UNIVERSITA' DEGLI STUDI DI PADOVA

**DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI
"M.FANNO"**

DIPARTIMENTO DI SCIENZE STATISTICHE

**CORSO DI LAUREA MAGISTRALE IN
ECONOMICS AND FINANCE**

TESI DI LAUREA

"The relation between news releases, price jumps and assets interconnection"

RELATORE:

CH.MO PROF. Massimiliano Caporin

LAUREANDO/A: Caselli Victoria

MATRICOLA N. 1189999

ANNO ACCADEMICO 2019 – 2020

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere.
Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell'elaborato sono stati indicati nel testo e nella sezione "Riferimenti bibliografici" e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo alla pubblicazione originale.

The candidate declares that the present work is original and has not already been submitted, totally or in part, for the purposes of attaining an academic degree in other Italian or foreign universities. The candidate also declares that all the materials used during the preparation of the thesis have been explicitly indicated in the text and in the section "Bibliographical references" and that any textual citations can be identified through an explicit reference to the original publication.

Firma dello studente

Victoria Corselli

Summary

1. Introduction

The crises we witnessed in the last decades led to various studies investigating what drives financial markets and their behaviour. One particularly studied topic is the occurrence of jumps in asset prices, which have been examined using various mathematical instruments and methods. Understanding jumps behaviour is extremely important, since a large drop in a market could cause, through contagion, similar drops in connected or adjacent markets, and has implications for asset pricing and risk management. Moreover, comprehending how markets react to news announcement could have policy implications, since different announcements could have different impact on separate markets.

Recent studies, both empirical and theoretical, have proved the existence of jumps and their effect on financial decision making, starting from Merton (1976)[1]. Following Black and Scholes methodology to price options, we assume prices to be the result of a stochastic process characterized by continuous trajectories. Merton (1976) introduces the idea that jumps could be a component of the price itself, an idea that could explain the discrepancies observed between option's market prices and the values predicted using the Black-Scholes methodology. This way, the price is the combination of a continuous part and a jump part: the first accounts for normal variation in prices, while the second represents abnormal changes. The last component is the one usually investigated, which is proven to be related to news arrival.

The main difficulty in identifying jumps, once a change is found empirically in asset's return, is to distinguish whether it comes from a proper jump or it is simply a realization of the Brownian process which is supposed to describe the price movement. To address this problem, many techniques have been used, including threshold analysis, which is the most applied in the literature.

The following Thesis aims to further study the probability of incurring in a jump in prices, given a set of information measures and previously observed jumps, for the price series of 88 constituent of the S&P 100, over the period between February 2005 and February 2015. In previous literature, return jumps have been linked to news measures and to jumps in other sectors, separately. In this elaborate, the regressors include both news measures and jumps in the stock considered and in the other of the dataset, to inquire the interconnection between jumps, and also the influence of the jumps occurred in the past. This is a topic that interest mostly financial investors, since spillover effects may be hurtful for international diversification

[1] MERTON, R.C. (1976). *Option pricing when underlying stock returns are discontinuous*. Journal of Financial Economics. Vol. 3 Issue 1-2, pp 125-144.

performed in portfolios. The methodology implemented is the penalised logistic regression, to highlight the most significant news measures in affecting the probability of incurring in a jump.

The elaborate is divided as follows: the first chapter involves literature review, which summarizes the results reported in the current literature, both regarding the connection between jumps and news arrival and jumps interconnection; in the second section an overview on logistic regression is provided, the statistical instrument used to perform the empirical part of the analysis; the subsequent chapter contains the description of the dataset composition and preliminary analysis. In the following chapter, the results of the empirical analysis are reported. At last, conclusions reached through this work are described.

2. Literature review

In this chapter we will revisit the main papers that focus on the topic of price jumps, differing in their methodologies, some of which are computationally more advanced than others. Depending on the techniques utilised, they either find a link between news announcements, or the contrary.

