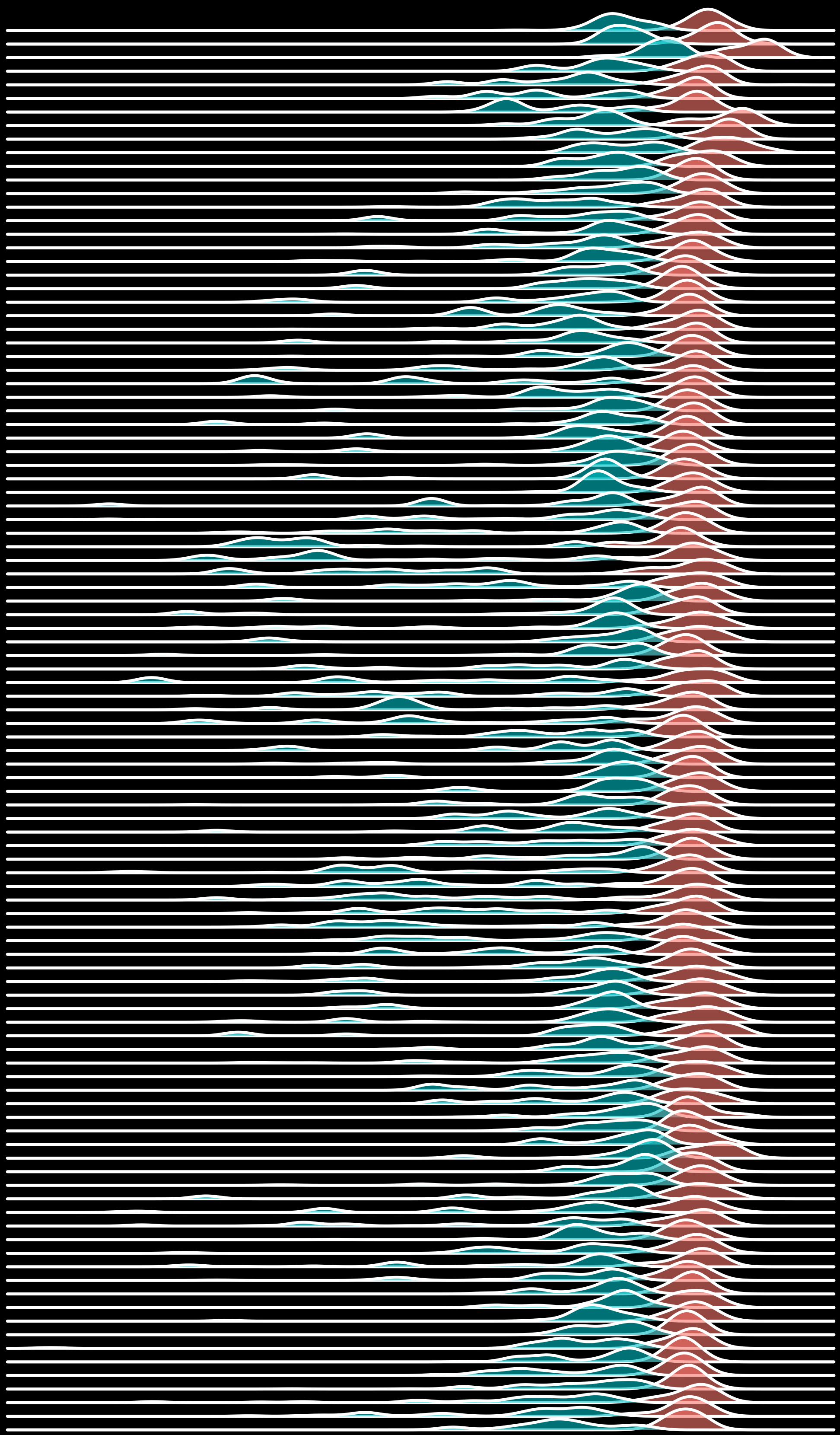
LONDON REPORTING HOUSE



GC Division: classifying trades in the repo market using machine learning techniques

Dr Marian Priebe - Quantitative Analyst

Cover page: Stylised ridgeline plot of the UK gilts repo market for trades with a tomorrow-next (TN) tenor between 1st January and 1st June, 2024. Each data point has been labelled as either 'GC' or 'special' by applying the Head/Tails machine learning method to each day's data. Daily density distributions of GC (red) and specials (blue) are plotted in descending order.

Abstract

London Reporting House (LRH) has access to Securities Financing Transactions Regulation (SFTR) data, which is otherwise only available to regulators. This allows us to build a clearer picture of activity in the European repo market than is available to individual participants. One particular point of interest is the division of the repo market into cash-driven and collateral-driven segments, that is, between so-called "general collateral" or GC repo and repo against "specials." Since GC repo is cash-driven, it should trade at or around the same repo rate - the GC repo rate. Special repo is securities-driven and trades at a wide range of different rates which depend on the intrinsic values of the collateral securities. This intrinsic value drives special repo rates below the GC repo rate.

To determine which repos are GC, so that we can then identify those that are trading special, we have applied simple Machine Learning (ML) methods to an anonymised set of repo rates from a sample of actual transactions reported under SFTR. We have characterised and compared the performance of these methods to one another, as well as to a simple, percentage-based cut-off method that is used in the market. All the ML methods implemented - Jenks Natural Breaks, Head/Tails, and Gaussian Mixture Models (GMM) - can classify trades in a manner that is robust to varying proportions of GC and specials in different market conditions. Head/Tails consistently outperforms current benchmarking methods, adapting to daily market conditions.

Introduction

GC and special repo

A repo is a sale of an asset and a simultaneous commitment to subsequently repurchase the same or a similar type of asset at a different price. The typical asset in repo is a domestic government security. The security repurchased will typically have the same ISIN as the security sold but does not need to be the same parcel of securities that were delivered in the sale. During the term of a repo, the seller will have the use of the sale proceeds, and the buyer will have the use of the securities, which they can use as collateral if the seller defaults or which they can trade (with the aim of replacing those securities in time to return to the seller). Repo can therefore act like a secured loan (that is, credit secured by collateral). The difference between the sale and repurchase prices represents a return to the buyer for effectively lending cash to the seller and is quoted as the price of the repo in the form of an annualised percentage interest rate called the “repo rate.” The difference between the settlement dates of the two legs of a repo is referred to as the tenor. The most common tenor is one day, which can be subdivided into overnight (ON), tomorrow-next (TN) and spot-next (SN), depending on whether settlement of the front leg is today or in one or two business days.

When used to borrow and lend cash, the repo rate will reflect the supply of and demand for cash, so we can call this type of transaction “cash-driven.” In cash-driven repos, the buyer will not be concerned by the particular security issue provided as collateral, provided it is from a creditworthy issuer and is liquid. In fact, the buyer will be willing to accept many, if not all, of the issues of a suitable issuer. The set of acceptable securities is called “general collateral” or GC. Because it is a cash-driven interest rate in which risk premia have been minimised by collateralisation, the GC repo rate is taken to represent the (near) risk-free cost of money. As such, it is a vitally important benchmark for the pricing of repo against non-government collateral, as well bonds and derivatives, and to interpreting monetary conditions.

