École polytechnique

PSC

Final Report

Exploring the use of AI techniques for "intelligent anomaly detection" in data and its application to risk management

Authors:
Imane FARHAT
Toby Johnstone
Adam HALMI
Robin MICHARD
Joachim SASSON

Supervisors:
Martijn Vandervoort
Marcos Carreira

June 10, 2019

BNP PARIBAS

Contents

Introduction

Banks today rely a great deal on data to evaluate their positions. Extracting useful information from data often uses complex calculations, meaning that the metrics used by decision makers are often far removed from the raw data making it hard for them to spot anomalies. We plan to help BNP implement an algorithm to help detect errors in the calculation of such metrics. The idea is that our algorithm should use past data to recognize what is "normal" and then in real time flag potential erroneous results.

The area that BNP requested we work on is counterparty exposure profiles. The bank trades large amounts of derivatives with many different counterparties. For each counterparty there exists a risk that it might default, in which case it would not be able to fulfill its obligations under the derivative contracts. The bank manages this risk based on exposure profiles which measure the potential amount that can be lost in the future, should a counterparty default. Such computations are hugely complex as they require projecting potential future values of all derivatives traded with a given counterparty, sometimes far into the future. Moreover, the calculations depend on static data from many different systems within the bank. As a result, there are occasional errors in the computation of these exposure profiles. The goal of this project is to help the bank automatically identify such errors.

Accurately calculating counterparty credit risk is very important. The main reason for this is that regulators nowadays demand that banks closely monitor their exposure and make sure they have sufficient capital on hand. Indeed, one of the reasons that the financial crisis of 2008 was so economically dangerous is that there was much lending between institutions that were considered "too big to fail" meaning that when the Lehman Brothers went under, many other financial institutions threatened to follow, forcing governments to step in. Regulators now want banks to monitor their exposure to prevent a repeat of such a situation from inducing a domino effect in the financial industry.

BNP sent us the counterparty credit risk exposure profiles they had calculated from October 2013 to October 2018. This amounted to approximately 50GB worth of data. The objective of this project is to examine the data, find anomalies, and develop an algorithm that checks the banks data points as they are calculated, flagging those that seem suspicious.

1 Understanding Risk Management

We believe it is important before delving into the data provided to have a decent understanding of what it represents, giving us insights into what techniques could be the most effective for detecting anomalies. In this section, we give a brief outline of how and why the data are generated in an attempt to understand what kind of motifs can be expected.

Counterparty risk is a sub-class of credit risk. It measures the risk associated with the default of a counterparty that owes the bank money on derivative contracts. In the case of BNP, the data provided show that the bank is exposed to risk based on a wide variety of asset classes including but not limited to interest rate derivatives, foreign exchange derivatives and equity derivatives. A counterparty is said to default when it fails to meet the obligations under a contract. Such defaults can result in large losses for the bank, making it important to measure the bank's exposure to such risks. The standardized approach for measuring counterparty credit risk exposure (SA-CCR) is detailed by the Basel Committee on Banking Supervision in a report from 2014 [1]. The exposure under the SA-CCR consists of two components: replacement cost (RC) and potential future exposure (PFE). The mathematical formula which gives this exposure is:

$$\text{Exposure At Default Under } SA = 1.4 \cdot (RC + PFE)$$

The replacement cost is the immediate loss if the counterparty defaults. It is therefore calculated as the maximum of the value of the contract for the bank and zero.

The potential future exposure is a necessary term since the main reason that it is difficult to calculate the exposure at default is that the amount of exposure is uncertain due to the random nature of the contract's pay-offs. The potential future exposure is the credit exposure on a future date modelled with a specified confidence interval (for example 95%). The idea is to calculate the probability distribution of the value of the contract at a future date which can then be used to find an upper bound on the exposure at that date with a high level of confidence. The models used to calculate the potential future exposure are often complex, relying on large quantities of historical market data to study the profit/loss distribution of the asset class over the desired time-period while taking into account potential differences between the past and present.

The 1.4 coefficient is set by the Basel Committee on Banking Supervision

However, to give us an idea of how PFE is expected to evolve over time, we can take a look at a stochastic model. A relatively simple model that is often used in finance is the the Geometric Brownian Motion (GBM) stochastic process. It combines Brownian motion with volatility and drift components. It's stochastic differential equation is:

$$\frac{dS_t}{S_t} = \mu \cdot dt + \sigma \cdot W_t$$

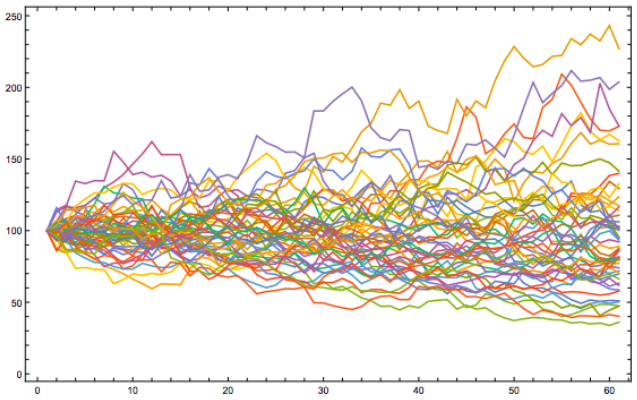


Figure 1: Simulations of Geometric Brownian Motion

Where S_t is the stochastic process modelling the asset price, μ is the percentage drift, σ is the percentage volatility and W_t is a Wiener process. In this simple case, with σ and μ constant, the explicit solution is well known[2] . However, in models used by BNP's risk department it is likely that σ and μ vary and other parameters come into play, hence the complexity of the bank's calculations. Using this equation we can calculate potential paths of an asset price. Then with a large number of simulations Monte Carlo methods can be used to estimate the distribution of probability of the price at a certain time as well as a high confidence upper-bound (illustrated by previous figure).

In practice we find the following shape for the 95% confidence interval:

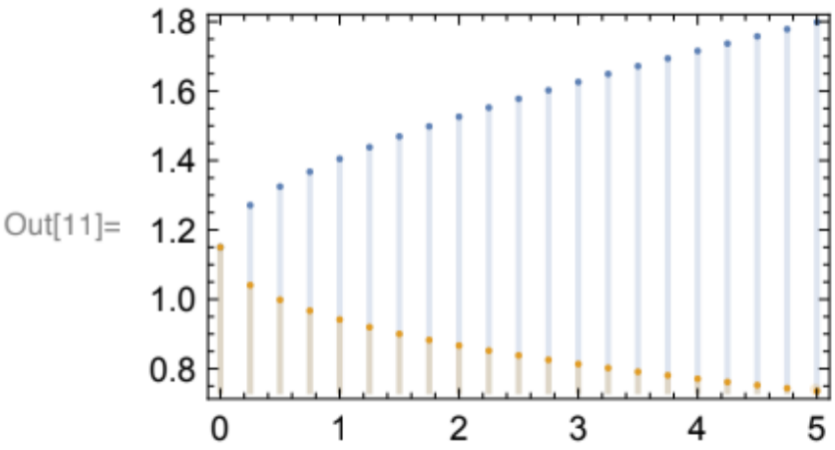


Figure 2: 95% confidence interval for GBM over time

We would expect the PFE of a single trade to follow a path similar to that of the upper bound of the envelope in figure 2.

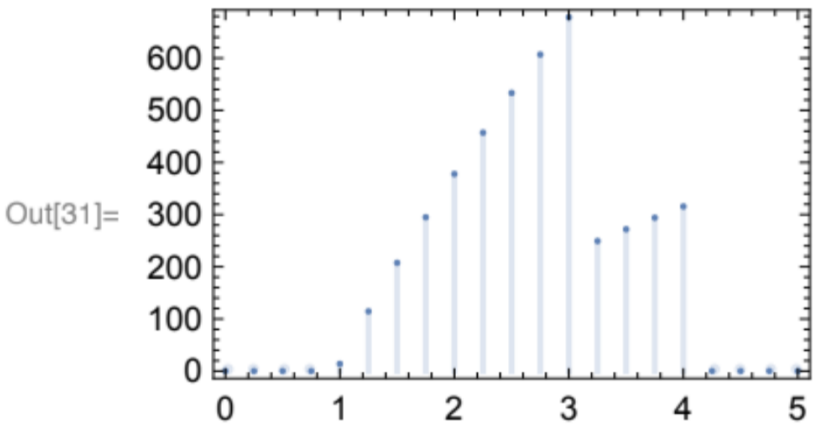


Figure 3: Expected shape of potential future exposure

If a contract matures, we expect a drop in the PFE and therefore in the exposure profile. If more than one contract is open with a single counter-party, the exposure to the counter-party is the sum of the exposures for each contract. Hence, figure 3 illustrates the expected potential future exposure of the bank to a counterparty with which it has two trades, one maturing at time 3 and the other maturing at time 4.

However, in some cases, different contracts with a single counter-party can offset one another. This happens when they are in the same hedging set and represent bets that the market will move in opposite directions. Hedging sets are groups of derivatives based on the same fundamentals and therefore expected to be highly correlated. For example, all derivatives that reference interest rates of the same currency in the same maturity category are in the same hedging set. Partial offset is permitted on derivatives of interest rates of the same currency but different maturity categories.

With BNP often having many trades with a single counterparty, each trade with different characteristics and some partially offsetting each other, it is likely that the exposure profiles we observe will be much more complex than those the simple model provided above. However, from our simplistic understand of exposure profiles we can extract some important information. For one, it is clear that it is important, when looking for anomalies, to keep exposure profiles based on different types of contract separate. This is because important factors in the calculation of exposure profiles are the volatility and the drift of the asset, as they make a big difference to the calculation of the potential future exposure. Moreover, the behavior of the couterparties trading with BNP is likely to differ depending on the asset class, leading to different trading patterns and contracts of different tenor[1].

[1] the amount of time left until a financial contract expires

2 Anomalies in Trade Counts

Originally, the data we received gave us access to exposure profiles calculated by BNP for a variety of counterparties over a variety of dates. We decided to request to be given access to the number of trades with each counterparty at each date as we believed it would help us find problems in the credit risk data, since one of the potential sources of errors is forgetting trades in the calculation. As well as potentially helping us detect outliers, the tradecount is a way to explain some anomalies; when a sudden drop in the exposure profile coincides with a sudden drop in the trade count it is likely that trades have been omitted. BNP kindly sent over the requested data. It covers the period of time between 2013 and 2018 (same as the exposure profiles), and is relative to over 100 sites (contract types, for example foreign exchange derivatives, bonds, etc...). We decided to first attempt to find anomalies in the number of trades, which should allow us to hone in on problematic counterparties and exposure profiles. To find the anomalies, one potential technique is to attempt to predict the today's trade count based off data from yesterday and before and then compare the actual value to the predicted value. To do this we use an auto-regressive model.

2.1 Auto-regressive model

A time series is a sequence of measurements of the same variable made over time. Usually the measurements are made at evenly spaced times - for example, monthly or yearly. Let us first consider the problem in which we have a variable y measured as a time series. For example y could measure global temperature, with measurements taken each year. An autoregressive model is when a value from a time series is regressed on previous values from that same time series. for example, y_t on y_{t-1}: $y_t = \beta_0 + \beta_1 \cdot y_{t-1} + \epsilon_t$.

In this regression model, the response variable in the previous time period has become the predictor (or covariate) and the errors have our usual assumptions about errors in a simple linear regression model. We call ϵ_t the residual. The order of an autoregression is the number of immediately preceding values in the series that are used to predict the value at the present time. So, the preceding model is a first-order autoregression, written as AR(1).

If we want to predict y this year (y_t) using measurements of global temperature in the previous two years (y_{t-1},y_{t-2}), then the autoregressive model for doing so would be: $y_t = \beta_0 + \beta_1 \cdot y_{t-1} + \beta_2 \cdot y_{t-2} + \epsilon_t$. This model is a second-order autoregression, written as AR(2), since the value at time t is predicted from the values at times $t-1$ and $t-2$. More generally, a kth-order autoregression, written as AR(k), is a multiple linear regression in which the value of the series at any time t is a linear function of the values at times t−1,t−2,...,t−k.

Autocorrelation and partial autocorrelation

The coefficient of correlation between two values in a time series is called the autocorrelation function (ACF) For example the ACF for a time series yt is given by: $Corr(y_t, y_{t-k})$.

This value of k is the time gap being considered and is called the lag. A lag 1 autocorrelation (i.e., k = 1 in the above) is the correlation between values that are one time period apart. More generally, a lag k autocorrelation is the correlation between values that are k time periods apart.

The ACF is a way to measure the linear relationship between an observation at time t and the observations at previous times. If we assume an AR(k) model, then we may wish to only

measure the association between y_t and y_{t-k} and filter out the linear influence of the random variables that lie in between (i.e., y_{t-1},y_{t-2},. . . ,y_{t-k+1}), which requires a transformation on the time series. Then by calculating the correlation of the transformed time series we obtain the partial autocorrelation function (PACF).

The PACF is most useful for identifying the order of an autoregressive model. Specifically, sample partial autocorrelations that are significantly different from 0 indicate lagged terms of y that are useful predictors of yt.

Graphical approaches to assessing the lag of an autoregressive model include looking at the ACF and PACF values versus the lag. In a plot of ACF versus the lag, if there are large ACF values and a non-random pattern, then likely the values are serially correlated. In a plot of PACF versus the lag, the pattern will usually appear random, but large PACF values at a given lag indicate this value as a possible choice for the order of an autoregressive model. It is important that the choice of the order make sense.

2.2 Application to number of trades

In this report we chose a single site[2] to focus on: G-FXD, because it is the biggest in size. For this particular site, we have over 12 million entries over more than 5 years (from 2013 to 2018).

