UNIVERSITY OF GOTHENBURG
SCHOOL OF BUSINESS, ECONOMICS AND LAW

Master Thesis

Boom, Bust and Betrayal

Business cycles and securities fraud in the US between 1996-2019

Graduate School

Viktoria Carshaw and Rinske de Vries

Supervisor: Heather Congdon Fors

Table of Contents

Special thanks go to Heather Congdon Fors and Marcin Zamojski for their supervision and help on this thesis.

List of Abbreviations

ADF	augmented Dickey-Fuller test
ASTAR	adaptive spline threshold autoregression
AIC	Akaike information criterion
CEO	Chef Executive Officer
CBOE	Chicago Board Options Exchange
CI	Confidence Indexes
DM	Diebold-Mariano test
EDof	effective degrees of freedom
FRED	Federal Reserve Bank of St. Louis
GAM	Generalised Additive Models
GCV	Generalized Cross-Validation
GDP	Gross Domestic Product
GLM	Generalized Linear Regression Models
GRSQ	generalized R^2
GSV	generalization error score
IPO	initial public offering
MARS	Multivariate Adaptive Regression Splines
MAD	mean absolute forecast error deviation
MCDR	modified-cash-deposits-ratio
MCI	monetary conditions index
MSE	mean standard error
OLS	Ordinary Least Squares
OOSR^2	Out of sample R^2
OVB	omitted variables bias
PCE	personal consumption expenditures
PSLR	Private Securities Litigation Reform Act
R^2 / RSQ	Coefficient of Determination
RMSE	Root Mean Square Error
RSS	Residual Sum of Squares
SEC	Securities and Exchange Commission
SCAC	Securities Class Action Clearinghouse
TSS	total sum of squares
VIF	variance inflation factor
VIX	Volatility Index

List of Figures

List of Tables

Abstract

We study the relationship between macroeconomic factors and the number of settled securities fraud cases through the proxy of SEC class action lawsuits. We perform an empirical study of all SEC class action convictions between 1996-2019 and their relationship with business cycles as proxied by a housing price index and GDP. Furthermore, our dynamic model looks at the effect of market volatility and unemployment as well as the persistence in the time series of lawsuits. Alongside this, we study the extreme circumstances of the dot com bubble of 1999. We use non-parametric models such as GAM and MARS to better account for business cycle fluctuations. We find that the most significant effects seem to be coming from the dot com bubble, which was the only extreme outlier in our dataset, and the one month lag of the number of fraud cases. There is also an indication that both market volatility and business cycles may be correlated with the prevalence of securities fraud in our MARS model. However, the robustness checks did not find these significant for GAM.

1 Introduction

From the 300 year old fraud scandal of the British South Sea Bubble of 1720 to the contemporary financial crisis of the housing bubble, securities fraud undermines confidence and trust in the financial system (Stout, 2002). It incurs a large cost on securities markets, companies, investors and the wider economy (Kedia and Philippon (2005), Du and Wei (2004), Jain et al. (2008), Utset (2013)). The problem is also pervasive, 5.5 % of U.S. exchange-listed companies in 2019 faced fraud filings (Bulan and Simmons, 2020).

The purpose of this paper is to investigate, based on the existing literature, the relationship between securities fraud and macroeconomic variables. Most of the empirical papers on the subject focus on microeconomic factors, so that the ties with macroeconomic factors remain mainly theoreticised. Our paper attempts to bridge that gap through an empirical study of the macroeconomic variables and the prevalence of securities fraud litigations in the USA between 1996-2019. We will look at non-parametric regressions which offer a better fit than traditional models to determine the link between macroeconomic variables such as the business cycle, unemployment and market volatility with the number of fraud cases in the USA in the modern day.

The paper investigates the relationship between various macroeconomic variables and their relationship to securities fraud class actions under the SEC in the US between 1996-2019. Some frauds will go undetected or unproven, therefore we use securities fraud class action filings in the US as a proxy for the number of securities frauds. The data which was available to us comes from the Securities Class Action Clearinghouse (SCAC) database. The claims covered by the SCAC are listed in the table in the appendix (7.1).

Securities class actions are cases brought before the court on behalf of a number of investors who would otherwise find it too costly to bring this case on their own. When a case is brought before the court this is known as the filing date, when the fraud is alleged to have occurred is known as the class period. These dates differ greatly, in our data the difference is on average 601 days or just over one and a half years. We study the class action period to determine when the fraud has reportedly been committed to gain a more accurate timing of the fraud than the one given by the filing date. We use only those cases which were successfully convicted, known as settled cases, to try to remove spurious allegations from our data.

We find that fraud is correlated with the business cycle through the growth of GDP. Also, the dot com crisis affects the number of settled fraud allegation. The robustness analysis adds to the internal validity of our findings.

Du and Wei (2004) suggest an increase in insider trading increases market volatility. Baker (2016) suggests an increase in market volatility may obscure fraud, especially during financial crises, resulting in too many convictions. We only find a connection between market volatility and the number of fraud cases in our MARS model, not our GAM model. The results indicate firm specific effects may become a market wide phenomenon which affect the number of securities fraud cases, however the link is more tenuous as it is only present in one of our models.

The contributions of this thesis are twofold. Firstly, the study is the first scholarly attempt to connect macroeconomic factors to the prevalence of securities fraud using class action lawsuits, which capture a far wider range of securities fraud types than any prior study to our knowledge. Moreover, we look at a wide range of macroeconomic factors that have not been considered together before. This is the only empirical study of market volatility and its connection with all types of securities fraud. The article empirically proves the impact of business cycles on the criminal activity in the financial sector of the US economy and gives credit to some theoretical frameworks on fraud, alongside expanding prior evidence of empirical studies on only certain fraud to a broad range of securities fraud violations. This includes previously untested securities fraud types associated with macroeconomic factors, such as price manipulation and fraudulent reporting, in order to determine if there is a market wide rise in fraud as a result of changing macroeconomic factors, which has not been tested before.

Secondly, our thesis broadens the range of econometric tools applied in macroeconomics in a systematic way to non- and semi-parametric approaches. Generalized additive model (GAMs) as well as multivariate adaptive regression splines (MARS) are applied to a set of issues relevant to academics in several sectors as well as parties involved in the department of justice. The MARS and GAM models better explain the change in the number of fraud cases in the USA conditioned on different macroeconomic specifications by more effectively capturing the shape of the business cycle, and the non-linearity of our variables, than traditional models.

The paper is structured in the following way: we begin with an account of the theoretical background on the subject of financial fraud and macroeconomic factors (section 2). From there on, the dataset is presented and linked to the literature (section 3). A discussion of the empirical method and it's connection to the assumptions of the models follows in section 4, while preliminary findings can be found in section 5, alongside the robustness analysis. Further investigation of the results is located in section 6 together with the limitations of the paper. Section 7 concludes this thesis and offers some ideas for future work.

2 Literature Review and Theory

In this thesis, we focus on securities fraud and more specifically on the macroeconomic effects which affect the number of securities fraud cases. Studies on securities fraud have often given a microeconomic explanation of the reason behind fraud, these explanations include equity incentives of managers and supervisors (Hass et al., 2016; Goldman and

Slezak, 2006) and how sensitive the CEO's option portfolio is to the price of the firm's stock (Burns and Kedia, 2006). Yu (2013) and van Driel (2019) give a good overview of the securities fraud literature. This thesis however will focus on only macroeconomic effects which best predict securities fraud.

As Buell (2011) notes, how to define securities fraud and fraud in general is difficult, the definition cannot be too narrow for fear of not encompassing all fraudulent behaviour. Here, we simply assume all class action lawsuits constitute possible fraud and those cases which are settled are definitively fraud. Our paper focuses on all securities class action lawsuits which encompass a range of securities fraud laws and we do not differentiate between the different types of fraud. In this respect, the closest paper to ours is that of Dyck et al. (2010) who investigate the prevalence of whistle-blowing between 1996 and 2004.

Convicted class action lawsuits occur on two grounds, one that fraud has occurred and two that the fraud has been discovered. Regarding the first, the accounting literature provides a model of fraud known as the fraud triangle. The fraud triangle is made up of three points which help explain why fraud occurs: opportunity, pressure and rationalisation (Schuchter and Levi, 2016; Albrecht et al., 2008).

Opportunity can be influenced by trust, for example investment in the stock market is based on the trust of investors. When prices rise for a continued period, as in a boom, Stout (2002) argues that trusting investors do not consider fundamental values, but instead believe this behaviour will continue regardless of fundamentals. If too many investors are trusting, then they are more easy to fool and fraud is more easy and profitable to commit (Mayer, 2002; Stout, 2002).

Pressure can occur as a result of market wide practices or economic environments. Akerlof and Shiller (2015) discuss manipulation as market practice in their book *Phishing for Phools*, two relevant examples they give are reputation mining and phishing equilibria. Reputation mining occurs when firms undermine the reputation they have built in order to benefit themselves, when a person or firm commits securities fraud they effectively mine their good reputation to profit themselves. This is especially relevant for our study as our data covers the 2008 financial crisis and the dot com bubble when banks, ratings agencies and auditors mined their reputation in exchange for profit (Akerlof and Shiller, 2015; Wheale and Amin, 2003; Mayer, 2002; Ljungqvist and Wilhelm, 2003). Unfortunately, although they offer many examples, they don't offer any theoretical models or empirical studies for their arguments. However, Basu (2018) builds a game theoretic model of fraud based on the phishing equilibria mentioned in Akerlof and Shiller (2015). Phishing equilibria occur when individuals who adopt a corrupt strategy have non-negative profits and payoffs. If individuals who adopt this strategy gain more than those who do not, in a competitive market, this will push out those who do not pursue a corrupt strategy. The rate of this pushing out effect increases with how competitive the market is. The result is an equilibria in which all firms become corrupt. Indeed Mishina et al. (2010) find that relative, not absolute performance, is a greater incentive for firms to commit fraud, suggesting increased competition between firms is a key factor in manipulation and fraud incentives. Thornton (2016) however, finds Akerlof and Shiller (2015)'s examples to be cherry picked and argues that the free market as described by Akerlof and Shiller (2015) does not exist in reality, thanks to numerous government regulations.

Utset (2013) argues that securities fraud may be a form of corporate signalling which at its worst can exacerbate market bubbles and is very costly for both investors and the economy. Similarly, Kedia and Philippon (2005) look to the relationship between firm employment and investment and fraud, those firms which commit fraud over-invest and over-employ to conceal their fraud and imitate non-fraudulent firms and once the fraud is revealed they will drop the excess jobs and investment. Therefore, fraud may exacerbate the recession after a boom once it is revealed.

The professional recommendations and the academic literature disagree on when frauds are committed, many professional auditors and accountancy firms give warnings for fraud during contractionary periods (Deloitte, 2009; Cooper, 2009; PWC, 2009). However, many academic papers find securities fraud is more likely to occur during a boom rather than a bust. Cooper (2009), chief audit executive for MCI, in an interview in 2009 argued fraud occurs after a bubble to hide the excessive risk taking during a bubble. If companies can see the end of the bubble, this may explain the observations in the data that fraud occurs at the end of the bubble. However, if firms are surprised by the end of a bubble, we would expect fraud to correlate with the beginning of a contraction. She draws parallels between the 2008 financial crisis and the dot com bubble arguing that both stemmed from the same lack of regulatory oversight and firms taking on too much risk. She also offers an explanation of the rationalisation of those who commit fraud as justifying it to save their departments and colleagues whom they respect (Cooper, 2009).

Discovery of fraud is naturally related to regulation and oversight. Dias et al. (2005) find a clear relationship between regulatory oversight and improvements in the stock market price for Enron alongside a reduction in market volatility. Class action lawsuits also seem to have a greater effect on investor beliefs and stock market prices than SEC-only investigations (Choi and Pritchard, 2016).

Regulation and oversight may fluctuate across the business cycle. Povel et al. (2007) build a model which considers investors beliefs on the number of fraudulent firms in the economy and their private monitoring costs. If more firms commit fraud to attract investment, investors will monitor more and fraud will be low. If monitoring costs are high, investors will monitor less and fraud will increase. Their model suggests, at the very end of a boom, investor beliefs of the economy will be overoptimistic and investors will not monitor as strongly. Thus, the incentive to commit fraud decreases, therefore fraud will peak just before the end of a boom. Hertzberg (2003) argues that fraud and the business cycle are linked and fraud is most prevalent during a boom. Similarly, Blanqué (2003) argues fraud increases during a boom, can be exacerbated when firms scramble to cover their fraud, and finally contributes to the crash when the frauds are revealed. In line with Hertzberg (2003) and Blanqué (2003), Povel et al. (2007) predict that fraud is most likely during a boom, however unlike Hertzberg (2003) they predict fraud will decrease at the very end of a boom. Blanqué (2003)'s suggestion is dependent on if firms can recognise the end of the boom before it takes place or not, if they can (like Povel et al. (2007)) Blanqué (2003)'s theory predicts fraud decreases before the end of a boom, if not the fraud will remain high at the end of a boom. Their models are backed by the findings of Wang et al. (2010), who study IPO fraud, and Fernandes and Guedes (2010) who study accounting fraud. Both empirical studies find that fraud increases when investor beliefs are high and decreases when investors' beliefs are low, this is because of the cost of investor monitoring (Wang et al., 2010) and competition arising from investor's expectations (Fernandes and Guedes, 2010). Wang et al. (2010) find frauds increase during a boom, but decrease at the end of a boom which is most in line with the model of Povel et al. (2007). Lohse and Thomann (2015) study the effect of the business cycle on regulation and find that the SEC's funding increases when the stock market is doing badly and vice versa, which is a strong indication that even the SEC follows the same monitoring cycle as investors.

Many papers find a link between fraud and stock price - or stock market - volatility, for example Jain et al. (2008) find that an increase in the number of frauds significantly decreased liquidity in the market and raised market volatility. They also study the Sarbanes-Oxley Act of 2002 and find that it has had a long term positive effect on market liquidity through a reduction in fraud. Morris et al. (2019) find that market regulation, specifically SEC regulations and Sarbanes Oxley provisions, have a positive effect on firm's stock price and decrease market volatility. This link is again noted by Schenk (2017) who find a link between volatile markets and rogue trading at Lloyds Bank International. Du and Wei (2004) study the relationship between insider trading and market volatility in different markets, controlling for other factors which may differ between these markets they find that countries where insider trading is more prevalent also have higher average stock market volatility.

Another angle to the relationship between market volatility and class action lawsuits comes from Baker (2016) who suggests single-firm event studies perform less well in the presence of excessive market volatility and may over estimate the number and size of frauds. Event studies (used in class action lawsuit cases) find it difficult to differentiate between market and stock price volatility during periods of high market volatility. This may make false convictions of fraud more common during periods of high market volatility and, if this is the case, we should expect the number of successful cases to increase with greater market volatility.

Fox et al. (2015) show that every major downturns since the 1920s has been accompanied by a spike in idiosyncratic risk as measured by market volatility. This gives us a good indication that, certainly in respect to major downturns, we can expect business cycles and market volatility to be correlated, which may cause multicollinearity issues (see section 4.7 for a discussion on multicollinearity).

Fraud may also be linked to market confidence. Blanqué (2003) suggests that fraud is correlated with market speculation during a boom, and Cretarola et al. (2020) suggest that market sentiment and speculation are correlated. Schrand and Zechman (2012) find that financial misreporting is more likely when executives are overconfident and Scheinkman and Xiong (2003) find that overconfidence creates disagreements over the true value of assets and, if the overconfidence is persistent and pervasive enough, market bubbles can be created as a result. Investor and Business confidence can be measured by Shiller's investor and institutional confidence indexes (Yale, 2019). We attempted to analyse the Yale Shiller Confidence Indexes **CI** as a proxy for executives confidence. Unfortunately, due to missing observations in the variable, we weighted the loss of observations before 2000 against the insights we could get from including the business confidence variable in the regression. Our dataset is not large enough to allow for that kind of data loss. An attempt was also made to impute the missing values in the Shiller indexes but the results were unsatisfactory. For one, we believe that if the observations aren't missing at random, using a model to fill them in would eventually lead to us fitting imputational models to our non-parametric models, creating too much of a bias. Moreover, this is a part of a systematic problem in the data which would influence the regressions to a large extent. Due to that, we decided to leave the research on business confidence affecting the number of fraud cases in the USA to future studies. See Baneshi and Talei (2010) for an elaboration on imputation of missing values.

3 Data

See the description of the variables (table 1) and the appendix (section 7.1) for references to all the data mentioned here. We used the following data:

- The Securities Class Action Clearinghouse (SCAC) database of securities class action lawsuits,
- The National Bureau's Business Cycle Dating Committee which determines business cycles in the US,
- The CBOE Volatility Index [VIX] from the FRED (Federal Reserve Bank of St. Louis),
- Unemployment Rate [UNRATE] retrieved from the U.S. Bureau of Labor Statistics, through FRED,

- Gross Domestic Product (GDP), Normalised for the United States [USALORSGPNOSTSAM], retrieved from the Organization for Economic Co-operation and Development, Leading Indicators OECD. Can be found at FRED,
- S&P/Case-Shiller U.S. National Home Price Index (HPI) [CSUSHPISA] found on S&P Dow Jones Indices LLC, FRED,
- Trimmed Mean PCE Inflation Rate [PCETRIM12M159SFRBDAL] from the Federal Reserve Bank of Dallas, retrieved from FRED.

We found 5160 possible cases between 1996 and 2021. Of which, 2696 were dismissed, 551 ongoing and 2462 settled. There were 12 observations that lacked Class Action Start Period, which were dropped at the very start. We also removed the cases from 2020 and 2021 as we were worried that the limited number of cases which were convicted and not still ongoing in this time period may bias our results. The bias might be coming from the COVID-19 crisis and the fact it takes on average two years between filing and settlement, so cases settled dated 2020/2021 would have been registered around the start of the COVID-19 outbreak. Moreover, those settled cases which were settled immediately may be biased to a particular type and will not represent a complete picture of all settled cases. In total, our sample includes 2448 settled cases, summed on monthly basis, which equals 287 observations. All data on the right-hand side variables was seasonally adjusted by the publishing body.

A discussion on confidence indexes and their removal due to missing data can be found in the literature review (section 2).

3.1 Construction and Definition of Variables

Several frameworks were designed to capture the suspected relationships between the US securities fraud prevalence and its possible explanatory variables. The table below presents a short overview of the variables that were considered for the models and whether they were included in the final version of the paper:

Variable	Included Y/N	Note	Units
settled fraud cases	Y	main dependent variable	number of cases
one month lag of the settled fraud cases	Y	making the regression dynamic and accounting for persistent time series	number of cases
business cycle dummies	Y	strong indication from the literature	months
GDP growth rate	Y	strong indication from the literature, high correlation values	unit-free
Housing price index (HPI) growth rate	Y	strong indication from the literature, high correlation values.	unit-free
volatility index	Y	strong indication from the literature. Also, lags are relevant.	unit-free
unemployment growth rate	Y	good proxy for the business cycle, high correlation values	unit-free
Business confidence index	N	weak correlations, missing observation in data	unit-free
Equity to GDP	N	contributes with multicollinearity issues, microeconomic focus	unit-free
inflation	N	lack of resources	unit-free
globalisation	N	difficult to proximate or instrument	n/a
auditory skills of the lawyers	N	lack of data	n/a
institutional framework	N	lack of data	n/a
modified-cash-deposits-ratio (MCDR)	N	lack of sources	unit-free
money supply	N	lack of data and time limitation	n/a
interest rate	N	lack of time	unit-free
crisis dummy	Y	to account for the dot com bubble	months

Table 1: Summary of all the variables.

The panel data constructed from the sources described in 3 allows for construction of the final model variables in the following manner :

3.1.1 Dependent variables:

- *The number of successfully convicted securities fraud cases.* Those cases which are successfully convicted are unlikely to be false allegations as 'settling the current frivolous suit may prevent other litigation with greater merit' (Coffee Jr., 2019, section A, 2nd paragraph).

  Convicted fraud cases will give us a more accurate picture of the relationship between fraud and our macroeconomic factors, not just fraud allegations and our macroeconomic factors. The time frame of the data used is explained in section 3. See section 4.5 for the model.

3.1.2 Independent variables:

- *Business cycles.* Although our time period is small it contains the dot com bubble, the market boom pre-2008, the great recession of 2007-2008 and the boom post 2008. Other studies have found a connection between fraud and business cycles, for example Lawal et al. (2017) study fraud and business cycles in Nigeria and find that there is an identified range of frauds that may increase during busts. By contrast, Povel et al. (2007) find that fraud peeks at the end of a boom and this relationship increases with a reduction in fraud monitoring. Kluger and Slezak (2018) conduct in lab experiments on fraud and find that environment can determine the likelihood of fraud. There is also evidence of reverse causality, not only do business cycles cause fraud, but fraud can contribute to business cycles. The large scale mortgages fraud committed by banks could well have contributed to the the 2008 financial crisis (Griffin, 2019). If securities fraud is linked to business cycles, this knowledge can be used by investors so that they can be more suspicious of potential securities fraud during relevant business cycle periods.

For the regression, it is possible to approach the business cycle variable from two ends. For one, we could model the phases of the cycle by dummy variables such as *peak, trough, contraction, expansion*. For the other, we could be using proxy variables such as *unemployment, GDP & Housing Price Index (HPI)*. The time series of real **Gross Domestic Product** (GDP) and Real Output Growth, measured by the annual percentage change of GDP per capita adjusted for inflation would be used for that purpose (inspired by a study conducted by Valle e Azevedo et al. (2006); Mustafa and Khan (2020)).

Unemployment growth rate and crime have long been known to have a positive relationship (Freeman, 1983; Long and Witte, 1981). Carmichael and Ward (2001) found a significant and positive relationships between fraud and forgery and total crime and both adult and youth unemployment in England and Wales. In Sweden, the result was confirmed in a paper written by Edmark (2005) where a single percent increase in unemployment rate was estimated to increase fraud by 0.22 percent. An important connection between bank fraud proxy and unemployment during different phases of the US business cycle was made by Stewart (2016). Moreover, Mustafa and Khan (2020) explicitly relates a rise in unemployment with an increase in the number of accounting frauds. With those indications in focus, unemployment rate was adapted for the regressions.

Leamer (2015, p. 43) made a claim in his paper that 'housing is the single most critical part of the U.S. business cycle'. The theory was further tested and extended in a paper written by Huang et al. (2020) where housing factors were found to predict long-term variations in the macroeconomies of the OECD countries. 'At the national level, housing appears to be an important driver of cyclical fluctuations' (Ghent and Owyang, 2010, p.16), in line with the theoretical predictions of Davis and Heathcote (2005). Taking those facts into consideration, **HPI growth rate** (see appendix 7.1) was chosen as an additional proxy for business cycles. It is arguably also a measure of investor confidence in finance (Yılmaz, 2019).

The **dummy variables** modelling the business cycle were used to contain the information on the special periods in the recessions and booms. Since the *contraction and expansion* dummies contain the *trough and peak* dummies, we have chosen to only include the latter which specifically focus on our recessions and booms. In order to avoid a dummy variable trap, only one of those variables, *trough*, was included in the final regression (see section 4.7 for closer discussion on the issue).