On occasion, however, the buyer will seek to transact a repo into order to acquire a particular security issue (e.g. a benchmark government bond), often to fulfil a delivery commitment or to take a short position. The security in this type of repo is referred to as “specific collateral.” Because the specific collateral has value to the buyer, they will often have to offer the seller cheaper cash, i.e. a lower repo rate, the difference reflecting the intrinsic value attached to the issue. This type of transaction can be called “securities-driven”, and such repo are analogous to cash-collateralised securities loans. If a specific repo rate falls materially below the GC repo rate, the collateral security will be said to have become “special.” Trading specials is a lucrative business for securities dealers and a means of yield enhancement to investors holding specials in their portfolios.

We can picture GC repo rates as a cluster of closely correlated interest rates (where the GC repo rate is defined as either the highest rate in the cluster or the median). Specials are the transactions with repo rates clearly below the base of the cluster.

The challenge

Dealers often form a view of where the GC rate is for a particular tenor using judgment. Our aim is to identify GC more rigorously and objectively, as a measurable property of the repo market.

To do that, we need repo rate data. Such data is normally restricted to dealers or trading venues and tends to be partial (e.g. coming from one trading venue). However, there is an independent, comprehensive and exact source: SFTR data. SFTR (Securities Financing Transaction Regulation) was the response of the EU-28 to calls by the G-20 Financial Stability Board (FSB), after the Global Financial Crisis, for greater transparency over the markets in securities financing transactions (SFT), including repo. Both the UK and EU-27 versions of SFTR require market participants to report more than 90 fields of detailed information on their SFTs, including repos, to an authorised trade repository daily.

Plotting the repo rates for a subset of overnight (ON) repo against UK gilts in 2024, we can intuitively recognise the cluster of rates at the top of each day’s data which correspond to GC, as well as the spread of specials trading at lower rates.

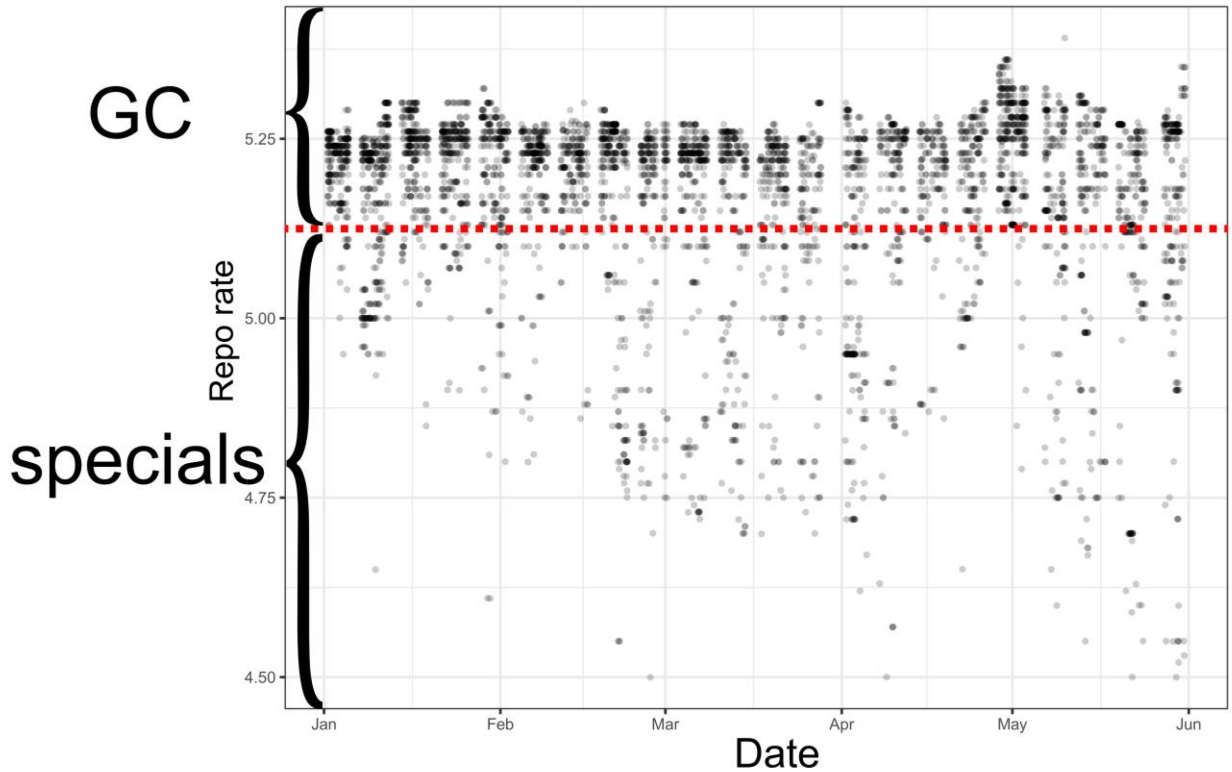


Figure 1. Rates for ON UK gilts repo for the first half of 2024. Overlaid are markings representing the intuitive division between GC and specials.

Our objective is to find a method that can consistently distinguish the GC cluster for a particular day, market and tenor and do this in a manner that is robust to variations in the volume and distribution of data points.

Methods

Static thresholds

A simple method for dividing the distribution of rates into GC and specials is to apply a fixed-percentage threshold to any distribution of data. Trades with repo rates above this threshold are labelled GC and are used to calculate a representative daily GC repo rate.

This method is used by the CME and MTS in the calculation of their Repo Fund Rates (RFR). RFR consists of all CCP-cleared trades in the GC facilities of the CME BrokerTec and MTS Repo platforms, plus the top 75% of rates traded in their “specials” facilities (in fact, these are specifics and not necessarily specials). The static threshold of 75% was chosen by measuring the change in the volume-weighted mean of the GC segment for different thresholds – it was observed that for stricter thresholds the repo rate did not considerably change. The volume-weighted mean of this segment is then calculated to produce the RFR. A similar method is employed by the FRBNY to calculate the SOFR benchmark – also using the 25th volume-weighted percentile of DVP repo transactions as a threshold to trim specials (though this method is currently under review by the FRBNY).