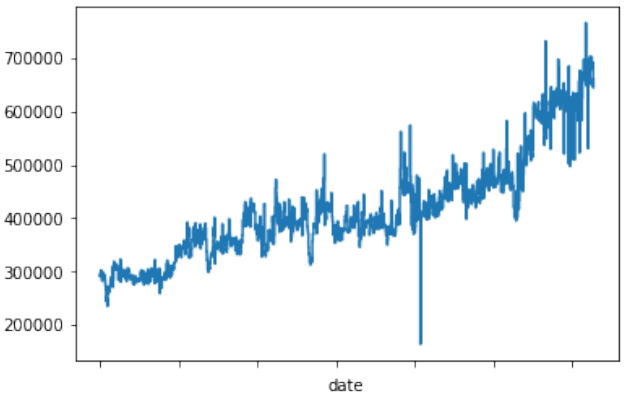


Figure 4: Plot of the trade counts for site G-FXD with respect to dates

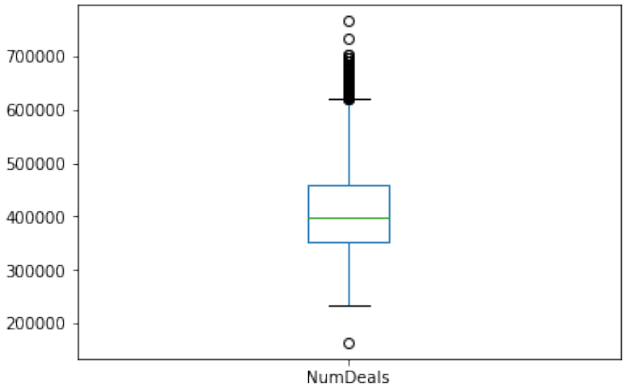


Figure 5: Boxplot of the number of deals for the G-FXD site

We then compute the relative differences of the data set to get an idea of how much the data is dispersed:

[2]A site here refers to a category of contracts traded by a BNP division. G-FXD is most likely Global Foreign Exchange Derivatives

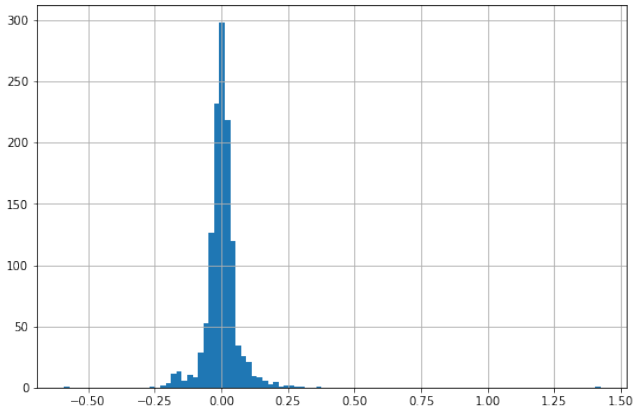


Figure 6: Histogram of the relative differences of the data on the G-FXD site

The histogram shows that for a large portion of the data, there is no significant relative variation between two consecutive dates. There are however outliers that diverge considerably from the previous date's value.

For instance, we have found that there are 12 dates on which the value of relative difference is greater than 3 times the standard deviation of the relative differences time series.

We then plotted the correlation of the lagged data, which means the correlation between the values of y_t and y_{t-1}.

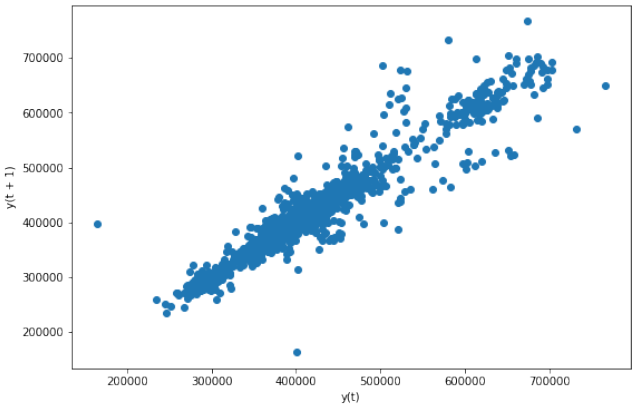


Figure 7: Correlation plot of the number of trades between two consecutive dates

This plot shows that the data is well centered around the line $y_t = y_{t-1}$, which means the number of trades on a certain day is very correlated to the number on the previous day, and that correlation is exactly : $\text{corr}(y_t, y_{t-1}) = 0.95677$. It is worth noting that the two outliers in the bottom left of the chart most likely correspond to the sudden trough seen in figure 4.

Auto correlation plots are often used for checking randomness in time series. This is done by computing auto correlations for data values at varying time lags. If time series is random, such auto correlations should be near zero for any and all time-lag separations. If time series is non-random then one or more of the auto correlations will be significantly non-zero. The horizontal lines displayed in the plot correspond to 95% and 99% confidence bands. The dashed line is 99% confidence band.

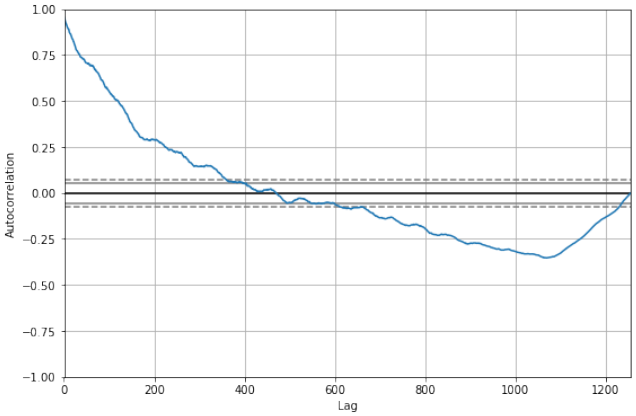


Figure 8: Auto correlation function with respect to the lags

The auto correlation function shows that confidence intervals are drawn as a cone. By default, it is set to a 95% confidence interval suggesting that correlation values outside of the cone are very likely a correlation and not a statistical fluke (an outlier).

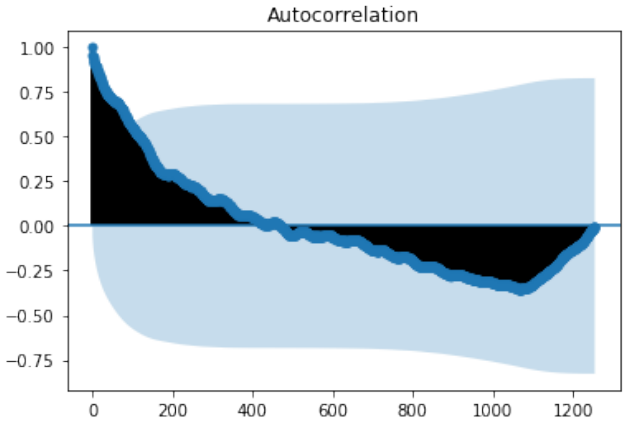


Figure 9: Auto correlation plot

By default, all lag values are printed, which makes the plot noisy. We can limit the number of lags on the x-axis to 50 to make the plot easier to read.

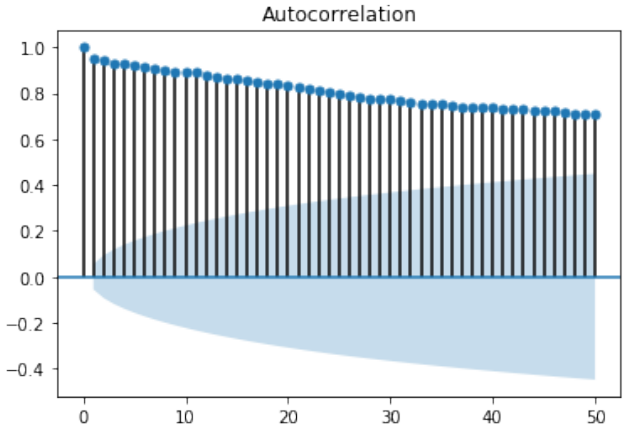


Figure 10: Auto correlation plot with 50 lags

2.3 Improving the regression

Based on the observation of the good correlation between y_t and y_{t-1} in the previous section, we decide in this section to stick to first order auto-regressions. In order to predict the number of trades on a day t, we use the number on the day t-1. We train the model on a certain period of time preceding the date t, for example the 21 days, in order to estimate the best coefficients β_0 and β_1 such that in theory: $y_t = \beta_0 + \beta_1 \cdot y_{t-1}$ with a very good approximation.

We choose to set the length of the training set to 21 days. This means that for each day we use the previous 21 days as a training set and we have a sliding window of 21 days for each new day.

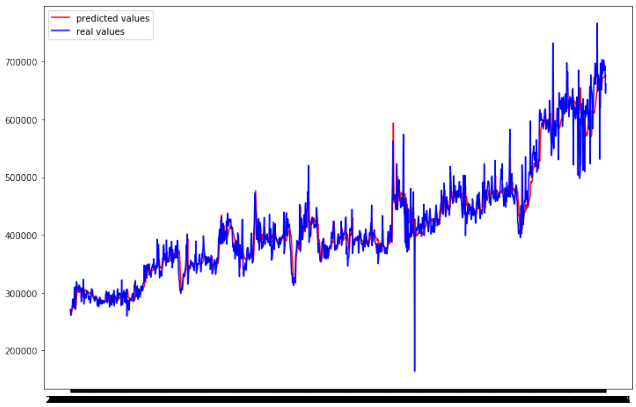


Figure 11: The predicted values and real values of number of trades using a first-order autoregression model with a training set of 21 days

On figure 11 we observe the predicted values (in red) and the real values (blue) of the number of trades using a first-order auto-regression model with a training set of 21 days.

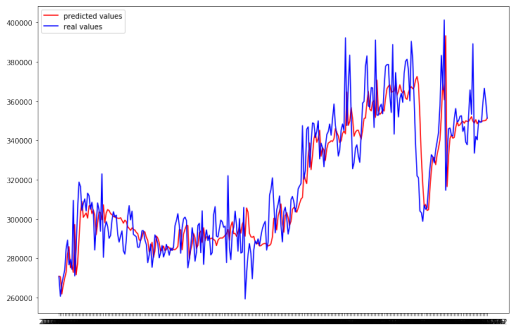


Figure 12: Reduced Timeframe from figure 11

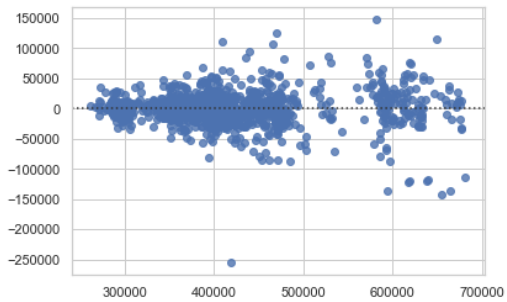


Figure 13: Relative residuals of the regression

2.3.1 Data Visualisation

Trade Count Time Series Since the fundamentals underlying the data are from a vast range of financial products, the exposure profiles of the different sites have very different behaviors. This was to be expected as counterparties don't trade foreign exchange derivatives the same way they trade interest rate derivatives. Hence the difference in volatility that we observe on the trade count charts of G-FXD and G-IRD in figure 11. Moreover, it seems that there are many anomalies where the trade count drops suddenly from one day to the next, only to return to its previous value the day after. These sudden troughs are most likely due to trades that were forgotten in the count. Our proposed technique to find which trades were forgotten

is explained in section 2.4. Here we concentrate on improving the regression to get the best possible predictions and hence accurately determine the dates where problems arise.

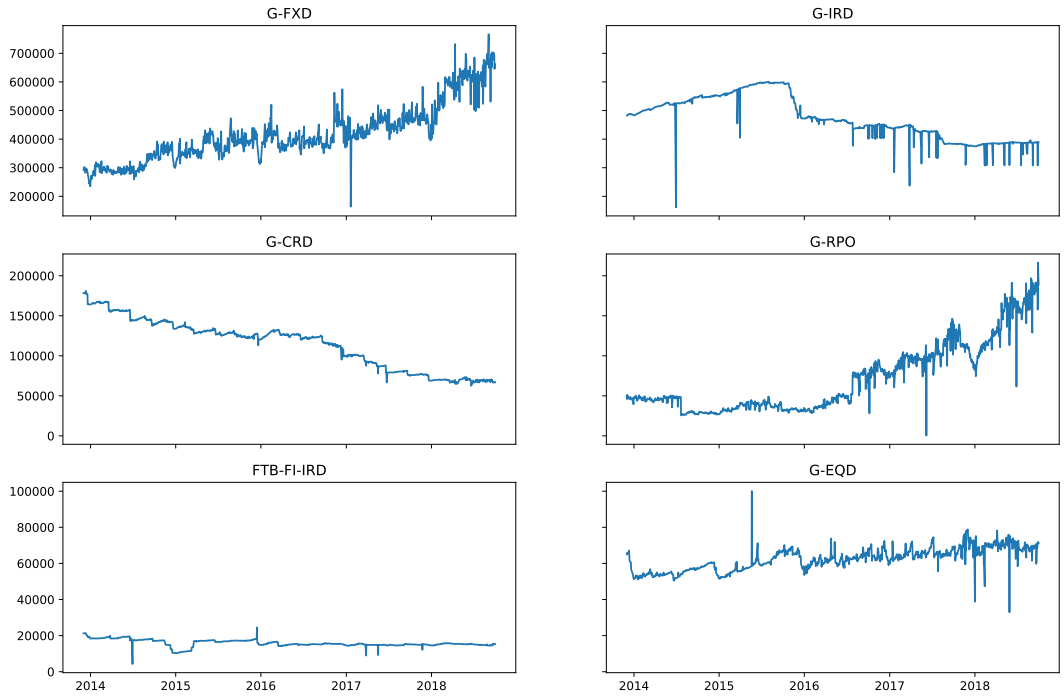


Figure 14: Number of trades over time for several different counterparties

Analysis of the values at time t with respect to those at time $t-1$ Figure 15 illustrates the correlation between TC(t) (the Trade Count at time t) and TC(t-1) that we calculated earlier and makes a linear regression predictor of TC(t) seem inviting. Hence we attempt to find α and β such that the following formula holds:

$$TC(t) = \alpha + \beta \cdot TC(t-1)$$

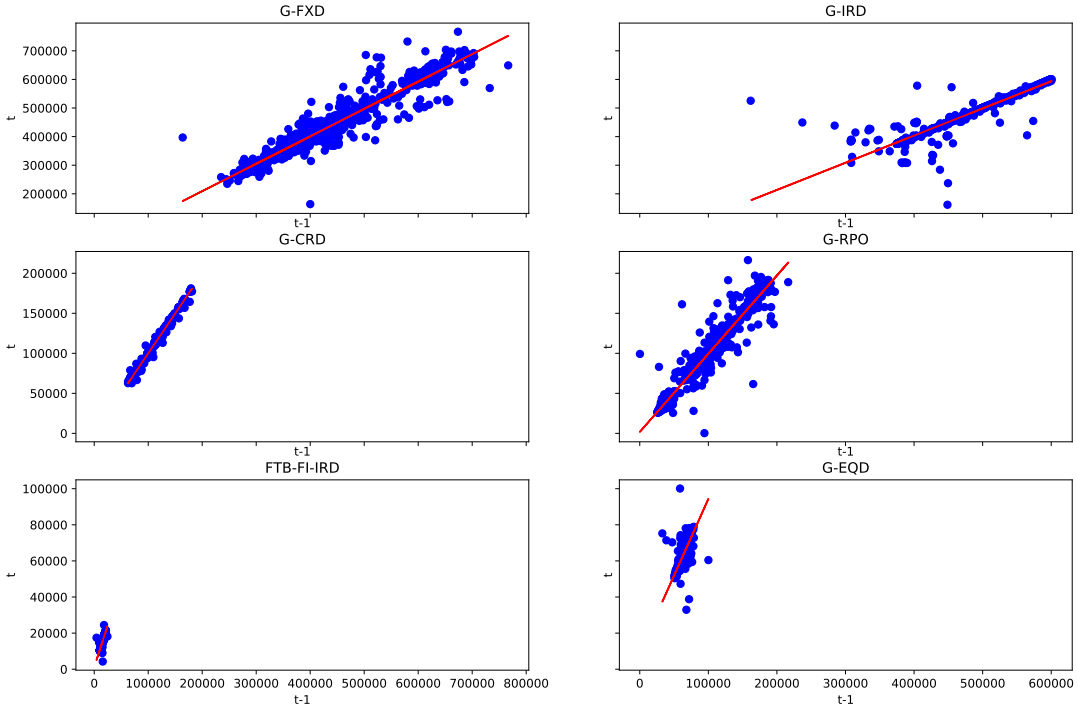


Figure 15: Plot of NT(t+1) with respect to NT(t)

2.3.2 Implementation of simple autoregression techniques

The first technique we attempt it trying to predict TC(t+1) using TC(t) having refined the model using $(TC(t), TC(t-1)), \ldots, (TC(t+1-(n-1)), TC(t-(n-1)))$ ie the n most recent measurements. For the first algorithm, n is given as a parameter and is known as the *sample size*.

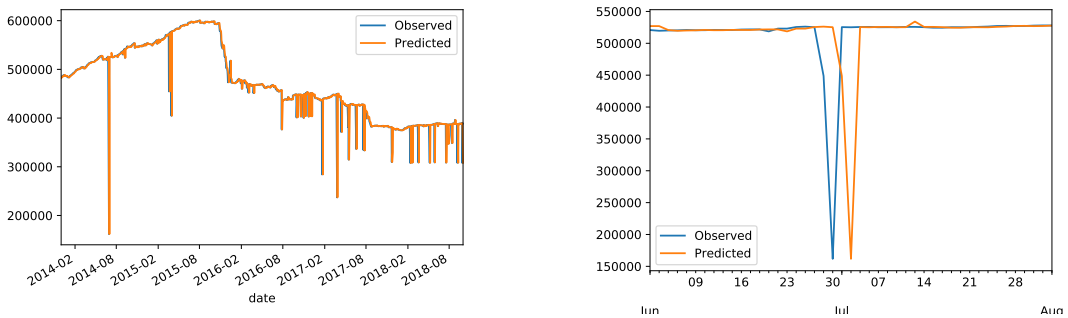


Figure 16: $n = 2$

As can be seen on figure 16 with $n = 2$ we get more or less $TC(t+1) = TC(t)$ this allows us to detect anomalies, but also gives many false positives.

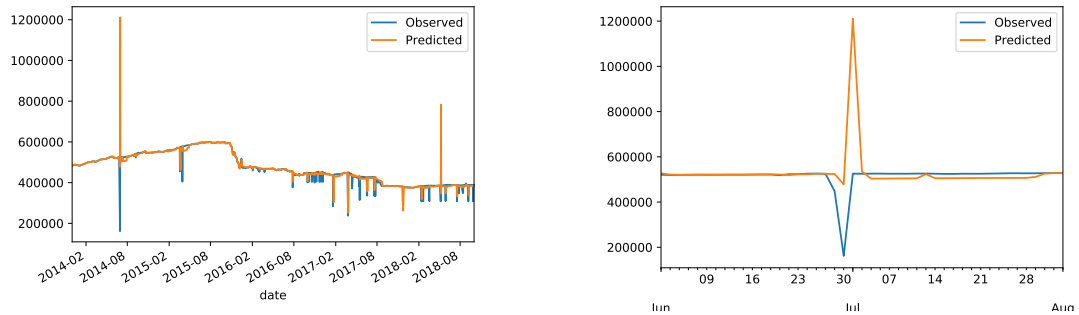


Figure 17: $n = 21$

Figure 17 shows that with $n = 21$ the predicted values seem closer to what they should be, but there are still a great deal of false positives.

2.3.3 Looking for a better method

One of the big sources of mistakes in the algorithm comes from the fact that outliers are left in the data and they skew the predictions that follow them; both because they alter the regression (figure 18) and because predicted value of $TC(t + 1)$ is off when the regression is right and $TC(t)$ is an outlier.

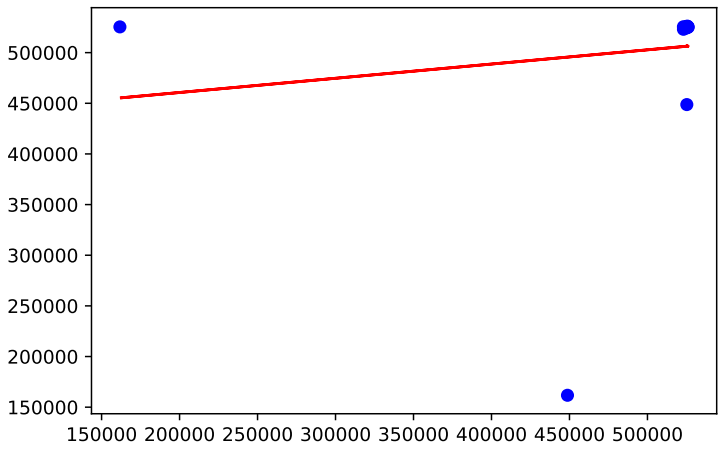


Figure 18: Regression skewed by an outlier

Basic Method To counter this problem, we implemented an algorithm that works on the same principles as the previous one, but removes outliers from the data as it detects them. For this algorithm, outliers are defined as points which differ from the previous by more than 5 times the standard deviation observed over the previous n days.

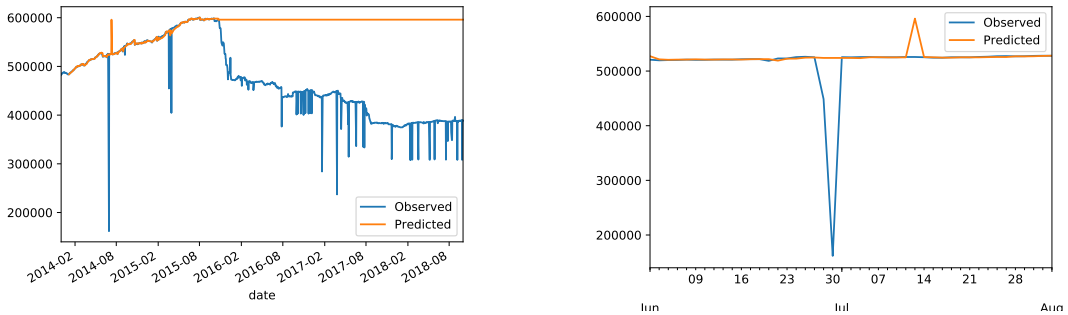


Figure 19: Basic method

While this method seems to be an improvement on the previous ones, it can be thrown off by sudden sustained alteration in the trend as seen on figure 19.

2nd version of the Basic Method In this method, we work under the hypothesis that outliers are sufficiently rare for us to consider that there is no chance of 2 successive outliers. This should allow us, in the case of a sudden change, to only return 1 false positive before re-adapting to the new situation. To do this, whenever an outlier $TC(t)$ is detected, $TC(t+1)$ is given a free pass and considered to not be an outlier.

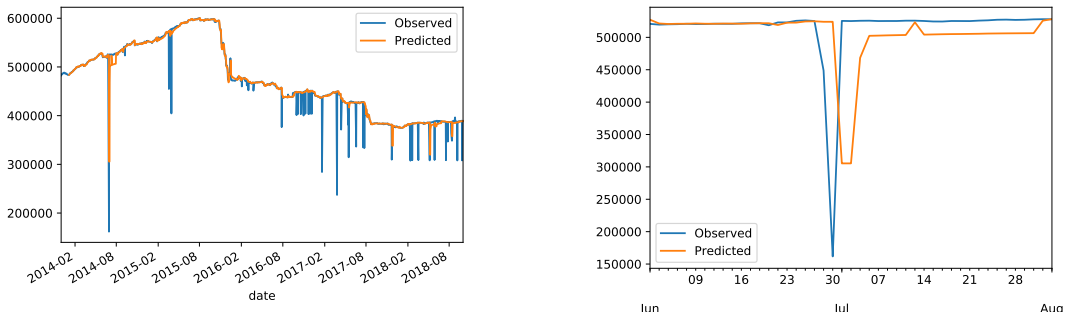


Figure 20: Basic method V2

This greatly improves our algorithm, though problems still remain, especially when there are two anomalies in succession in the data.

3rd version of the Basic Method In this algorithm, we no longer make the assumption that the outliers are isolated. The threshold for outliers is still 5 times the standard deviation of the n previous days, but when an outlier is detected, it goes up to 10 times the standard deviation for the following point, then up to 15 if that is also an outlier, and so on and so forth so long as outliers are detected. This method works well, as shown by figure 21.

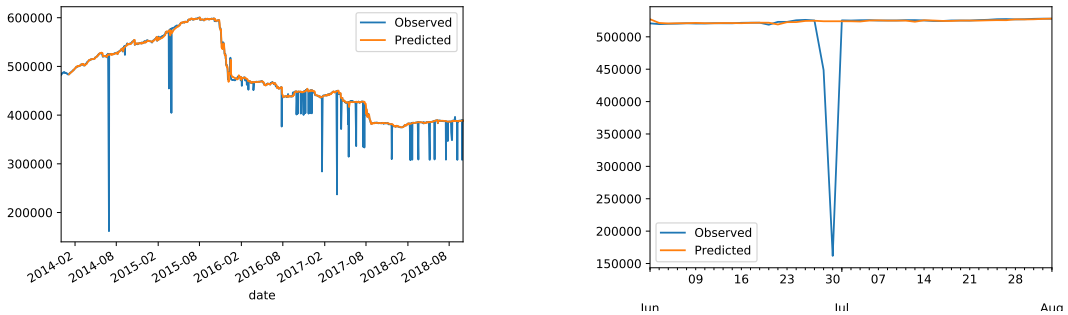


Figure 21: Basic method V3

However, despite the quality of this algorithm, it can run into trouble as illustrated by figure 22.

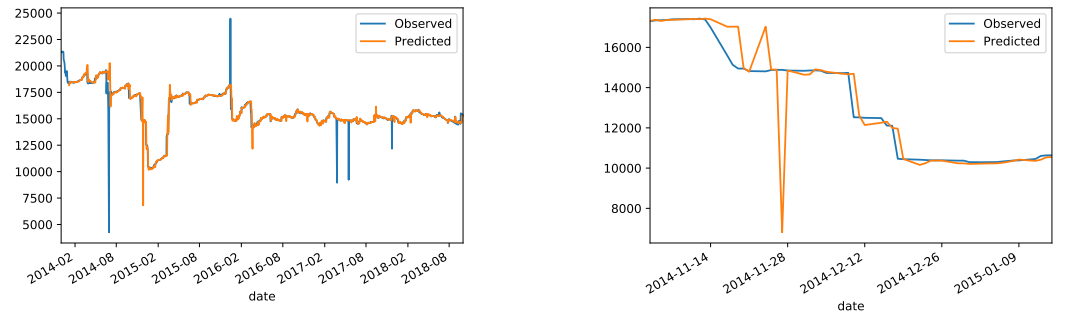


Figure 22: Basic method V3

To avoid this kind of problem we attempt to bound the coefficients of the linear regression.