Among the strand of literature that investigates the response of prices to news releases, the study by Ryan and Taffler (2004) analyses the effect firm-specific news have on returns and trading volumes. In their paper, which is centred on 215 constituents of FTSE 100 and FTSE Mid-250 for the period 1994-1995, they compute abnormal return as the difference between the actual return, obtained as the change in the logarithm of prices, and the expected one. The abnormal returns are classified as significant if they exceed the average abnormal return. The same computations are made for the trading volume. The resulting measures are ranked from the highest absolute value to the lowest, then matched to firm-related news contained in three different information sources over a 7 days window. Through this simple matching analysis, the authors report that 65% of the abnormal price movements are due to firm-specific news releases, and the same could be said for a similar proportion of the trading volume movements. Drawbacks of this analysis are the narrow information set, which may not guarantee the caption of all available sources of information, and the methodology of the matching, which has been done manually.

A demonstration that on the discussed topic have been used a variety of statistical and econometric methodologies can be found in Sidorov et al. (2014), which extends the application of GARCH models with jumps, previously introduced, to account also for information flow in the volatility-generating process. Their work proposes a GARCH model where the volatility of the log returns consists in two terms, the first related to unobservable normal information, and the second caused by unexpected news events. The peculiarity of the model is that it adopts jump intensity changing over time, and it is assumed to be linearly dependent on news intensity, computed as the number of company news per day. Sidorov et al. test their model on a sample of daily closing prices of ten UK stocks of FTSE 100, over the period between 2005 and 2008. The data on news releases were restricted to the most relevant. From their estimation, they get coefficients predominantly negative, meaning that jumps are associated with negative movements of prices. The authors prove their model performs better than models not taking into account news intensity, while it still remains a GARCH/ARCH effect in all the companies analysed, which indicates further study can improve the results.

Dumitru et al. (2012) present a summary of renowned nonparametric tests implemented in jumps identification are:

- Barndorff-Nielsen and Shepard (2006) test proposing realised bi-power variation as an estimator for variance, which, by multiplying adjacent returns, has the advantage of reducing the effect of jumps' returns on volatility estimates. The test detects a jump by cross-checking realized volatility with bi-power variation.
- Andersen et al. (2007), along with Lee and Mykland (2008), build tests that compare standardized intraday returns to a threshold. The null hypothesis of both tests is the absence of jumps in the process' observations at a certain time. Once the standardized return is computed with a robust estimate of volatility, Andersen et al. (2007) exploit the fact that the standardized statistics is asymptotically normal and choose a normal threshold. On the contrary, Lee and Mykland state that the upper quantiles of the normal distribution cause an over-rejection of the null, so they set the critical values of the threshold looking at the limit distribution of the maximum of the statistics.
- Aït-Sahalia and Jacod (2008) test is constructed as the ratio between two realized power variation estimates on samples of the observations, with different scales. The computed ratio, in the limit, is proportional to a power of the scales, in the absence of jumps. The statistic based on the ratios follows a standardized normal distribution.
- Jiang and Oomen (2008) test, which uses as indicator the difference between arithmetical and logarithmic returns. In case of no jumps, in the limit the delta between the indicator and the realized volatility is null, while in case of jumps, they are detected in exponential form. The statistic becomes larger in case of great returns, enabling jumps detection.
- Corsi et al. (2010) employing corrected realized threshold bi-power variation, instead of realized volatility. This test is the one used also further on in this elaborate, so it will be explained in more detail in the next chapters.

In the paper presenting their work, Lee and Mykland validate their thesis through a series of Monte Carlo simulations, on which they compute their test: the effectiveness of their methodology is evaluated looking at the probability of success in detecting jumps, and it is compared to two major nonparametric tests already existing. The reported results show that the newly introduced approach outperforms the previously known ones, which also allows to proceed with the analysis and search for variables that are relevant for explaining jump processes. In fact, an empirical analysis trying to do that, is presented in the last section of the paper: Lee and Mykland find that individual jumps of three US stocks are driven by earning announcements and other company-specific news release. Furthermore, jumps in singular stocks are greater with respect to the ones observed in the index containing them.

The tests described above have been implemented in various researches on the topic, with respect to different financial instruments, not only stocks: for example, Kapetanios et al. (2019), focus on option prices. The authors exploit Lee and Mykland's test to find whether the drivers of options' price jumps are exclusively due to changes in the underlying, or if also other factors, like market liquidity, should be considered. This is to confirm that Lee and Mykland's test could, indeed, be applied to several instruments.