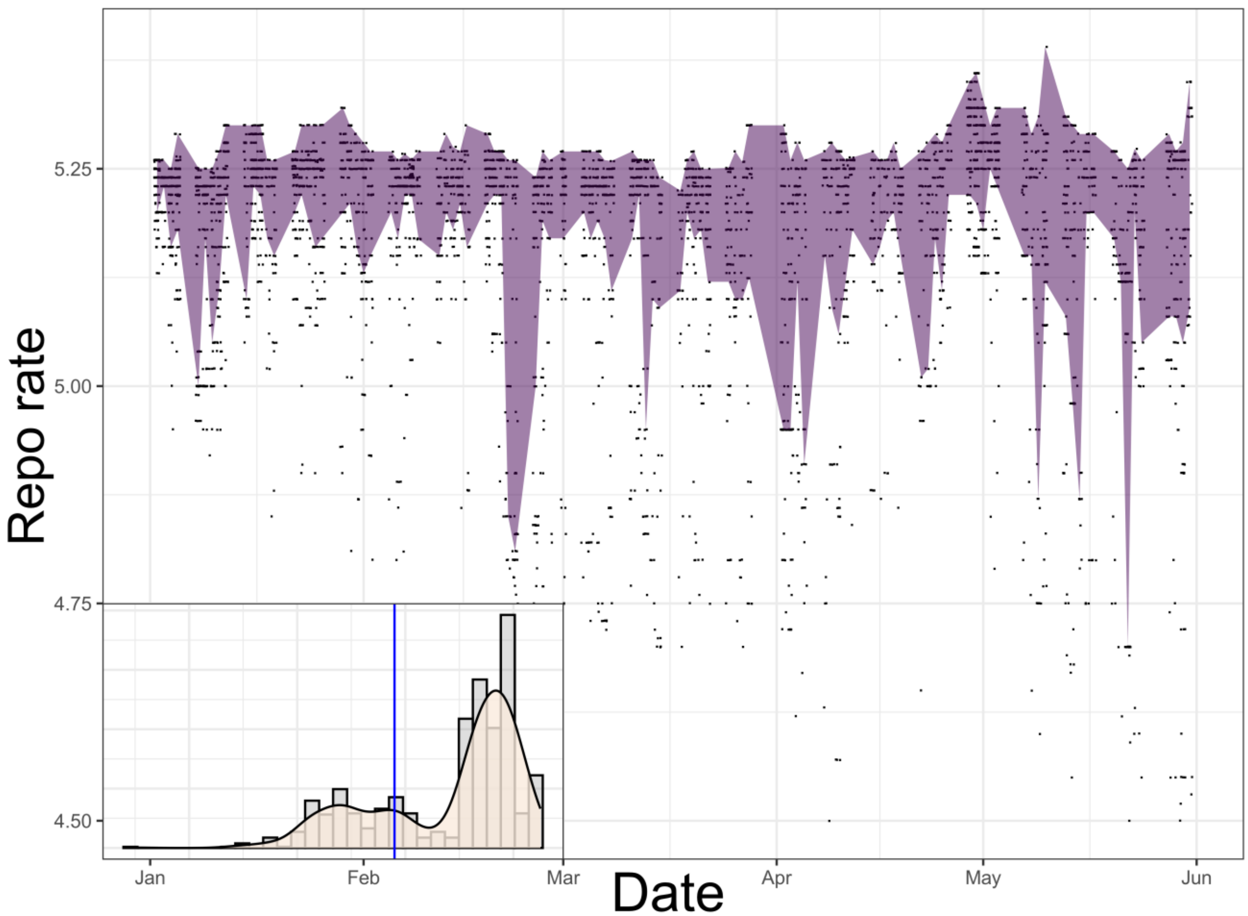


Figure 2. A visual representation of the 75/25 cutoff applied to the overnight (ON) repo against UK gilts in 2024 (main) as well as to a sample distribution of a single day's data (insert bottom left).

However, when we apply the 75/25 threshold method to our data (Figure 2), we can see that the boundary where this method splits the data does not consistently follow the lower edge of the visible GC cluster at the top of each day's distribution of rates. This means that some special rates are included in the calculation of the benchmark rate, making it lower than it should be. The problem with this simple method is that the threshold is not adjusted to the proportion of data found within the GC cluster. The measured width of the GC band will vary more depending on the ratio of GC/special trades than on differences in the width of the underlying distribution of GC repo.

Jenks Natural Breaks

The Jenks Natural Breaks method is a data classification technique designed to determine the optimal arrangement of a set of values into different classes. This method minimizes the variance within each class while maximizing the variance between classes, by iteratively adjusting class boundaries along a single dimension or variable to reduce the sum of squared deviations from the class means. This process ensures that the data within each class is as homogeneous as possible. By focusing on natural groupings inherent in the data, Jenks Natural Breaks provides a more accurate and insightful classification compared to methods using a fixed threshold.

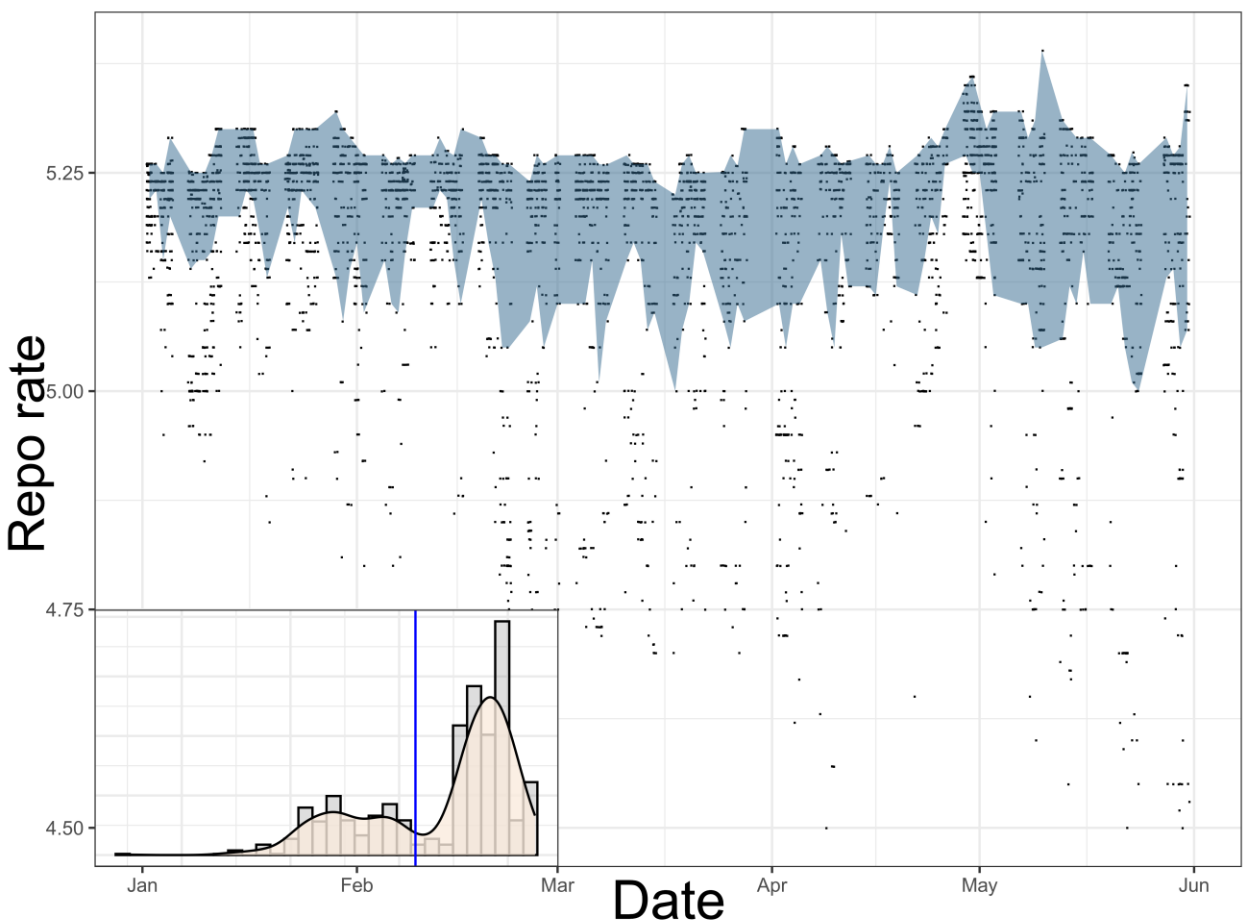


Figure 3 (page 6). A visual representation of Jenks Natural breaks applied to the overnight (ON) repo against UK gilts in 2024 (main) as well as to a sample distribution of a single day's data (bottom left).