4th version of the Basic Method Our solution to this problem is to implement the same algorithm as the 3rd version but instead of using a classic linear regression, we use a Ridge linear regression [3] which minimises the following loss function:

$$||TC(t+1) - \frac{\begin{pmatrix}1 & TC(t)\end{pmatrix}}{||\begin{pmatrix}1 & TC(t)\end{pmatrix}||} \cdot \begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix}||^2 + \alpha \cdot ||\beta||^2$$

In such a loss function, α represents the importance of bounding the coefficients of β with respect to the importance of minimising the residuals. We chose $\alpha = 1$. This method allows us to limit the extent of the errors in the prediction but, as can be seen on figure 23, the predictions can be thrown off in other areas.

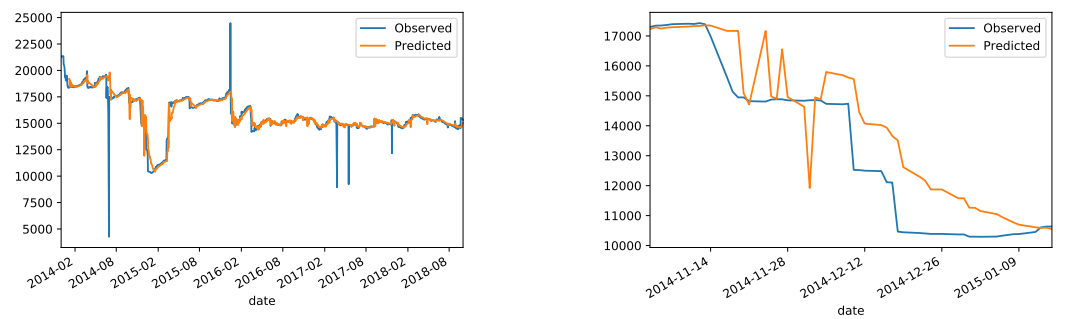


Figure 23: Basic method V4

2.3.4 Comparing the 4 methods

Here we compare and contrast the 4 different methods with different values of k ranging from 10 to 63.

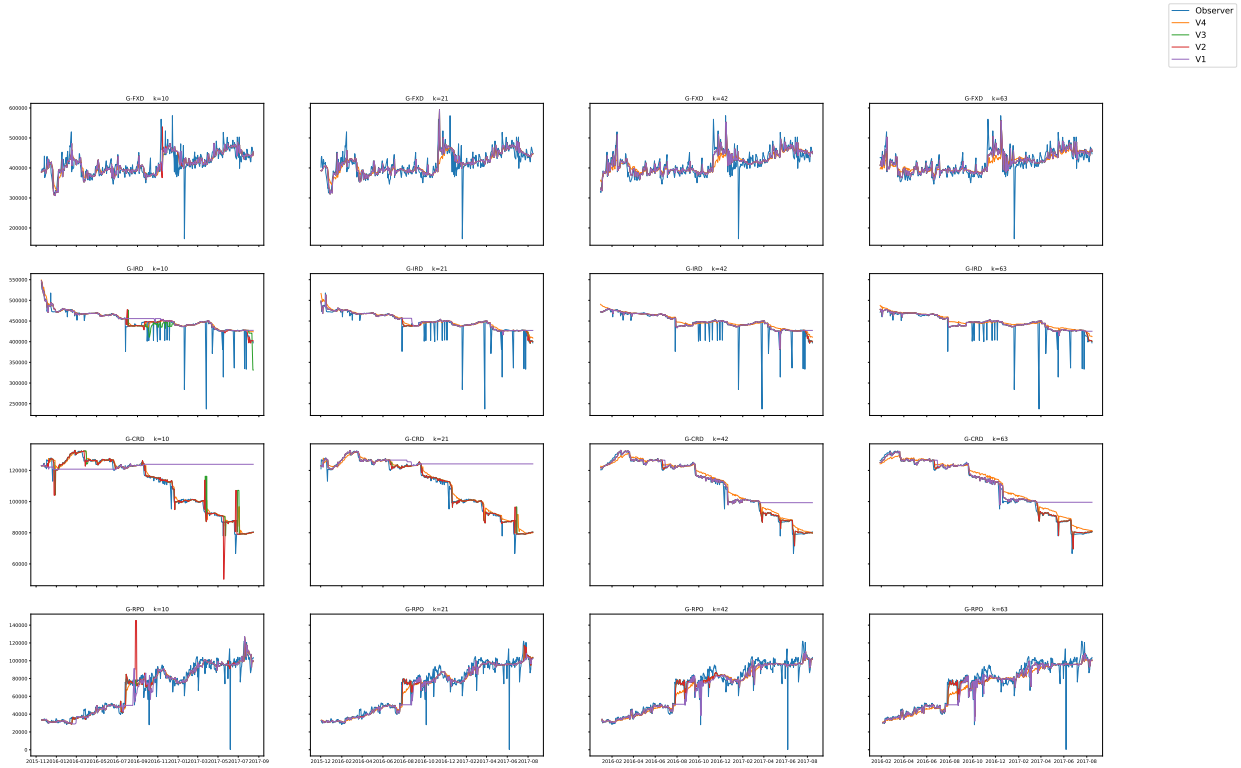


Figure 24: Observed and predicted for the different methods; over various sites and values of n

Site	k	Basic V4	Basic V3	Basic V2	Basic
G-CRD	10	105	101	61	1218
	21	44	36	24	1157
	42	30	19	13	1157
	63	17	10	7	1091
G-FXD	10	24	22	21	31
	21	9	8	8	9
	42	5	2	2	2
	63	3	3	3	3
G-IRD	10	115	90	64	867
	21	49	48	44	781
	42	42	37	35	329
	63	42	38	34	329
G-RPO	10	55	57	52	283
	21	22	25	21	239
	42	12	10	9	192
	63	5	8	7	191

Table 1: Number of problematic dates detected

On figure 24 and tables 1 and 2, we observe that version 3 with $n = 42$ seems to give a good compromise: keeping a low mean square error while only returning dates that seem truly problematic. That being said, the latter criterion is particularly difficult to measure.

2.3.5 Limitations of our method

The lack of labelled data Not having any points which were confirmed anomalies or normal made measuring the performance of the different algorithms we used difficult as we could never be 100% sure whether the values they returned were anomalies or not.

The removal of NaNs We made the decision to remove the NaN (Not a Number) elements from the data in this part. This may skew the results in that, in some cases, NaNs should be returned as outliers.

The calculation of the error The error calculated in table 2 excludes the problematic dates, such that as the number of problematic dates increases, the error decreases. We also need to take into account the number of problematic dates detected (see table 1).

Use of a limited number of sites We only used the 10 sites with the most data due to runtime considerations.

Site	k	Basic V4	Basic V3	Basic V2	Basic
G-CRD	10	1980	2411	2645	1492
	21	1892	1502	1511	518
	42	2232	1424	1436	486
	63	2534	1558	1557	1602
G-FXD	10	23519	23304	23743	23017
	21	25021	24413	24413	24189
	42	28078	25946	25946	25946
	63	29234	25854	25854	25854
G-IRD	10	4978	5515	17855	3932
	21	4581	2692	13721	3859
	42	6875	3389	15028	3818
	63	8848	2963	15899	3293
G-RPO	10	6262	6071	6564	6839
	21	7024	6487	6477	6878
	42	7993	7179	7178	8044
	63	9034	7691	7690	8297

Table 2: Means square error between the observed data points and those predicted, with the problematic dates removed

2.4 Pinpointing anomalies in the trade counts

Having located a date with a suspicious trade count, we attempt here to find if there are any counterparties that appear to be missing trades.

2.4.1 Finding problematic dates

In the previous section, we refined out auto-regression model to get the best predictions possible for TC(t). Once the prediction is made, there are several ways to define an anomaly. One is to, as previously, classify a point as an anomaly is it deviates from its predicted value by more than 5 times the standard deviation of the n previous days. Another possible classification technique is to define an anomaly as a point with relative residual between the predictive value of trade counts given by the auto regression and the effective value of the trade counts is greater than 15%. The objective is to write a method based on the observation of the data which gives the list of the dates and the counterparty ids for which we detect an anomaly. In all this section and below in the document, we will call these dates and counterparties problematic dates and problematic counterparties.

Before writing our methods, it was essential to define a criterion for both problematic dates and counterparties. Thus, we first decided to consider problematic the dates for which the relative residual between the predictive value of trade counts given by the auto regression and the effective value of the trade counts was above 15%. This decision was essentially based on observation of the data at different dates.

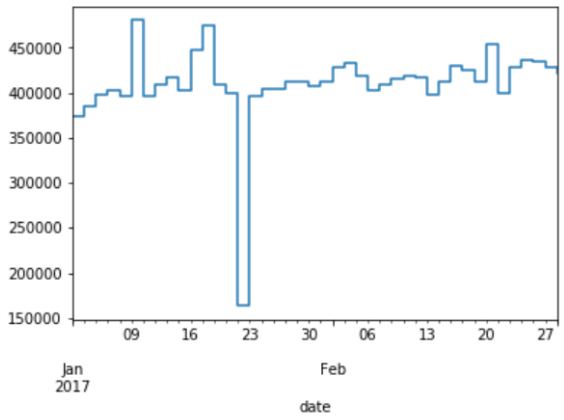


Figure 25: Example of gap in the Trade Counts temporal series

Considering the chart from figure 25, we expect that our general method applied to the data frame corresponding to the site G-FXD would return a set of dates containing the date-time type element 2017-01-20. Once that method was written and the detection of problematic dates was done for each site, we synthesized this information in a matrix with lines indexed by the dates contained in the data set and columns indexed by the name of the different sites.

Before moving to the next step of the analysis, we attempted to verify the results we had. In order to check the quality of our intermediate results, we first verified if the particular problematic dates that we had recognized during the qualitative phase were returned by our function. Then, in greater detail at the results, we noted that the proportion of problematic dates was low. This was to be expected suggesting that the algorithm is not completely off.

2.4.2 Finding problematic counterparties

Finding problematic dates basically consisted in finding dates at which the total number of trades was abnormal according to our criterion. Finding problematic counterparties was a more subtle task. It consisted in finding which counterparties were responsible for the abnormal relative residuals on the trade counts at a given problematic date.

The choice we made was first to find the most obvious anomalies. Given a problematic date t_0 for a given site, we looked at the number of trades for each counterparty at this date for the given site. By doing this with particular problematic dates for the site G-FXD, we noticed that a lot of counterparties had a trade count equal to zero at this date. Of course, finding a zero does not necessarily indicate that a certain counterparty is problematic. Indeed, it is entirely possible that a contract between the bank and the given counterparty matured before the given date. In this case, it makes sense for the counterparty to have a zero trade count.

However, in the case of a zero trade count, we thought that the behavior of the trade counts time series at the dates preceding and following t_0 could be a relevant information. For the sake of clarity, we will call these dates respectively $t_0 - 1$ and $t_0 + 1$ even though those dates could be separated by more than one day. The different behaviors we were likely to encounter were the following:

- The number of trades is steady at zero before the date t_0.
- The number of trades is steady at zero after the date t_0.
- The number of trades is strictly positive at t_0-1, equals to zero at t_0 and strictly positive at t_0+1.

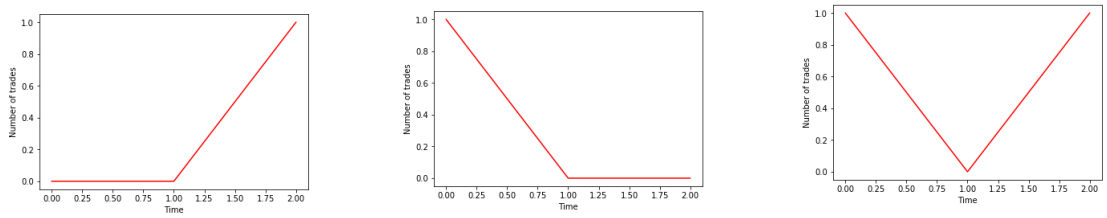


Figure 26: Typical behaviors of the trade counts around a problematic date

The two first cases are not problematic. Indeed, the first case could indicate that a previous contract had matured before t_0-1 and a new one was signed at t_0+1. Regarding the second case, it could indicate that a contract matured at t_0. However, the third case arouses our suspicion and may be a symptom of missing trades in the data.

Hence, we focused our attention on counterparties of the third type. The first step of this mission was to generate the table of the trade counts at t_0-1, t_0 and t_0+1 for a given site, where t_0 is a problematic date for this site. We thus wrote a method which takes as an argument the name of a site and a date and returns the table described above.

cpty_id	t0-1	t0	t0+1
000ad475a319418206dc0dd9ea23ce46	80.0	0.0	69.0
000d78b9a9574e0d0bce0ec67f177778	1.0	1.0	1.0
001bc7036565fbefe0b326b0dd5d3375	1.0	1.0	1.0
001d4279c24380246237a4104fa5c4f5	2.0	1.0	1.0
001f53cb78f850a6be09d91a6df8d0ed	NaN	NaN	2.0
0020e8e4f51fd3fc894c0ca71f3eee51	3.0	2.0	2.0
00217efcd18cb3f8bb70ba339868f6ab	1.0	1.0	1.0
002a397cdb798711dd5ea3230a3c865c	3.0	3.0	3.0
002ebaa521faecb6037f7690c9c38ef9	0.0	NaN	NaN
0031dc655f65324036b49417f55990d8	23.0	0.0	22.0
0040522620c943524f2d460598d75aeb	2.0	2.0	2.0
004f14b83337b6e05a36366303588292	9.0	9.0	9.0
0050a8bc9fc55580fd4b1289f5178e16	2.0	2.0	3.0
0059a1daa87719f47a9a5cf272742fa1	1.0	1.0	1.0

Figure 27: Comparison of the trade counts

Note: We can see in figure 27 that some counterparties have NaN (Not a Number) as number of trades. It is worth noting that NaNs are different from zeros. The presence of a NaN in a column indicates that the value of the trade count does not exist at the date given by the index of the column for the counterparty id given by the index of the line. This could be due to the fact that the regression has a training set of 21 days, thus the 21 first dates do not have any predicted value. It could also be caused by holes in the data, where for one reason or another the number of trades was not recorded by the bank for that counterparty on that particular day.

