illiquidity periods. Building on these findings, we also demonstrate how LLMs can further support this process by sifting through large text datasets, such as financial news (or supervisory texts, internal bank documents etc), to help identify "needles in a haystack" related to periods of forecasted illiquidity. This can further support financial stability monitoring by providing signals to authorities that might otherwise miss.

The performance of current models can improve further. Finding model configurations that perform well in practice is an empirical exercise: the model parameter space is wide and multi-dimensional, and there is scant guidance on configurations that work well in practice in financial data. In addition, the architecture of neural networks, ie how their layers are set up and connected to each other, is usually very flexible; choices around data transformations are also plentiful. This requires the econometrician to test out a large number of configurations. We hope that our work will contribute to inform this exercise, by

offering insights into an architecture and data transformations that works in practice, providing a basis for further improvements.

References

Adrian, T, D Covitz and JN Liang (2015): "Financial stability monitoring", Annual Review of Financial Economics, v.7, n.1, p.957-395.

Adrian, T and H S Shin (2010): "Liquidity and leverage", Journal of Financial Intermediation, v.19, n.3, p.418-437.

Almosova, A and N Andresen (2023): "Nonlinear inflation forecasting with recurrent neural networks", Journal of Forecasting, v.42, p.240-259.

Americo, A, Araujo, D K G, J Damp, S Nilsen, D Rees, R Schmidt, C Schmieder (forthcoming): "Transforming nowcasting with transformers".

Araujo, D K G (2023): "gingado: a machine learning library focused on economics and finance", BIS Working Papers, no 1122.

Araujo, D K G, S Doerr, L Gambacorta, B Tissot (2024): "Artificial intelligence in central banking", BIS Bulletin, no 84.

Arik, S Ö and T Pfister (2021): "TabNet: attentive interpretable tabular learning", AAAI.

Aquilina, M, M J Lombardi, A Schrimpf and V Sushko (2024): "The market turbulence and carry trade unwind of August 2024", BIS Bulletin, no 90.

Augustin, P, M Chernov, L Schmid, D Song (2024): "The term structure of covered interest rate parity violations", Journal of Finance, v.79, n.3, p.2077-2114.

Bahdanau, D, K Cho and Y Bengio (2014): "Neural machine translation by jointly learning to align and translate", ICLR.

Bali, T G, H Beckmeyer, M Mörke and F Weigert (2023): "Option return predictability with machine learning and big data", Review of Financial Studies, v.36, p.3548-3602.

Brandão-Marques, L, B González-Hermosillo, A Bouveret, F Han, D Jones, J Kiff, F Lambert, W Monroe, O Khadarina, L Valderrama, K Yan (2015): "Market liquidity – resilient or fleeting", in: Global Financial Stability Report, Chapter 2, International Monetary Fund, October.

Brunnermeier, M K and L H Pedersen (2009): "Market liquidity and funding liquidity", Review of Financial Studies, v.22, n.6, p.2201-2238.

Cao, Y, Z Chen, Q Pei, F Dimino, L Ausiello, P Kumar, K P Subbalakshmi and P M Ndiaye (2024): "RiskLabs: predicting financial risk using large language model based on multi-sources data", *arXiv:2404.07452 [q-fin.RM]*.Chen, T and C Guestrin (2016): "XGBoost: a scalable tree boosting system", KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Chen, J, L Song, M Wainwright and M Jordon (2018): "Learning to explain: an information-theoretic perspective on model interpretation", Proceedings of Machine Learning Research, v.80, p.883-892.

Davis, S J (2016): "An index of global economic policy uncertainty", Macroeconomic Review, October.

Dammon, R M, K B. Dunn and C. Spatt. "The Relative Pricing of High-Yield Debt: The Case of RJR Nabisco Holdings Capital Corporation" The American Economic Review 83, no. 5 (1993): 1090–1111.

Domanski, D, I Fender and P McGuide (2011): "Assessing global liquidity", BIS Quarterly Review, December.

Drehmann, M and V Sushko (2022): "The global foreign exchange market in a higher volatility environment", BIS Quarterly Review, December.

Du, W, A Tepper and A Verdelhan (2018): "Deviations from covered interest rate parity", Journal of Finance, v.73, n.3, p.915-957.

Elahi, A and F Taghvaei (2024): "Combining financial data and news articles for stock price movement prediction using large language models", *arXiv:2411.01368 [cs.IR]*.