Since we want to classify data into two groups based on variation in a single dimension – in this case the repo rate – this method is appropriate for partitioning GC and specials through a single natural break in the distribution of the data. Applying it to the example distribution of rates for ON gilt repo shows a break close to the local minima between the two statistical modes of the distribution (Figure 3). This method dynamically adjusts the threshold for distributions with different relative volumes of GC and specials trades. However, in cases where the trades below the GC cluster do not form a cluster of their own, this method underperforms since the homogeneity of trades labelled as special is considered when deciding on the optimal threshold. Only one of the two classes of trades – the GC portion – can be expected to consistently form a cluster, while we have no assumptions about what shape the distribution of specials trades should take.

Head/Tails

The Head/Tails algorithm is a data classification method designed to handle data with a heavy-tailed distribution. As with Jenks, this method iteratively splits the data into two parts: the "Head," containing values above a boundary along a single variable or dimension, and the "Tail," containing values below this threshold. The process repeatedly adjusts the threshold until a stopping criterion is met, typically when the head portion becomes sufficiently small or homogeneous.

The primary advantage of the Head/Tails algorithm is its ability to classify data points with a single value (or dimension) in a manner that is robust to distributions with significant skewness or outliers. By focusing on the heavy-tailed nature of the data, this method provides a more nuanced classification that can reveal underlying patterns and trends that might be missed by other methods.

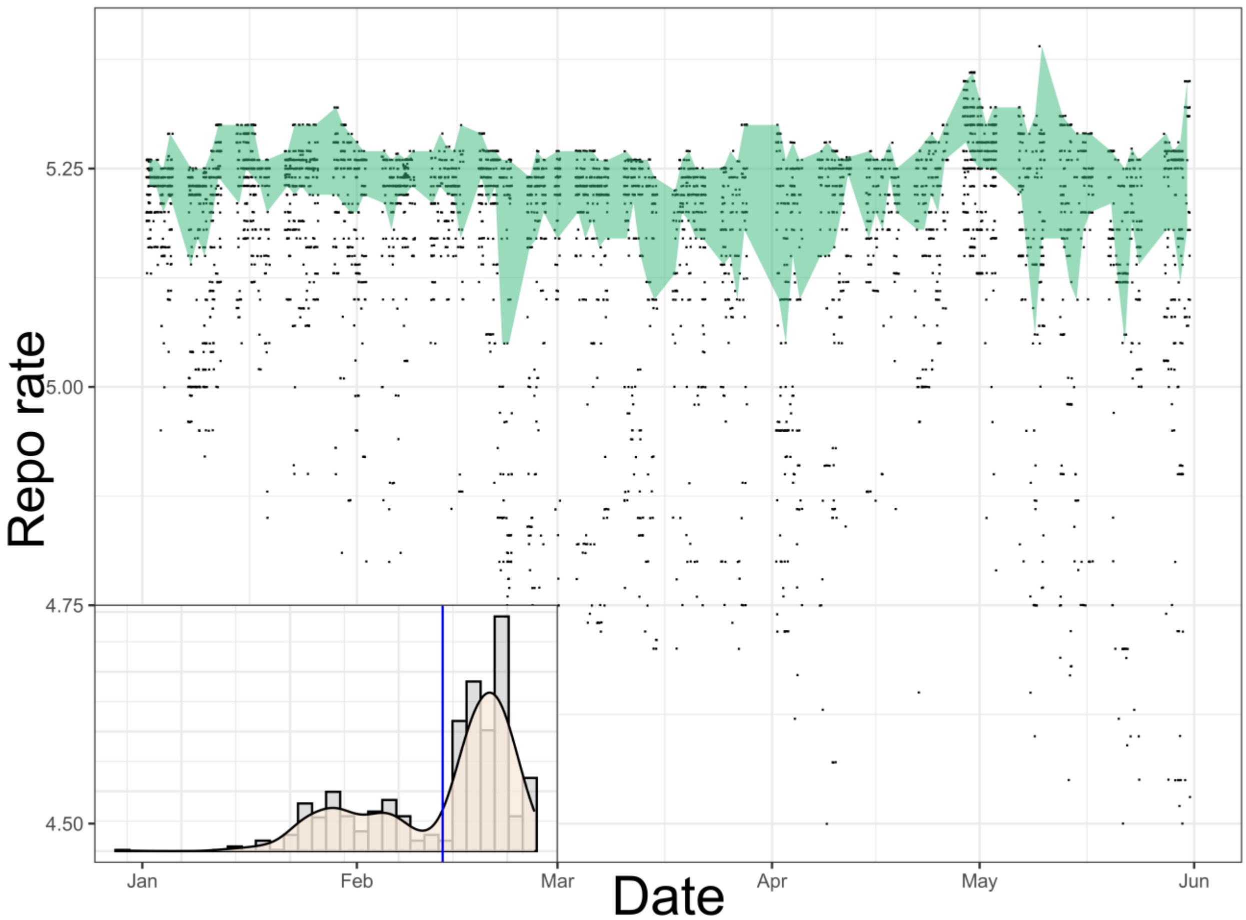


Figure 4. A visual representation of Head/Tails applied to the overnight (ON) repo against UK gilts in 2024 (main) as well as to a sample distribution of a single day's data (bottom left).

The skewed nature of rates data within our sample distribution – where the majority of datapoints are found within the GC cluster – meets the assumptions that underpin a Head/Tails clustering. The threshold that Head/Tails applies to this distribution is slightly higher than that chosen by Jenks, partitioning the data with a stricter definition of what is part of the GC head of the distribution (Figure 4). As a result, this method, compared to Jenks, is likely to produce less false positives (special trades labelled as GC), at the cost of more false negatives (GC trades incorrectly labelled as special and consequently excluded from further analysis of the GC cluster), this trade-off on the whole reduces the noise in the GC cluster.

Applied to the range of distributions of the ON gilt repo market, these observations hold true as a general trend. The threshold set by Head/Tails is consistently higher than that of Jenks, resulting in a narrower GC band. This method is more robust to the daily variations in how the specials trades are distributed since it is only looking to optimise the clustering of the GC Head of the data. It therefore aligns with our lack of assumptions about how specials should be distributed.

Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are probabilistic models that assume data points are generated from a mixture of several Gaussian distributions with unknown parameters. Unlike clustering algorithms such as Jenks Natural Breaks and Head/Tails - which are deterministic and partition data, based on a threshold which aims to minimise within-class variance or balance class sizes - GMMs provide a flexible approach by estimating the probability that each data point belongs to each group. This probabilistic nature allows GMMs to capture more complex data distributions and overlapping clusters, making them particularly useful for identifying nuanced patterns in multidimensional data. However, they can also be applied to classify data points along a single variable.

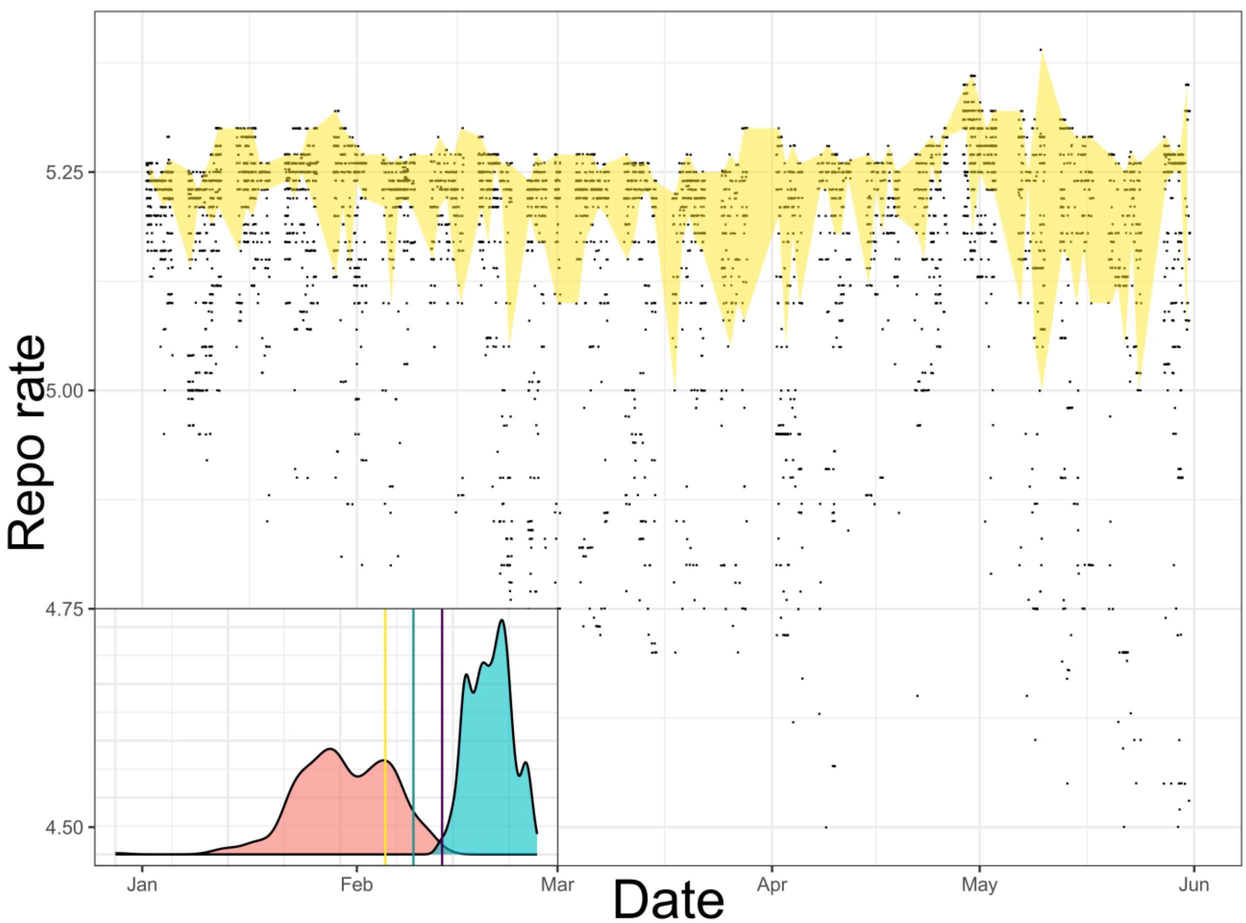


Figure 5 (page 9). A visual representation of a GMM applied to the overnight (ON) repo against UK gilts in 2024 (main) as well as to a sample distribution of a single day's data (bottom left). Overlaid on this sample distribution are the thresholds chosen by (left to right) 75/25, Jenks and Head/Tails methods as vertical lines.

Applied to the sample distribution of rates, the GMM, set to produce exactly two clusters, mirrors the results we see from the other models - labelling the cluster of GC trades at the top of the distribution together into GC and everything below it into specials. Unlike the other methods, the GMM does not explicitly partition the data based on a threshold. Instead, we can imply a threshold where the probabilities of belonging to both clusters are equal, rates higher than this value will have a greater probability of being labelled as GC, while those lower are more likely to be labelled as special. For the sample distribution, this implied threshold is closest to the one set by Head/Tails (Figure 5). However, when applied to the entire data for the ON gilt repo market in 2024, this method appears to be very sensitive to daily differences in the distributions of rates data. In cases where the two Gaussian distributions fitted by the GMM to a day's data differ highly in their variance, we see some points with particularly high rates labelled as special, since the probability of belonging to the narrow GC cluster is less than the probability of belonging to the tails of the wider specials cluster at extreme values.

Application to other markets and tenors

Within a single market, we found high variance between the distributions of rates on any given day, but these distributions should also be expected to vary between markets and tenors. Therefore, to test these models further, we extended our analyses to the TN tenor of gilt repo (the second largest tenor within this market), as well as to the US ON market and the SN trades for three European markets (DE, FR, and IT).