- *Market Volatility.* Frivolous securities fraud filings have been been an issue for defendants in securities fraud cases, to prevent them the Private Securities Litigation Reform Act (PSLRA) was introduced in 1995 (Pritchard (1999)). Securities frauds may also be harder to detect in a volatile market as business cycles and market volatility are often correlated (Hamilton and Lin (1996)) and increased market volatility may lead to miss-identification of fraud or an increase in the number of frivolous securities allegations (Baker, 2016). For the purpose of the regression, CBOE Volatility Index (**VIX**) is used as a proxy for true market volatility measure since it is a real-time market index representing the market's expectations for volatility over the coming 30 days.

  Since most of the variables in our regressions are time series, controlling for lagged values might reduce the issue of autocorrelation arising from miss-specification of the model.

  We also include the monthly lag of the volatility index. From the figure in the appendix 14b, it can be seen that there is a delay in the behaviour of the fraud cases in relation to the volatility index. Since most of the variables in our regressions are time series, controlling for lagged values might reduce any autocorrelation which results from miss-specification of the model. We believe that past value of VIX affects mostly VIX itself and not our dependent variables since the correlation between the respective variables are low for the all cases and settled cases but high for volatility index itself (see heatmap 2a). This is partly reflected by covariation heatmap 2b.

3.2 Descriptive characteristics of the dataset.

The extended summary statistics included in the table below has been performed on the final dataset and is meant to give an idea of how the independent variables and the dependent variable compare. From there, it can be observed that on a monthly average, there were about 9 fraud cases settled. Moreover, the variation among the monthly number of cases was as high as their mean, which indicated fluctuations. Kurtosis value for the settled cases indicated that their distribution was more heavily tailed than the distribution of the other variables, and that the settled cases distributions was skewed to the right, more than the continuous independent variables. The dummy variables are mentioned at the end of the table for consistency. The information about settled cases can be also inferred from the boxplot below (see figure 1).

variable	settled cases	volatility index (VIX)	growth rate of GDP	growth rate of unemployment	lag VIX	growth rate of HPI	lag settled cases	trough	crisis
mean	8.589474	20.093404	0.068203	-0.001487	20.361825	0.003362	8.189474	0.007018	0.084211
std	8.286033	7.691934	0.860414	0.027743	8.063034	0.007233	8.411463	0.083623	0.278192
skew	2.719871	1.669316	-0.122420	0.115969	1.675849	-0.729854	2.690873	11.873897	3.010354
kurt	11.085490	4.378207	0.103280	0.305094	4.106235	1.033102	10.799591	139.971657	7.112095
median	7.000000	18.530000	-0.011071	0.000000	18.600000	0.003966	6.000000	0.000000	0.000000

Table 2: Descriptive statistics.

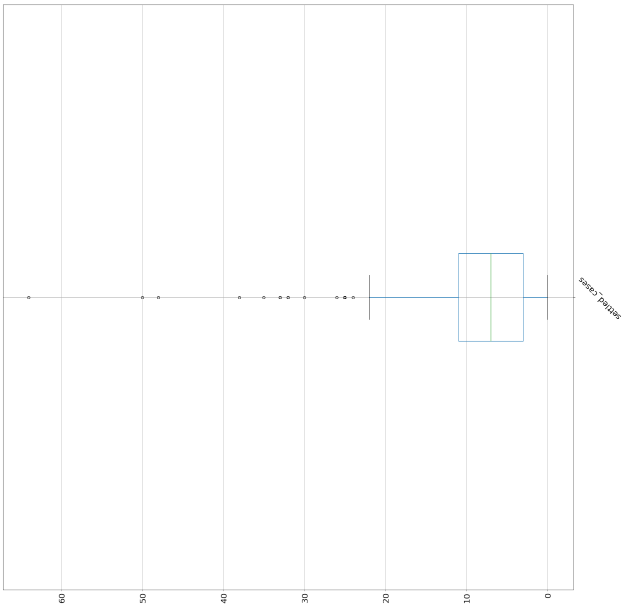
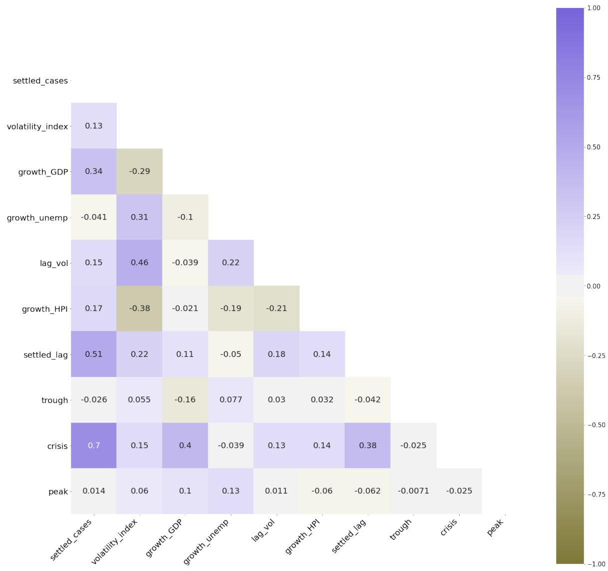


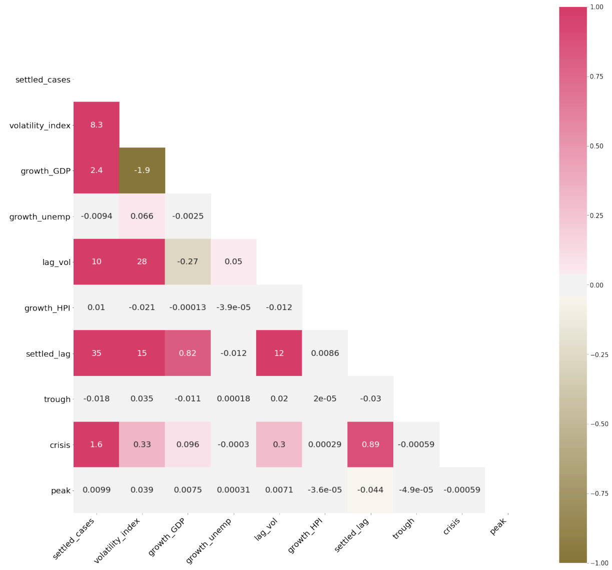
Figure 1: Boxplot of the dependent variable : number of monthly settled fraud cases.

Further, volatility index had a relatively small variance and was on average 20 units. There was a positive growth in GDP over the investigated years, however it varied by a relatively large amount. Unemployment decreased during those same years and its growth fluctuated less than that of GDP. The growth of HPI was on average small but positive, with a moderate variation.

For the business cycle dummies trough and peak, there was around 0.7% observations that occurred during trough and 0.7% that occurred while there was a business cycle peak. Around 8% fraud litigations were started according to their Class Action Period during the dot-com bubble crisis.



(a) Heatmap of pearson correlation between the model variables.



(b) Heatmap of covariation of the variables.

Figure 2: Heatmaps of dataset variables relationships.

The correlation analysis can be useful to give a first impression of what possible relationships might exist between the variables. Moreover, in the final analysis, the potential of non-linear modelling comes forward as some of the variables that exhibit low correlation with the settled fraud litigations nevertheless still have some significant impact on those lawsuits.

The correlation heatmap indicates that there should be positive relationships between the settled cases and the dummy variable indicating the dot-com bubble, as well as the dependent variable's one month lagged value. Weaker positive correlations were found with the growth rates of the GDP and HPI as well as the volatility index and its lag.

Settled cases showed stronger positive covariation with its lagged value, dot com bubble dummy, the lag of the volatility index and VIX itself, as well as the growth rate of GDP. For a closer graphic representation of the relationships between variables, see appendix 15.

4 Discussion of empirical techniques.

4.1 Non-parametric and semi-parametric models.

Parametric models refer to a class of regressions whereby assumptions play a key role. The method of investigation relies on theory as the bases for the formulation of the function that is meant to describe the relationship between the response and explanatory variables.

In non-parametric modelling, it is the form and shape of the observed data that allows for the exploration of those relationships, often in a localised manner (through looking at nearby observations in a chunk-like manner). Roughly speaking, as compared to parametric modelling, in semi- or non-parametric models the relationships between variables are extracted from the very form of the empirical data rather than assumed on the basis of theoretical assumptions. Such data-driven models are used to a minor extent in economics, but have largely gained popularity in machine learning over the past decades.

Semi-parametric models are a wide class of models that combine the characteristics of both types of regressions. By using a fixed number of parameters, as well as methods where a flexible number of parameters are included, final models are build.

As pointed out by Yatchew (1998), there are many advantages of using non- or semi-parametric models in economics. Apart from the fact that regressions are no longer restricted to rigid functional forms and their assumptions, they are also more likely to produce reliable coefficients since most implications of economic theory do not conform to the traditional parametric models. For example, there is a significant body of science indicating that business cycle variables might not be adequately formalised in linear terms, which suggests that any attempts at modelling it as such, or including it in linear regressions in the above manner, would dim the casual relationships which are the objectives of the researchers (Ghysels, 1997; Poměnková et al., 2010).

At the same time, non-parametric methods remain an underdog in the econometrics race due to the computational intensity required to actually discover relationships from neighbouring observations, as well as the sheer complexity of the methods (Yatchew, 1998). On the other hand, the models we will use in this thesis rely on well-established mathematical mechanisms such as minimizing the least squares sum. The complexity is therefore limited to the novelty of the application of the methods. By examining the residuals, instead of imposing constraints on estimators, hypothesis testing ought to be a mandate procedure rather than an obstacle.

As the parametric regressions offer a poor fit of the data, we have decided to focus on two types of semi-parametric models: GAM and MARS, which are described in the sections below. Due to the nature and character of the relationships between the variables of interest, we were unable to fit a suitable OLS regression to our data, even after the dependent and independent variables were transformed in different ways (see appendix for output of such regressions 7.1).

Moreover, we see the real-life implementation of the semi-parametric modelling of macroeconomic data as a further contribution of this thesis to the body of econometric research in the field national statistics and outside of it.

In the sections below, we avoid going into detail of the mechanics behind both GAM and MARS since the main objective of this thesis is to investigate the impact of the macroeconomic indicators on the number of settled fraud litigations, instead of investigating the econometric methods. Thus, the explanations are given at an introductory level and we recommend the more curious readers to follow up on the details of both approaches through the references used in the following sections.

4.2 GAM

Generalised additive models (GAMs) are a broad, widely encompassing class of semi-parametrics models. The more famous GLM (generalised linear model) and OLS (ordinary least squares) regressions are both sub-classes of GAMs type regressions. As compared to those traditional regression techniques, GAMs offer more flexibility in that the relationships between relevant variable are not assumed to be linear per se.

GAMs were first applied by Trevor Hastie and Robert Tibshirani in 1986 (Hastie and Tibshirani, 1990; Buja et al., 1989). In GAMs, the effect of the independent variables is captured through smoothing functions which can be linear or non-linear and are decided upon based on the underlying patterns in the data itself (Larsen, 2015).

A generalised GAM equation can be written down as (Larsen, 2015, p.2):

$$g(E(Y)) = \alpha + s_1(x_1) + ... + s_p(x_p)$$

where Y is the dependent variable, *E(Y)* denotes its expected value, and *g(Y)* stands for the link function that links the expected value to the explanatory variables $x_1, ..., x_p$. The terms $s_1(x_1), ..., s_p(x_p)$ denote smooth, non-parametric functions where the shape of predictor functions are determined based on the data itself. As compared to parametric functions, where a small set of parameters would determine the shapes, non-parametric estimation becomes more flexible, and the underlying predictive patterns can be detected without á priori knowledge of what they actually look like (Larsen, 2015).

GAM is a universal model that is relatively straightforward to interpret. In addition, hidden patterns in the data can be discovered using flexible predictor functions. Using regularised, non-parametric functions can diminish the risks that arise while using higher order polynomial terms, common in linear models. Relaxing the linearity assumption is therefore appropriate for regressing variables such as business cycle (Larsen, 2015).

The flexibility of GAM comes at a cost however. The smoothing terms and the degree of smoothing must be determined. In other words, the strength of the regularisation penalty on each explanatory variable should be carefully optimised through the use of λ parameter. In the extreme cases, $\lim_{\lambda \to +\infty}$ results in a straight line whereas $\lim_{\lambda \to 0}$ leads to an unpenalised estimate. Another issue results form the need of a choice of a standard base for regression splines. No matter the choice, only one smooth per variable can be used (Köhn, 2008). A researcher might choose between a variety of splines such as cubic or B-splines. In our case, `pyGAM` has a calibrated spline setup. Nevertheless, the use of free parameters matching the size of data lowers computational efficiency. The main disadvantage of GAM is therefore its computational complexity. Modern computers can handle the strain with success and this was not considered a limitation to the paper. Other limitations of GAM are discussed in section 4.7.

4.3 MARS

MARS stands for multivariate adaptive regression spline, which was invented and subsequently patented by Jerome Friedman (1991). MARS produces a series of connected linear regressions to better capture non-linear relationships where underlying functional forms are unknown. In our paper, the assumed relationships come from theoretical studies but according to heatmaps, any hypothetically obvious causality is at best questionable.

Basic equation for MARS can be summarised in the following:

$$\widehat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

where the model is a weighted sum of basis functions $B_i(x)$, each multiplied by c_i, a constant function (Zeng, 2018). For all possible variable and covariates MARS creates a pair of hinges. A hinge function is a pair of functions which is defined as $\max(0, x - c)$ or $\max(0, c - x)$. The basic function can be an intercept, a hinge function, or a product of several hinges. Since zero is a part of the hinge functions, the algorithm dividing the original data into sub-parts can be used to appropriately evaluate the shape of each part of the data. Several hinges are connected in the MARS model by a corresponding number of knots (constants). In such a way, piecewise linear functions can be used to describe an otherwise non-linear relationship (Friedman, 1991; Koc and Bozdogan, 2015). MARS can be understood as a generalization of the recursive partitioning regression strategy (Lewis and Stevens, 1991a; Friedman, 1991). The first stage of MARS is the forward stage, during which the algorithm acts in a greedy manner, being with a simple intercept, and adding a mirrored hinge function at each step. Each side of the hinge is a simple function which contains one of the model variables multiplied by a new hinge called a knot. This continues until the error term is deemed too small or the number of iterations reaches a pre-set limit. The second stage is the backward stage. The first stage makes a model which is very overfitted, so the backward stage prunes the least useful terms. The pruning usually removes the less useful side of the hinge, which reduces the overfitting and allows for the creation of a generalisable MARS model.

A GSV (generalization error score) is used in a like manner to a RSS (residual sum of squares) of a model. This implies that the importance of variables can be measured by summing the reduction in the GSV (or SSE or other metric) whenever a term with the variable is added (or equivalently the amount it increases as all terms with containing a given variable are removed) (Humphrey, 2017; Koc and Bozdogan, 2015).

Among advantages of modelling with MARS is that it can be used for small data samples such as 300 observations ($50 \leq \mathrm{N} \leq 1000$). Moreover, it is adjusted for moderate to high dimensions ($3 \leq \mathrm{n} \leq 20$) which ought to be appropriate regarding the nature of explanatory variables of our choice. Furthermore, multicollinearity issues arise to a lesser extent, which makes it possible to use many independent variables in the regression. See section 4.7 for longer discussion on multicollinearity. The usage of hinges brings another benefit to the analysis: little or no data preparation is needed (as the hinge functions split the dataset accordingly), as well as that hinges are more appropriate for numeric values, compared to recursive partitioning such as regression trees (Friedman, 1991).

Moreover, the available degrees of freedom are used in a more judicial way since each individual basic function added to the model is more likely to make a positive contribution (Li et al., 2009). More complex interactions can be modelled by MARS as the 'parent' basis functions is kept at all times while basis functions involving a variety of interaction levels can be added at any time. For further discussion of the mathematical assumptions of the regression see Li et al. (2009); Friedman (1991).

Besides speed, there is a question of the MARS model finding global vs.local optimum. As stated above, non-parametric methods (such as decision trees), use a greedy fitting process. That way, only the best basis function given the current model is added/removed. One way to solve the issue of finding global optima is to consider all possible basic hinge functions at once and then run a Lasso or similar penalized regression to find the best ones. This method is however complex computationally, in proportion to an increase in data size (Seitz, 2018). Furthermore, the solutions are unique so that the issue of obtaining good starting points does not arise (Li et al., 2009).

Although the method does not have any a priori assumptions about data distribution, it is still important to choose the right form of the dependent variable. In his original work, Friedman (1991) suggested selecting the set of basis functions that is best across all responses. The reason for this approach is that a wide spread of locations and scales among variables can cause instabilities that would affect the quality of the final model. MARS does not change with the locations and scales of the input variables, except for numerics (Friedman, 1991).

Many regression methods can be unreliable when dimensionality becomes high, a phenomenon known as the *curse of dimensionality* (Muñoz and Felicísimo, 2004). Solutions to the problem were proposed by Koc and Bozdogan (2015, p.37), who maintains that 'evaluation and selection of relevant subset of predictor variables with corresponding proper knots' is the most relevant for the MARS class of models.

4.4 Model choice

Regressions are a well-established kind of predictive modelling. It is important to distinguish between predictive modelling itself and econometric modelling. Both terms refer to the process of using real data to generate insights about different phenomena. While predictive modelling mainly aims at making accurate predictions (compared to time series econometrics), economic modelling focuses on identifying and quantifying causal relationships, or comparing theory with the real world.

Among the vast population of non-parametric methods, GAM and MARS stand out due to their applicability in different fields of research, from econometrics to chemical engineering (Muñoz and Felicísimo, 2004). We have implemented both GAM and MARS because of their supreme performance and relatively untapped potential when it comes to macroeconomics (Leathwick et al., 2006).

MARS (Friedman, 1991) is an automated, flexible data-mining method that is often used in machine learning, or predictive modelling problems. MARS regressions can therefore be used in constructing feasible models adequate for reasonably large datasets, which makes them suitable for a paper like this one. Apart from that, even if the independent variables take only one method in our analysis, they may take on different values for different conditions under the MARS method, which increases the possibility of obtaining accurate results (Yüksel and Adalı, 2017). As compared to linear models, MARS tend to be slower in computational terms, and they do not produce explicit confidence intervals or in-built checks. However, the benefits mentioned above qualify the method as highly relevant for our study.

GAMs (Hastie and Tibshirani, 1990) are known as flexible algorithms describing non-linear relationship between variables, and are a class of economic models that have a well established credibility. Wang et al. (2005) found that GAM showed the best fit to extreme data among other nonlinear regression model and non-parametric algorithms of that class. The fact that the operations in GAM imply simply summing the non-linear functions, the model is still very interpretable as compared to, for example, neural networks where the functions are multiplied and then transformed again (Marín, 2018).

Both models have also been independently praised for their interpretability (Balshi et al., 2009; Larsen, 2015). A comparative study was made by Moisen and Frescino (2002) (p.209) where five predictive models were assessed on 'mapping tasks given multiple objectives and logistical constraints'. During simulation processes, MARS and GAM performed only slightly better in predicting the objective characteristics than the other models tested. The two models however turned out to be much more superior when real data was used, which is a strength we considered important for the internal validity of our study. Furthermore, Leathwick et al. (2006) compared the implementation of GAM and MARS and found that there was little difference in the performance of the two.

To sum up, MARS models can be regarded as a promising predictive modelling strategy, at the same time as GAMs ought not to be neglected when it comes to the pursuit of casual inferences. This paper includes both models to increase the potential of correctly identifying potential relationships between variables as well as making accurate forecasts.

The table below summarises the important features that the models complement each other on and highlights their most significant characteristics. The methods can also be compared on their computational mechanics summarised in the flowcharts 3.

Regression	Multivariate Adpative Regression Splines (**MARS**)	Generalized Additive Model (**GAM**)
Type	non-parametric regression	non-parametric regression
Assumptions	data driven	the functions are additive, the components are smooth, GLM assumptions minus linearity
Method	linear regressions with hinges	smooth functions of predictor variables
Benefits	easy to interpret does not make assumptions on variable properties pruning phase limits overfitting,	can choose to build a grid for parameter tuning, or use intuition and subject knowledge to find optimal smoothing penalties for the model
Drawbacks	correlation is problematic, prone to muticollinearity might create complex hinges,	propensity to overfitting and multicollinearity, model can't be fitted unless NA are interpolated, bad fit for dummy variables
Implementation	`pyEarth`	`pyGam`

Table 3: Comparison of MARS and GAM

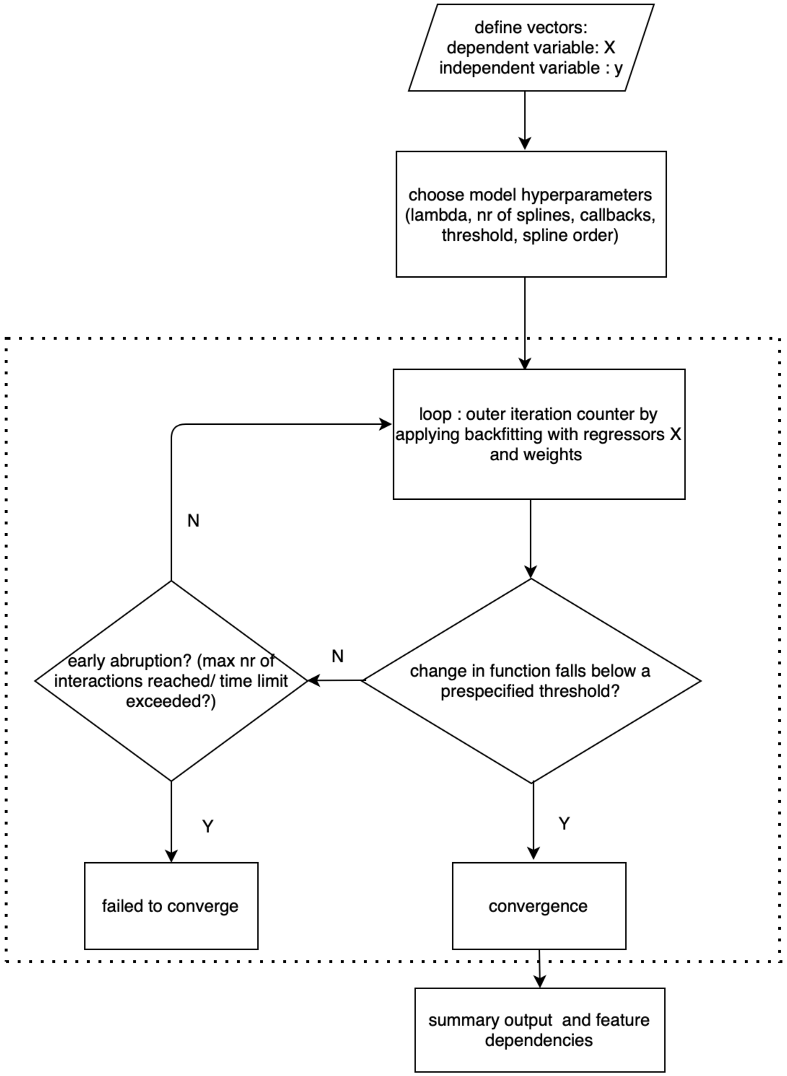


Figure 3: Flowchart of a GAM regression model. Based on Hastie and Tibshirani (1990).