Favara, G, S Gilchrist, K F Lewis and E Zakrajšek (2016): "Updating the recession risk and the excess bond premium", FEDS Notes, Board of Governors of the Federal Reserve System, 6 October.

Fenn, D J, S D Howison, M McDonald, S Williams and N F Johnson (2009): "The mirage of triangular arbitrage in the spot foreign exchange market", International Journal of Theoretical and Applied Finance, v.12, n.8.

Fleckenstein, M, F A Longstaff, and H Lustig. "The TIPS-Treasury Bond Puzzle" The Journal of Finance 69, no. 5 (2014): 2151–97.

Foucault, T, R Kozhan and W.W. Tham, (2017): "Toxic Arbitrage", The Review of Financial Studies, Volume 30, Issue 4, Pages 1053–1094.

Foucault, T, M Pagano and A Roell, (2013), ""Market Liquidity", Oxford University Press

Gabaix, X, R S J Koijen, R Richmond, M Yogo (2024): "Asset embeddings", SSRN.

Garcia, L, U Lewrick and T Sečnik (2021): "Is window dressing by banks systemically important?" BIS Working Papers, no 960.

Gârleanu, N and L H Pedersen (2011): "Margin-based asset pricing and deviations from the Law of One Price", Review of Financial Studies, v.24, p.1982-2022.

Gelb, A (Ed) (1974). Applied Optimal Estimation. MIT Press.

Giannone, D, M Lenza and G E Primiceri (2022): "Economic predictions with big data: the illusion of sparsity", Econometrica, v.89, n.5, p.2409-2437.

Gilchrist, S and E Zakrajšek (2012): "Credit spreads and business cycle fluctuations", American Economic Review, v.102, n.4, p.1692-1720.

Goodfellow, I, Bengio, Y, and Courville A. (2016). Deep Learning. MIT Press.

Gradojević, N, D Erdemlioglu and R Gençay (2020): "A new wavelet-based ultra-high-frequency analysis of triangular currency arbitrage", Economic Modelling, v.85, p.57-73.

Graves, A, S Fernández, F Gomez, and J Schmidhuber (2006): "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". Proceedings of the 23rd International Conference on Machine Learning (ICML '06), 369-376.

Graves, A (2013): "Generating sequences with recurrent neural networks", arXiv.

Gu, S, B Kelly and D Xiu (2020): "Empirical Asset Pricing via Machine Learning", The Review of Financial Studies, Volume 33, Issue 5, May 2020, Pages 2223–2273, https://doi.org/10.1093/rfs/hhaa009

Haddad, V and T Muir (2021): "Do intermediaries matter for aggregate asset prices?", Journal of Finance, v.76, n.6.

He, Zhiguo and Arvind Krishnamurthy (2013): "Intermediary asset pricing", American Economic Review, v.103, n.2, p.732-770.

He, Zhiguo and Arvind Krishnamurthy (2018): "Intermediary asset pricing and the financial crisis", Annual Review of Financial Economics, v.10, p.173-197.

Hochreiter S and Schmidhuber, J (1997) "Long short-term memory", Neural Computation, v.9, n.8, p.1735-1780.

Hoerl, A E and R W Kennard (1970): "Ridge regression: biased estimation for nonorthogonal problems", Technometrics, v.12, p.55-67.

Iacovides, G, T Konstantinidis, M Xu and D Mandic (2024): "FinLlama: LLM-based financial sentiment analysis for algorithmic trading", in *Proceedings of the 5th ACM International Conference on AI in Finance*, pp 134–141.

IMF, (2015), "Market Liquidity – Resilient or fleeting?", *Chapter 2 of the Global Financial Stability Report*.

Jin, M; S Wang, L Ma, Z Chu, J Y Zhang, X Shi, P Chen, Y Liang, YF Li, S Pan and Q Wen (2024): "Time-LLM: Time Series Forecasting by Reprogramming Large Language Models"

Jordà, O (2005): "Estimation and inference of impulse responses by local projections", American Economic Review, v.95, n.1, p.161-182.

Kelly, B T and D Xiu (2023): "Financial Machine Learning", Foundations and Trends in Finance.

Kelly, B T, S Malamud and K Zhou (2023): "The virtue of complexity in return prediction", Journal of Finance, v.79, n.1, p.459-503.

Koijen, R S J, M Yogo (2019): "A demand system approach to asset pricing", Journal of Political Economy, v.127, n.4.