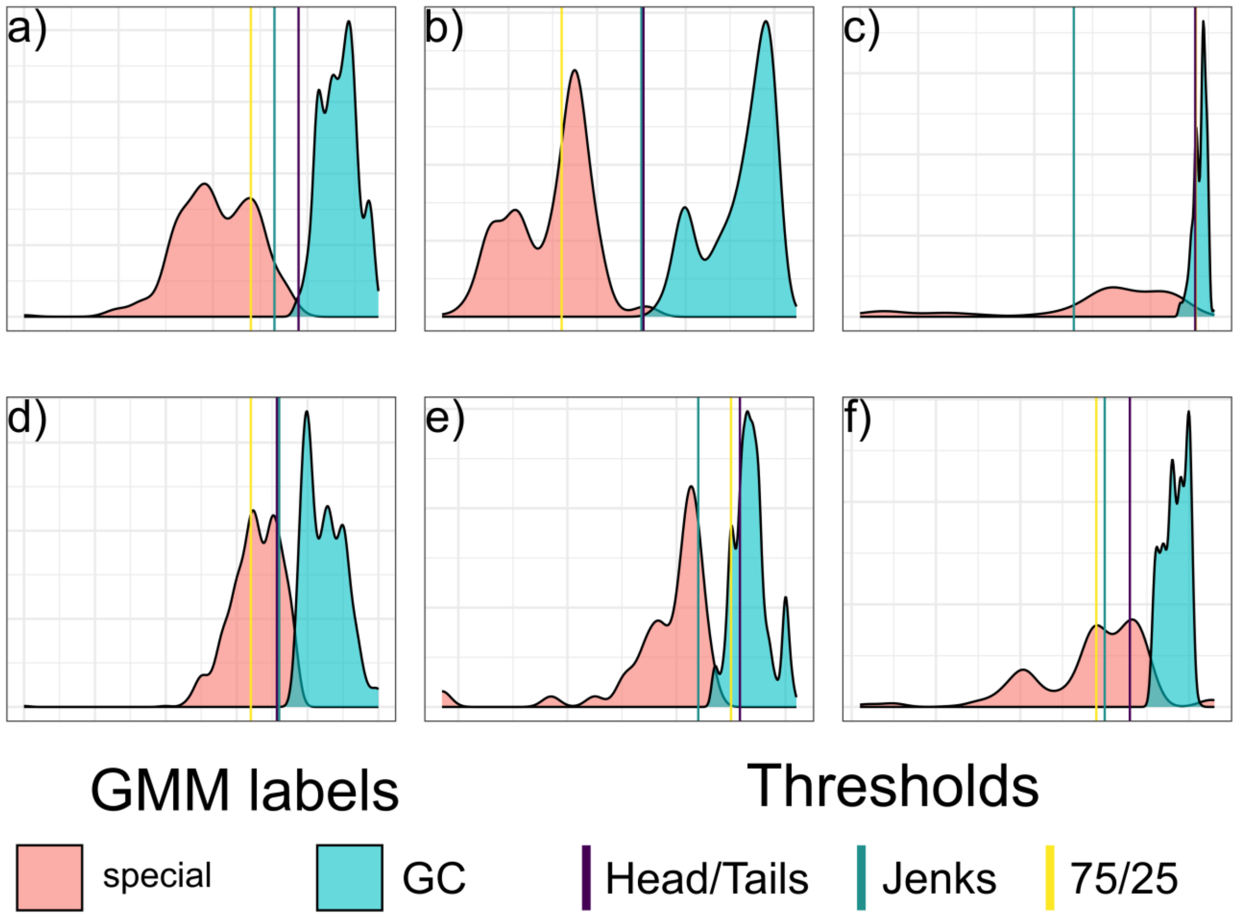


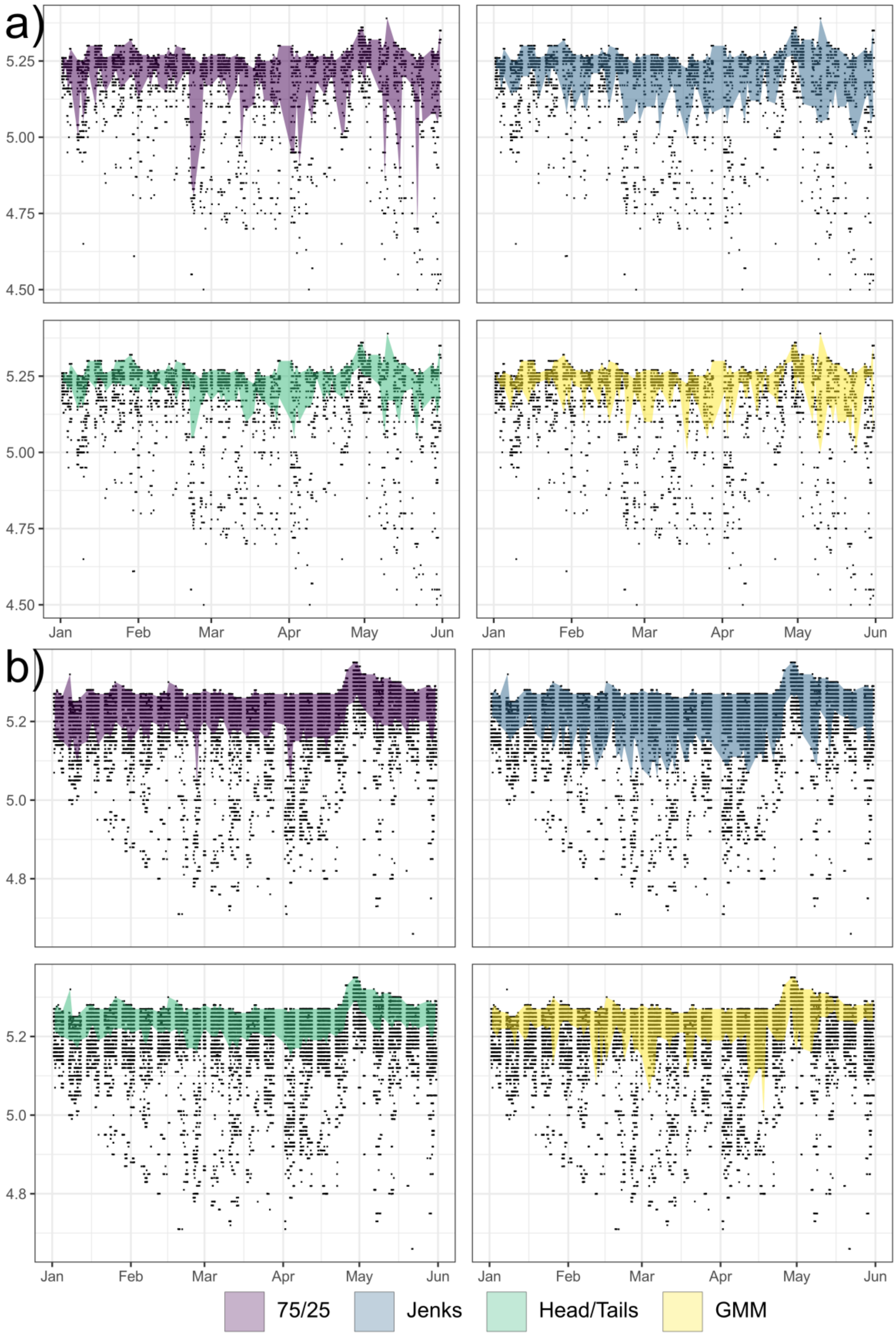
Figure 6. Comparison of how the different methods partition GC and specials when applied to representative, sample distributions from six combinations of tenor and market; **a)** ON and **b)** TN UK gilt repo, **c)** ON US Treasury bond repo as well as the SN repo for **d)** German Bunds, **e)** French OATs and **f)** Italian government bonds. In each case the two densities represent distributions of GMM labelled data while the thresholds used by the 75/25, Jenks and Head/Tails methods are overlaid as vertical lines.