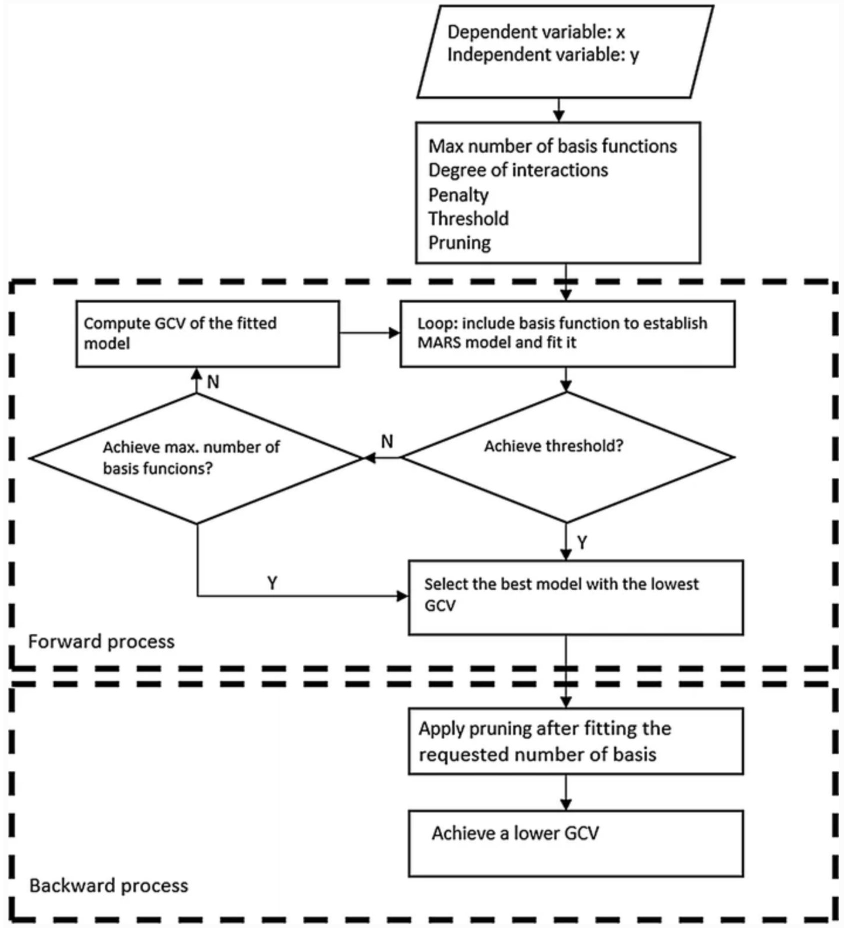


Figure 4: Flowchart of MARS model. Adopted from Erdik and Pektas (2019).

4.5 The empirical specification.

The following model was estimated:

$$\begin{aligned}\text{settled cases}_t = \beta_0 + \beta_1\text{settled cases}_{t-1} + \beta_2\text{VIX}_t + \beta_3\text{VIX}_{t-1} + \beta_4\text{change unemp}_t \\ + \beta_5\text{growth HPI}_t + \beta_6\text{growth GDP-100}_t + \beta_7\text{trough}_t + \beta_8\text{crisis}_t + \epsilon. \quad (1)\end{aligned}$$

Where *growth* stands for the once differentiated and logged variable, *GDP* stands for the gross domestic product, *VIX* stands for volatility index, *HPI* stands for the House Price Index, *unemp* is short for the unemployment rate, *trough* stands for the recession period dummy and *crisis* stands for the dot com period dummy, while ϵ captures the residual variation. For closer description of the variables, see section 3.1.2.

The GDP variable we have is already a growth variable and was stationary from the start, therefore we did not need to transform it as we did the other trending data. It is the growth with the year 1985 as the base year which is set to 100. As we would like the value to centre around 0, rather than 100, we subtracted 100 from this variable to account for this. The variables on HPI and unemployment were clearly non-stationary and therefore we took their derivative to remove the possibility of spurious correlations. We also logged reported HPI and unemployment to remove the impact of outliers, as we had a very large peak in fraud cases around 2000 as a result of the dot-com bubble. Logs also help mitigate heteroskedacticity and improve our interpretations as we are more interested in their change, the relative level, in these variables than their objective level. Finally, we added the monthly lags of *VIX* and *settled cases* to account for some of the potential omitted variable bias (see section 4.7 for a discussion on OVB) and potential autocorrelation.

The convicted fraud cases were sorted so that only cases that were convicted in court were included. In the regressions, this specification is referred to as settled cases. All securities fraud cases would include allegations, not only convictions, of fraud such as ongoing and dismissed cases.

4.6 Implementation

All the coding was performed in Python or R using Jupyter lab.

The MARS model was obtained with package `pyEarth` (Rudy and Cherti, 2017). The GAM regression was performed with several packages such as: `pyGAM, statsmodels` (Servén and Brummitt, 2018; Seabold and Perktold, 2010). Statistical tests were mainly performed in `sklearn`, `dm_test` (Tsang, 2017) and with R package `bootUR` (Friedrich et al., 2020). The analysis was supported by the material from Zamojski (2020a,b).

4.7 General issues

Both models that were chosen for this paper, MARS and GAM, can contribute with advantages and disadvantages to the investigation. The particular strengths and weaknesses of both models will be handled in another section of this paper (see section 4). Some concerns however stretch across the dimensions of a single model and are more likely to refer to the underlying economic issues, which is why we have chosen to pick them out and highlight them in the following section.

1. **Multicollinearity.** With panel data there may be some multicollinearity concerns among independent variables. If there is a high correlation between two explanatory variables there might a problem in identifying the individual causal effects of the variables on the securities fraud prevalence. However, according to Verbeek (2008); Daoud (2017), as long as collinear variables are only used as control variables, and they are not collinear with the variables of interest, there's no problem. Multicollinearity also does not affect goodness-of-fit statistics.

   Multicollinearity takes on different forms depending on which model it is permeating. In a nonparametric settings, concurvity can impact the estimates, making them more likely to be unstable (Ramsay et al., 2003) [1]. In the rest of the paper, the terms multicollinearity and concurvity will be exchangeable.

   In MARS, the dependent variable y_i is modelled as a polynomial of a degree of choice. Usually, individual dependent variables do not have a previously known, fixed polynomial degree (denoted d), even if the relationship between y_i and x_i can be investigated by looking at scatter plots. In general, increasing the highest allowed d tends to increase the presence of multicollinearity (Boehmke, 2020). This problem (i.e. high correlations) is reduced in MARS by introducing a penalty on the added variables according to the GCV criterion, as well as increasing the number of interaction terms in the model (Friedman, 1991). A more extensive discussion of possible solutions to multicollinearity problems with regard to MARS can be found in (De Veaux and Ungar, 1994; Muñoz and Felicísimo, 2004). The ground-breaking methods of circumventing multicollinearity concerns proposed by the authors can be an alternative for research which unfortunately stretches beyond the scope of this very study. We believe however that MARS is better suited to model data that is characterised by multicollinearity as the model, according to Muñoz and Felicísimo (2004), has been shown to perform relatively well even in its purer form.

   In the model class GAM, on the basic level, regularisation implies that GAMs enjoy some level of built-in immunity against concurvity (Larsen, 2015). However, high values of concurvity may lead to instability of the estimated coefficients in GAMs.

[1] Concurvity can be seen as a generalisation of collinearity which can affect interpretation in the same way (Ramsay et al., 2003)

In our dataset, variance inflation factor (VIF) was used to detect multicollinearity. VIF is calculated as the ratio of the overall model variance divided over the variance of a model regressed on a single independent variable (Staff, 2021). A high VIF indicates collinearity with some other variables in the model (Potters, 2021; Daoud, 2017). Using `statsmodels`, we were able to conclude that there might be some minor concerns when it comes to volatility index variable, see table in 7.1. It might be that the number of observations we could work with did not provide sufficient variation. However, the test values were all below 10 which would usually indicate no serious concerns regarding multicollinearity. The rule 10 has its limitations (see O'brien (2007)), but for the purpose of this thesis we consider all variables to be reasonably safe from multicollinearity.

2. **Heteroscedasticity.** Homoscedasticity requires that the model errors are independent, identically distributed with mean zero and have a constant variance (Verbeek, 2008; Lang, 2016). Homoscedasticity, or absence of heteroscedasticity, is one of the basic assumptions a large number of regression models make in their attempt to establish causality. This is because it is a condition for constructing valid prediction intervals of the modelled estimates.

   MARS is an entirely data driven model which makes no assumptions on the distribution of the predictor variables of interest (Menon et al., 2014). Gelfand (2015) finds that MARS has a stable predictive ability when it comes to modelling linear heteroscedastic data, as compared to other modern regression methods. By looking at the predicted and actual observations in the analysis, we could conclude how well our MARS models performed on dealing with heteroscedasticity, and therefore, outliers.

   GAMs stems from a generalised linear model framework which can model any data as drawn from different distributions. In `pyGam`, our models were fitted with a Gaussian conditional distribution and an identity link, both of which were most appropriate for the data. Due to that however, the model assumes conditional normality, even if the general class of GAMs does not, as one can choose any appropriate distribution for the response.

   If homoscedasticity is violated, the interpretation of the results from those models would at hazard as the errors of the coefficients would no longer be heteroscedasticity-adjusted. Because of that, we added extra tests to our methodology according to the python package `heteroscedasticity-tests`. We have decided to look at Park and Glejser methods because they are commonly used detection measures in econometrics (Glejser, 1969). Due to the small sample size in the study, note that Park and Glejser tests should be considered as strictly suggestive.

   There are several ways of performing the Park method. We have chosen the most popular version whereby the linear form of a Park test is the same as a Breusch-Pagan test. The idea is to regress the natural log of squared residuals against the independent variable. If the coefficient of the variable turns out to be significant, the variable itself might be heteroscedastic (Burgess, 2004).

Glejser method tests whether the size of random error is increasing proportionally to changes in one or more dependent variables. It is done by running three linear regressions where the outcome variable has three different forms (absolute value, square root of the absolute value and an inverse of the absolute value. Heteroscedasticity is tested, similarly to Park test, by comparing the p-value of the resulting regressions to the threshold value ($\alpha = 0.001$ or $\alpha = 0.05$) (Burgess, 2004).

We were able to conclude that some of the variables exhibit some degree of heteroscedasticity as the residuals were not found to have constant variance in the diagnostic statistics (see appendix 7.1). To deal with this issue, we used logarithm of the values unemployment, GDP and HPI so that the outliers would not interfere with the regressions. Looking at the scatter plots of the residuals (see appendix 7.1), it can be gauged that most continuous variables exhibit some homoscedastic properties despite the indications from the statistics highlighted in the paragraph above. The dummy variables make the residuals form somewhat vertical lines. Their use in multiple regression should not introduce heteroscedasticity, but rather reduce it, by resolving overlapping groups of residuals into separate ones.

The remaining heteroscedasticity concerns can be seen as the limitation of our work.

3. **Non-stationarity.** Time series may be non-stationary, which means that when modelled, the variables could show spurious correlation (Granger et al., 1974), in that the relationships found between the non-stationary variables do not really exist. In other words, non-stationary time series shows seasonal effects, trends, and other structures that depend on the time index. This is undesirable when it comes to estimating causality as non-stationarity might obscure the true causal relations and impair the judgement on statistical tests. Lewis and Stevens (1991b) find that MARS-based models such as ASTAR (adaptive spline threshold autoregression) where the predictor variable was a lagged value of a time series, work well even in non-normal situations. Furthermore, Serinaldi and Kilsby (2015) points out that a generalised models such as GAMs allow for dynamically varying variance, skewness and kurtosis which not only allow for a time varying mean, but also for changes in the full shape of a cumulative density distribution, making GAMs powerful tools with regard to non-stationarity.

   Nevertheless, to ensure that the variables we are using are suitable for the purpose of the analysis, we are looking at the graphs of the respective time series and perform several diagnostics tests. Among those, the Dickey-Fuller (ADF) test as well as

Augmented unit root tests were included, since those are considered well suited for shorter time series such as our data (Fedorová et al., 2016).

ADF test is a kind of a unit root test which determines how strongly a time series is defined by a time trend. As an autoregressive model, the variable of interest is used to optimise on information criterion across multiple different lag values of the said variable. The null hypothesis of the test is that the time series can be represented by a unit root. The alternate hypothesis is that the time series is stationary. Rejecting the null hypothesis with respect to the threshold of choice ($\alpha = 0.05$) means that the variable has non-stationary characteristics (Fedorová et al., 2016).

Without taking the first difference of the time series on unemployment and HPI, there were concerns about them being non-stationary. The findings from initial ADF told us that indeed, that could be the case. In table 8, the output on the differentiated variables from an ADF test specification are included. Because the regular ADF that is implemented in the package provided in `sklearn` has been known for low power (Elliott et al., 1992; Perron and Qu, 2007; Cavaliere et al., 2015; Palm et al., 2008) we have decided to control for stationarity by the implementation of a bootstrap unit root test (Friedrich et al., 2020). To perform the tests on multiple time series simultaneously, resampling-based bootstrap methods were used [2]. The results can be found in table 8. To account for multiple testing, the function controls for the false discovery rate (FDR) [3] (Romano et al., 2008; Moon and Perron, 2012). According to the Smeekes and Wilms (2020) tests, all the variables included in the final regression, some of them differentiated, were stationary. HPI, upon closer investigation, is assumed to be a false positive. See model 4.5 specification for details.

4. **Endogeneity.** In general, endogeneity refers to the obstacles that emerge when the variables that determine the outcome are also influenced by the modelled system.

   One kind of endogeneity problem that might occur when it comes to business cycle proxy and fraud filings is *incremental predictability* (Engle and Granger, 1987). We will look at Granger test with help of `statsmodels` to see if there are serious concerns regarding autocorrelation of the variables in the regression. The intuition of the Granger statistic is to check whether in a situation where x_t causes y_t, the forecast of y_t based on y_{t-1} and x_{t-1} results in the same as a forecast of y_t based on y_{t-1} only (Granger, 1969).

[2] ADF bootstrap refers to that the package tweaks the original version of the test to the user preference. While constructing p-values, Smeekes and Wilms (2020) allows for repeated draws from the sample of the studied time series so that distribution of the test would not be affected by size properties. Also, extensions of the original test such as quasi-differenced or sieve variant are available in the R-package.

[3] FDR can be defined as the expected proportion of false rejections divided over to the total number of rejections (Romano et al., 2008)

Furthermore, endogeneity might be also caused by *reverse causality* and *simultaneity*. The former means that when x_t affects y_t, y_t also affects x_t (Verbeek, 2008). This might be the case for fraud cases and volatility index, since investors become more uncertain about their prospects due to a large number of frauds reported in an economy, at the same time as fraud cases are likely to be more numerous in the periods of increased volatility (see section 2 for theory). When it comes to the latter, or *simultaneity*, x_t and y_t would simply be simultaneously determined in the system of equations (Verbeek, 2008). The explanatory variables become correlated with the residual term of the regression and thus become endogenous.

The f-statistic tests for Granger-causality does not reflect the concept of the "true" causality, but they can be indicative of some issues in the data. Using the Granger test however requires all of our variables to be stationary, which is the case (see paragraph above). The results can be seen in figure 7 in the appendix. Our dependent variable settled cases tested to be Granger-caused by its monthly lag, growth in GDP and the recession period modelled by the dummy.

Another issue that we will have to tackle is the *omitted variable bias* (OVB), which occurs when pertinent variables are left out from the model (Verbeek, 2008). The issue with observable and unobservable omitted factors is that they might be obscuring the causal relationship between fraud and the regressors we will model. A possible solution to OVB is to conduct an extensive research on the subject and try to identify as many factors as possible and in case they are not measurable or lacking, resort to proxies (Verbeek, 2008). Omitted variables may include microeconomic variables, as mentioned in the literature review (2), which have often been considered important factors in explaining securities fraud. Sadly, due to time constraints we were not able to control for those here, so it is likely that we will have some omitted variable bias as a result. Other possible omitted variables include those variables we mention in our future research section 7.1, sadly time constraints again prevented us from including these. We hope that the lag of the cases will help alleviate some of the problems with omitted variable bias.

Finally, *measurement error* can become a source of endogeneity. In our case, we are taking for granted that the reported statistics from section 3 were correctly collected. Moreover, we are looking in particular at settled cases since those might be the only category of fraud cases that truly fits the definition.

We will return to those limitations in the analysis and discussion of the results. A general measure of how well our models explain the variation in the data is GCV (generalised cross-validation) scores in MARS and GAM. Comparable to R^2, the higher the statistics, the better the model was in generating an approximately unbiased estimate of the prediction error.

5. **Identification.** The concept can be understood accordingly to the context when it comes to economics. In general, a system of equations is exactly identified if it can be solved with unique values of the equation parameters. In other words, identification problems occur when the data generating process cannot produce unique solutions to the optimisation problem. When it comes to non-parametric models of our choice, GAM and MARS, both of them are based on algorithms which, upon, success, converge to the 'correct' set of solutions. Hastie and Tibshirani (1990) provides three justifications of the correctness of the solutions produced by GAM :'finding projections in L^2 function spaces, minimising certain criterion with solutions from reproducing kernel Hilbert spaces and finding the solution to penalised least squares' (see Hastie and Tibshirani (1990); Buja et al. (1989) for details).

   When it comes to MARS, since no numerical optimisation is required, the solutions to the optimisation problem are assummed to be unique (Li et al., 2009).

6. **Overfitting.** This issue refers to situations where model performance on the training dataset is improved at the cost of a worse performance on data outside of the training dataset. The notion of overfitting is largely connected to the curse of dimensionality. In machine learning this is interchangeable with the peaking phenomenon, which states that the more variables (dimensions) are added the smaller the error and the larger the predictive power gets until a peak is reached. After this peak is hit adding variables will increase the error again and decrease the predictive power. The model becomes too complex so that the test sample error is higher than the train sample error. Such phenomena can lead to generalisation issues. Predictability outside of the data range is exchanged for precision within the data range. Schaffer (1993) points out that avoiding overfitting itself can also lead to a bias, depending on how it is achieved. The paper claims that the avoidance strategies should not be classified according to being better or worse, but rather more or less appropriate to specific application areas. There are several ways of dealing with overfitting. One of the most relevant methods is regularisation, or brute-force simplification of a model.

   Like other nonparametric methods, GAM and MARS have certain propensity for overfitting. There are several ways of dealing with overfitting. Overfitting refers to a situation where the performance of a model on a training dataset is improved at the expense of a poorer performance on non-training data. Among the solutions against overfitting, some popular methods from machine learning involve using cross-validation, reducing complexity of the model and regularisation.

   For the instance of GAM, the inbuilt optimisation mechanism automatically seeks the lowest GCV score. In `pyGAM`, the gridsearch 'fit' function uses GCV as its objective, which is computationally more efficient than doing manual cross validation through train and test subset available through `sklearn` package. Furthermore, regularisation of predictor functions helps avoid overfitting (Larsen, 2015), in that it can create sparsity in the solution, which implies that noise terms will be zero and that the regression becomes more robust to outliers (Das, 2019). In our regression, we model the strength of the regularisation penalty on each explanatory variable by lam or λ. It is extra important that the parameter is chosen correctly and individually.

In the context of MARS, there is another inbuilt solution to overfitting. After that the independent variables with potential hinges are added step by step. Then, all the basis functions above a penalty number, usually 2 or 3, are deemed not to contribute to the accuracy of the model and are pruned. In that way, the mechanism ensures that only the most efficient predictors are left in the final version (Machine Learning Catalogue).

A test that looks at the classifier efficiency splits the dataset into training and test to train the model on the larger portion of the data (train subset) and then cross-validate the model on the test subset. An illustration of how the principle works can be found in the appendix (see part 12). Ideally, both train and test errors turn out to be low, but if the training error is much lower than the testing error, overfitting is a problem. If both error are high, underfitting is the issue. If the training error is high but testing error is low, there might be an error in the implementation or it might indicate that the dataset was too small from the start. There are several other strategies for conducting the overfitting tests. We look at point statistics of mean squared error (MSE) and at stability of the models. Also, Diebold-Mariano statistic is evaluated on different criteria.

Due to resource and time constraints, we refer to the other possibilities as future research suggestions (Chouldechova and Hastie, 2015).

The results of the overfitting tests can be found in the appendix section 7.1. It can be seen that when it comes to the point accuracy, MARS is performing relatively well as the difference between MSE in train and test sub-samples was small (around 0.779). GAM is performing slightly worse than MARS on that point since the discrepancy between MSE is higher (around 2.08). Both values are however low and indicate a small tendency to point overfitting.

When it comes to the stability of the model, from graphs 13 it can be seen that MARS is overfitting more as it is tested closely outside of the sample. The further out of sample however, the smaller the MSE difference gets and the value seems to converge at around 3.5, which is close to the point train and test MSE overfit estimation. Meanwhile, GAM tends to overfit a lot when it is tested straight out of sample, however if it is tested far enough from it the difference in MSE converges downwards, indicating the propensity to overfit diminishes.

We also run Diebold-Mariano test (DM test) on the predicted values of GAM and MARS evaluated on the actual values. The DM test looks at the significance in differences of predictive accuracy of two models (Harvey et al., 1997). Given the predictions of GAM and the MARS, the test evaluates the null hypothesis that the mean of the loss differential of GAM is lower or equal to the mean of the loss differential function of MARS. If null hypothesis is rejected, it implies that the forecasts made by MARS are significantly more accurate than those made by GAM. The statistic can be evaluated on different criterion such as MSE or MAD (mean absolute forecast error deviation). The results of the test indicate that when judged on either MAD or MSE, on $\alpha = 0.05$, there is not enough evidence to reject null hypothesis of that MARS is better at predicting settled cases than GAM.

Finally, it can be added that making the model less complex is counterproductive when it comes to the attempts to avoid the omitted variable bias. We have performed several robustness checks where variables were transformed, as this could be considered an additional measure of playing it safe with overfitting risks (see section 16).