Once that auxiliary method was written, we wrote a general method which takes the name of a site and a date as arguments and which returns the names of all problematic counterparties for this given site and that given date. That method applied to the site G-FXD at the problematic date 01-20-2017 returns the following table.

cpty_id	t0-1	t0	t0+1
000ad475a319418206dc0dd9ea23ce46	80.0	0.0	69.0
0031dc655f65324036b49417f55990d8	23.0	0.0	22.0
00778f224e6607b5ad5921e1001cdb55	21.0	0.0	9.0
00a163030b88f88eb87bb80da43c7538	23.0	0.0	24.0
00e12d9f6a3275335ec352cb29a722fa	16.0	0.0	12.0
01773264bb18cfaa3c1ed929b0b1913c	245.0	0.0	302.0
01e134f4e59f58fefe11a65054f72784	9.0	0.0	8.0
01e2fb2ac5e5851672e8b40f3f53aa91	10.0	0.0	10.0
0204db4e702f84ae3d474f0b467ffb9c	9.0	0.0	8.0
025630b179cf1abe5b73c917b60d5e32	312.0	0.0	308.0
02a31488c632e1ece618085cd8e5372a	18.0	0.0	18.0
02bfb3e83075274fc85bfbfdfc6d573e	1.0	0.0	1.0
02ef94e6e8accd17de0ea0e246df4360	17.0	0.0	17.0
03053e4fc98b340c343310e6d64c5353	366.0	0.0	389.0

Figure 28: Table of some problematic counterparties for the site G-FXD

This table contains 14% of all the counterparties of G-FXD on the 20th of January 2017. However, this method may select a range of anomalies that is too wide. Indeed, some counterparties fit with the third behavior described previously but have a very small number of trades on the dates before and after the problematic date. For example, one of the counterparties selected by the method has one trade on the day before the problematic date and four on the date after which is not particularly shocking.

Thus, we decided to narrow the detection by considering as problematic the counterparties which have more than 4 trades on the day before the problematic date, 0 trades on the problematic date and more than 4 trades on the day after the problematic date. By doing so, we detected a set of 1158 problematic counterparties which reduced the proportion of problematic couterparties to 10,8%.

Hence we produced an algorithm capable of identifying dates on which the trade count seems suspicious, and identifying counterparties which may have been omitted on those days.

3 Anomalies in Exposure Profiles

3.1 Processing the data

Identifying anomalies in the trade count is all very well, but the real objective is to find problems in the counterparty credit risk exposure profiles (PE). BNP provided us with the results of its counterparty credit risk calculations from 2013 to 2018. The data was sent as tables where each line represents the exposure profile of a certain counterparty at a certain site at a certain date. The exposure of profile here is, for a given site, date and counterparty an estimate how the bank's exposure will vary over time. Figure 29 shows the exposure profile of a counterparty of the site G-IRD as calculated by BNP on the 28th of September 2018.

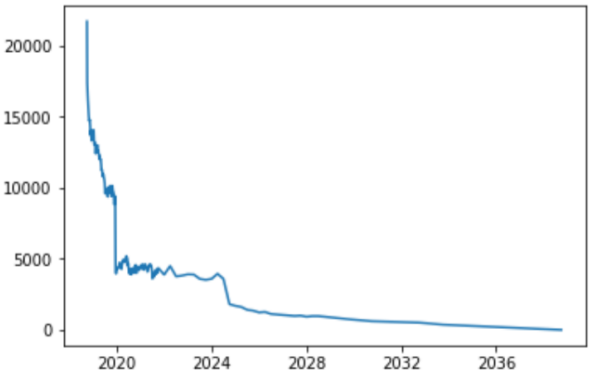


Figure 29: Example of exposure profile for a counterparty of the site G-IRD

The data we have does not contain estimates of the exposure for all days. It starts giving a value a day for the first two weeks, then a value a week for the next 3 years, and then a value every 3 months. This data configuration is not ideal for comparing PE's, so we decided to regroup the dates in four groups:

- The first group contains 14 days, for which we have data on each day. For the rest of the report this group shall be known as $2w$.
- The second group contains an entry once a week for 3 months. For the rest of the report this group shall be known as $3m$.
- The third has an entry each week over 3 years. For the rest of the report this group shall be known as $3y$.
- The fourth has long dates, ranging from an entry each 6 months to an entry each 3 years or more. For the rest of the report this group shall be known as *long*.

For each group, we computed the PEWA (Profile Exposure Weighted Average), using the following formula:

$$PEWA = \frac{<PE, dates>}{mean(dates)}$$

Where *dates* is for example the vector $[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]$ for the first group, containing days for which we have exposure values. The PEWA for the same G-IRD counterparty is shown in figure 30.

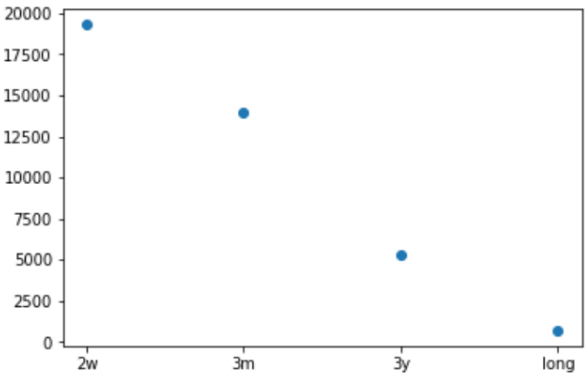


Figure 30: Profile Exposure Weighted Average from Figure 29

We simultaneously processed the data to group together in files by counterparty and calculate the PEWAs. After examining how the PEWAs evolved over time for several counterparties, we decided there was too much noise to attempt to detect anomalies directly. We therefore decided to sum the PEWAs of all the counterparties in each site to give the total PEWA of each site at each date, hoping that it would be easier to detect problematic dates in such data. Then, we merged the trade counts by sites and the PE exposures by sites (cf. table 3), so that we could include the Trade Counts when doing the linear regressions and looking for outliers.

site	date	2w	3m	3y	long	NumDeals
G-FXD	2016-01-11	4.505196e+07	2.617260e+07	4.474090e+06	153961.499042	383824
G-FXD	2016-01-12	6.608705e+07	3.321617e+07	4.748509e+06	149315.726384	387098
G-FXD	2016-02-11	4.370629e+07	2.563208e+07	4.452375e+06	153936.255703	474487
G-FXD	2016-02-12	6.386954e+07	3.233310e+07	4.714682e+06	147637.088124	469989
G-FXD	2016-03-10	3.752535e+07	2.311194e+07	4.001759e+06	121395.521426	447642
G-FXD	2016-03-11	4.285596e+07	2.536561e+07	4.422220e+06	153427.036470	444263
G-FXD	2016-04-11	4.210144e+07	2.499281e+07	4.383022e+06	152016.509971	358164
G-FXD	2016-05-10	3.877207e+07	2.316853e+07	3.972850e+06	121274.027792	375855
G-FXD	2016-05-12	5.968765e+07	3.091262e+07	4.536246e+06	145159.511604	380317
G-FXD	2016-06-10	4.074996e+07	2.375540e+07	4.013760e+06	121011.167536	410912

Table 3: First lines of the table for G-FXD site.

Before attempting any analysis, we checked that the dates were the same when merging the files. We also checked that the counterparties matched. Finally, we had to check if the appearance of NaNs (Not a Number) in the first '2w' columns for the PE exposures matched that of a NaN value in the trade counts for a given date and counterparty.

Problem 1: (Dates Matching) The dates didn't quite match. Indeed there were discrepancies between the number of dates in the Exposure Profile file versus the Trade Count file. In fact, the Trade Count files had more dates than the PE exposures, thus we took the intersection of dates which consisted of 1205 different dates. Though we lost some data doing this, it is worth noting that each date has a lot of transactions, so 1205 dates easily becomes 1 million lines of data for some bigger sites such as G-FXD.

Problem 2: (Counterparty Matching) Next step was to check if the counterparties do also appear in both files. Unfortunately, BNP Paribas did not use the same hash code when anonymising the counterparties for the exposure profiles as it did when anonymising them for the trade counts, so for a given date and site we can not check if the counterparties are the same. However, we came to the conclusion that we would not lose too much in dropping that level of detail and contenting ourselves to using the combined data to detect anomalies at the

site level over the aggregate values by dates especially since we intending on working this way anyway, at least at first. However, we coded the function that return the problematic counterparties so that BNP could use it since they have the original data.

Problem 3: (NaNs) Fortunately, no NaNs were found on one of the files and not the other. The opposite would have been illogical and in that case, we would have dropped those lines.

3.2 Regressions

3.2.1 Profile Exposure Weighted Average

Next we ran 2 types of auto regressions on the files we computed above. For each type, we ran regressions over the columns: 2w, 3m, 3y, long. Then, as we did for the Trade counts previously, we predicted the future value then computed the residuals allowing us to return the problematic dates for each column. What characterizes each regression are the explanatory variables (covariates).

Type 1: This one is the most basic auto regression as we try to explain each column using the same variable at t_1 to predict the values:

$$2w : (2w)_t = \alpha_1 \cdot (2w)_{t-1} + \beta_1$$

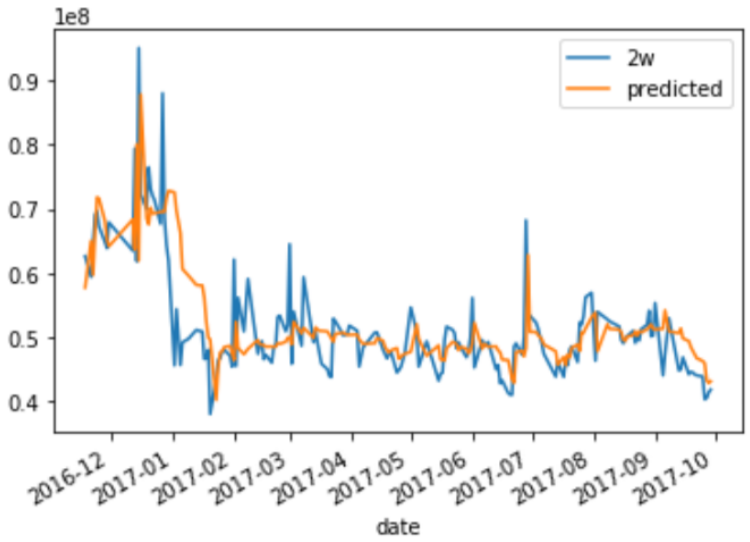


Figure 31: WPE G−FXD 2weeks Type 1 Regression

Type 2: In this type of regression, we include the other columns (without including Trade Counts) as covariates when trying to predict values instead of doing our prediction using only one column. For example, when predicting the value at time t for the 2w column we will be using the value t-1 from 2w column but also the t data from the other columns. Hence, as an example the formula for the 2 weeks prediction is:

$$2w : (2w)_t = \alpha_1 \cdot (2w)_{t-1} + \alpha_2 \cdot (3m)_t + \alpha_3 \cdot (3y)_t + \alpha_4 \cdot long_t + \beta_1$$

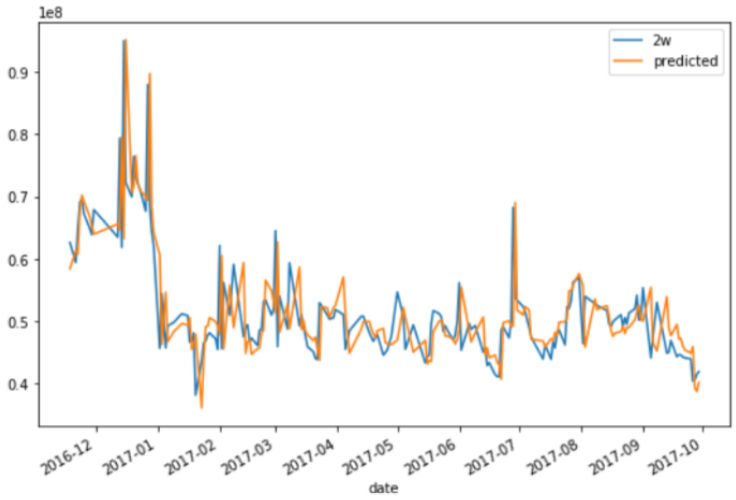


Figure 32: WPE G−FXD 2 weeks Type 2 Regression

After doing these 2 Types of regression for every column (2w, 3m, 3y & long), we now compare them by computing the mean of relative errors as well as the adjusted R^2 following these formulae:

$$MeanErrors = \sum_{i=1}^{n} \frac{predict_i - true_i}{n \cdot true_i} \qquad R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad \widetilde{R^2} = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1} \tag{1}$$

where SS_{res} is the sum of squares of residuals, SS_{tot} is the total sum of squares, n is the number of training data and k is the order of the autoregression. The results can be found in the following tables:

Table 4: $\widetilde{R^2}$

Type	2w	3m	3y	long
Type 1	48%	31%	53%	49%
Type 2	48%	28%	39%	27%

Table 5: Mean Errors

Type	2w	3m	3y	long
Type 1	7.5%	6.9%	7.3%	16%
Type 2	7.4%	6.6%	6.6%	16%

At first, we were slightly disappointed with the amount of noise in the PEWA which made it difficult to be sure which points were and were not anomalies. But when we tested it on more data, we were pleased to note that it was capable of picking up some of the more blatant anomalies. For example, in figure 33 it picks up the two spikes in early November 2015.