A different application of the tests described can be found in Huang (2008), which implements the Barndorff-Nielsen and Shepard (2006) test to divide continuous volatility from jump part, and studies how they respond to macroeconomic news announcements. The paper uses market forecasts in the form of surveys and economic derivative data, in addition to news measures, while it employs high-frequency futures data from both equity and bond markets as market responses. The author detects a significant connection between news announcements and jumps. Also, the other measures of market forecast influence the market, although with different effects on volatility and jumps.

Similar works have been conducted to find relations between news and bond price [Jiang et al. (2011)], futures on stock indexes, on bonds and on exchange rates [Lahaye et al. (2011)].

The strand of literature analysing the reaction of jumps to news announcements is vast. Kanneiainen and Yue (2019), focuses on return jumps of Nasdaq Nordic and its reaction to macroeconomic and firm-specific news. It is found that some macroeconomic news, such as FOMC, have a contribution to jumps dynamics. Firm-specific pre-scheduled announcements have an impact as well. The way jumps react to information differs with respect to the market considered: some type of news could cause a jump on the Copenhagen market, but not on the Stockholm one.

A similar relation results in the study of Evans (2011), which works with intraday jumps and macroeconomic news, and reports that nearly one third of jumps can be related to macroeconomic news, even when they are considered news-related if they occurred close to the news announcement (5 minutes). Furthermore, the author demonstrates that news announcements have significant statistic and economic impact on jumps size, but this effect varies depending on the market considered. A further contribution of the paper is the analysis of the effect of news-related jumps and non-related jumps on return predictability. The first category is empirically larger than the second, on average. However, there is no evidence that jumps could predict returns, while they are generally able to pre-empt volatility persistence.

Advances in technology and IT created the opportunity to study financial markets adopting new instruments: it is, for example, the case of neural networks and machine learning applied to the textual analysis. Through several methodologies different from one another, large texts are analysed by algorithms to capture the sentiment embodied in it, then included in empirical studies as a variable. Differences in approaches might come both from divergences in methods and in sources of text. Given the influence that news and social media have on financial markets nowadays, having the opportunity to better understand this effect could have an impact on portfolio management. The main literature comprises Tetlock (2007), Loughran and McDonald (2011), Garcia (2013), Caporin and Poli (2017).

Tetlock (2007) initially defines pessimism using Harvard Psychosocial Dictionary, while later he updates his study including negative words and weak words most highly linked to pessimism. The author finds a high effect of negative words in the Wall Street Journal on stock returns, which lowers after a negative news is released.

Loughran and McDonald (2011) analyse a sample of 10-Ks from the EDGAR website, checking the existence of the words contained in it and classifying them as negative when they belonged to a list. The authors constructed the list of negative words starting from the Harvard-IV-4 TagNeg (H4N), a list containing 2.004 words considered negative, and expanding it by including also conjunctions that retain the same root of the original word. They also create word lists for positive, uncertain, litigious, strong and weak modal words. The paper reports that many words normally considered negative are not classified as such when used with a financial sense. Loughran and McDonald also confirm the result given by Tetlock (2007): stock returns are lower for companies who experience a higher number of negative words in their financial reports.

Caporin and Poli (2017) construct a database collecting firm-specific and macro-economic information from two news providers, FactSet-Street Account and Thomson Reuters Thompson One, along with Google Trends. The first two mentioned providers are assumed to represent the information that professional investors more likely look to, while Google Trends could be representative of searching behaviour of retail investors, approximated by the times a certain term is searched. Using these news providers, the authors gather news indicators for S&P 100 companies, over the years 2005-2015, divide them for relevance and discard the less important items. The sentiment of news stories is detected referring to the list developed by Loughran and

McDonald (2011). As a further contribution to the literature, news measures are created based on sentiment and later included as variables to forecast volatility. The results of the analysis highlight the role played by EPS announcements and news stories, which appear to be the main drivers in volatility, followed by macroeconomic news and Google Trends. The authors also conclude that including news measures increments the forecasting of volatility.