5 Preliminary findings

5.1 MARS

#	**Basis Function**	**Pruned**	**Coefficient**
	(Intercept)	No	15.5681
1	crisis	No	-299.68
2	h(settled_lag-26)	No	-6.97454
3	h(26-settled_lag)	No	-0.402461
4	h(growth_GDP-1.65897)*crisis	No	698.422
5	h(1.65897-growth_GDP)*crisis	No	-37.6133
6	h(volatility_index-22.64)	No	-0.176811
7	h(21.91-volatility_index)*h(22.64-volatility_index)	No	-0.0357709
8	h(growth_GDP+0.163888)*h(21.91-volatility_index)*h(22.64-volatility_index)	No	0.0285782
9	crisis*h(volatility_index-22.64)	No	-69.9847
10	volatility_index*crisis	No	8.6357
11	h(growth_GDP-1.65897)*crisis*h(volatility_index-22.64)	No	-1808.81
12	h(1.65897-growth_GDP)*crisis*h(volatility_index-22.64)	No	8.94166
13	lag_vol*crisis	No	18.629
14	lag_vol*lag_vol*crisis	No	-0.493142
15	lag_vol*crisis*h(volatility_index-22.64)	No	2.0795
16	h(22.05-volatility_index)*h(volatility_index-21.91)*h(22.64-volatility_index)	No	7363.83
17	volatility_index*h(settled_lag-26)	No	0.267573
18	h(volatility_index-18.55)*h(21.91-volatility_index)*h(22.64-volatility_index)	No	-0.374059
19	growth_unemp*crisis*h(volatility_index-22.64)	No	1822.27
20	volatility_index*growth_unemp*crisis*h(volatility_index-22.64)	No	-72.0109
21	volatility_index*h(-0.163888-growth_GDP)*h(21.91-volatility_index)*h(22.64-volatility_index)	No	0.0124798
	MSE: 13.0098, GCV: 19.7178, RSQ: 0.8098, GRSQ: 0.7138		

For the settled cases from the Securities Class Action Clearinghouse, it can be seen that MARS was able to produce a model correctly predicting as much as roughly 81% of the observations from the given data. The MSE statistics (the measure of the error the model can make in predicting data) was at around 13 which means that there is some probability our trained model on the settled fraud cases dataset would produce predictions localised 13 cases away from the model mean by chance. MSE is however mostly valuable in comparing different models and selecting the most appropriate one. A very low MSE, such as MSE of 0, might indicate overfitting, so low, non-zero values are preferable. In this model, we consider the fit to be good since the R^2 value is high while the MSE statistic is moderately high and our test train split performs well, see appendix 13.

Among the coefficients that were not pruned, several variables were used for the construction of the final hinges. *The dot com crisis*, interaction terms with *growth of GDP* and *volatility index* as well as the included *lag of VIX* can be seen in the output from MARS model specification.

As stated before, MARS is most advantageous when it comes to predictive modelling. However, causal interpretations are not impossible and an example of a cautious interpretation for the variable on volatility index VIX would be that on its own, if VIX lies between 21.91 and 22.64, the number of settled fraud cases would decrease by -0.176811. At the same time, when the lag of settled cases approaching the value 26, the corresponding value of VIX would increase the number of settled cases themselves by 0.267573.

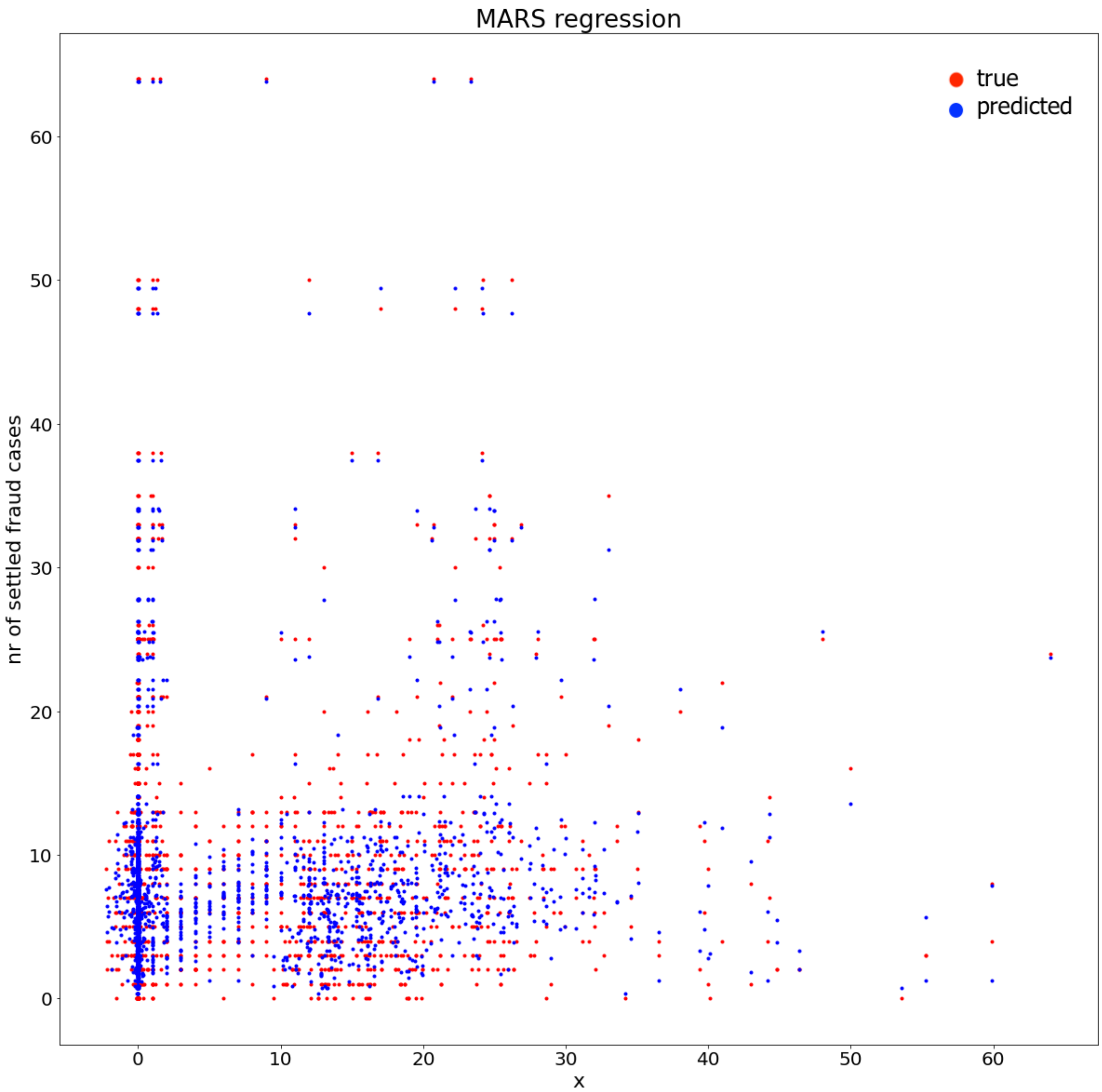


Figure 5: Graphic representation of MARS: settled cases.

Figure 5 presents the fit of the pyEarth model to the data. The vertical axis presents the dependent variable while the horizontal axis presents the vector of a matrix of the input independent variables. By looking at the graph, it can be seen that the distance between the two groups of points in not large and that the mapping of the forecasted values to the actual observations is overall what could be expected from the RSQ of the model.

5.2 GAM

Distribution:	NormalDist	Effective DoF:	48.8617		
Link Function:	IdentityLink	Log Likelihood:	-1183.7887		
Number of Samples:	285	AIC:	2467.3009		
Pseudo R-Squared:	0.6975	AICc:	2488.9637		
Scale:	24.9812	GCV:	35.8371		
Feature Function	**Lambda**	**Rank**	**EDoF**	**P > x**	**Sig. Code**
volatility index	[35.5629]	35	12.8	2.67e-01	
growth GDP	[18.881]	35	10.4	2.55e-01	
change unemployment	[22.3216]	35	9.4	1.70e-01	
lag volatility	[7.0986]	35	6.9	8.59e-01	
growth HPI	[13.4386]	35	5.1	9.91e-01	
lag settled cases	[1.6943]	35	3.7	9.93e-06	***
trough	[3.1918]	35	0.3	9.73e-01	
crisis	[4.1389]	35	0.1	1.11e-16	***
intercept	1		0.0	1.67e-07	***
Significance codes:	***' 0.001 '**' 0.01 * 0.05 '.' 0.1				

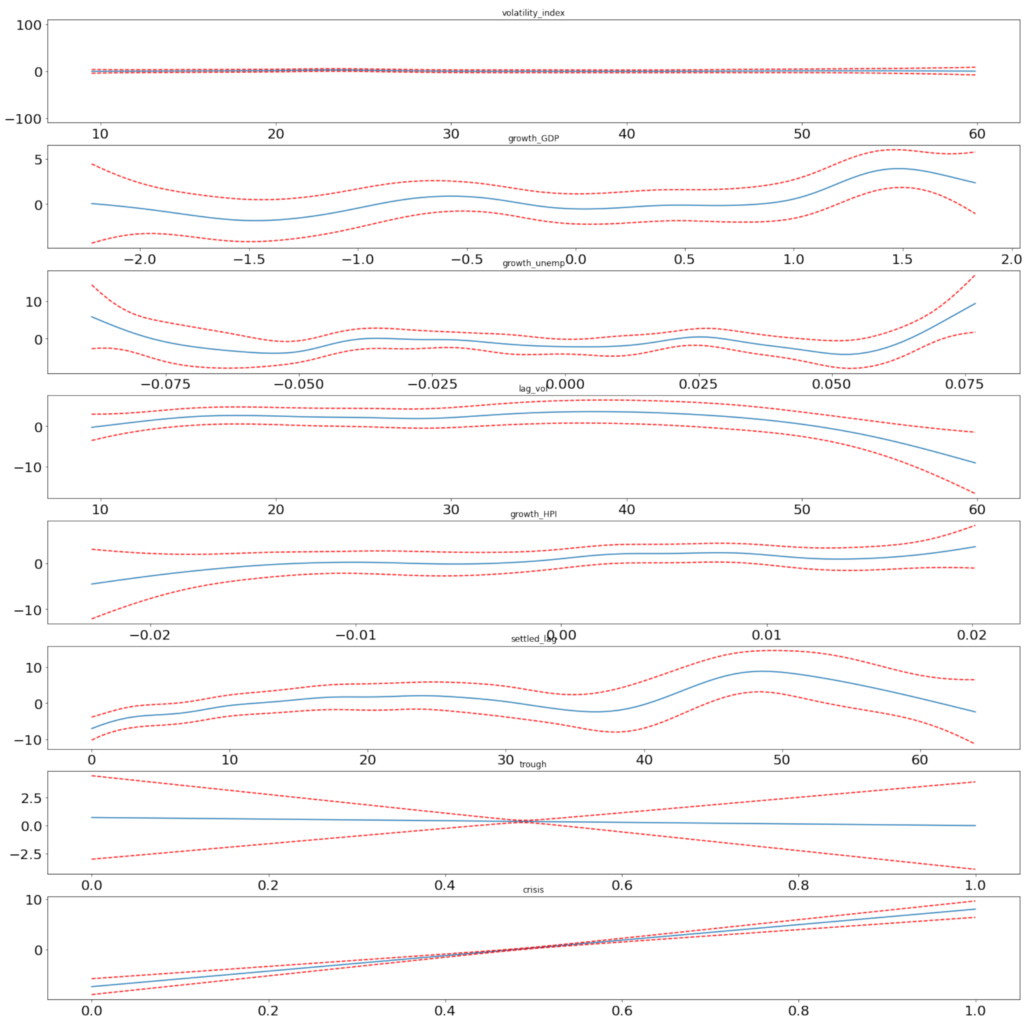


Figure 6: Graphic representation of the fit through GAM regression: settled cases

Figure 6 presents partial dependence plots of our explanatory variables and their predictions modelled in the Linear GAM. The blue lines stand for smoothed terms while the dotted red lines visualise the 95% confidence intervals of the prediction functions. It can be seen that all the inputs were contained within the confidence intervals, albeit the fit for *trough* dummy is not as tight as the fit of the other variables.

GAMs are a class of linear models which means that they can be interpreted as such. However, there is a difference between making inference based on OLS and GAM since the latter has smooth terms. Therefore, there will be no single coefficient that can be interpreted in terms of magnitude or direction. Partial effects of the smooth terms can be interpreted visually however, or it is possible to make inferences from the predicted values.

In our model, `pyGam` creates 35 splines per variable, which in total produces 280 coefficients (one per each spline in each feature, 281 including the intercept).

Looking at the partial derivatives graphs, we can see that some variables are more 'wiggly' then the others. There is a trade-off between increasing the number of splines (better fit) and R^2 (Perperoglou et al., 2019). The more splines which are included, the closer the modelled relationship gets to the true data. However, this does not guarantee a good exploration of causal relationships due to overfitting issues. Therefore, a penalty for wiggliness is boiled down to lambda which is individually parametrised for each variable. From the graphs and the output table it can be seen that all variables lie within their 0.95 confidence intervals. *The dot com period dummy* and *lag of settled cases* are statistically significant at 0.001 level.

5.3 Robustness checks

Different robustness tests aim at establishing whether the results obtained in the paper are valid in an internal and external way. Internal validity refers to the results withstanding the alteration of the functional form of the model, of the variables or other specifications in the study. External validity allows for the application of the results of a single paper to other fields or disciplines. A robust model would not produce extremely different results when tested for altered specifications or assumptions.

Among the robustness checks we included:

I *regression form checks*: different modelling possibilities were explored, starting with the classical linear regressions such as robust OLS and a GLM with Gamma identity function link.

II *variable form checks*: dependent and independent variables were all taken a log transform of or had it removed, to compare the performance of the models on those specifications

From the histograms in the appendix and the summary statistics that can be seen in section 3, it can be seen that the variables used in models 4.5 are not free from outliers and have skewed distributions. For that reason, and for the sake of interpretation, we have decided to model HPI, GDP and unemployment as logarithms rather than linear differentiated time series.

For the robustness, we have transformed volatility index and its lag as well as settled cases and its lag into natural logarithm [4]. The results were as follows: the growth rate of the GDP was consistently showing up in the hinges constructed by MARS. The same was true for the dummy variable defined for the dot com crisis. In GAM, the lag of the settled cases was significant on $\alpha = 0.05$.

Once we tried removing all the logarithms of the variables, MARS found consistently the variable on GDP as well as the dot com crisis dummy to be significant. GAM on the other hand found the lagged value of settled cases to be statistically viable.

Running the original dataset on OLS and GLM, variables such as the dummy on dot com crisis, the monthly lag of settled cases as well as the growth rate of GDP were confirmed to be statistically significant.

From the robustness tests described above, we could conclude that the variables found to be significant in GAM and MARS, such as *crisis* and *lag of the settled cases*, are significant in the parametric regressions too. However, many more variables are considered significant in the OLS which might indicate that it is not fully capturing the shape of the data, also due to the assumptions of linearity not being full-filled. Despite relatively high R^2 values of the non-parametric models and GAM and MARS run on transformed variables, together with the overfitting test, we conclude that the most robust models were the ones included in the main results. When it comes to the internal validity of the findings however, there are indications of that the *dot com crisis* and the *number of the settled cases during the previous month* indeed are causally related to the number of settled fraud cases in the USA in the relevant years. Growth in GDP, captured by MARS but not GAM, was also significant in several specifications, see figure 7.1.

6 Analysis and Discussion

6.1 Final results GAM

For 35 splines times 8 variables plus the intercept, the model in `pyGAM` created 281 coefficients. Those coefficients of the features in the decision function for $\widehat{settled\_cases}$ can be found in the appendix (see section 7.1). Looking to the graphs we can clearly see that GAM finds, in general, a rise in GDP will increase the number of securities fraud

[4] In the cases where there were zero settled cases per month, we used the transformation $log(n+1)$ where n is the number of cases.

cases. This is confirmed by our variable trough which has a negative relationship with the number of securities fraud cases. Both volatility and the lag of volatility seem to have a negative relationship with securities fraud at low and high volatilities according to GAM. This indicates that volatility may have a delayed relationship with the number of securities fraud cases according to GAM. These results must be taken into context with our coefficients and the table of results which clearly shows the only variables our GAM model finds significant are the lag of settled cases and our crisis variable. This indicates that, outside of the dot com bubble, GAM suggests there are no variables which help to significantly explain the number of securities fraud cases. However, our GAM model over fits more than our MARS model, therefore our MARS results are probably more reliable.

6.2 Final Results MARS

The estimated $\widehat{settled\_cases}$ becomes:

$$\begin{aligned}\widehat{settled\_cases} &= 15.5681 - 299.68 * H_1 - 6.97454 * H_2 - 0.402461 * H_3 + 698.422 * H_4 \\ &- 37.6133 * H_5 - 0.176811 * H_6 - 0.0357709 * H_7 + 0.0285782 * H_8 + -69.9847 * H_9 \\ &+ 8.6357 * H_{10} - 1808.81 * H_{11} + 8.94166 * H_{12} + 18.629 * H_{13} - 0.493142 H_{14} + 2.0795 * H_{15} \\ &+ 7363.83 * H_{16} + 0.267573 * H_{17} - 0.374059 * H_{18} + 1822.27 * H_{19} - 72.0109 * H_{20} + 0.0124798 * H_{21}\end{aligned} \tag{2}$$

For an overview of the hinges see the appendix 7.1.

Hinges with no level mentioned such as volatility_index always have an effect, while those with a level mentioned such as h(volatility_index-22.64) are having an effect above 22.64 and h(22.64-volatility_index) has an effect below 22.64. There are also some limitations with these functions as is demonstrated by H_13 and H_14 which interact with each other to mitigate each others effects. In order to determine the overall effect, especially in complicated time periods such as the dot com bubble, we combine the effect of the hinges to study the overall relationship.

The hinge at the lag of settled cases at 26 only affects the dot com bubble which is an extraordinary outlier, well known for major accounting scandals such as the Enron affair. This seems to indicate that, in spite of its large coefficient, the lag of settled cases is not significant in predicting the number of securities fraud cases outside of the dot com bubble, which was an exceptional event.

For growth of GDP the hinges lie predominantly during the dot com bubble and the 2008 financial crisis, but there are also spikes from 2004-2008, during the summer of 2014 and during the years 2012, 2018 and 2019. Growth of GDP is prevalent in H4, H5, H8, H11, H12 and H21. Although H4, H5, H11 and H12 only have an effect during the dotcom bubble, H8 and H21 have an effect outside of the dot com bubble indicating that growth of GDP is relevant in general for predicting securities fraud. This indicates that business cycles, as proxied by GDP, are relevant in explaining the number of securities fraud cases. As the economy expands, the number of securities fraud cases rise and then falls just before the end of a boom. The overall effect of a change in GDP however is very small.

Volatility has the largest number of affected time periods. The hinges were principally related to the dot com bubble, the 2008 financial crisis and the years 2016, 2018 and 2019. However there were also spikes around the early 2000s and spikes in the years 2009, 2010, 2011 and 2012. Volatility is present in H6, H7, H8, H9, H10, H11, H12, H15, H16, H17, H18, H19, H20 and H21. While H9, H10, H11, H12, H15, H17, H19 and H20 are only prevalent during the dot com bubble, H6, H7, H8, H16, H18 and H21 are relevant across our time period and this indicates that volatility has an effect on securities fraud cases outside of the dot com bubble. As market volatility increases, the number of securities fraud cases also increases. Interestingly, this effect was much larger than any of the other variables, as can be judged by the coefficients of the relevant hinges.

The other variables mentioned in the interaction hinges are: crisis, unemployment and lag of volatility. Crisis impacts only during the dot com bubble, we have used it to isolate the effect of the dot com bubble rather than to explain the prevalence of securities fraud across time. Unemployment is prevalent in H19 and H20 and lag of volatility is prevalent in H13, H14 and H15, all of these hinges only have an effect during the dot com bubble, so again these variables are not useful for explaining the prevalence of securities fraud cases outside of the dot com bubble.

Thus, according to the MARS model, the variables which appear to be relevant for explaining a change in the number of securities fraud cases outside of the dot com period are GDP and volatility. Of these, volatility has by far the largest effect.

6.3 Combining results

Our GAM model overfits more than our MARS model, as can be seen in our graphs in the appendix 13, and both models find the lag of settled cases and crisis are significant. As MARS over fits less than GAM, it is likely that market volatility and GDP are also significant. MARS's findings that market volatility may have an effect on the number of securities fraud cases lends credence to the work of (Hastie and Tibshirani, 1990), in the way that an increase in market volatility, common during expansions, would contribute to an increase in fraud. In that way, expansions in business cycle would affect lawsuits through market speculation.

It also suggests that the theoretical models of Povel et al. (2007), Blanqué (2003), Stout (2002) and Hertzberg (2003) do appear to hold for settled securities fraud in general. Our results lend greatest credence to Povel et al. (2007)'s predictions as there appears to be a decrease in the number of securities fraud cases just before the end of a boom, which Hertzberg (2003) or Stout (2002) do not predict. Our findings agree with those of Wang et al. (2010) and Fernandes and Guedes (2010), there seems to be a significant relationship between the business cycle and the number of class action lawsuits based on the class action period, this would suggest that fraud or at least its detection is more prevalent during booms.

Basu (2018)'s game theoretic model based on Akerlof and Shiller (2015), and the work of Mishina et al. (2010), suggest fraud increases with competition, however in a boom there is often less competition and investments and raises are easier to obtain, especially at the end of a boom. This suggests that competition may not be an explanation for the relationship between securities fraud in general and business cycles during a boom. However, Basu (2018)'s model in particular assumes there is no cost to fraud and it may be the role of regulatory fluctuations across the business cycle which is responsible for the discovery of fraud during booms. Sadly, we could not test either market confidence, as explained in our literature review section, or the effect of market regulation as we did not focus on this specifically in our thesis. However, the suggestion that market regulation may have an impact on fraud's connection with the business cycle and that therefore frauds are more prevalent during booms is plausible based on our findings. We recommend specific research into this to determine if the correlation with business cycles has more to do with optimism, regulation or other factors.

Utset (2013)'s suggestion that fraud may exacerbate market bubbles seems to be possible, sadly we cannot differentiate if fraud is contributing to bubbles or if bubbles contribute to fraud, however there is a clear relationship between the two. It also seems that academic predictions of fraud appear to be more well founded than professional recommendations that fraud is more likely during a bust than a boom.

We find that market volatility has the largest significant relationship with the number of securities class action lawsuits and this suggests (Baker, 2016)'s concerns of convictions during more volatile market periods may be well founded. Sadly, we cannot test each security to the market volatility of the market on which it is traded, however firms often trade on multiple exchanges, therefore we would expect the market volatility variable to reflect the effect of the prevalence of fraud. It seems this effect is quite considerable and may call into question some of the present practices and models used as evidence in securities fraud cases. Our results back concerns that frauds may be wrongfully convicted and there are genuine concerns to be addressed with regards to market volatility and the number of securities fraud convictions. Our results also confirm the link between volatility and fraud found by (Jain et al., 2008; Schenk, 2017). We can also confirm Fox et al. (2015)'s findings that major downturns are accompanied by spikes in market volatility, this is clear in our data for both models.

We do not find that unemployment (outside of the dot com bubble) or HPI to be related to the number of securities class action lawsuits. This suggests that, at least for securities fraud, there does not appear to be a statistically significant effect between the unemployment rate or HPI on the number of securities fraud cases. It may be that unemployment as a result of mass fraud is not large enough to affect the unemployment across the US. However, this does suggest the predictions of Kedia and Philippon (2005) and Mustafa and Khan (2020) may not be applicable to all types of securities fraud and, although GDP may be affected by fraud, the relationship with unemployment is not strong enough to alter unemployment in general.

Finally, the dot come bubble appears to be a unique time period in our data, during which all our variables except HPI become relevant. During this period an increase in the number of convicted fraud cases accompanies an economic expansion (measured through trough and GDP), an increase in unemployment and an increase in market volatility.