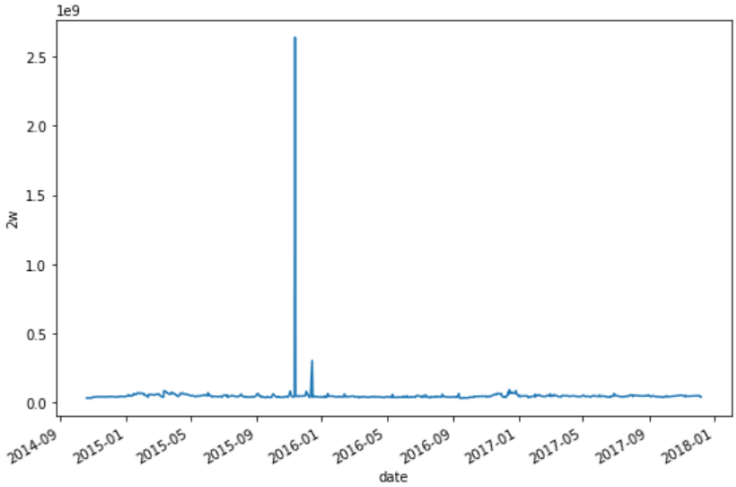


Figure 33: PEWA G−FXD 2w

From here, we decided to try to understand the link between the anomalies in the Trade Count and the anomalies in the PEWA. In figure 34, we observe the residuals of the Trade Count autoregression plotted with respect to the residuals of the 2 week PEWA autoregression for the site G-FXD. The timeframe is the same as that in figure 33. The problematic dates returned by the PEWA 2 week autoregression are in blue and the 20th of January 2017, a known anomaly in the Trade Count, is in red.

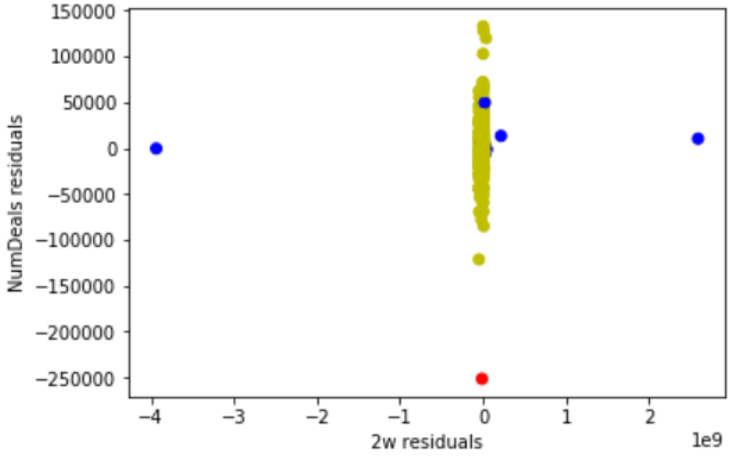


Figure 34: PEWA G−FXD 2w

It is striking that the types of anomaly here seem not to coincide. It would appear that early November there was problem with the calculation of the PE despite there not being a problem with the Trade Count and vice versa on the 20th of January. Note: While only the PEWA 2w chart is shown in figure 34, we found a similar lack of correlation in the other groups.

3.2.2 Profile Exposure Weighted Maturity

While we have detected anomalies with the previous techniques, our algorithms have returned some potential problematic dates that are neither clear cut false positives nor anomalies. For this reason, we decided to introduce a different set of data points derived from the exposure profiles. Hence, we calculated for each of the 4 groups the PEWM (Profile Exposure Weighted Maturity) using the following formula:

$$PEWM = \frac{<PE, dates>}{mean(PE)}$$

Instead of being closely linked to the average value of the exposure profile in each group as the PEWA is, the PEWM characterises the shape of the profile over the days in the group.
After experimenting with several types of auto regression, we found that we obtained the best results with the Type 1 autoregression from section 3.2.1.

Table 6: Type 1 autoregression on PEWM

Type	2w	3m	3y	long
$\widetilde{R^2}$	75,46%	83,8%	88%	94%
Mean Errors	2,1%	0,9%	1,0%	1,3%

We also note that the auto regression is better at predicting the 'long' values than the shorter term ones. This is illustrated by figure 35 below.

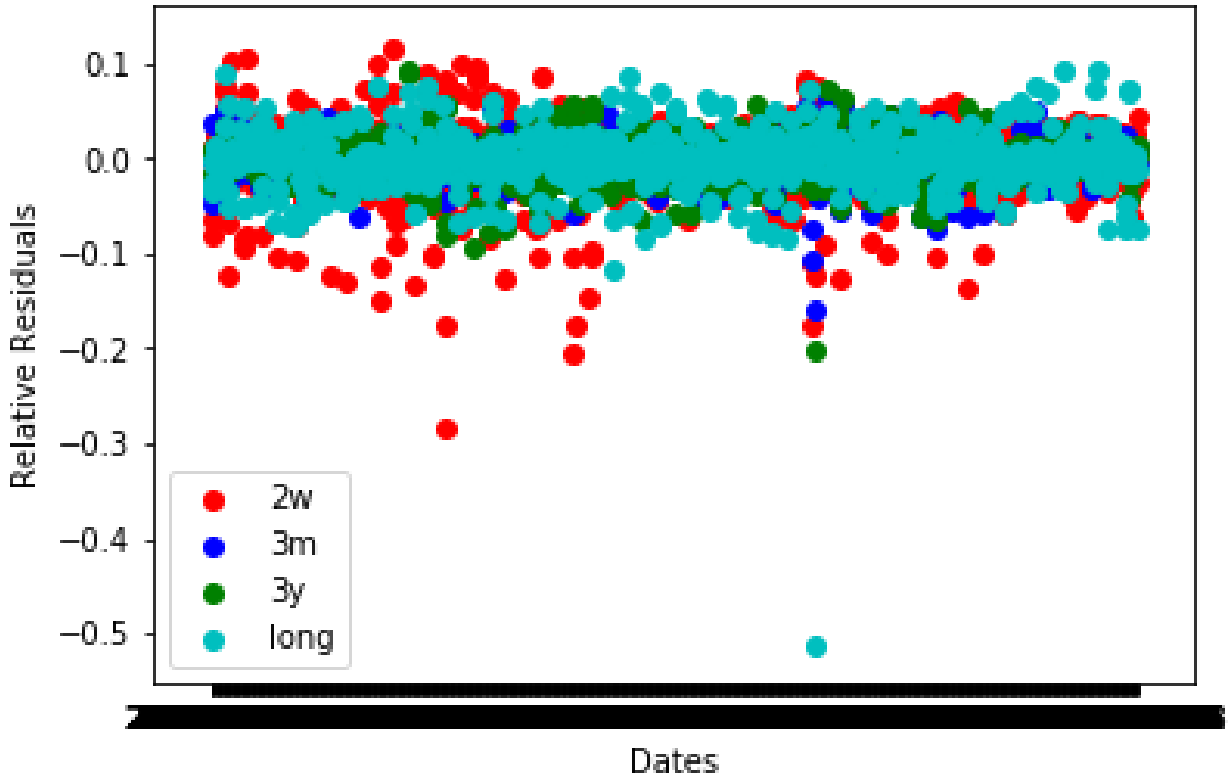


Figure 35: Relative Residuals over time for the Four Categories of G-FXD PEWM

3.2.3 A closer look at exposure profile anomalies

In the following tables (7 & 8), we have the problematic dates found for each time series of the G-IRD and G-FXD sites PEWM using an autoregression with a training period of 21 days and a single explanatory covariate - the variable on the previous day. Here, the problematic dates are defined, as has previously been the case, as those with a prediction error greater than 5 times the standard deviation on the training set. The algorithm returned more errors than are in the G-IRD table for Trade Count (NumDeals) but they were cut for reasons of legibility. With much satisfaction we note, that the blatant anomalies that we can see on the graphs in figures 36 & 37 (for example 2017-01-20 in G-FXD) have been picked up by the algorithm with a minimal number of false positives.

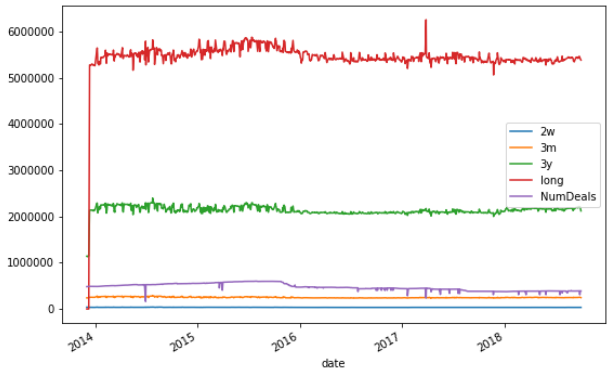


Figure 36: G−IRD over time

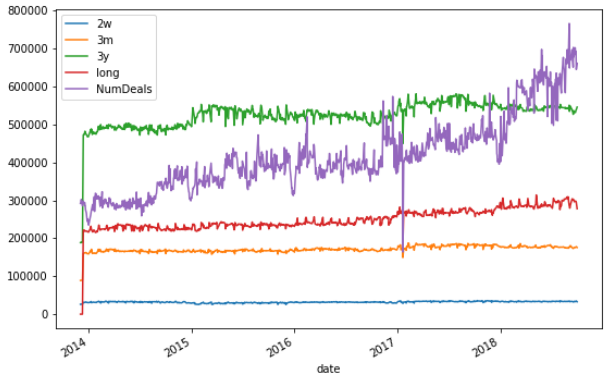


Figure 37: G−FXD over time

2w	3m	3y	long	NumDeals
2014-06-30	2017-03-07	2017-03-07	2014-05-16	2014-06-27
2014-07-01	2017-04-14	2017-10-03	2016-06-01	2014-06-30
2016-04-01	2017-06-07	2018-05-01	2016-11-01	2014-07-01
NaN	NaN	NaN	2017-11-23	2015-03-23
NaN	NaN	NaN	NaN	2015-10-07
NaN	NaN	NaN	NaN	2017-01-20

Table 7: G−IRD problematic dates

2w	3m	3y	long	NumDeals
2015-11-03	2014-07-01	2015-01-05	2016-01-11	2014-10-13
2017-05-01	2015-06-03	2015-05-01	2017-01-20	2016-11-10
2017-11-01	2015-11-03	2017-01-20	2018-05-09	2016-12-15
NaN	2017-01-20	2017-11-07	2018-07-09	2017-01-20
NaN	2017-08-02	NaN	NaN	NaN
NaN	2018-02-07	NaN	NaN	NaN

Table 8: G−FXD problematic dates

We note that the 20th of January 2017 does not appear on the list of problematic dates for the G-FXD 2 weeks timeseries. This is most likely because with a large volume of short term contracts, the relative volatility of the PEWM 2 weeks time series is high, drowning out the error.

We expected correlation between anomalies in the number of deals and those in the exposure profile data. This appears to sometimes be the case - probably indicating that some trades were forgotten in the calculation. Examples of this include 2017-01-20 for G-FXD (red in table 8) when the longer term trades appeared affected and the end of June/ beginning of July for G-IRD (blue in table 7) when the short term exposure profile values were affected. This could be due to fundamental differences in the types of contracts traded for the different sites or rather simply due to the types of contracts which were omitted. However, the correlation between Trade Count anomalies and exposure profile anomalies is not always present. Indeed, there are several instances of anomalies in the Trade Counts without corresponding anomalies in the PE data and vice-versa.

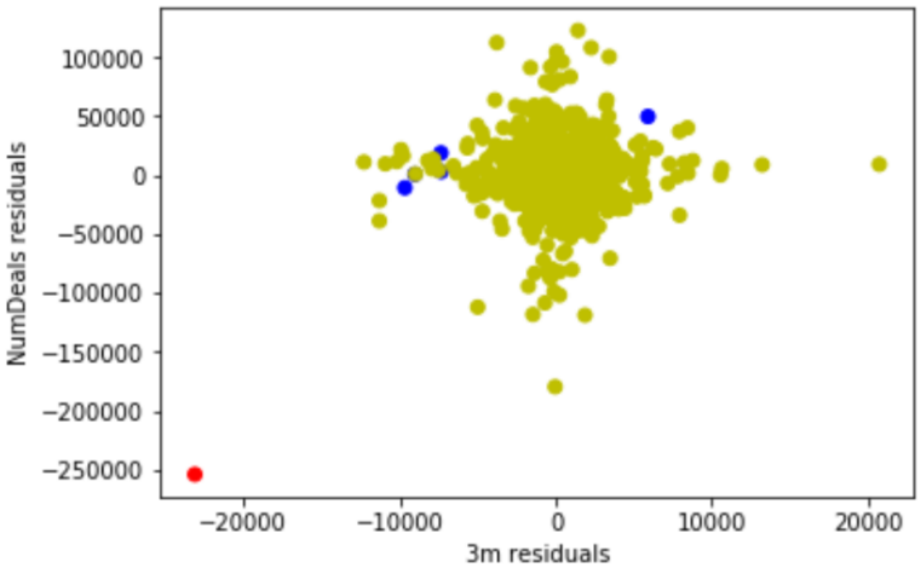


Figure 38: Graph of 3 month PEWM residuals of G−FXD site

To get a better understanding of how anomalies in the PEWM are linked to those in the number of trades, we plotted the relationship between the two. On the graph of figure 38, each point represents a date, with the 20^{th} of January 2017 in red. Similarly to the chart in figure 34, there seems to be little correlation between the two residues. However there is one point, 2017-01-20, where the number of trades and the PEWM seem correlated. The other anomalies detected in the 3 month data appear in blue and don't seem to to be linked to variation in the number of trades, with the exception perhaps of the one to the upper right. The others may be false positives or have other explanations; with the data available it is difficult to tell. A look at equivalent graphs for other sites gives similar results: some of the anomalies appear to be related to the trade count but not all of them.