Regarding financial contagion, it is worth mentioning Audrino and Tetereva (2019), who study the relation between news and returns. While many studies investigate spillovers among markets, the cited one also examines spillover effects among industries, based on news data from Thompson Reuters. The authors are interested in finding whether stock prices are influenced by news related to other sectors or countries, employing a graphical Granger model to visualize the causality between series. They implement Adaptive Lasso to shrink the number of regressors in their model. The empirical analysis emphasizes the importance of news of different sectors, that have effects on other industries. The strength of this influence is changing depending on the period: before times of financial and economic distress, the spillover effect is larger, and have its maximum during crises. Furthermore, results show that the link between news and returns is stronger for US than for European market.

An analysis more connected to the topic of jumps can be found in Asgharian and Bergtsonn (2006). In fact, they build a model focused on event risk on equity indices of several countries, to identify the time of jumps in each index, and study the spillover effects it has on the others. The spillover effect is examined following two perspectives: the first is based on simultaneous jump intensity between two indices, and the second focuses on conditional jumps spillover probability of jumps causing negative returns in other countries' markets. The study of simultaneous intensities points out that jumps' intensities are large for couple of countries in the same region, with similar market capitalizations. This means that markets with these characteristics show a higher degree of jumps spillover. Considering the results of the work carried on on conditional probabilities, the authors find that those are greater than the ones that could be expected under the hypothesis of no international spillover. There also appears to be a lagged effect from US to other countries, probably due to differences in trading hours.

Another work reporting studies on the correlation among equity markets of European countries and US equity market is Asgharian and Nossman (2011). Correlation, variance spillover and jump spillover are investigated. The authors identify three sources of shocks using a stochastic volatility model, allowing jumps both in volatility and returns, granting the possibility of analysing variance spillover and contagion of extreme events. From this analysis it is ground that the major contribution to country variance is coming from other regional countries, secondly from the US.

Jawadi et al. (2015) carry on a study where jump contagion is modelled across international markets, using a nonparametric methodology. The aim of the paper is to investigate the hypothesis of contagion between jumps in different markets, both during overlapping and non-overlapping hours. The authors first compute the realized volatility, that is the sum of intraday squared returns [Andersen and Bollerslev (1998)], then use the bi-power variation, i.e. the scaled sum of the product of adjacent intraday returns in absolute values to identify the jumps component of said volatility. This last component is then tested to highlight only significant jumps, through the Z-statistic employing the realized tri-power quarticity, which is demonstrated to be jump-robust. What characterizes this work is the implementation of threshold autoregressive models (TAR), which are a class of nonparametric models specifying different regimes, whose activation is linked to a threshold. This way, there is a nonlinear relationship in the entire period, but a linear one in the specification of the single regime. The main steps implemented after model specification are to specify the initial value of the threshold, estimate the model by least square method, and specify a new threshold and a new TAR model.

Jawadi et al. (2015) applies the previous procedure to data, provided by Bloomberg, concerning four international markets (S&P 500, FTSE 100, CAC 40, DAX 30), for the period between 2004 and 2009, to test the contagion hypothesis. Since the considered markets have different trading hours, the sample is divided in overlapping and non-overlapping hours. The authors conclude that it exists interdependence between international stock markets. In particular, European markets are shown to be dependent from the US market, and a significant simultaneous jump occurrence is stressed. This linkage varies depending on the market, the regime and the trading hours.

The main article on which this Thesis is based is Caporin and Poli (2018), in which the probability of observing a jump, given a set of news measures is investigated. The authors base their analysis on 88 stocks from the S&P 100, over ten years, starting from 2005. News were gathered from two news providers, then sentiment has been detected and measures have been built following the procedure of a previous study; in addition, jumps are detected using the Corrected Threshold Multipower Variation (C-TMPV). The importance of news variables has been retrieved using Elastic Net and Adaptive Lasso, correcting for the problem of imbalanced sample with 5 machine learning techniques. Although with a preliminary analysis the majority of jumps did not seem to be related to the news considered, FOMC rate decisions, EPS announcements and some topic from the news providers are found to be important determinants of jumps. The regularized approach results in the exclusion of news measures occurring rarely, while the most important news measures in explaining jumps are shown to be FOMC rate decisions and stories released by StreetAccount. The authors also confirm that jumps' size is linked to the sentiment of news stories.