6.4 Limitations

Given that we look to all sectors and all class action lawsuits, we cannot confirm or deny the results for particular securities fraud types, we sadly did not have enough time to collect and process this data within the given time frame of a master's thesis. However, our results do seem to back the models of Hertzberg (2003) and Povel et al. (2007). Expanding on Wang et al. (2010) and Fernandes and Guedes (2010)'s findings, who only study IPO fraud and accounting fraud respectively, we find that our data agrees with the theoretical literature and also suggests that their findings are applicable to all securities fraud, not only the particular frauds they look at. Based on our finding that GDP is most significant during the dot com bubble and its effect at other times is smaller although it is still significant, it may be that Wang et al. (2010)'s significance with regards to IPO fraud is so great because they focus on the dot com bubble. Had they used a larger or different time period their effect may have been weaker. We would suggest research that stretches further back in time than our sample, and specifically looks at IPO fraud, which could better confirm or refute this suggestion. However, our work does seem to suggest that the dot com bubble should not be treated as a typical case in terms of securities fraud.

7 Conclusion

In our thesis we aimed to study the effect of business cycle relevant macroeconomic variables (housing price index, GDP, and a dummy for peak and trough), market volatility and unemployment, and analyse their effect on securities fraud cases (proxied by all convicted SEC class action lawsuits from 1996-2019). We find that the macroeconomic trends which have an effect on the number of class action lawsuits outside of the dot-com bubble are growth of GDP and market volatility. The number of convicted securities fraud cases increases with an increase in the growth of GDP and also with an increase in market volatility. However, growth of GDP may have a small effect on securities fraud prevalence. We have expanded the work of Wang et al. (2010) and Fernandes and Guedes (2010) to incorporate all types of securities fraud and also find that market volatility is linked with the number of fraud litigations.

Our study has tested securities fraud across a wider time period and a wider range of securities fraud types than any other study we are aware of and we also include previously untested securities fraud types. Alongside this, our inclusion of semi-parametric models offers a more accurate fit to securities fraud and business cycle data to both fields of research than many of the models currently used.

We recommend the continued and expanded use of GAM and MARS in regards to business cycles to capture the clear non-linear pattern of the data without the worry of the curse of dimensionality. There appeared to be a clear overfitting in the OLS robustness check with regards to some of the variables and the inclusion of semi-parametric models which are not affected by the curse of dimensionality, as OLS is, has helped to mitigate the false significance values we would have obtained had we run only parametric models. Our research confirms the claims of Leathwick et al. (2006) in that GAM and MARS do perform on similar levels in empiric studies. It also indicates that what Wang et al. (2005) claimed about GAMs ability to deal with extreme values checks out, since the dot com bubble was correctly modelled by our linear GAM and in stability evaluation, the model performed well when it comes to out-of-sample convergence of MSE.

Finally, we hope that our work has added empirical backing to the relationships between securities fraud and the business cycle, and market volatility. We hope our work can help introduce better econometric tools which more accurately capture non-linear variables such as business cycle, and are now more computationally feasible thanks to advances in modern computational technologies.

7.1 Future research

Because of the resource and time constrains, we were forced to limit ourselves to a certain angle in our paper. However, there is much space for further work and some suggestions are listed below:

- Create a *unique business cycle indicator* similar to the ones constructed by Ferrara (2003); Valle e Azevedo et al. (2006) and incorporate it in the regressions or use modified-cash-deposits-ratio (MCDR) (Lee and Wang, 2012; Lee et al., 2008; Yuliang, 2015; Matkowski, 2019).

- Create a *globalisation proxy* and incorporate it in the regressions, since globalisation has been found to increase exposure to fraud in many different ways (Button and Cross, 2017).

- Include extra *variables* in the regression that we have left out according to table 1.

- Account for *special events* or control for most influential cases, for example by exclusion of the potential influential points identified in the literature and news such as the dot come bubble. As a result of the dot com bubble, between 2001-2002 there were many securities fraud scandals in the US including Enron, WorldCom and Tyco which were a result of fraudulent accounting. This, alongside publicity with regards to securities fraud cases, may have encouraged many more securities fraud filings (Kluger and Slezak (2018)). Indeed, in the data we see a spike in securities fraud cases in 2001. As a result, changes were made to securities fraud legislation under the Sarbanes Oxley Act of 2002 which may have increased the number of cases which were successfully identified by the courts and we would expect the number of prosecutions to rise (Kluger and Slezak (2018)). The COVID-19 recession may be of future interest, it was not accompanied by sector wide corruption, but instead an exogenous, global shock to the world economy which contrasts it with many prior recessions and is an interesting test of securities fraud and recession types.

- Include a *measure of liquidity* such as money supply into the regressions, since, according to Friedman and Schwartz (1965), the underlying cause of the business cycles are the fluctuations in the growth rate of money supply.

- Look at particular sectors/industries and investigate how fraud cases behave according to their specific indexes and cycles. There will be specific effects on particular sectors of the economy as there are different and relevant characteristics which we were sadly unable to capture, for example tech sectors are often growth firms (Frankenfield, 2021) and the finance sector often exhibits high levels of volatility (Davis, 2020). This will require tracking the size of these sectors in relation to the market to more accurately portray which effect is as a result of sector growth or fraud specific growth, the organisation of the data for this fell beyond the scope of this study, but we do hope it will be picked up by others at a later date. See the appendix 7.1 for a greater discussion on this.

- Perform the regressions using different non-parametric models that do not assume normality or stationarity: GEEs - generalised estimating equations, GLMMs - generalised linear mixed models, or GAMMs - generalised additive mixed models, ASTAR - adoptive spline threshold autoregression.

- A regression can be performed on the 'alleged' fraud reports stemming from the difference between the reported and convicted securities fraud claims during the investigated years. A calculation similar to this could be performed for the relevant years:

$$alleged fraud_i = reported fraud_i - convicted fraud_i$$

- In a like manner, a study on a dataset based on the *filing date* of the fraud cases claims instead of Class Period Start could be performed whereby a regression similar to model 4.5 would be performed, with the cases sorted on different date.
- We also hope that there will be a continued and expanded use of MARS and GAM and its derivatives in future research which we believe will help alleviate many of the current problems in trying to capture the non-linearity of the business cycle and plenty of other variables.

References

Unit root testing, 2019. URL https://arch.readthedocs.io/en/latest/unitroot/unitroot_examples.html.

R. Agnew, N. Piquero, and F. Cullen. General strain theory and white-collar crime. In S. Simpson and D. Weisburd, editors, *The Criminology of White-Collar Crime*, chapter 3, pages 35–60. Springer, New York, 2009. URL https://doi.org/10.1007/978-0-387-09502-8_3.

G. A. Akerlof and R. J. Shiller. *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton University Press, 2015. ISBN 9780691173023. URL http://www.jstor.org/stable/j.ctvc777w8.

W. S. Albrecht, C. Albrecht, and C. C. Albrecht. Current trends in fraud and its detection. *Information Security Journal: A Global Perspective*, 17(1):2–12, 2008. doi: 10.1080/19393550801934331. URL https://doi.org/10.1080/19393550801934331.

L. Allen and A. Saunders. Bank window dressing: Theory and evidence. *Journal of Banking & Finance*, 16(3):585–623, 1992.

A. Atkinson. Regression diagnostics, transformations and constructed variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):1–22, 1982.

M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research*, 20(1):40–49, 2011.

A. C. Baker. Single-firm event studies,securities fraud, and financial crisis: Problems of inference. *Stanford Law Review*, 68:1207–1262, 2016. doi: https://heinonline.org/HOL/LandingPage?handle=hein.journals/stflr68&div=33&id=&page=.

M. S. Balshi, A. D. McGUIRE, P. Duffy, M. Flannigan, J. Walsh, and J. Melillo. Assessing the response of area burned to changing climate in western boreal north america using a multivariate adaptive regression splines (mars) approach. *Global Change Biology*, 15(3):578–600, 2009.

M. Baneshi and A. Talei. Impact of imputation of missing data on estimation of survival rates: an example in breast cancer. 2010.

K. Basu. Markets and manipulation: Time for a paradigm shift? *Journal of Economic Literature*, 56:185–205, 2018. doi: http://dx.doi.org.ezproxy.ub.gu.se/10.1257/jel.20161410.

M. S. Beasley, J. V. Carcello, D. R. Hermanson, and P. D. Lapides. Fraudulent financial reporting: Consideration of industry traits and corporate governance mechanisms. *Accounting Horizons*, 14(4):441–454, 2000. doi: https://doi.org/10.2308/acch.2000.14.4.441.

P. Blanqué. Crisis and fraud. *Journal of Financial Regulation and Compliance*, 11:60–70, 2003. ISSN 1358-1988. doi: https://doi.org/10.1108/13581980310810417.

B. Boehmke. Hands-on machine learning with r, Feb 2020. URL https://bradleyboehmke.github.io/HOML/mars.html.

M. Boumans and J. B. Davis. *Economic methodology: Understanding economics as a science*. Macmillan International Higher Education, 2015.

C. N. Braaten and M. S. Vaughn. Convenience theory of cryptocurrency crime: A content analysis of u.s. federal court decisions. *Deviant Behavior*, 0(0):1–21, 2019. doi: 10.1080/01639625.2019.1706706. URL https://doi.org/10.1080/01639625.2019.1706706.

I. E. Brick and N. Chidambaran. Board meetings, committee structure, and firm value. *Journal of Corporate Finance*, 16(4):533–553, September 2010. URL https://ideas.repec.org/a/eee/corfin/v16y2010i4p533-553.html.

S. W. Buell. What is securities fraud? *Duke Law Journal*, 61(3):511–581, 2011. ISSN 00127086. URL http://www.jstor.org/stable/41353728.

A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

L. T. Bulan and L. E. Simmons. Securities class action settlements—2019 review and analysis. Technical report, 2020. URL https://www.cornerstone.com/Publications/Reports/Securities-Class-Action-Filings-2019-Year-in-Review.

M. Burgess. *Transformation techniques in data mining*. PhD thesis, University of East Anglia, 2004.

N. Burns and S. Kedia. The impact of performance-based compensation on misreporting. *Journal of Financial Economics*, 79(1):35–67, 2006. ISSN 0304-405X. doi: https://doi.org/10.1016/j.jfineco.2004.12.003. URL https://www.sciencedirect.com/science/article/pii/S0304405X05001133.

M. Button and C. Cross. Technology and fraud. *The Routledge Handbook of Technology, Crime and Justice*, page 78, 2017.

F. Carmichael and R. Ward. Male unemployment and crime in england and wales. *Economics Letters*, 73(1):111–115, 2001. ISSN 0165-1765. doi: https://doi.org/10.1016/S0165-1765(01)00466-9. URL https://www.sciencedirect.com/science/article/pii/S0165176501004669.

G. Cavaliere, P. C. Phillips, S. Smeekes, and A. R. Taylor. Lag length selection for unit root tests in the presence of nonstationary volatility. *Econometric Reviews*, 34(4):512–536, 2015.

Y. S. Chang and D. Pak. Warren buffett value indicator vs. gdp size – is the relationship superlinear? *International Journal of Economics and Business Research*, 15(2):223–235, 2018. ISSN 1756-9869. doi: 10.1504/IJEBR.2018.089688.

Chicago Board Options Exchange. CBOE Volatility Index: VIX (VIXCLS). URL https://fred.stlouisfed.org/series/VIXCLS. Accessed: January 1, 2021.

H. H. Chin and L. Fenuttyi. Applying digital analysis to detect fraud: an empirical analysis of us marine industry. *Applied Economics*, 45(1):135–140, 2013. doi: 10.1080/00036846.2011.605759. URL https://doi.org/10.1080/00036846.2011.605759.

S. J. Choi and A. C. Pritchard. Sec investigations and securities class actions: An empirical comparison. *Journal of Empirical Legal Studies*, 13(1):27–49, 2016. doi: https://doi.org/10.1111/jels.12096.

A. Chouldechova and T. Hastie. Generalized additive model selection. *arXiv preprint arXiv:1506.03850*, 2015.

J. C. Coffee Jr. The changing character of securities litigation in 2019: Why it's time to draw some distinctions, Jan 2019. URL https://clsbluesky.law.columbia.edu/2019/01/22/the-changing-character-of-securities-litigation-in-2019-why-its-time-to-draw-some-distinctions/.

C. Cooper. Cynthia cooper discusses fraud, ethics, and standing up for what's right. *Journal of Health Care Compliance*, 11:37–42, 2009. ISSN 1520-8303.

Cosine. Detecting multicollinearity with vif - python, Aug 2020. URL https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/.

A. Cretarola, G. Figà-Talamanca, and M. Patacca. Market attention and bitcoin price modeling: theory, estimation and option pricing. *Decisions in Economics and Finance*, 43:187 – 228, 2020. doi: https://doi.org/10.1007/s10203-019-00262-x.

C. E. Crutchley, M. R. H. Jensen, and B. B. Marshall. Climate for scandal: Corporate environments that contribute to accounting fraud. 42:53–73, 2007. doi: https://doi-org.ezproxy.ub.gu.se/10.1111/j.1540-6288.2007.00161.x.

J. I. Daoud. Multicollinearity and regression analysis. In *Journal of Physics: Conference Series*, volume 949, page 012009. IOP Publishing, 2017.

A. Das. The game of regularization, Jun 2019. URL https://towardsdatascience.com/the-game-of-regularization-91442b3be862.

M. Davis. The 8 most volatile sectors, Aug 2020. URL https://www.investopedia.com/financial-edge/0712/the-8-most-volatile-sectors.aspx.

M. A. Davis and J. Heathcote. Housing and the business cycle. *International Economic Review*, 46(3):751–784, 2005.

R. D. De Veaux and L. H. Ungar. Multicollinearity: A tale of two nonparametric regressions. In *Selecting models from data*, pages 393–402. Springer, 1994.

Deloitte. Managing the business risk of fraud: New guidance for a new risk environment. 2009. URL https://www2.deloitte.com/content/dam/Deloitte/in/Documents/risk/Corporate%20Governance/Audit%20Committee/in-gc-managing-the-business-risk-of-fraud-noexp.pdf.

Y. Deng, C. Chang, M. S. Ido, and Q. Long. Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific reports*, 6(1):1–10, 2016.

T. Dias, I. S. R., M. Tardivo, and G. Tardivo. Liberalisation and capital market performance: lessons from the enron affair. *Journal of Financial Regulation and Compliance*, 13:229–253, 2005. doi: https://doi-org.ezproxy.ub.gu.se/10.1108/13581980510622090.

T. Dietterich. Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR)*, 27(3):326–327, 1995.

G. Dionne and K. D. Wang. Journal of risk and uncertainty. 47:67–92, 2013. doi: https://doi.org/10.1007/s11166-013-9171-y.

S. Dolgopolov. Risks and hedges of providing liquidity in complex securities: The impact of insider trading an options market makers. *Fordham Journal of Corporate & Financial Law*, 15:387–438, 2010. URL https://heinonline.org/HOL/P?h=hein.journals/fjcf15&i=391.

J. Du and S.-J. Wei. Does insider trading raise market volatility? *The Economic Journal*, 114(498):916–942, 2004. ISSN 00130133, 14680297. URL http://www.jstor.org/stable/3590244.

A. Dyck, A. Morse, and L. Zingales. Who blows the whistle on corporate fraud? *The Journal of Finance*, 65(6):2213–2253, 2010. ISSN 00221082, 15406261. URL http://www.jstor.org/stablCB/23324409.

K. Edmark. Unemployment and crime: Is there a connection? *Scandinavian Journal of Economics*, 107(2):353–373, 2005.

I. Ehrlich. Crime, punishment, and the market for offenses. *The Journal of Economic Perspectives*, 10(1):43–67, 1996. ISSN 08953309. URL http://www.jstor.org/stable/2138283.

G. Elliott, T. J. Rothenberg, and J. H. Stock. Efficient tests for an autoregressive unit root. Technical report, National Bureau of Economic Research, 1992.

R. F. Engle and C. W. Granger. Co-integration and error correction: representation, estimation, and testing. *Econometrica: journal of the Econometric Society*, pages 251–276, 1987.

T. Erdik and A. O. Pektas. Rock slope damage level prediction by using multivariate adaptive regression splines (mars). *Neural Computing and Applications*, 31(7):2269–2278, 2019.

D. Fedorová et al. Selection of unit root test on the basis of length of the time series and value of ar (1) parameter. *Statistika*, 96(3):3, 2016.

N. Fernandes and J. Guedes. Keeping up with the joneses: A model and a test of collective accounting fraud. *European Financial Management*, 16:72–93, 2010. doi: https://doi.org/10.1111/j.1468-036X.2009.00494.x.

L. Ferrara. A three-regime real-time indicator for the us economy. *Economics Letters*, 81 (3):373–378, 2003.

Fidelity Investments. Sectors & amp; industries - weighting recommendations. URL https://eresearch.fidelity.com/eresearch/markets_sectors/sectors/si_weighting_recommendations.jhtml?tab=sirecommendations.

S. I. Foodman. Predatory lending and mortgage fraud. *Banking LJ*, 126:254, 2009.

E. G. Fox, M. B. Fox, and R. J. Gilson. Economic crisis and the integration of law and finance: The impact of volatility spikes. *Columbia Law Review*, 116:325–408, 2015. doi: http://dx.doi.org/10.2139/ssrn.2401712.

J. Frankenfield. Inside the technology sector, Mar 2021. URL https://www.investopedia.com/terms/t/technology_sector.asp.

FRED. Number of listed companies for united states, Oct 2019. URL https://fred.stlouisfed.org/series/DDOM01USA644NWDB.

R. Freeman. Crime and unemployment. *Crime and public policy*, 1983.

J. H. Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991.

M. Friedman and A. J. Schwartz. Money and business cycles. In *The state of monetary economics*, pages 32–78. NBER, 1965.

M. Friedrich, S. Smeekes, and J.-P. Urbain. Autoregressive wild bootstrap inference for nonparametric trends. *Journal of Econometrics*, 214(1):81–109, 2020. doi: 10.1016/j.jeconom.2019.05. URL https://ideas.repec.org/a/eee/econom/v214y2020i1p81-109.html.

J. Gee and M. Button. The financial cost of fraud 2018: The latest data from around the world. 2018.

S. J. Gelfand. Understanding the impact of heteroscedasticity on the predictive ability of modern regression methods. 2015.

A. C. Ghent and M. T. Owyang. Is housing the business cycle? evidence from us cities. *Journal of Urban Economics*, 67(3):336–351, 2010.

E. Ghysels. On seasonality and business cycle durations: A nonparametric investigation. *Journal of econometrics*, 79(2):269–290, 1997.

H. Glejser. A new test for heteroskedasticity. *Journal of the American Statistical Association*, 64(325):316–323, 1969.

E. Goldman and S. L. Slezak. An equilibrium model of incentive contracts in the presence of information manipulation. *Journal of Financial Economics*, 80:603–626, 2006.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

C. W. Granger, P. Newbold, and J. Econom. Spurious regressions in econometrics. *Baltagi, Badi H. A Companion of Theoretical Econometrics*, pages 557–61, 1974.

J. M. Griffin. Ten years of evidence: Was fraud a force in the financial crisis? 2019. doi: http://dx.doi.org/10.2139/ssrn.3320979. URL https://ssrn.com/abstract=3320979.

J. D. Hamilton and G. Lin. Stock market volatility and the business cycle. *Journal of Applied Econometrics*, 11(5):573–593, 1996. ISSN 08837252, 10991255. URL http://www.jstor.org/stable/2285217.

A. Harel and K. N. Hylton. *Research Handbook on the Economics of Criminal Law*. Edward Elgar Publishing, 2012.

D. Harvey, S. Leybourne, and P. Newbold. Testing the equality of prediction mean squared errors. *International Journal of forecasting*, 13(2):281–291, 1997.

L. H. Hass, M. Tarsalewska, and F. Zhan. Equity incentives and corporate fraud in china. *Journal of Business Ethics*, 138:723–742, 2016. ISSN 1573-0697. doi: 10.1007/s10551-015-2774-2. URL https://doi.org/10.1007/s10551-015-2774-2.

T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

A. Hertzberg. Managerial incentives, misreporting, and the timing of social learning: A theory of slow booms and rapid recessions. *working paper*, 2003.

Y. Huang, Q. Li, K. H. Liow, and X. Zhou. Is housing the business cycle? a multiresolution analysis for oecd countries. *Journal of Housing Economics*, 49:101692, 2020.

K. Humphrey. Using reinforcement learning to personalize dosing strategies in a simulated cancer trial with high dimensional data. 2017.

In Choi and Bhum Suk Chung. Sampling frequency and the power of tests for a unit root: A simulation study. *Economics Letters*, 49(2):131–136, 1995. ISSN 0165-1765. doi: https://doi.org/10.1016/0165-1765(95)00656-Z. URL https://www.sciencedirect.com/science/article/pii/016517659500656Z.

International Center of Finance, Yale School of Management, Yale University. United States Stock Market Confidence Indices. URL https://som.yale.edu/faculty-research-centers/centers-initiatives/international-center-for-finance/data/stock-market-confidence-indices/united-states-stock-market-confidence-indices. Accessed: February 14, 2021.

Investor.gov. Blue sky laws, 2020. URL https://www.investor.gov/introduction-investing/investing-basics/glossary/blue-sky-laws. Accessed: 31 March 2021.

P. K. Jain, J. Kim, and Z. Rezaee. The sarbanes-oxley act of 2002 and market liquidity. *The Financial Review*, 43:361–382, 2008. doi: https://doi.org/10.1111/j.1540-6288.2008.00198.x.

D. D. P. Johnson and J. H. Fowler. The evolution of overconfidence. *Nature*, 477:317–320, 2011. doi: https://doi.org/10.1038/nature10384.

J. M. Karpoff, A. Koester, D. S. Lee, and G. S. Martin. A critical analysis of databases used in financial misconduct research. *SSRN Electronic Journal*, 2012.

S. N. Kasturi. Cross-validation techniques to assess your model's stability, Jun 2020. URL https://medium.com/swlh/cross-validation-techniques-to-assess-your-models-stability-3a4d55d90409.

S. Kedia and T. Philippon. The economics of fraudulent accounting. *Review of Financial Studies*, 22, 09 2005. doi: 10.2139/ssrn.676241.

W. Kenton. How the stock market capitalization-to-gdp ratio is used, Jan 2021. URL https://www.investopedia.com/terms/m/marketcapgdp.asp.

B. D. Kluger and S. L. Slezak. Signal jamming models of fraudulent misreporting and economic prospects: An experimental investigation. *Journal of Economic Behavior & Organization*, 151:254 – 283, 2018. ISSN 0167-2681. doi: https://doi.org/10.1016/j.jebo.2018.04.008. URL http://www.sciencedirect.com/science/article/pii/S0167268118301082.

K. Knorr Cetina and A. Preda. *The Oxford handbook of the sociology of finance*. Oxford University Press, 2012.

E. K. Koc and H. Bozdogan. Model selection in multivariate adaptive regression splines (mars) using information complexity as the fitness function. *Machine Learning*, 101(1): 35–58, 2015.