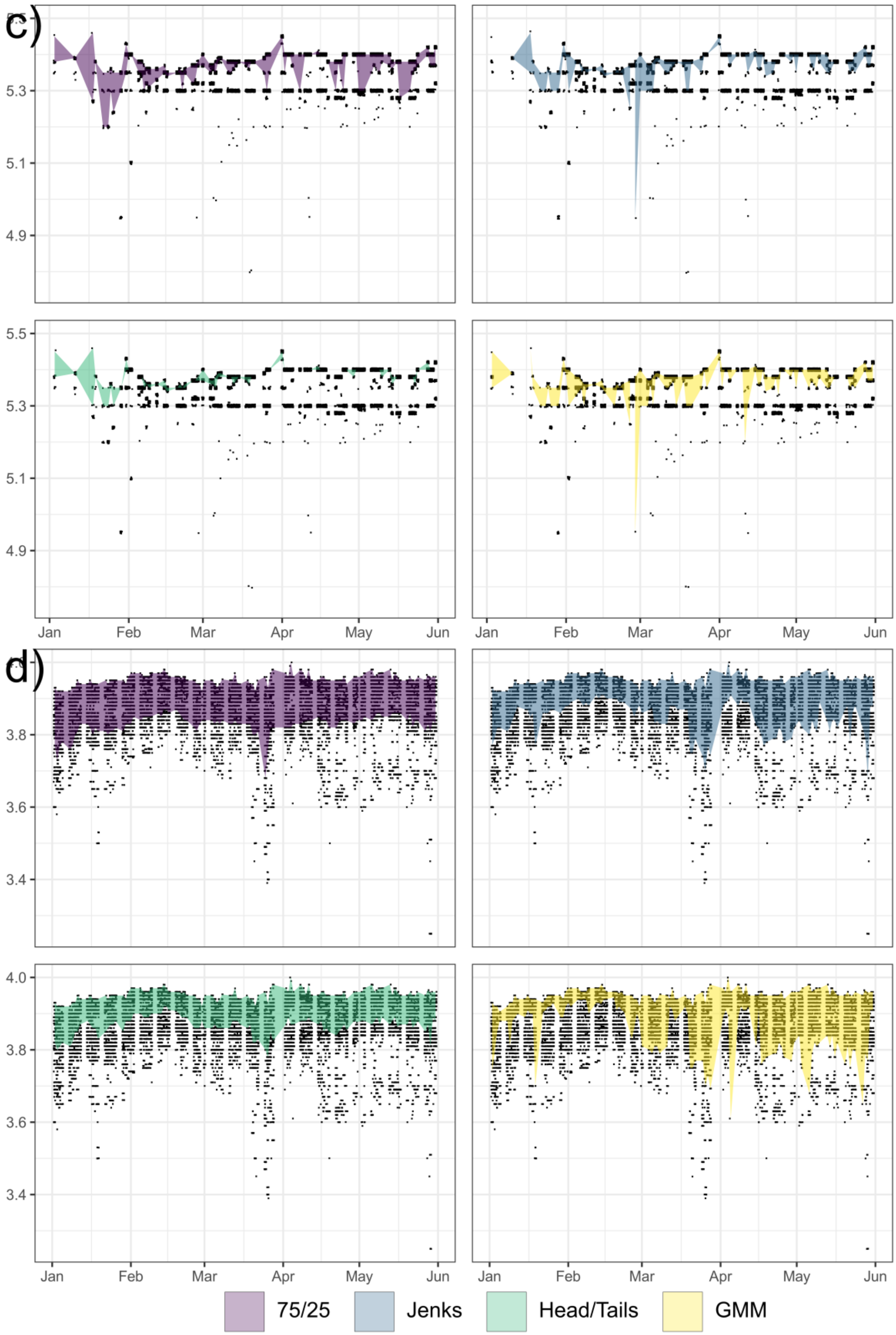
Differences between markets and tenors in the distribution of a day's data dramatically affect the performance of the various methods. In cases where the relative proportions of GC and specials differ from our expectation (such as in the TN UK gilts in Figure 6b), the 75/25 threshold does not provide a meaningful division of the distribution and only cuts out the most extreme specials from GC calculations. The performance of the Jenks methods is sensitive to wide distributions of special trades, as can be seen from the US, French and Italian examples (Figure 6c, e & f). In these cases, the requirements of this method to try and optimise for a well-clustered special group impacts the tightness of the resulting GC clusters.

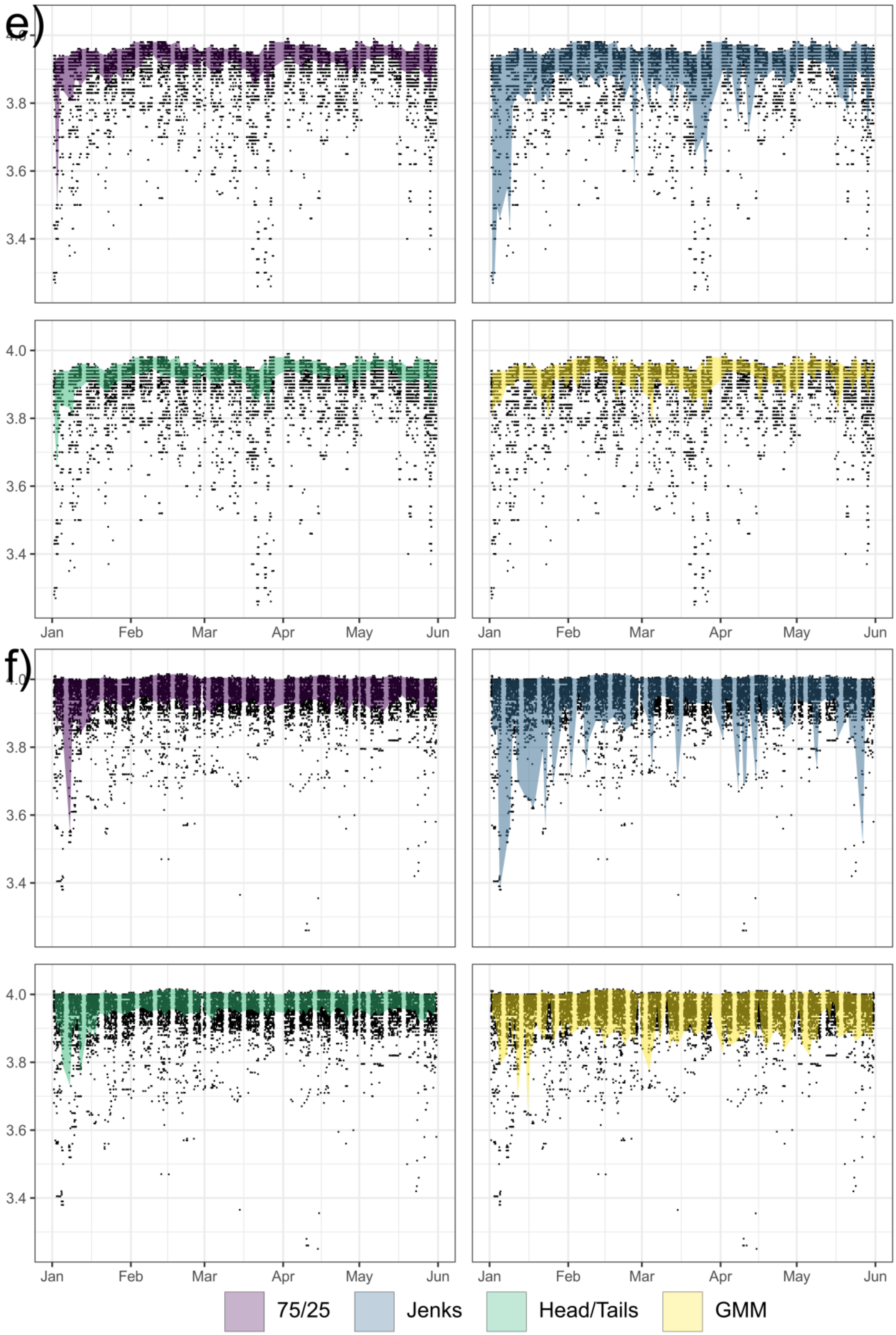
Throughout these examples, both Head/Tails and GMM produce consistent results, resulting in tight and well-defined clusters of GC trades, while being robust to larger proportions and wider distributions of specials trades.