3. Logistic regression

3.1. Introduction to binary logistic regression

Nowadays, it is common to use regression methods to examine and describe the relationship between a response variable and a set of explanatory variables, also called regressors. Regression models have the aim to study which variables determine the response of the variable elected as the dependent one, and to find the most fitting way to describe the outcome. The model most commonly used is the linear regression, usually implemented when the outcome variable is of the continuous type. Their wide popularity is due to the simplicity of their implementation.

Linear regression analysis allows to test whether the considered variables show a linear relationship, described by the model:

$$Y = \alpha + \beta X$$

Where Y is the dependent variable, X is a vector or a matrix of covariates, used to predict Y, α and β are the population parameters to be estimated through the regression. This class of regression can be implemented following a set of assumptions: measurability of the independent variables, homoskedasticity and normality of the errors, no autocorrelation and no correlation between the errors and the independent variables.

The type of regression used in the context of this elaborate is logistic regression. The main difference between the two specifications is the relationship between the dependent and the independent variable: in a linear regression, the key quantity is the conditional mean of the outcome variable, given the covariates ($E(Y|x)$), which is assumed to be a linear equation in x, and can take any value. In a logistic regression with dichotomous variables, which assume only two possible values, the conditional mean has to remain inside the interval between zero and one $[0 \leq E(Y|x) \leq 1]$. In the latter case, the conditional mean can be interpreted as the predicted probability that an observation of the dependent variable will be the higher of the two categories the variable can take (i.e., one or Yes). Furthermore, while the simplest way to estimate a linear regression is computing the ordinary least square (OLS), this method is inefficient when dealing with dichotomous variables, since the assumptions of homoskedasticity, linearity and normality are violated.

A dependent dichotomous variable can be regressed also with a linear model, which will be called linear probability model. However, it can be shown that this model is less efficient: since the variable to be predicted can take only two values, computing linear conditional probability (taking a mean of 0s and 1s) can result in predicted probabilities infinitely large or small, which differ substantially from the observed conditional probability. A graphical interpretation is reported in Figure 1: with the linear probability model, we have a small conditional probability for small values of the regressors, way below 0, and vice versa for great values of X. A nonlinear model could be more appropriate to perform the analysis on a dichotomous variable. For this reason, the Logistic Curve Model, with its "S-shaped" pattern, is preferred.

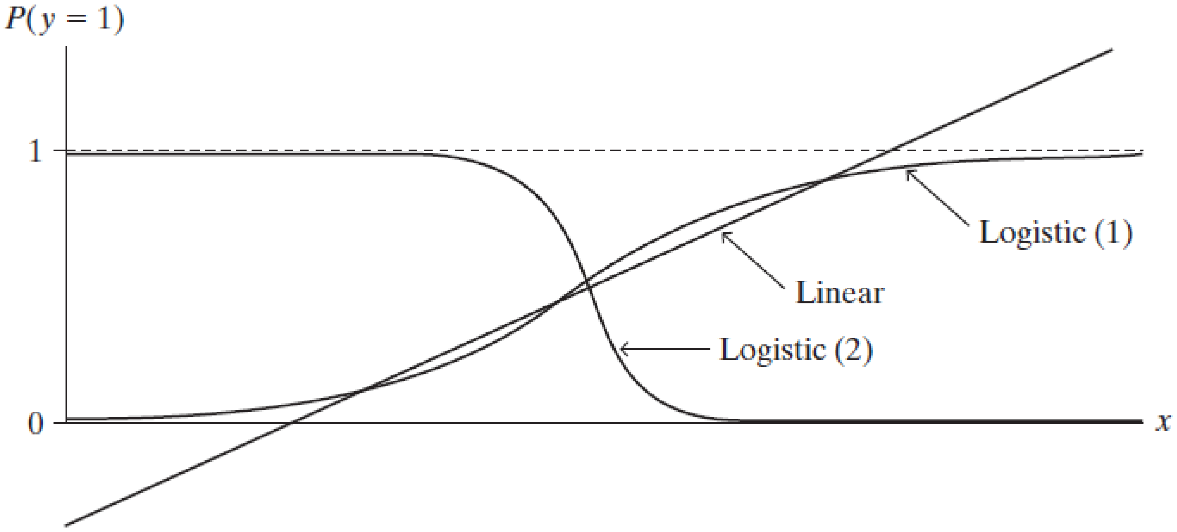

Figure 1 – A representation of the Linear Probability Model and the more appropriate logistic function.