S. Köhn. *Generalized additive models in the context of shipping economics*. PhD thesis, University of Leicester, 2008.

I. Kuznetsova. Heteroscedasticity-tests, Park and Glejser methods, Aug. 2020. URL https://github.com/kuzn137/Heteroscedasticity-tests.

H. Lang. Elements of regression analysis. *Stockholm: KTH Mathematics*, 2016.

K. Larsen. Gam: the predictive modeling silver bullet. *Multithreaded. Stitch Fix*, 30: 196–223, 2015.

A. I. Lawal, E. O. Amogu, J. O. Adeoti, and M. A. Ijaiya. Fraud and business cycle: Empirical evidence from fraudsters and fraud managers in nigeria. *Studies in Business and Economics*, 12(1):110 – 128, 2017. doi: https://doi.org/10.1515/sbe-2017-0009. URL https://content.sciendo.com/view/journals/sbe/12/1/article-p110.xml.

E. E. Leamer. Housing really is the business cycle: what survives the lessons of 2008–09? *Journal of Money, Credit and Banking*, 47(S1):43–50, 2015.

J. Leathwick, J. Elith, and T. Hastie. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological modelling*, 199(2):188–196, 2006.

Y.-M. Lee and K.-M. Wang. Searching for a better proxy for business cycles: With supports using us data. *Applied Economics*, 44:1433–1442, 04 2012. doi: 10.1080/00036846.2010.543073.

Y.-M. Lee, K.-M. Wang, and T. Thanh-Binh Nguyen. A common-use proxy for economic performance: Application to asymmetric causality between the stock returns and growth. *International Journal of Business & Economics*, 7(2), 2008.

M. Levi, A. Doig, R. Gundur, D. Wall, and M. Williams. Cyberfraud and the implications for effective risk-based responses: themes from uk research. 67:77 – 96, 2017. ISSN 1573-0751. doi: https://doi.org/10.1007/s10611-016-9648-0.

P. A. Lewis and J. G. Stevens. Nonlinear modeling of time series using multivariate adaptive regression splines (mars). *Journal of the American Statistical Association*, 86 (416):864–877, 1991a.

P. A. W. Lewis and J. G. Stevens. Nonlinear modeling of time series using multivariate adaptive regression splines (mars). *Journal of the American Statistical Association*, 86 (416):864–877, 1991b. ISSN 01621459. URL http://www.jstor.org/stable/2290499.

B. Li, B. Bakshi, and P. Goel. 3.13 - other methods in nonlinear regression. In S. D. Brown, R. Tauler, and B. Walczak, editors, *Comprehensive Chemometrics*, pages 463–476. Elsevier, Oxford, 2009. ISBN 978-0-444-52701-1. doi: https://doi.org/10.1016/B978-044452701-1.00062-4. URL https://www.sciencedirect.com/science/article/pii/B9780444527011000624.

A. Ljungqvist and W. J. Wilhelm. Ipo pricing in the dot-com bubble. *The Journal of Finance*, 58(2):723–752, 2003. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/3094556.

T. Lohse and C. Thomann. Are bad times good news for the securities and exchange commission? *European Journal of Law and Economics*, 40:33–47, 2015. ISSN 1572-9990. doi: 10.1007/s10657-014-9455-y. URL https://doi.org/10.1007/s10657-014-9455-y.

S. K. Long and A. D. Witte. Current economic trends: implications for crime and criminal justice. *Crime and criminal justice in a declining economy*, pages 69–143, 1981.

Machine Learning Catalogue. Multivariate adaptive regression splines algorithm. URL https://machinelearningcatalogue.com/algorithm/alg_multivariate-adaptive-regression-splines.html.

D. S. Marín. pygam: balancing interpretability and predictive power using generalised additive models, Aug 2018. URL https://www.youtube.com/watch?v=XQ1vk7wEI7c.

Z. Matkowski. 23 business cycles in poland. In *Social and Structural Change: Consequences for Business Cycle Surveys-Selected Papers Presented at the 23rd Ciret Conference, Helsinki*, page 407. Routledge, 2019.

M. Mayer. Comments on lynn a. stout's the investor confidence game. *Brook. L. Rev.*, 68:449, 2002.

R. Menon, G. Bhat, G. R. Saade, and H. Spratt. Multivariate adaptive regression splines analysis to predict biomarkers of spontaneous preterm birth. *Acta obstetricia et gynecologica Scandinavica*, 93(4):382–391, 2014.

D. Miron. EEG-processing. Granger causality matrix, Mar. 2021.

Y. Mishina, B. J. Dykes, E. S. Block, and T. G. Pollock. Why "good" firms do bad things: The effects of high aspirations, high expectations, and prominence on the incidence of corporate illegality. *The Academy of Management Journal*, 53(4):701–722, 2010. ISSN 00014273. URL http://www.jstor.org/stable/20788788.

G. G. Moisen and T. S. Frescino. Comparing five modelling techniques for predicting forest characteristics. *Ecological modelling*, 157(2-3):209–225, 2002.

D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*, volume 821. John Wiley & Sons, 2012.

H. R. Moon and B. Perron. Beyond panel unit root tests: Using multiple testing to determine the nonstationarity properties of individual series in a panel. *Journal of Econometrics*, 169(1):29–33, 2012.

B. C. L. Morris, J. F. Egginton, and K. P. Fuller. Return and liquidity response to fraud and sec investigations. *Journal of Economics and Finance*, 43(2):313–329, 2019. doi: 10.1007/s12197-018-9445-y. URL https://doi.org/10.1007/s12197-018-9445-y.

J. Muñoz and Á. M. Felicísimo. Comparison of statistical methods commonly used in predictive modelling. *Journal of Vegetation Science*, 15(2):285–292, 2004.

S. Mustafa and F. Khan. The Relationship between Accounting Frauds and Economic Fluctuations: A Case of Project Based Organizations in UAE. *Journal of Economics and Public Finance*, 6(1):2377–1046, 2020.

I. Myung. Computational approaches to model evaluation. In N. J. Smelser and P. B. Baltes, editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 2453–2457. Pergamon, Oxford, 2001. ISBN 978-0-08-043076-8. doi: https://doi.org/10.1016/B0-08-043076-7/00589-1. URL https://www.sciencedirect.com/science/article/pii/B0080430767005891.

NBER. Us business cycle expansions and contractions, 2020. URL https://www.nber.org/research/data/us-business-cycle-expansions-and-contractions. Accessed: 31 March 2021.

O. Olsson. *Essentials of advanced macroeconomic theory*, volume 17. Routledge, 2013.

R. M. O'brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690, 2007.

M. Pakaluk. Reply to karpoff et al. (2012), 'a critical analysis of databases used in financial misconduct research'. *SSRN Electronic Journals*, 2012. doi: http://dx.doi.org/10.2139/ssrn.2127460.

F. C. Palm, S. Smeekes, and J.-P. Urbain. Bootstrap unit-root tests: comparison and extensions. *Journal of Time Series Analysis*, 29(2):371–401, 2008.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid. A review of spline function procedures in r. *BMC Medical Research Methodology*, pages 46–19, 2019. ISSN 1471-2288. doi: 10.1186/s12874-019-0666-3. URL https://doi.org/10.1186/s12874-019-0666-3.

P. Perron and Z. Qu. A simple modification to improve the finite sample properties of ng and perron's unit root tests. *Economics letters*, 94(1):12–19, 2007.

J. Poměnková et al. An alternative approach to the dating of business cycle: Nonparametric kernel estimation. *Prague Economic Papers*, 19(3):251–272, 2010.

C. Potters. Variance inflation factor (vif), Mar 2021. URL https://www.investopedia.com/terms/v/variance-inflation-factor.asp.

H. R. Pourghasemi and M. Rossi. Landslide susceptibility modeling in a landslide prone area in mazandarn province, north of iran: a comparison between glm, gam, mars, and m-ahp methods. *Theoretical and Applied Climatology*, 130(1):609–633, 2017.

P. Povel, R. Singh, and A. Winton. Booms, busts, and fraud. *The Review of Financial Studies*, 20(4):1219–1254, 2007. ISSN 08939454, 14657368. URL http://www.jstor.org/stable/4494801.

A. C. Pritchard. Markets as monitors: A proposal to replace class actions with exchanges as securities fraud enforcers. *Virginia Law Review*, 85(6):925–1020, 1999. ISSN 00426601. URL http://www.jstor.org/stable/1073966.

PWC. Fraud in a downturn: A review of how fraud and other integrity risks will affect business in 2009. *PricewaterhouseCoopers International Limited*, 2009. URL https://www.pwc.com/gr/en/publications/assets/fraud-in-downturn.pdf.

T. O. Ramsay, R. T. Burnett, and D. Krewski. The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, 14(1):18–23, 2003.

J. P. Romano, A. M. Shaikh, and M. Wolf. Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, 17(3):417–442, 2008.

J. Rudy and M. Cherti. Py-earth: a python implementation of multivariate adaptive regression splines, 2017.

C. Schaffer. Overfitting avoidance as bias. *Machine learning*, 10(2):153–178, 1993.

J. A. Scheinkman and W. Xiong. Overconfidence and speculative bubbles. *Journal of Political Economy*, 111(6):1183–1220, 2003. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/10.1086/378531.

C. Schenk. Rogue trading at lloyds bank international, 1974: Operational risk in volatile markets. *Business History Review*, 91(1):105–128, 2017. doi: 10.1017/S0007680517000381.

C. M. Schrand and S. L. Zechman. Executive overconfidence and the slippery slope to financial misreporting. *Journal of Accounting and Economics*, 53(1):311 – 329, 2012. ISSN 0165-4101. doi: https://doi.org/10.1016/j.jacceco.2011.09.001. URL http://www.sciencedirect.com/science/article/pii/S0165410111000644.

A. Schuchter and M. Levi. The fraud triangle revisited. *Security Journal*, 29(2):107–121, 04 2016. URL https://search-proquest-com.ezproxy.ub.gu.se/scholarly-journals/fraud-triangle-revisited/docview/1778894315/se-2?accountid=11162.

S. Seabold and J. Perktold. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, volume 57, page 61. Austin, TX, 2010.

U. Securities. Securities exchange act of 1934, 1934.

S. Securities Class Action Clearinghouse. Securities class action clearinghouse filings database, 2020. URL http://securities.stanford.edu/filings.html. Accessed: 1 January 2021.

S. Securities Exchange Commission. Manual of publicly available telephone interpretations: Tender offer rules and schedules, 1997. URL https://www.sec.gov/interps/telephone/cftelinterps_tender.pdf. Accessed: 31 March 2021.

S. Securities Exchange Commission. Laws and rules, 2020. URL https://www.sec.gov/investment/laws-and-rules. Accessed: 31 March 2021.

Securities Lawyer's Deskbook. Securities act of 1933, 2020a. URL https://lawblogs.uc.edu/sld/the-deskbook-table-of-contents/the-securities-acts-statutory-law/the-securities-act-of-1933/. Accessed: 31 March 2021.

Securities Lawyer's Deskbook. Securities exchange act of 1934, 2020b. URL https://lawblogs.uc.edu/sld/the-deskbook-table-of-contents/the-securities-acts-statutory-law/the-securities-exchange-act-of-1934-15-usc-%c2%a7-78a-et-seq/. Accessed: 31 March 2021.

S. Seitz. Non-greedy mars regression, Sep 2018. URL https://numbersandcode.com/non-greedy-mars-regression.

F. Serinaldi and C. G. Kilsby. Stationarity is undead: Uncertainty dominates the distribution of extremes. *Advances in Water Resources*, 77:17–36, 2015.

D. Servén and C. Brummitt. pygam: Generalized additive models in python, Mar. 2018. URL https://doi.org/10.5281/zenodo.1208723.

S. Smeekes and I. Wilms. bootur: An r package for bootstrap unit root tests. *arXiv preprint arXiv:2007.12249*, 2020.

I. Staff. Variance inflation factor (vif), May 2021. URL https://www.investopedia.com/terms/v/variance-inflation-factor.asp.

R. T. Stewart. Bank fraud and the macroeconomy. *Journal of Operational Risk*, 11(1), 2016.

L. A. Stout. The investor confidence game. *Brooklyn Law Review*, 68(2):407–437, 2002. URL https://scholarship.law.cornell.edu/facpub/762.

Svenska Vetenskapsrådet. Good research practice, 2017.

P. Taylan, G.-W. Weber, and F. Y. Özkurt. A new approach to multivariate adaptive regression splines by using tikhonov regularization and continuous optimization. *Top*, 18(2):377–395, 2010.

M. Thornton. Review essay: Phishing for phools: The economics of manipulation and deception george a. akerlof and robert j. shiller princeton, n.j.: Princeton university press, 2015, 272 pp. *The Quarterly Journal of Austrian Economics*, 19:85–100, 2016.

J. Tsang. Diebold mariano test, 2017. URL https://github.com/johntwk/Diebold-Mariano-Test.git.

U.S. Bureau of Labor Statistics. Unemployment Rate [UNRATE]. URL https://fred.stlouisfed.org/series/UNRATE. Accessed: February 14, 2021.

M. A. Utset. Fraudulent corporate signals: Conduct as securities fraud. *Boston College Law Review*, 54(2):645–710, 2013. URL https://lawdigitalcommons.bc.edu/bclr/vol54/iss2/6.

J. Valle e Azevedo, S. J. Koopman, and A. Rua. Tracking the business cycle of the euro area: A multivariate model-based bandpass filter. *Journal of Business & Economic Statistics*, 24(3):278–290, 2006.

H. van Driel. Financial fraud, scandals, and regulation: A conceptual framework and literature review. *Business History*, 61(8):1259–1299, 2019. doi: 10.1080/00076791.2018.1519026. URL https://doi.org/10.1080/00076791.2018.1519026.

M. Verbeek. *A guide to modern econometrics*. John Wiley & Sons, 2008.

T. Y. Wang, A. Winton, and X. Yu. Corporate fraud and business conditions: Evidence from ipos. *The Journal of Finance*, 65(6):2255–2292, 2010. ISSN 00221082, 15406261. URL http://www.jstor.org/stable/23324410.

Y. Wang, F. Raulier, and C.-H. Ung. Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods—a case study of lodgepole pine productivity. *Forest Ecology and Management*, 214(1-3):201–211, 2005.

G.-W. Weber, İ. Batmaz, G. Köksal, P. Taylan, and F. Yerlikaya-Özkurt. Cmars: a new contribution to nonparametric regression with multivariate adaptive regression splines supported by continuous optimization. *Inverse Problems in Science and Engineering*, 20(3):371–400, 2012.

P. R. Wheale and L. H. Amin. Bursting the dot.com "bubble': A case study in investor behaviour. *Technology Analysis & Strategic Management*, 15(1):117–136, 2003. doi: 10.1080/0953732032000046097. URL https://doi.org/10.1080/0953732032000046097.

Wilshire Associates. Wilshire 5000 Total Market Index (Wilshire 5000). URL https://www.wilshire.com/indexcalculator/. Accessed:.

Yale. United states stock market confidence indices, Jun 2019. URL https://som.yale.edu/faculty-research-centers/centers-initiatives/international-center-for-finance/data/stock-market-confidence-indices/united-states-stock-market-confidence-indices.

A. Yatchew. Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36(2):669–721, 1998. ISSN 00220515. URL http://www.jstor.org/stable/2565120.

B. Yılmaz. Housing market dynamics and advances in mortgages: option based modelling and hedging. 2019.

C. Yin, X. Cheng, Y. Yang, and D. Palmon. Do corporate frauds distort suppliers' investment decisions? *Journal of Business Ethics*, 2020. doi: https://doi.org/10.1007/s10551-019-04369-4.

X. Yu. Securities fraud and corporate finance: Recent developments. *Managerial and Decision Economics*, 34:439–450, 2013. doi: https://doi.org/10.1002/mde.2621.

C. Yu-liang. Business cycle turning points of the var threshold model and empirical analysis: Based on quarterly data 1970˜ 2012 of 12 countries and regions. *On Economic Problems*, (6):6, 2015.

S. Yüksel and Z. Adalı. Determining influencing factors of unemployment in turkey with mars method. *International Journal of Commerce and Finance*, 3(2):25–36, 2017.

M. Zamojski. Lecture notes in financial econometrics, February-March 2020a.

M. Zamojski. Lecture notes in advanced data analysis, October-November 2020b.

N. Zeng. Multivariate adaptive regression splines in a nutshell, Jun 2018. URL https://blog.zenggyu.com/en/post/2018-06-16/multivariate-adaptive-regression-splines-in-a-nutshell/.

A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. Glm and gam for count data. In *Mixed effects models and extensions in ecology with R*, pages 209–243. Springer, 2009.

Appendix A

Securities Claims Table

Class Action Claims		
Act	**Section**	**Concerns**
Securities Exchange Act 1934	Section 9	Security price manipulation
Securities Exchange Act 1934	Section 10b	Manipulation regarding securities-based swaps
Securities Exchange Act 1934	Section 13d	Reports from acquirers of >5% for particular securities classes
Securities Exchange Act 1934	Other Section 13	Fraudulent reporting
Securities Exchange Act 1934	Section 14a	Executive compensation must be regularly approved by shareholders
Securities Exchange Act 1934	Section 14d	Fraud regarding tender offers
Securities Exchange Act 1934	Section 14e	Misleading statements or omissions regarding tender offers
Securities Exchange Act 1934	Section 16	Filing statements of directors, officers, and stockholders holding >10%
Securities Exchange Act 1934	Section 18	Misleading statements
Securities Exchange Act 1934	Section 20a	Insider trading
Securities Act 1933	Section 11	false registration statement civil liabilities
Securities Act 1933	Section 12a1	Prospectus and communication civil liabilities under section 77e
Securities Act 1933	Section 12a2	Prospectus and communication civil liabilities not under section 77e
Securities Act 1933	Section 15	Controlling persons' liability under section 11 and 12
Securities Act 1933	Section 17a	Fraud concerning interstate transactions
Investment Advisers Act of 1940	n/a	Investment Advisers
Investment Company Act of 1940	n/a	Conflicts of interest of limited liability, mutual funds, investment, and securities trading companies
Other Federal Securities Claims	n/a	Securities claims not otherwise specified
Other Federal Non-Securities Claims	n/a	Non-Securities claims not otherwise specified
States' Blue Sky Laws	n/a	State specific securities laws
Other States' Claims	n/a	State claims not otherwise specified

Based on the following sources: Securities Lawyer's Deskbook (2020a), Securities Lawyer's Deskbook (2020b), Securities Exchange Commission (2020), Investor.gov (2020), Securities Exchange Commission (1997) and Securities Class Action Clearinghouse (2020).

Title	Author	Date	Relevance	Method	Relevant Results
Phishing for Phools	Akerlof and Shiller	2015	Strong	Theoretical	Phishing equilibrium, Reputation mining
Current Trends in Fraud and its Detection	Albrecht, Albrecht, and Albrecht	2008	Medium	Review	Discusses the nature of fraud, gives and overview of the major accounting scandals from 1998-2008
Single-Firm Event Studies, Securities Fraud, and Financial Crisis: Problems of Inference	Baker	2016	Strong	Empirical	Single-firm event studies perform less well in the presence of excessive market volatility and may over estimate the number and size of frauds.
Impact of imputation of missing data on estimation of survival rates: an example in breast cancer	Baneshi and Talei	2010	Weak	Empirical	Disussion of imputation
Markets and Manipulation: Time for a Paradigm Shift?	Basu	2018	Strong	Theoretical Lit. Review	Gives a game theoretical model of Akerlof and Shiller's phishing equilibrium
Crisis and fraud	Blanque	2003	Strong	Theoretical	Fraud is more prevalent during booms
What is Securities Fraud?	Buell	2011	Weak	Theoretical	Discusses the legal definition of securities fraud
The impact of performance-based compensation on misreporting	Burns and Kedia	2006	Weak	Empirical	The effect of CEO's option portfolio sensitivity on the price of the firm's stock on misreporting
SEC Investigations and Securities Class Actions: An Empirical Comparison	Choi and, Pritchard	2016	Medium	Empirical	Class-action-only lawsuits have a greater impact on investors beliefs and therefore stock market prices than SEC-only investigations
Cynthia Cooper Discusses Fraud, Ethics, and Standing Up for What's Right.	Cooper	2009	Medium	Interview	Cooper draws parallels between dot-com and 2008: Companies take risks during bubbles thanks to a lack of regulation and oversight, when the bubble bursts they use fraud to hide the excessive risk they took on
Market attention and Bitcoin price modeling: theory, estimation and option pricing	Cretarola Figà-Talamanca and Patacca	2020	Weak	Empirical	Market sentiment and speculation are correlated
Managing the business risk of fraud: New guidance for a new risk environment	Deloitte	2009	Medium	Theoretical	Discusses prevalence of fraud and business cycles

Table 4: Literature Overview Part 1

Title	Author	Date	Relevance	Method	Relevant Results
Liberalisation and capital market performance: lessons from the Enron affair	Dias, Soares and Tarvido	2005	Medium	Empirical	Positive relationship between regulatory oversight and improvements in stock market prices alongside a reduction in market volatility
Financial fraud, scandals, and regulation: A conceptual framework and literature review	van Driel	2018	Medium	Lit. Review	Gives an overview of the literature on fraud
Does Insider Trading Raise Market Volatility?	Du and Wei	2004	Strong	Empirical	Countries where insider trading is more prevalent also have higher average stock market volatility.
Who Blows the Whistle on Corporate Fraud?	Dyck, Morse and Zingales	2010	Weak	Empirical	Investigate the prevalence of whistle-blowing (1996-2004)
Keeping Up with the Joneses: A Model and a Test of Collective Accounting Fraud	Fernandes and Guedes	2010	Strong	Empirical	Accounting fraud increases during a boom, but decreases at the end of a boom.
Economic Crisis and the Integration of Law and Finance: The Impact of Volatility Spikes	Fox, Fox, and Gilson	2015	Medium	Review Empirical	Every major downturns (1920-2008) has been accompanied by a spike in idiosyncratic risk as measured by market volatility
An equilibrium model of incentive contracts in the presence of information manipulation	Goldman and Slezak	2006	Weak	Theoretical	Agency model of stock-based incentives and propensity to commit fraud
Equity Incentives and Corporate Fraud in China	Hass Tarsalewska and Zhan	2016	Weak	Theoretical	Equity incentives of managers and supervisors affect their propensity to commit fraud
Managerial Incentives, Misreporting, and the Timing of Social Learning: A Theory of Slow Booms and Rapid Recessions	Hertzberg	2003	Strong	Theoretical	Fraud is most prevalent during a boom

Table 5: Literature Overview part 2

Title	Author	Date	Relevance	Method	Relevant Results
The Sarbanes-Oxley Act of 2002 and Market Liquidity	Jain, Kim and Rezaee	2008	Medium	Empirical	An increase in the number of frauds significantly decreases liquidity in the market and in turn raises market volatility. The Sarbanes-Oxley Act (2002) reduced fraud and had a long term positive effect on market liquidity
The Economics of Fraudulent Accounting	Kedia and Philippon	2005	Strong	Empirical	Firms which commit fraud over-invest and over-employ to conceal their fraud and imitate non-fraudulent firms and once the fraud is revealed they will drop the excess jobs and investment
IPO Pricing in the Dot-Com Bubble	Ljungqvist, and Wilhelm, Jr.	2003	Weak	Empirical	Under pricing during the dot com bubble
Are Bad Times Good News for the Securities and Exchange Commission?	Lohse and Thomann	2014	Medium	Empirical	Find that the SEC's funding increases when the stock market is doing badly
Comments on Lynn A. Stout's The Investor Confidence Game	Mayer	2002	Medium	Theoretical	Discusses Stout(2002)
Why 'Good' Firms Do Bad Things	Mishina, Johnson Dykes, Block and Pollock	2010	Strong	Empirical	Relative, not absolute performance, is a greater incentive to commit fraud
Return and liquidity response to fraud and sec investigations	Morris, Egginton, and Fuller	2019	Medium	Empirical	Market regulation, specifically SEC investigation announcements, have a positive effect on firm stock prices and decrease volatility
Booms, Busts, and Fraud	Povel, Singh and Winton	2007	Strong	Empirical	At the very end of a boom investor beliefs of the economy will be overoptimistic and will not monitor as strongly, so the incentive to commit fraud decreases, therefore fraud will peak just before the end of a boom.
Fraud in a Downturn: A review of how fraud and other integrity risks will affect business in 2009	PWC	2009	Medium	Theoretical	Accountancy analysis of risk and business cycles
Overconfidence and Speculative Bubbles	Scheinkman and Xiong	2003	Medium	Theoretical	Overconfidence creates disagreements over the true value of assets persistent, pervasive overconfidence can create market bubbles

Table 6: Literature Overview part 3

Title	Author	Date	Relevance	Method	Relevant Results
Rogue Trading at Lloyds Bank International, 1974: Operational Risk in Volatile Markets	Schenk	2017	Medium	Empirical	Shows the important links between operational and market risk
Executive overconfidence and the slippery slope to financial misreporting	Schrand and Zechman	2012	Medium	Empirical	Financial misreporting is more likely when executives are overconfident
The Fraud Triangle revisited	Schuchter and Levi	2016	Medium	Theoretical	Discussion of the fraud triangle
The Investor Confidence Game	Stout	2002	Strong	Theoretical	Trusting investors do not consider fundamental values too many trusting investors make fraud easier to commit
Phishing for Phools: The economics of manipulation and deception, Review Essay	Thornton	2016	Medium	Review	Critiques Shiller and Akerlof's Phishing for Phools
Fraudulent Corporate Signals: Conduct as Securities Fraud	Utset	2013	Medium	Theoretical	Securities fraud may be a form of corporate signalling which at its worst can exacerbate market bubbles and is very costly for both investors and the economy.
Corporate Fraud and Business Conditions: Evidence from IPOs	Wang, Winton and Yu	2010	Strong	Empirical	Frauds increase during boom, but decrease during the end of a boom. Fraud increases when investor beliefs are high and vice versa due to monitoring costs
Bursting the dot.com "Bubble': A Case Study in Investor Behaviour	Wheale, and Amin	2010	Medium	Empirical	Studies investor behaviour before and after the dot com bubble
United States Stock Market Confidence Indices	Yale School of Management	2020	Weak	Empirical	Shiller Confidence Indexes
Securities Fraud and Corporate Finance: Recent Developments	Yu	2013	Strong	Lit. Review	Literature review of research in corporate finance concerning financial securities frauds.