Kyle, A S (1985): "Continuous auctions and insider trading", Econometrica, v.53, n.6, p.1315-1335.

Lamont, OA, and RH. Thaler. "Can the Market Add and Subtract? Mispricing in Tech Stock Carve-outs." Journal of Political Economy 111, no. 2 (2003): 227–68.

Lim, B, SÖ Arik, N Loeff and T Pfister (2021): "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting", International Journal of Forecasting, v.37, n.4, p.1748-1764.

Liu, F T, K M Ting and Z-H Zhou (2012): "Isolation-based anomaly detection", ACM Transactions on Knowledge-Discovery from Data (TKDD), 6.1, 3.

Lyons, R K and M J Moore (2009): "An information approach to international currencies", Journal of International Economics, v.79, n.2, p.211-221.

Olah, C (2015): "Understanding LSTM Networks", https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Papasotiriou, K, S Sood, S Reynolds and T Balch (2024): "AI in investment analysis: LLMs for equity stock ratings", *arXiv:2411.00856 [cs.LG]*.

Park, T (2024): "Enhancing Anomaly Detection in Financial Markets with an LLM-based Multi-Agent Framework", arXiv:2403.19735 [q-fin.RM].

Pasquariello, P (2014): "Financial market dislocations", Review of Financial Studies, v.27, n.6, p.1868-1914.

Persaud, A (2003): "Liquidity Black Holes", Risk Books

Petropoulos, F and S Makridakis (2020): "The M4 competition: Bigger. Stronger. Better", International Journal of Forecasting, v.36, n.1 p.3-6.

Ranaldo, A and P Santucci de Magistris (2022): "Liquidity in the global currency market", Journal of Financial Economics, v.146, n.3, p.859-883.

Roque, L. (2024): "Chronos: The Rise of Foundation Models for Time Series Forecasting", available at Towards Data Science.

Rossi, B (2013): "Exchange rate predictabilty", Journal of Economics Literature, v.51, n.4, p.1063-1119.

Sen, R and Y Zho (2024): "A decoder-only foundation model for time-series forecasting", available at https://research.google/blog/a-decoder-only-foundation-model-for-time-series-forecasting/.

Sirignano, J and R Cont (2019): "Universal features of price formation in financial markets: perspectives from deep learning", Quantitative Finance, v.19, n.9, p.1449-1459.

Smyl, S (2020): "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting", International Journal of Forecasting, v.36, n.1, p.75-85.

Tibshirani, R (1996): "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society B, v.58, n.1, p.267-288.

Tirole, J (2011): "Illiquidity and all its friends", Journal of Economic Literature, v.49, n.2, p.287-325.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention is all you need". Advances in Neural Information Processing Systems 30, 5998-6008.

Vuletić, M, F Prenzel and M Cucuringu (2024): "Fin-GAN: forecasting and classifying financial time series via generative adversarial networks", Quantitative Finance, v.24, n.2, p.177-199.

Wen, R, K Torkkola, B Narayanaswamy, D Madeka (2018): "A multi-horizon quantile recurrent forecaster", arXiv, 1711:11053.

Yoon, J, J Jordon and M van der Schaar (2019): "INVASE: Instance-wise variable selection using neural networks", ICLR

Yu, Y, H Li, Z Chen, Y Jiang, Y Li, D Zhang, R Liu, J W Suchow and K Khashanah (2023): "FinMem: a performance-enhanced LLM trading agent with layered memory and character design", *arXiv:2311.13743 [q-fin.CP]*.

Yu, X, Z Chen, Y Ling, S Dong and others (2023): "Temporal data meets LLM: explainable financial time series forecasting", *arXiv:2306.11025 [cs.LG]*.

Zhang, C, X Liu, Z Zhang, M Jin, L Li, Z Wang, W Hua, D Shu, S Zhu, X Jin, S Li, M Du and Y Zhang (2024a): "When AI meets finance (StockAgent): large language model-based stock trading in simulated real-world environments", *arXiv:2407.18957 [q-fin.TR]*.

Zhang, W, L Zhao, H Xia, S Sun, J Sun, M Qin, X Li, Y Zhao, Y Zhao, X Cai, L Zheng, X Wang and B An (2024b): "A multimodal foundation agent for financial trading: tool-augmented, diversified, and generalist", *arXiv:2402.18485 [q-fin.TR]*.

Zou, H and T Hastie (2005): "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society Series B: Statistical Methodology, v.67, n.2, p.301-320.