These observations are generally echoed in the application of these methods across a longer timescale (Figure 7). We observe that, compared to a static threshold, all these ML methods produce better defined and tighter GC clusters. However, in terms of the variation in distributions that we observe in our different markets, both Jenks and GMM underperform under certain conditions, while the Head/Tails method remains robust to these and aligns most consistently with the cluster of GC trades we intuitively observe at the top of each distribution.

Figure 7 (pages 13-15). Ribbon plots showing the changing width of the GC band identified by the four methods applied to repo data for the first half of 2024 for six combinations of tenor and market; **a)** ON and **b)** TN UK gilt repo, **c)** ON US Treasury bond repo as well as the SN repo for **d)** German Bunds, **e)** French OAT and **f)** Italian government bonds.


a)
5.25
5.00
4.75
4.50
Jan
Feb
Mar
Apr
May
Jun
b)
5.2
5.0
4.8
75/25
Jenks
Head/Tails
GMM


c)
5.5
5.3
5.1
4.9
Jan
Feb
Mar
Apr
May
Jun
d)
4.0
3.8
3.6
3.4
75/25
Jenks
Head/Tails
GMM


e)
f)
4.0
3.8
3.6
3.4
Jan
Feb
Mar
Apr
May
Jun
75/25
Jenks
Head/Tails
GMM

Discussion

The implementation of these simple ML clustering methods to this one-dimensional classification problem has produced promising results, improving upon the accuracy of classifying the same data using a static, percentage-based threshold. These improvements are realised without the need to set an assumption about the relative volumes of GC and specials. As a result, the width of the GC band dynamically adjusts to the data within the cluster rather than being affected by the volume of specials trades.

Most data-points that are within the GC cluster are consistently labelled as such across all the methods we have implemented, with differences between them occurring almost exclusively at the lower boundary of this cluster (there are a few instances where the GMM classifies outliers higher than the GC cluster as special). The inclusion or exclusion of these points from calculations of the GC median do not have an enormous impact on the resulting value, since this measure is inherently resilient to the extreme values of the input. However, aside from the marginal impact on calculating GC rate, changing the exact value of the threshold used to partition the GC cluster from the specials does have a greater impact on the width of the GC band. A visual inspection of the data shows that the width of the cluster of GC trades at the top of a given day's distribution is different between markets and tenors while also changing day-to-day within these. This property of the GC cluster is not naturally reflected by a traditional measure of GC – using a static threshold to generate a single value for the GC repo rate followed by a measure of 'specialness' as the distance from that rate. These new methods can instead give some context to this value - when data within GC has a wider distribution, a trade 10 basis points lower than the GC repo rate might still fall firmly within this cluster, while, in other circumstances, with a tighter GC cluster, this would already be considered special.

Of the three ML methods we implemented, the Head/Tails algorithm had the overall most consistent success in producing a narrow GC band. Compared to the other methods, it generally used a higher threshold to partition the data, minimising the number of special trades incorrectly labelled as GC. These stricter thresholds, compared to Jenks, result from this method's focus on the Head of the data (GC in our case) at the cost of the Tail (specials) being a less well-defined cluster, which is a benefit since we had no prior assumptions about how specials should be distributed.

The regulatory reporting data that we used for this investigation contains many more fields that could be used as input variables for more complex and sophisticated multi-dimensional models. In this study we have focused on clustering data using only the reported repo rate. In future work, this could be expanded to include other values such as the size of a transaction, the security used as collateral or even the time of day that a transaction occurred. Increasing the amount of signal in our input data in this way should allow us to classify these reports with even greater accuracy - both k-means clustering (a multidimensional expansion of the Jenks method) as well as GMMs could take advantage of this richer input data. As such, the methods that have underperformed compared to Head/Tails for the analyses presented here represent important first steps towards solving this classification problem.

In a wider context, this use of SFTR data to classify GC and specials not only improves upon existing methods for calculating a GC repo rate but is also the first step towards leveraging the full extent of this rich data source. Further analysis of SFTR data could provide further transparency to the historically opaque repo market giving us opportunity to understand markets where high-quality sources of data have been particularly scarce.

Conclusion

The novelty of rich, large datasets for the traditionally opaque repo market, such as the transaction-level SFTR data aggregated by LRH, provides an opportunity to reassess some of the existing methods that are used to measure the markets and calculate benchmarks. Even simple ML methods, such as those presented here, are able to leverage these sources to generate results that improve upon current industry standards. The Head/Tails method is able to consistently split a distribution of repo rates into a GC 'Head' and a special 'Tail' in a manner that is robust to variation in the distribution of rates within the market, as well as the differences between markets and tenors. We also present the Jenks and GMM models, that also outperform a simple threshold-based classification, though are less consistent for our single input variable than Head/Tails. These methods both represent a starting point for building more complex and sophisticated models.

London Reporting House

London Reporting House was founded by a team of repo specialists who saw an opportunity to create a truly data-driven repo market using regulatory reporting information.

Having access to granular, detailed and timely data is invaluable for making superior investment decisions. London Reporting House's repo analytics tools provide a deep and granular view of the market for front and back-office functions across the industry.

To enhance our offering and build out our platform, we've joined forces with Kaizen to leverage their IT infrastructure and legal framework for onboarding. Our products and analytics tools are available via the Kaizen Hub, a web based RegTech platform for accessing a range of compliance tools. The SFTR input data we use is also tested using Kaizen's ReportShield™ quality assurance services, ensuring that the underlying input data is high quality.

About the Author

Dr Marian Priebe has recently joined London Reporting House as a Quantitative Analyst. With a robust academic background in Genetics and Bioinformatics, including a PhD from Queen Mary University of London, Dr Priebe brings a wealth of analytical expertise to the team. His unique skill set, honed through years of research and data analysis, positions him to make significant contributions to the repo market and the broader financial landscape of the city. Dr Priebe is enthusiastic about leveraging his scientific and analytical skills to drive innovation and transparency in the financial sector.

Please email any questions to: marian.priebe@londonreportinghouse.com