Table 7: Literature Overview part 4

Appendix B

Multicollinearity checks (Daoud, 2017; Cosine, 2020)

variable	settled cases	volatility index (VIX)	growth rate of GDP	growth rate of unemployment	lag VIX	growth rate of HPI	lag settled cases	trough	crisis
VIF	4.720636	8.974166	1.42569	1.081054	8.269106	1.334035	2.810503	1.041518	2.153949

Stationarity checks (Smeekes and Wilms, 2020, Python and R)

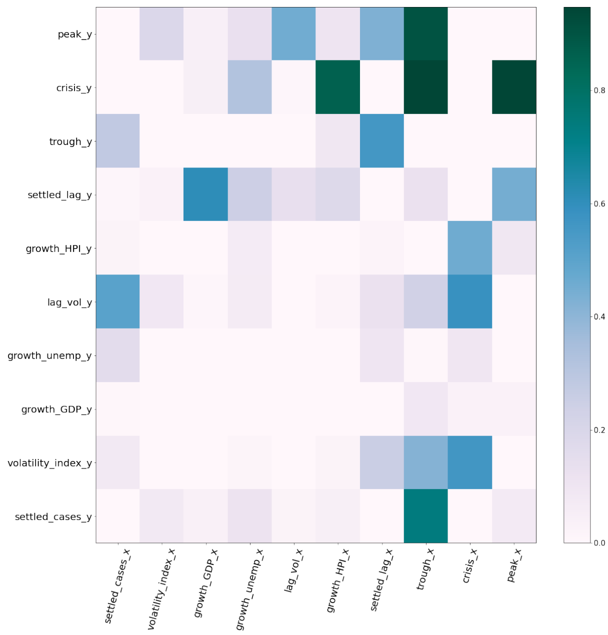
variable	**Augmented Dickey Fuller (ADF) Test with `adfuller`**	**Bootstrap FDR Controlling Tests with `bootUR`**	
	p-value	**test statistic**	**critical value**
settled cases	0.01892351024954988	-1.2044195	-1.1404713
volatility index (VIX)	3.59916801176108e-05	-1.1173858	-1.0768332
growth rate of GDP	0.03648401678627096	-1.2884915	-1.1821230
growth rate of unemployment	0.06289313405223051	-1.6020277	-1.3616217
lag VIX	0.0014171928338200697	-1.0951307	-1.0081979
growth rate of HPI	0.4016470056108172	-0.8804434	-0.8888802
lag settled cases	0.023673610580593814	-1.6020277	-1.3616217

Table 8: Augmented Dickey Fuller (ADF) Tests.

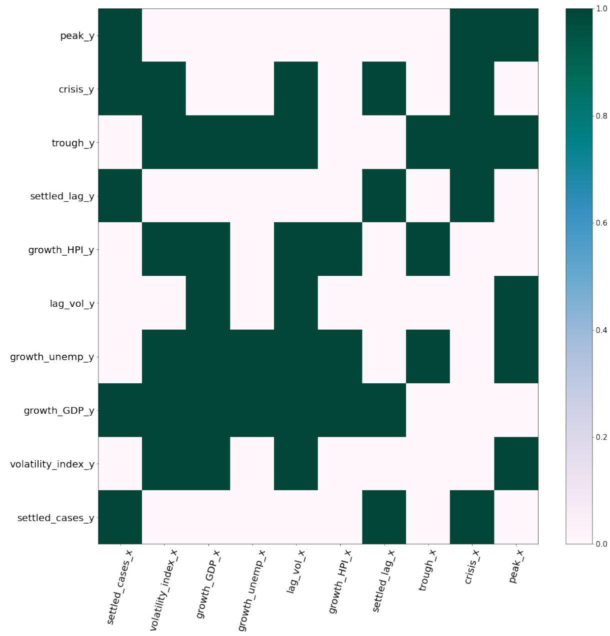
Heteroscedasticity checks (Kuznetsova, 2020)

variable	**settled cases**	
test	**Park test**	**Glejser test**
volatility index (VIX)	4.387791001329138e-16	2.2275100226169182e-103
growth rate GDP	1.6474824294697932e-71	5.133741413256327e-156
growth rate unemployment	NaN	0.0
lag VIX	1.09099304209301e-98	1.4763194251504409e-05
growth rate HPI	1.0366135040086333e-31	1.723540858439388e-144
lag settled cases	NaN	9.986695801279463e-85

Endogeneity checks (Miron, 2021)



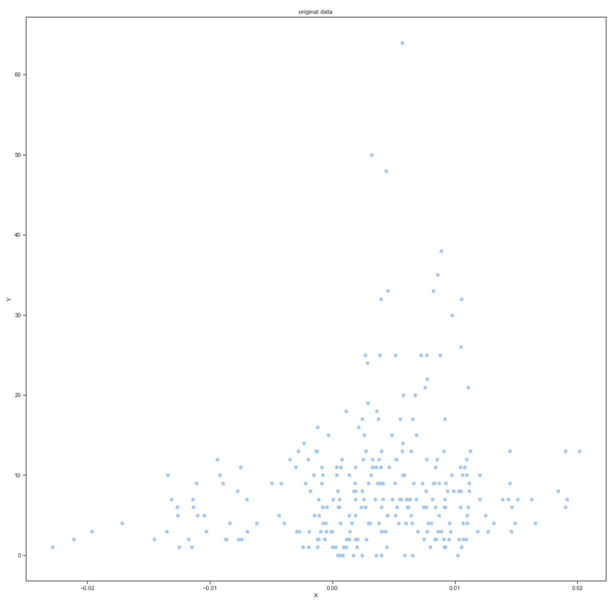
(a) Granger causality matrix



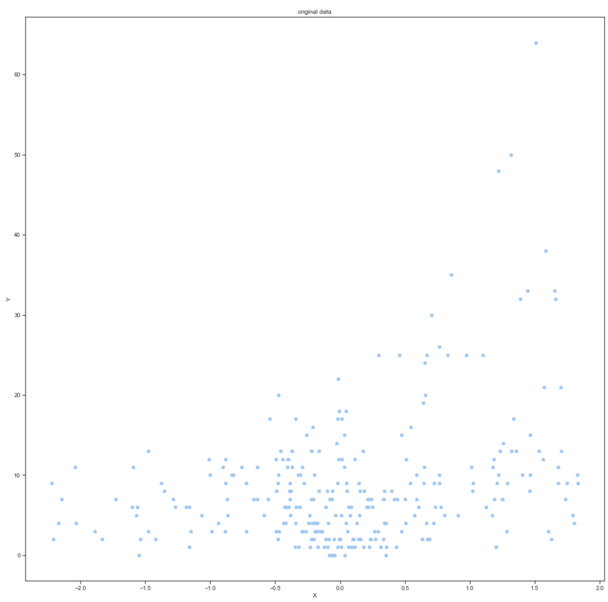
(b) Granger causality matrix, binary terms.

Figure 7: Granger causality tests.

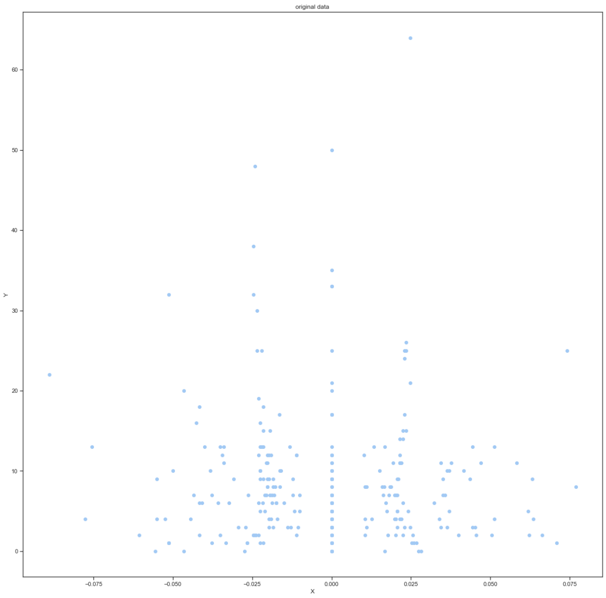
Residual plots



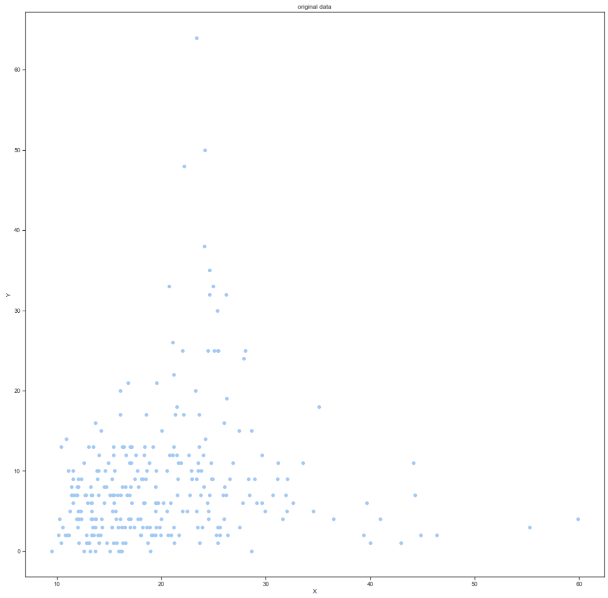
(a) Residuals of growth rate of HPI.



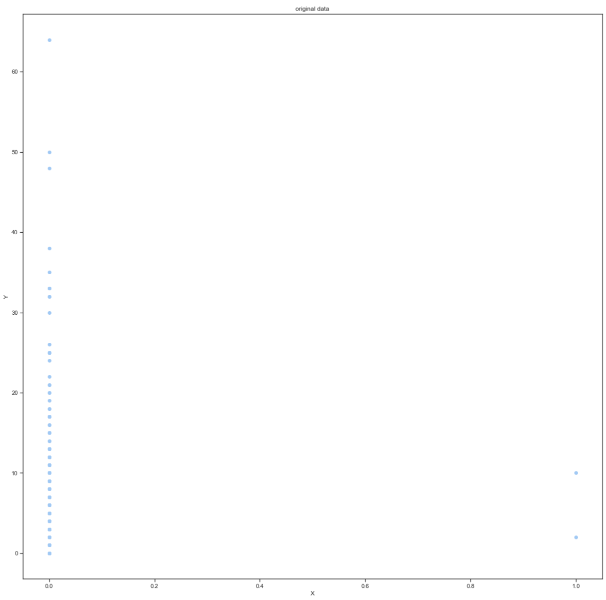
(b) Residuals of growth rate of GDP.



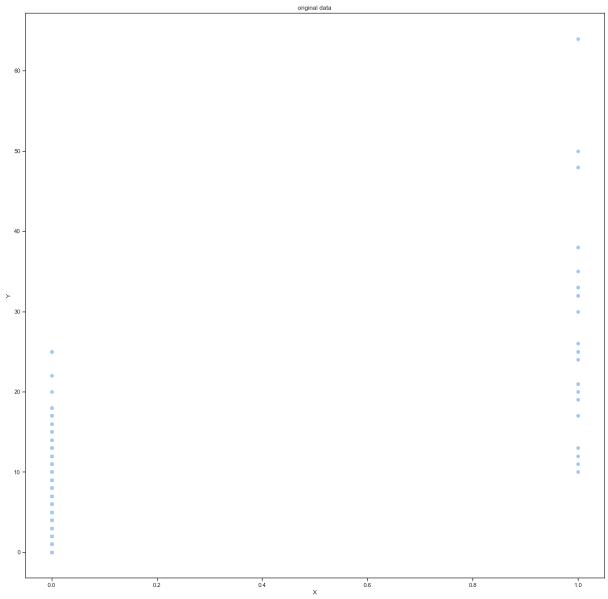
(a) Residuals of growth rate of unemployment.



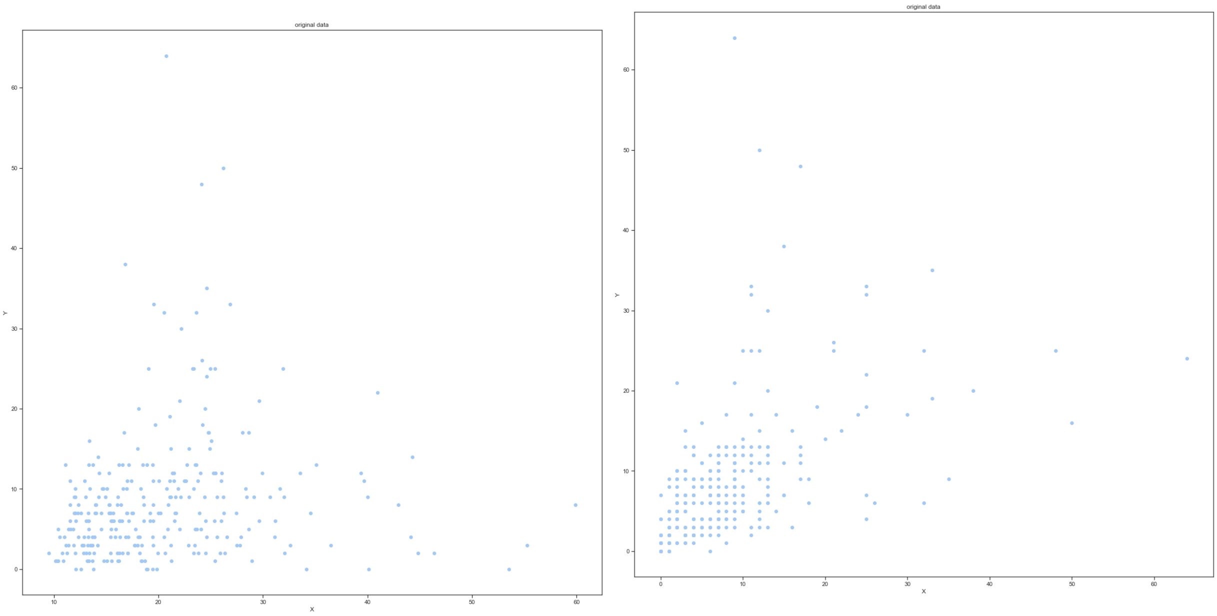
(b) Residuals of volatility index.



(a) Residuals of the dummy variable trough.



(b) Residuals of the dummy variable dot com crisis.



(a) Residuals of the single month lag of VIX.

(b) Residuals of the single month lag of the settled cases.

Figure 11: Residuals of the regression variables.

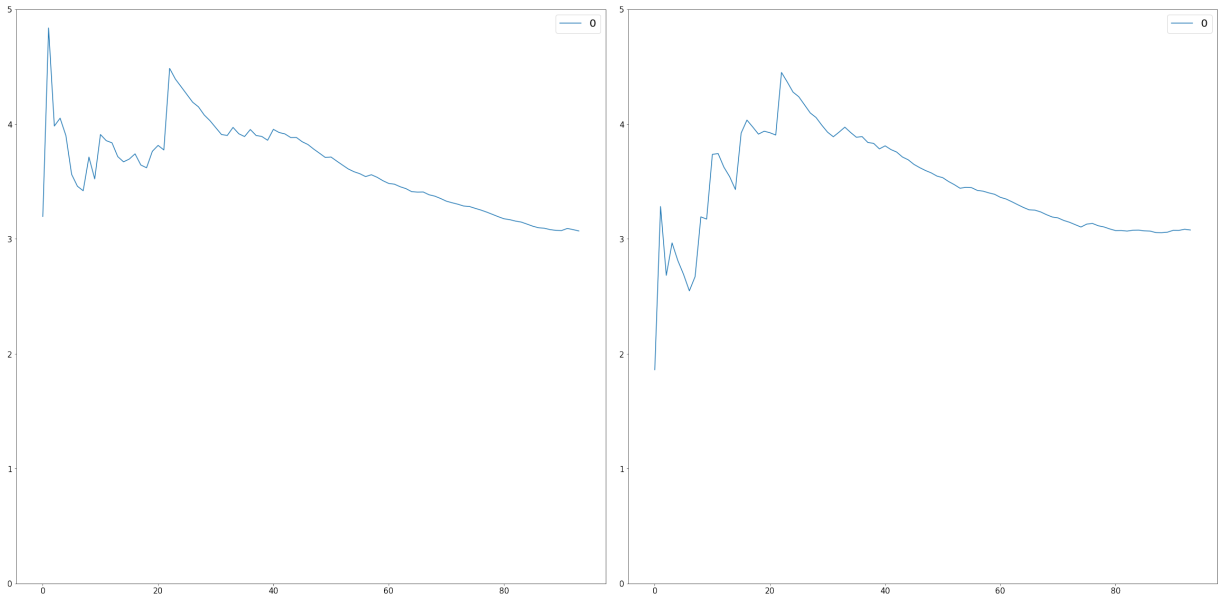
Overfitting tests

Cross validation

Figure 12: Illustration of the test train split used in the robustness analysis (Kasturi, 2020).

Diebold Mariano stat (MSE)	0.8282421952791612		p-value: 0.40822882705396235	
Diebold Mariano stat (MAD)	1.7552917198108071		p-value: 0.08028758140336661	
model	**GAM**		**MARS**	
sample	**train**	**test**	**train**	**test**
RMSE	5.148271476665032	3.1391615104622446	3.8479453125891196	3.0685457665411398

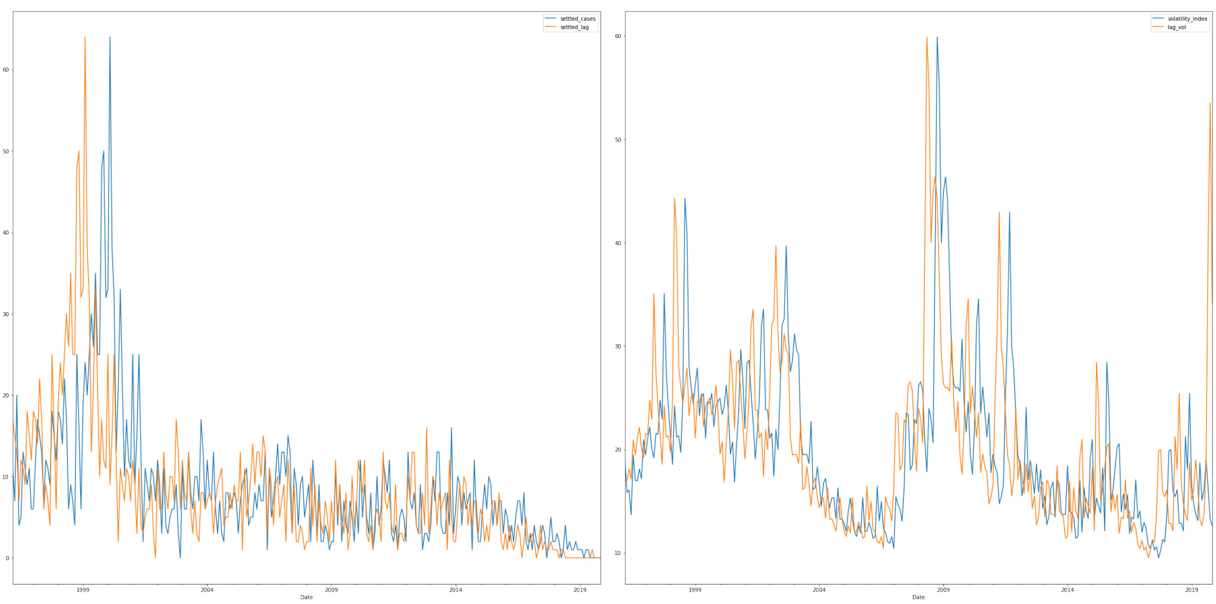
Table 9: Overfitting test results



(a) Model stability test: GAM (b) Model stability test : MARS

Figure 13: Out of sample MSE prediction.

Lags



(a) One month lag of settled cases. (b) One month lag of volatility index.