4 Statistical study of the anomalies

4.1 Reasons for the study

Once we had detected the anomalies due to missing trade counts or wrong PE calculation, we focused on the statistical regularities that could be found in the anomalies. The three major metrics examined in this section are the sizes of the anomalies due to wrong PE calculation, the total number of anomalies due to respectively missing trade counts and wrong PE calculations, and the time intervals which separate two anomalies. The core idea of this section is to use a simple technique of density estimation. We draw from the histogram methods detailed in [4] and the estimation methods presented in [5].

The main reason we did this was to estimate the total PE miscalculation. This quantity is important for the bank and having a probabilistic model for its evolution could help to anticipate the potential loss of money due to risk incorrect assessment. In order to calculate this quantity, we needed to make a probabilistic model for both the size of an anomaly and the number of anomalies per site. We define the size of an anomaly as the residual between the predictive and the effective value of the PE normalized by the maximal value of these residuals. This question is discussed in sections 4.2 and 4.3

We then examine the time interval between two problematic dates. The estimation of its probability law is important as it could help us, and the bank, anticipate when the anomalies will occur. This issue will be discussed in section 4.4.

4.2 The distribution of sizes

4.2.1 Empirical observations and hypothesis

In order to have an idea of the trends and the regularities in the anomalies, we observed the frequencies of the residuals between the predictive and the effective value of the PE normalized by their maximum value. As we mentioned in subsection 4.1, we will call this residual the size of the anomaly. The protocol is very simple and consists of calculating the size of each anomaly for a given site, counting their occurrences and making a histogram. The histogram obtained for the site G-FXD is shown below (Figure 39).

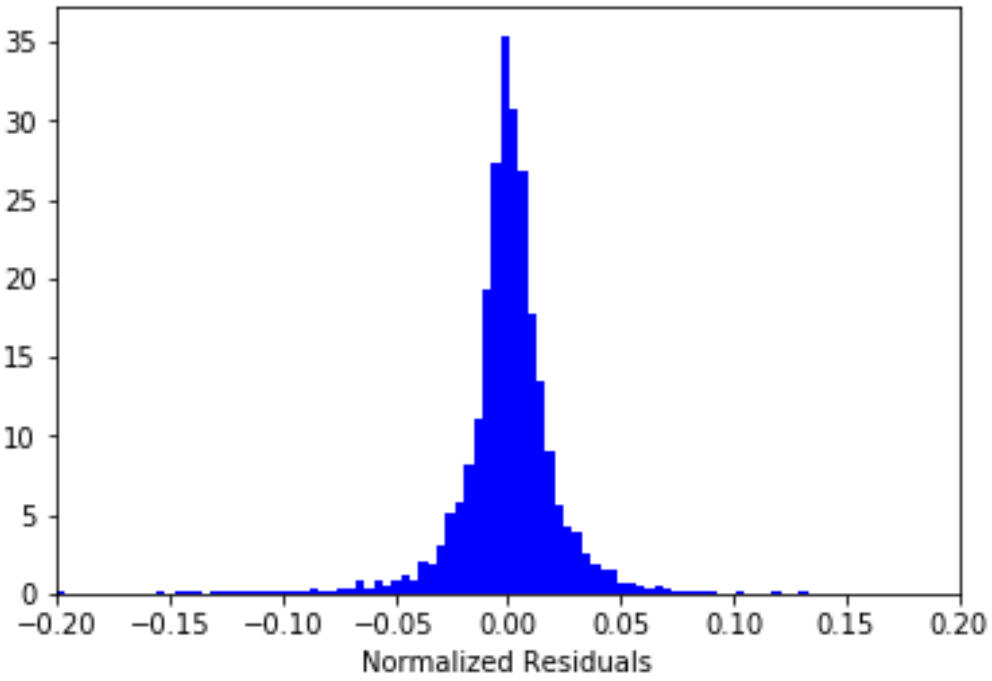


Figure 39: Histogram of the sizes of the anomalies for the site G−FXD

Based off observation of Figure 39, we can reasonably assume that the size of an anomaly follows a Gaussian distribution. This hypothesis is of course a simplification but it is experimentally realistic and consistent with the hypothesis of linear regression. More importantly, we assume that the respective sizes of the anomalies are independent and identically distributed.

4.2.2 Density estimation

To formalize the problem, let $Y_{i,Site}$ be the size of the i^{th} anomaly for the site $Site$ and n_{Site} be the number of anomalies due to wrong PE calculation for the same site. According to the hypothesis made in section 4.2.1, $(Y_{i,Site})_{1\leq i\leq n_{Site}}$ is a family of independent and identically distributed random variables which follow a Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$. In this subsection, we will focus on estimating the parameters μ and σ of this distribution. For the sake of clarity, we will henceforth consider the observations made on the site G-FXD. We will call n the number of anomalies for this site and $Y_1,...,Y_n$ their respective sizes.

A simple technique of estimation would be to calculate the value of (μ,σ) which maximizes the likelihood function. In the simple statistical model that we consider the likelihood function will be:

$$L((\mu,\sigma),Y_1,...,Y_n)=\Big(\frac{1}{\sqrt{2\pi}\sigma}\Big)^n\exp\Big(-\sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{2\sigma^2}\Big)$$

It is easier to solve the maximisation problem on the logarithmic likelihood function:

$$l((\mu,\sigma),Y_1,...,Y_n)=-\frac{n}{2}\ln\sigma^2-\sum_{i=1}^{n}\frac{(Y_i-\mu)^2}{2\sigma^2}+C$$

Where C is a constant independent of μ and σ. Finally, the maximum likelihood estimators are: $\hat{\mu}_n=\bar{Y_n}$ and $\hat{\sigma}_n^2=\bar{Y_n^2}-\bar{Y_n}^2$

4.2.3 Results

The estimation detailed in section 4.2.2 gives $\hat{\mu}_n = -1 \cdot 10^{-4}$ and $\hat{\sigma}_n = 2,3 \cdot 10^{-2}$. The following chart illustrates the quality of the estimation.

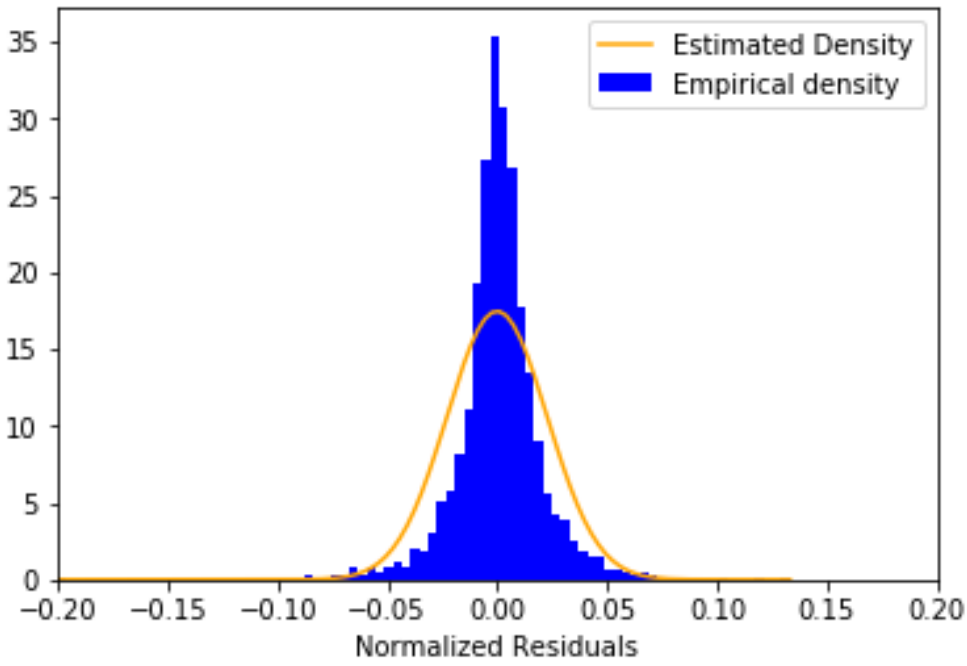


Figure 40: Estimated density of the size of PE anomalies for the site G-FXD

Despite the histogram and the estimated density not matching perfectly, we can reasonably assume this model is correct based on the shape of the histogram. However, on closer inspection we noticed that we could improve the estimation by assuming that the size of the residuals follows a Laplace distribution.

4.2.4 Improvements

The Laplace distribution is a distribution which has for density the function defined on R $f : x \mapsto \frac{1}{2\theta} \cdot \exp\left(-\frac{|x-\mu|}{\theta}\right)$. We actually guessed this distribution when we noticed that the histogram was an even function close to the graph of an exponential density on R^+.

The maximum likelihood estimation for the parameters μ and θ is similar to what is done in 4.2.2. Using the approach presented in [6], we find that $\hat{\mu}_n$ is the empirical median of the observations and $\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \mid Y_i - \hat{\mu}_n \mid$ where n is the number of observations. The computation of this estimation gives $\hat{\mu}_n = 5 \cdot 10^{-4}$ and $\hat{\theta}_n = 2,28 \cdot 10^{-2}$. This estimation is much better, as illustrated by figure 41.

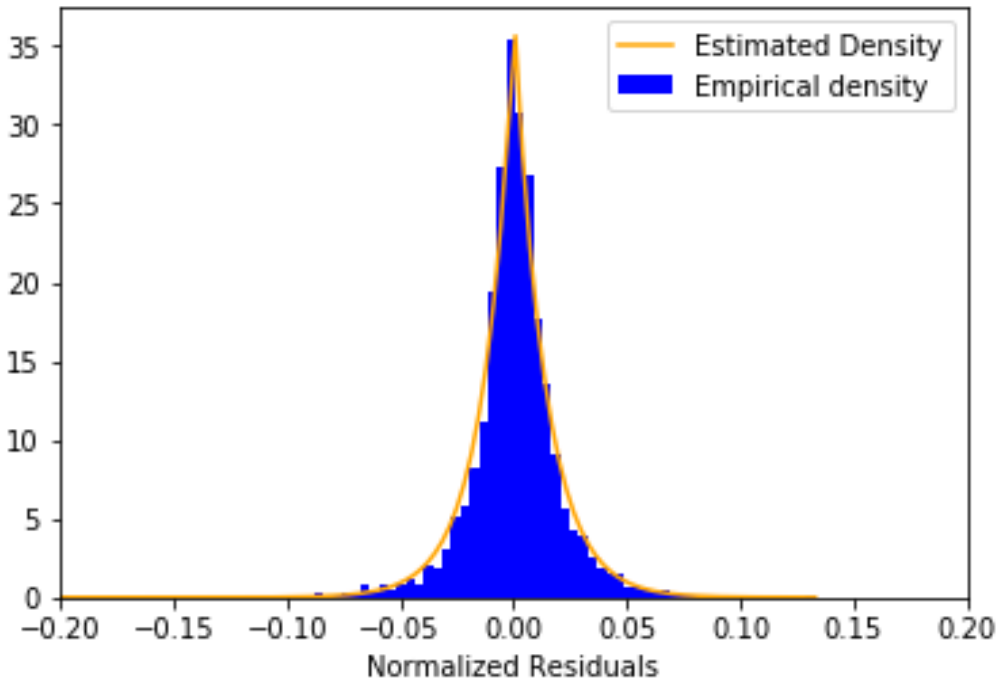


Figure 41: Improved estimation of the density of the size of PE anomalies for the site G-FXD

Hence, we choose to assume in our model that $Y_{i,Site} \sim \text{Laplace}(\mu_{Site}, \theta_{Site})$.

4.3 Number of anomalies distribution

Estimation the distribution of the number of anomalies due to wrong PE calculation per site at a given temporal horizon is could be useful as it would enable the bank to make an estimation of its total risk.

4.3.1 Estimation of the distribution

We will call N_T the random variable which gives the number of anomalies due to wrong PE calculation per site between the dates $t = t_0$ and $t = T$, where t_0 is the first date of the data set. In order to have an idea of the distribution of N_T, we followed the same protocol as for the distribution of $Y_1, ..., Y_n$. We calculated the number of anomalies due to wrong PE calculations for each site and then plotted the histogram presented in Figure 42.

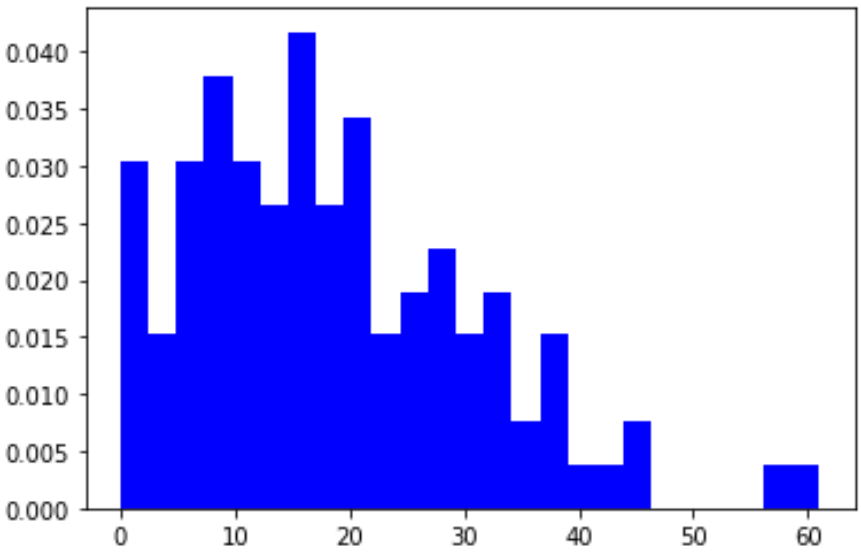


Figure 42: Frequency of the number of anomalies due to wrong PE calculation

This histogram looks like the graph of the probability mass function of a Poisson distribution. If we made this hypothesis, we would assume that there exists $\lambda > 0$ such that $P(N_T = k) = \frac{\lambda^k}{k!} \exp -\lambda$ for all $k \geq 0$. The maximum of likelihood estimator of the parameter λ is very simple to find. Following the same process as in section 4.2 we find that $\hat{\lambda} = \bar{N_n}$ that is to say the empirical mean of the number of anomalies per site. Thus, we establish that $N_T \sim P(6, 394)$.

4.3.2 Total miscalculation

Now that we have a probabilistic model for the size of an anomaly for a given site and the total number of anomalies due to wrong PE calculation in the data set, we can calculate the total algebraic miscalculation X_T at a given temporal horizon. With the same notations as in the previous sections, the expression of this quantity is:

$$X_T = \sum_{Site \in sites} \sum_{i=1}^{N_T} Y_{i,Site}$$

Calculating the mean of the random variable X_T gives an estimation of the consequences of PE miscalculations.

4.4 Problematic dates time interval

The purpose of this subsection is to find statistical regularities for the time intervals between two problematic dates. These quantities are important indicators. Indeed, if we achieve to make a coherent probabilistic model, the bank will be able to predict when the next anomaly is likely to happen knowing that there is an anomaly at a date t_0. This should help the search for outliers.

4.4.1 Empirical observations and assumptions

To formalize the problem, we will call $T_{i,Site,TC}$ the i^{th} problematic date taken in the chronological order due to missing trade counts for the site $Site$. In the same vein, we will consider the random variables $T_{i,Site,2w}$, $T_{i,Site,3m}$, $T_{i,Site,3y}$, $T_{i,Site,long}$.

In order to simplify the model, we will assume that all these variables do not depend on the site. Despite the fact this assumption is most likely unrealistic, we thought it was preferable to make simple assumption before delving into too complicated models.

More importantly, in what follows, we will suppose that for a given type of anomaly $type \in \{\text{TC,2w,3m,3y,long}\}$, and a given site $Site$ the variables $(T_{i,Site,type})_{0\leq i\leq n_{Site}}$ are stopping times, that is to say that they only depend on the past. For the sake of clarity, we will omit to mention the type of the anomaly in the index. We invite the reader must bear in mind that the anomalies from different categories are kept seperate.

The quantity of interest will be $T_{i+1,Site} - T_{i,Site}$ for $0 \leq i \leq n_{Site} - 1$. It is relevant to suppose that these random variables are independent if we consider $T_{i,Site}$ as a stopping time.

Given these hypothesis, we can measure the time intervals (in days) between consecutive problematic dates for all sites and all the problematic dates of these sites. The histograms of these quantities for the anomalies due to missing trade counts is given in 43.

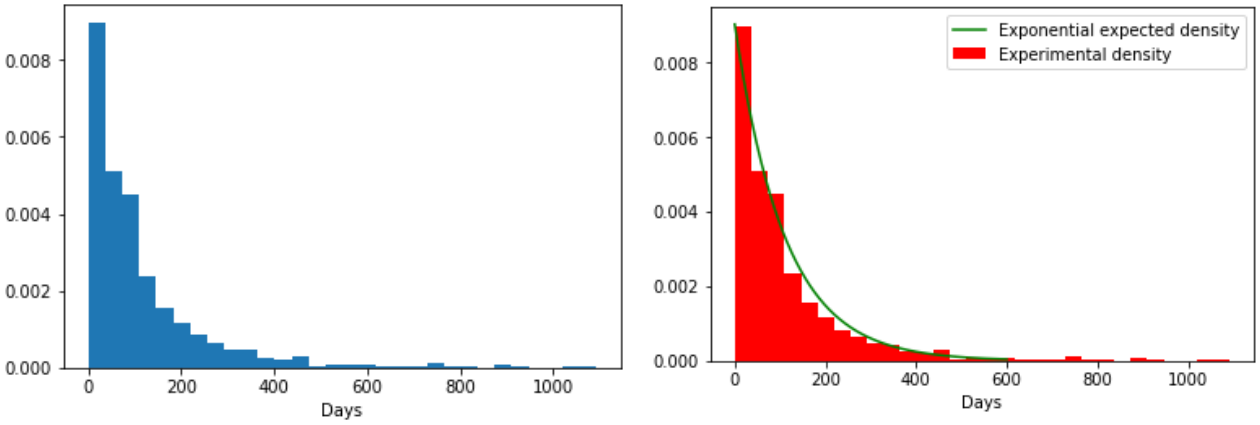


Figure 43: Frequencies of the intervals of time between two consecutive problematic dates due to missing trade counts

These histograms are satisfying as they show that it is reasonable to suppose that $T_{i+1} - T_i$ follows an exponential law. This result is consistent with the intuitive idea that these variables are memoryless.

4.4.2 Density estimation

The maximum likelihood estimator for the parameter λ of this law is $\hat{\lambda}_n = \frac{1}{\bar{S_n}}$ where S_n is the vector which contains all the lengths of the intervals of time between two consecutive problematic dates for a given type of anomaly and n is the total number of observations contained in S_n.

Let us check the quality of this estimation. For each type of anomaly, the number of observations n contained in S_n is quite big and largely superior to 30. Thus, we can apply the delta method to $\bar{S_n}$ with the function $g : x \mapsto \frac{1}{x}$ which is differentiable on R_+^* . As our random variables are almost certainly finite and independent, the central limit theorem ensures that:

$$\sqrt{n}(\frac{1}{\hat{\lambda}_n} - \frac{1}{\lambda}) \sim \mathcal{N}(0,\, \sigma^2)$$

where σ is the standard deviation of $T_{i+1} - T_i$. Then, the delta method gives that:

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \sim \mathcal{N}(0,\, \lambda^4\sigma^2)$$

As σ and λ are unknown, we will use their empirical estimators $\hat{\sigma}_n$ and $\hat{\lambda}_n$. Indeed, as these two estimators almost certainly converge toward respectively σ and λ, Slutsky theorem ensures that:

$$\frac{\sqrt{n}}{\hat{\lambda}_n^2 \hat{\sigma}_n}(\hat{\lambda}_n - \lambda) \sim \mathcal{N}(0,1)$$

Finally, we can calculate a confidence interval for a confidence level of 95%:

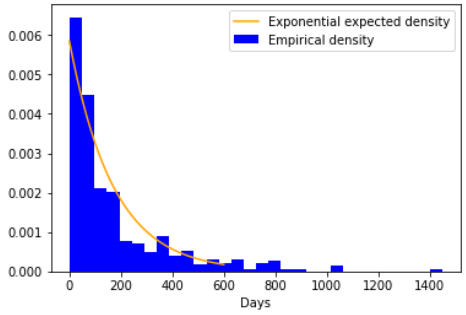
$$I_n = [\hat{\lambda}_n(1 - 1,96\frac{\hat{\lambda}_n \cdot \hat{\sigma}_n}{\sqrt{n}}), \hat{\lambda}_n(1 + 1,96\frac{\hat{\lambda}_n \cdot \hat{\sigma}_n}{\sqrt{n}})]$$

4.4.3 Results

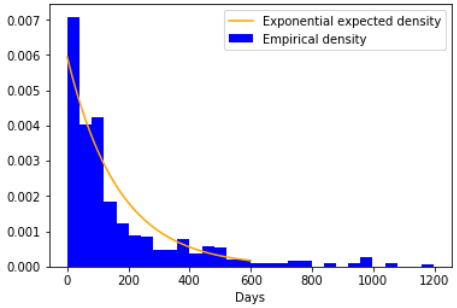
The final results of these estimations are presented in the table below.

	Trade Counts	2w	3m	3y	long
n	919	462	474	485	268
$\hat{\lambda}_n$	$0,009$	$0,006$	$0,006$	$0,006$	$0,006$
$\hat{\sigma}_n$	$140,04$	$210,88$	$212,43$	$197,31$	$205,48$
I_n	$[7,90; 10,1] \cdot 10^{-3}$	$[5,31; 6,69] \cdot 10^{-3}$	$[5,31; 6,68] \cdot 10^{-3}$	$[5,37; 6,63] \cdot 10^{-3}$	$[5,11; 6,89] \cdot 10^{-3}$

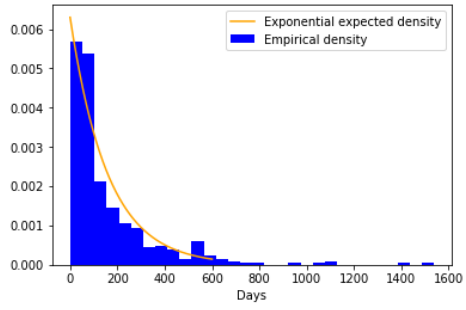
The following charts show the consistency of the model and the assumption made in 4.4.1.



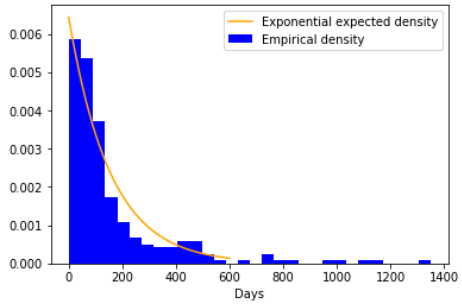
(a) Type 2w



(b) Type 3m



(c) Type 3y



(d) Type long

Figure 44: Densities of the time interval between two consecutive problematic dates for the four types of anomalies due to wrong PE calculation

Thus, the statistical analysis of the data set enables us to make a very simple and efficient model to predict the most probable times at which a future mistake will occur.

The previous result has two important consequences. The first one is that we can now state that the variables T_i follow Gamma distributions because they are sums of independent and identically distributed exponential variables. The second one is that the number of anomalies N_T which was the main subject of 4.3 follows an homogeneous Poisson Process for which we have calculated the intensity which confirms the empirical model of the section 4.3.

Conclusion

Financial institutions manage numerous portfolios associated with a risk that must be managed continuously, and the large amounts of data that has to be processed make this a considerable effort. We tried to develop, in partnership with BNP Paribas UK, a supervised system that autonomously detects anomalies in the counter party credit risk exposure profiles. In fact, in this paper, we developed algorithms capable of detecting many of the anomalies that may appear in real time, whether they affect the trade count, the average values of the exposure profile or the shape of the exposure profile. However, we struggle to know exactly how accurate our algorithms are, whether they flag false positives or miss anomalies. This could probably be improved in time with usage if, every time the programme detects an anomaly, a person reviews the data and determines whether the algorithm was right or not.

Regarding the team managing part of the project, we decided to distribute the work to make each member of the group responsible of an important part of the project, taking care to allow each member to contribute to the crucial statistical part of the project. The biggest technical difficulties stemmed from our lack of experience with the pandas library; we thought we had enough practice during September and October, but it was not the case. We lost valuable time in early January because of this lack of preparation. Keeping the team motivated and weeding out the bad parts was also very challenging. Fortunately, we managed to get familiarised with pandas and other machine learning libraries quickly.

Anomaly Detection is still a fledgling field of research. However, it is becoming increasingly important in today's risk management environment. We note that our conclusion from this paper is that there is no clear singular algorithm that can be called the "answer" to anomaly detection in finance. The key is in clever combinations of existing methods and pipelining them intelligently. The choices depend very much on the nature of the system and the need of the architect to accomplish specific goals. Interesting topics for further research include an extension of this study more sophisticated anomaly detection techniques such as Support Vector Machine-Based Anomaly Detection, Autoencoder, Local Outlier Factor.

Acknowledgements

We would like to thank Martijn Vandervoort who took a keen interest in this project and guided us, providing us with all the information necessary to achieve our objectives.

Our sincere thanks also go to Marcos Carreira who was always available to help us when we needed it. He allowed us enough autonomy to think about problems and take the initiative, but also steered us in the right direction whenever he thought we needed it. His teaching style and enthusiasm for statistics made a strong impression on us. We are extremely grateful and indebted to him for sharing his expertise and his sincere and valuable guidance.

We owe a debt of gratitude to our coordinator, Ms Flore Nabet, whose contribution, with stimulating suggestions and encouragement, helped us to coordinate our project and indeed write this report.

We would also like to thank everyone who, directly or indirectly, lent their hand in this venture.

References

[1] Basel Committee on Banking Supervision. The standardised approach for measuring counterparty credit risk exposures. *Bank for International Settlements.*, 2014.

[2] Luca Di Persio Luca Mammi et al. Bonollo, Michele. Estimating the counterparty risk exposure by using the brownian motion local time. *International Journal of Applied Mathematics and Computer Science.*, pages pp. 435–447, 2017.

[3] Wessel van Wieringen. *Ridge regression.* VU UniversityAmsterdam, 2018.

[4] Bernard W Silverman. Density estimation for statistics and data analysis. *Monograph on Statistics and Applied Probability.*, 1986.

[5] Eric Moulines Jaouad Mourtada Gersende Fort, Matthieu Lerasle. *Inférence statistique.* Applied Mathematics Department of the Ecole Polytechnique, 2018.

[6] Robert M. Norton. The double exponential distribution: Using calculus to find a maximum likelihood estimator. *The American Statistician*, 1984.