Figure 14: Lagged variables from the model 4.5

Relationships between variables: overview.

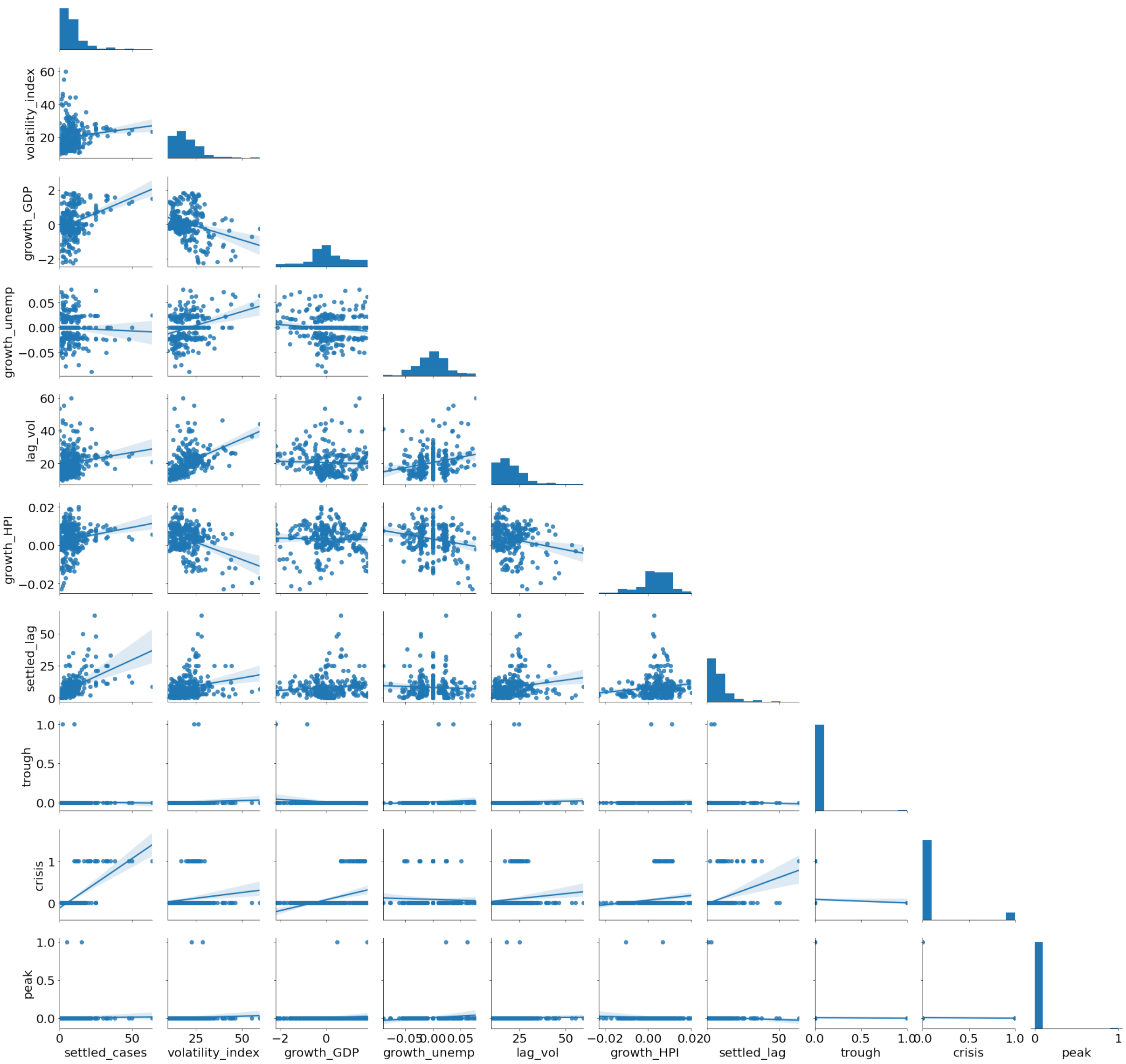


Figure 15: OLS regression plots of relationships between the variables.

Robustness tests

	Dependent variable:settled_cases	
	OLS w HAC errors	GLM
	(1)	(2)
Intercept		0.236***
		(0.022)
const	3.235**	
	(1.370)	
crisis	15.858***	-0.063***
	(4.760)	(0.008)
growth_GDP	1.091**	-0.021***
	(0.483)	(0.006)
growth_HPI	84.362	-3.709***
	(53.803)	(0.768)
growth_unemp	-0.038	0.041
	(12.968)	(0.115)
lag_vol	0.040	-0.002***
	(0.047)	(0.001)
settled_lag	0.264***	-0.003***
	(0.100)	(0.000)
trough	1.081	-0.022
	(2.912)	(0.088)
volatility_index	0.034	-0.000
	(0.062)	(0.001)
Observations	285	285
R^2	0.569	
Adjusted R^2	0.557	
Residual Std. Error	5.518 (df=276)	0.740 (df=276)
F Statistic	9.364*** (df=8; 276)	(df=8; 276)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figure 16: Different model forms.

MARS robustness checks: natural logarithm of the variables.

Basis Function	Pruned	Coefficient
(Intercept)	No	0.123148
crisis	No	0.826805
h(growth_GDP+0.0614208)*h(3.11972-vol_log)	No	1.26425
h(vol_log-2.40695)*h(3.11972-vol_log)	No	-4.12306
growth_GDP	No	0.17657
vol_lag_log*h(growth_GDP+0.078786)*h(-0.0614208-growth_GDP)*h(3.11972-vol_log)	No	-32709.1
h(0.00271741-growth_HPI)	No	-48.2326
vol_lag_log	No	0.677763
vol_lag_log*h(-0.0614208-growth_GDP)*h(vol_log-2.40695)*h(3.11972-vol_log)	No	128.871
h(vol_lag_log-2.62322)*h(2.67759-vol_lag_log)	No	856.811
h(vol_lag_log-2.5611)*h(vol_lag_log-2.48324)*h(3.11972-vol_log)	No	4.17399
vol_lag_log*vol_lag_log*h(-0.0614208-growth_GDP)*h(vol_log-2.40695)*h(3.11972-vol_log)	No	-26.6502
growth_GDP*h(vol_lag_log-2.62322)*h(vol_lag_log-2.48324)*h(3.11972-vol_log)	No	-2.77691
growth_GDP*h(2.67759-vol_lag_log)	No	-2.75147
growth_GDP*h(2.64048-vol_lag_log)*growth_unemp*growth_unemp*h(2.67759-vol_lag_log)	No	30177.5
growth_HPI*h(vol_lag_log-2.48324)*h(3.11972-vol_log)	No	-241.335
h(growth_HPI-0.00911704)*h(vol_lag_log-2.48324)*h(3.11972-vol_log)	No	839.273
growth_unemp*h(-0.0012369-growth_HPI)*h(vol_lag_log-2.48324)*h(3.11972-vol_log)	No	-29287.1
vol_log*h(-0.0614208-growth_GDP)*h(vol_log-2.40695)*h(3.11972-vol_log)	No	-47.0718
h(3.04927-vol_log)*h(vol_lag_log-2.48324)*h(3.11972-vol_log)	No	-6.26857
MSE: 0.2688, GCV: 0.3903, RSQ: 0.5723, GRSQ: 0.3833		

GAM robustness checks: natural logarithm of the variables.

Note that all missing variables were dropped.

Distribution:	NormalDist	Effective DoF:	48.8687	Independent variable:	log settled cases
Link Function:	IdentityLink	Log Likelihood:	-299.6378		
Number of Samples:	275	AIC:	699.0129		
AICc:	721.6493	Scale:	0.2967		
GCV:	0.4324	Pseudo R-Squared:	0.5665		
Feature Function	**Lambda**	**Rank**	**EDoF**	**P > x**	**Sig. Code**
growth rate GDP	[19.428]	35		1.11e-01	
growth rate unemployment	[38.2084]	35		2.01e-02	*
growth rate HPI	[5.6985]	35		8.42e-01	
trough dummy	[0.0212]	35		4.99e-01	
dot com crisis dummy	[16.0739]	35		2.60e-01	
log VIX	[29.1816]	35		4.55e-01	
log of the lag of VIX	[33.3166]	35		4.67e-01	
log of the lag of the settled cases	[3.9032]	35		3.55e-02	*
intercept	1			1.71e-11	***
Significance codes:	***' 0.001 '**' 0.01 * 0.05 '.' 0.1				

MARS robustness checks: no logarithms of the variables.

Basis Function	Pruned	Coefficient
crisis	No	68.904
h(HPI-101.466)*crisis	No	-14.9466
h(101.466-HPI)*crisis	No	-6.94403
h(HPI-183.701)	No	-48.7717
settled_lag*h(101.659-GDP)*h(183.701-HPI)	No	0.00286379
h(108.79-HPI)*h(183.701-HPI)	No	-0.390887
h(HPI-103.701)	No	48.6364
h(92.201-HPI)*h(39.39-volatility_index)*h(GDP-100.146)	No	19.0365
h(108.289-HPI)*h(HPI-92.201)*h(39.39-volatility_index)*h(GDP-100.146)	No	0.073895
settled_lag*h(HPI-101.466)*crisis	No	0.586668
h(134.167-HPI)*h(HPI-108.79)*h(183.701-HPI)	No	-0.00102787
h(95.363-HPI)*h(103.701-HPI)	No	-0.12193
h(GDP-101.531)*h(101.659-GDP)*h(183.701-HPI)	No	-92.0511
h(GDP-101.256)*settled_lag*h(101.659-GDP)*h(183.701-HPI)	No	-0.577263
h(GDP-101.358)*h(GDP-101.256)*settled_lag*h(101.659-GDP)*h(183.701-HPI)	No	4.77332
h(101.358-GDP)*h(GDP-101.256)*settled_lag*h(101.659-GDP)*h(183.701-HPI)	No	19.8947
h(19.47-volatility_index)*h(183.701-HPI)	No	0.0114759
h(HPI-143.963)*h(183.701-HPI)	No	-0.336517
h(143.963-HPI)*h(183.701-HPI)	No	0.343674
HPI*h(183.701-HPI)	No	0.337768
settled_lag*h(GDP-100.146)	No	0.20356
vol_lag*h(HPI-95.363)*h(103.701-HPI)	No	-0.0588881
h(HPI-99.155)*h(HPI-92.201)*h(39.39-volatility_index)*h(GDP-100.146)	No	1.49488e-05
MSE: 9.3218, GCV: 14.7589, RSQ: 0.8638, GRSQ: 0.7858		

GAM robustness checks: no logarithms of the dependent variables.

Distribution:	NormalDist.	Effective DoF:	54.6438	Independent variable:	settled cases
Link Function:	IdentityLink	Log Likelihood:	-1100.3177		
Number of Samples:	285	AIC:	2311.9231		
AICc:	2339.528	Scale:	18.5418		
GCV:	28.0021	Pseudo R-Squared:	0.781		
Feature Function	**Lambda**	**Rank**	**EDoF**	**P > x**	**Sig. Code**
GDP	[14.597]	35	14.4	4.49e-01	
unemployment	[8.3314]	35	13.4	2.17e-01	
HPI	[0.2554]	35	13.9	2.83e-12	***
trough dummy	[33.9204]	35	2.7	9.56e-01	
dot com crisis dummy	[30.8908]	35	3.4	9.63e-01	
VIX	[18.2405]	35	4.1	1.15e-01	
lag of VIX	[16.661]	35	2.1	5.41e-01	
lag of settled cases	[17.28]	35	0.7	4.32e-02	*
intercept	1	0.0		1.79e-02	*
Significance codes:	***' 0.001 '**' 0.01 * 0.05 '.' 0.1				

Securities fraud cases litigations characteristics

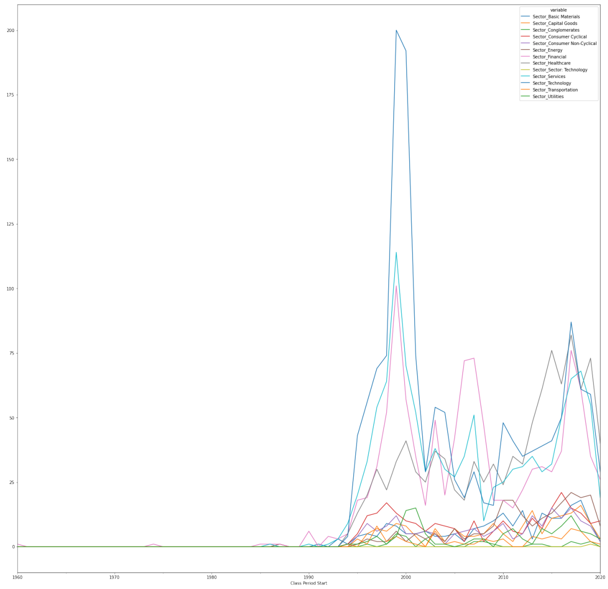


Figure 17: Sectorial focus of the securities fraud claims in absolute terms.

Figure 17 shows what percentage of new security fraud cases enter the market with a particular sectoral focus. It can be seen that technological sector accounts for as much as 25% of the claims, while services, finance and healthcare each amount to ca 15% of the total percentage of the fraud cases reported. A surge in the share of claims made against technology-oriented companies around year 2000 can be explained by the IT-bubble when the market crashed due to excessive speculation of Internet-related companies. This might partly be attributable to incentives for reputation mining perhaps. According to Foodman (2009), the dot com bubble influenced the number of fraud cases a little, while the 2009 mortgage crisis had a larger impact on the number of reported frauds.

Technologies fraud accounts for 25% of the share of claims which is just below the technology market weight which currently stands at 27.67% (Fidelity Investments), however this does not take into account the historical average market weight for the time period 1996-2019 or weight by number of companies. For example, services and finance are over weighted in the number of cases based on modern market weightings. Technologies and finance rank high on number of securities class action cases and also on market volatility of the industry.

The number of US listed companies on stock exchanges has decreased over time (FRED, 2019). We doubt that privately traded companies are sufficient in number to fill that gap, therefore there seem to now be more securities fraud cases per company than there used to be.

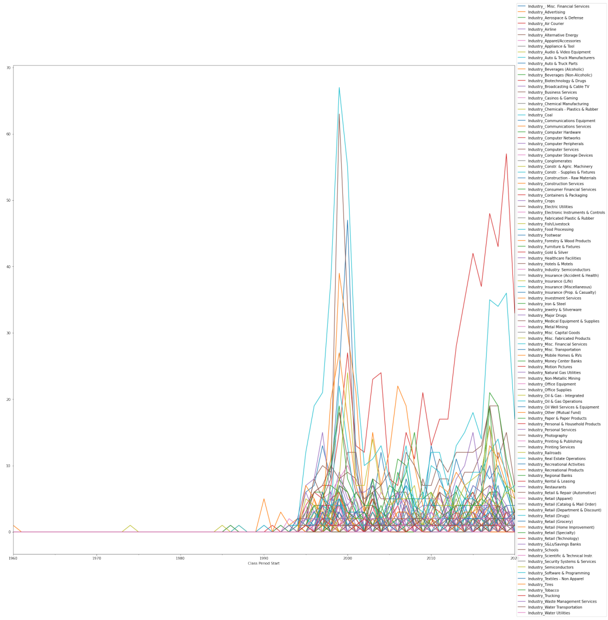


Figure 18: Industrial focus of the securities fraud claims in absolute terms.

Figure 18 shows what percentage of new security fraud cases enter the market with a particular industrial focus. It can be seen that around 10% of the claims occur within Biotechnology & Drugs, making it the most popular category, and around 8% in Software & Programming. Around 5% take place in regard to Medical Equipment & Supplies or Communications Equipment. A spike in certain industries before year 1990 and after that year may suggest trends.

Biotechnology & Drugs and Software & Programming are very research focused and depend heavily on earnings estimates based on the success of future research, which is often difficult to predict. The spikes can indicate a number of different things:

- the size of the industry has increased - therefore more cases for this industry
- the economy has increased and the industry has grown proportionally - therefore more case
- legislation has changed - making cases and fraud more/less likely
- perhaps a famous case in this industry which attracts more new cases (popularity factor)
- a boom in the economy in relation to this industry or frauds in this industry related to the boom.

Hinge	Variables	Time periods	only dot-com	mainly Financial crises	Other
H1	crisis	Jan 1999-Dec 2000	X		
H2	lag settled cases	Jun 1998-Aug 1999	X		
H3	lag settled cases	Jan 1996-Dec 2019			X
H4	crisis growth GDP	Jun-Jul 2000	X		
H5	crisis growth GDP	Feb 1999-Jan 2001	X		
H6	volatility	Aug 1997-Dec 2003 Jan 2008-Dec 2012 Jan-Feb 2016 Dec 2018		X	
H7	volatility	Aug 1997-Dec 2018			X
H8	growth GDP volatility	Jan 1998-Dec 2003 Jun 2007-Jun 2012 Jan-Feb 2016 2019		X	
H9	crisis volatility	Jul 1999-Oct 2000	X		
H10	crisis volatility	Feb 1999-Jan 2001	X		
H11	crisis growth GDP volatility	Jul 2000	X		
H12	crisis growth GDP volatility	Jul 1999 Sep-Oct 2000	X		
H13	crisis lag volatility	Jan 1999-Dec 2000	X		
H14	crisis lag volatility	Jan 1999-Dec 2000	X		
H15	crisis volatility lag volatility	July 1999-Dec 2000	X		
H16	volatility	Mar 1996-Nov 2019			X
H17	lag settled cases volatility	May 1998-Jul 1999	X		
H18	volatility	1996 Jan 2001-Dec 2019			X
H19	crisis volatility unemployment	Jul 1999-Aug 2000	X		
H20	crisis volatility unemployment	Jul 1999-Sept 2000	X		
H21	growth GDP volatility	Jan 1996-Dec 2001 Jan 2004-Dec 2008 Jan 2012-Dec 2019			X

GAM Coefficients		GAM Coefficients		GAM Coefficients		GAM Coefficients	
0	-0.023515	36	0.294038	72	4.254135	108	0.953469
1	-0.234959	37	-0.077281	73	1.868939	109	1.657343
2	-0.514462	38	-0.458834	74	-0.147135	110	2.208667
3	-0.884983	39	-0.860338	75	-1.684546	111	2.465044
4	-0.254358	40	-1.243336	76	-2.738497	112	2.434238
5	0.255115	41	-1.558065	77	-3.290787	113	2.272541
6	1.074597	42	-1.711689	78	-3.289991	114	2.032667
7	0.502974	43	-1.643272	79	-2.579343	115	1.870461
8	-0.478618	44	-1.410831	80	-1.185962	116	1.781049
9	2.823037	45	-1.069186	81	-0.035671	117	1.634060
10	4.403463	46	-0.656092	82	0.462683	118	1.566110
11	2.102902	47	-0.238826	83	0.549111	119	1.819067
12	-0.341148	48	0.095591	84	0.442189	120	2.238134
13	-0.498903	49	0.292031	85	-0.176942	121	2.600859
14	-0.697545	50	0.319919	86	-0.998770	122	2.844866
15	-0.507821	51	0.132535	87	-1.624197	123	2.973650
16	0.453968	52	-0.216834	88	-1.864396	124	2.997575
17	1.509653	53	-0.500108	89	-1.679459	125	2.934091
18	1.183411	54	-0.608843	90	-1.160372	126	2.798155
19	0.086817	55	-0.544075	91	-0.454558	127	2.576786
20	-1.161176	56	-0.364956	92	0.294493	128	2.262336
21	-1.884256	57	-0.221708	93	0.445814	129	1.848952
22	-0.968069	58	-0.150754	94	-0.393677	130	1.306555
23	0.285444	59	0.016816	95	-1.519000	131	0.620410
24	0.945464	60	0.349115	96	-2.506120	132	-0.211993
25	1.362599	61	0.852502	97	-3.301758	133	-1.192795
26	1.653814	62	1.657745	98	-3.550220	134	-2.322937
27	1.797531	63	2.642770	99	-2.705869	135	-3.585977
28	1.768393	64	3.433089	100	-0.841361	136	-4.944900
29	1.541042	65	3.761754	101	1.889477	137	-6.361261
30	1.092370	66	3.622545	102	5.053851	138	-7.798362
31	0.440228	67	3.274893	103	8.259683	139	-9.237058
32	-0.339028	68	2.911423	104	1.459882	140	-3.693713
33	-1.164188	69	2.549647	105	-1.332117	141	-3.195221
34	-1.992878	70	9.308993	106	-0.568605	142	-2.696670
35	0.665544	71	6.776296	107	0.195832	143	-2.197643

GAM	Coefficients	GAM	Coefficients	GAM	Coefficients	GAM	Coefficients
144	-1.705267	179	-3.742344	214	0.629614	249	-5.673915
145	-1.240683	180	-2.759059	215	0.610478	250	-5.208148
146	-0.825972	181	-1.720420	216	0.591344	251	-4.742381
147	-0.483855	182	-0.803187	217	0.572210	252	-4.276614
148	-0.236696	183	-0.030391	218	0.553078	253	-3.810847
149	-0.090296	184	0.621053	219	0.533947	254	-3.345080
150	-0.033998	185	1.097917	220	0.514818	255	-2.879313
151	-0.049805	186	1.346619	221	0.495690	256	-2.413546
152	-0.143559	187	1.432463	222	0.476565	257	-1.947780
153	-0.278941	188	1.349670	223	0.457441	258	-1.482013
154	-0.359806	189	1.025564	224	0.438319	259	-1.016246
155	-0.284360	190	0.544637	225	0.419200	260	-0.550479
156	-0.019752	191	0.034702	226	0.400082	261	-0.084713
157	0.387206	192	-0.379888	227	0.380966	262	0.381054
158	0.926881	193	-0.536407	228	0.361853	263	0.846821
159	1.533121	194	-0.348594	229	0.342741	264	1.312588
160	1.942868	195	0.230070	230	0.323631	265	1.778355
161	2.036830	196	1.176362	231	0.304523	266	2.244122
162	2.044728	197	2.339767	232	0.285417	267	2.709888
163	2.117650	198	3.536455	233	0.266312	268	3.175655
164	2.084672	199	4.582357	234	0.247209	269	3.641422
165	1.777529	200	5.296963	235	0.228107	270	4.107190
166	1.324591	201	5.587073	236	0.209007	271	4.572957
167	0.971765	202	5.501206	237	0.189908	272	5.038724
168	0.866558	203	5.099382	238	0.170809	273	5.504491
169	0.954503	204	4.439051	239	0.151712	274	5.970259
170	1.215673	205	3.577658	240	0.132615	275	6.436026
171	1.666998	206	2.572649	241	0.113518	276	6.901793
172	2.293288	207	1.481472	242	0.094422	277	7.367561
173	3.004813	208	0.359177	243	0.075326	278	7.833328
174	3.723486	209	-0.765511	244	0.056230	279	8.299096
175	-9.706878	210	0.706161	245	-7.536983	280	13.336916
176	-7.990292	211	0.687024	246	-7.071216		
177	-6.290723	212	0.667887	247	-6.605449		
178	-4.821655	213	0.648750	248	-6.139682		