When a nonlinear relationship is observed between variables, transformations are applied to the independent or dependent variable, in order to render the model "nonlinear in terms of its variables, but linear in terms of its coefficients" (Berry and Feldman, 1985, p.53). A nonlinear relation commonly used is the logistic distribution. In this case, the conditional mean of Y given x assumes the form

$$E(Y|x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

Rather than trying to predict the values assumed by the dichotomous variable Y, it can be more useful to predict the probability of a case falling in one category or the other. A possible solution is to replace the probability that $Y = 1$ with the *odds* that $Y = 1$. The *odds* of Y=1 is the ratio between the probability to fall in the first category and the complementary probability to fall in the second one.

$$odds(\mathrm{Y} = 1) = \frac{\mathrm{P}(Y = 1)}{[1 - P(Y = 1)]} = \frac{E(Y|x)}{[1 - E(Y|x)]}$$

The odds has a minimum value of 0, just like $P(Y = 1)$, but no maximum.

A further step is to apply the **logit transformation**, which means computing the natural logarithm of the odds

$$logit(Y) = \ln\left\{\frac{E(Y|x)}{[1 - E(Y|x)]}\right\} = \alpha + \beta x$$

As the odds decreases, the logit becomes negative, while it increases when the argument of the logarithm grows larger, from 1 to infinity. Using the $logit(Y)$ as dependent variable allows to overcome the problem of the estimated probability exceeding maximum and minimum possible values.

The logit transformation may be easily converted back to odds, by applying the exponential to both sides of the equation.

$$odds(Y = 1) = e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}$$

A change of one unit in the regressor X_i has an effect on the odds of e^{β_i}. Is it worth underlying that these three measures (odds, logit and probability) express the same thing, even if the logit form is the most utilized for the analysis of dichotomous variables.

Differently from linear regression, the estimation of α and β is computed through **maximum likelihood techniques**. Maximum likelihood implies the maximization of a function, the likelihood function, which indicates how likely it is to obtain the observed values of Y, given the values of the independent variables and the parameters. The estimate of the regression model is the value of β that maximizes the function. If the observations are assumed independent, in the context of logistic regression, the likelihood function has the form

$$l(\beta) = \prod_{i=1}^{n} [\pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}]$$

Where $\pi(x_i) = E(Y|x)$ is the conditional mean of the response variable given the predictors. Since it is mathematically easier to work with logarithms, it is common practice to retrieve estimates maximizing the *log-likelihood function*, defined as

$$L(\beta) = \ln [l(\beta)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i)\ln [1 - \pi(x_i)]\}$$

To maximize the log-likelihood function we have to differentiate it with respect to the parameters we want to estimate. With the OLS methodology in linear regression, we are able to solve directly for the parameters. Maximizing the log-likelihood function in logistic regression means dealing with nonlinear equations. For this reason, we are forced to start with a tentative solution, change the parameters and repeat the estimation, to see if the likelihood can be improved. This iterative process is repeated until the change from one step to the following is neglectable, and the solution converges. The two methods, OLS and maximum likelihood, give the same results, when the assumptions on which the OLS is based are met.

3.2. Evaluating the logistic regression model

When evaluating logistic regression, just like the analogue case with linear regression, we are interested in the goodness of fit of the model with reality (how well the model minimizes prediction errors) and the significance of all the independent values included in the specification. In addition to these two features, we could be interested in analysing, through Indicies of Predictive Efficiency, the frequency with which the model gives the correct (or incorrect) prediction of the exact value of the dependent variable. Considering this point of view, and the fact that we want to predict variables able to assume typically just two possible values, we should focus more on whether the predicted values are correct or not, than on the closeness of the predicted values with the observed ones.

The criterion used to select parameters in the context of logistic regression is **log likelihood**. Since the log likelihood is a negative quantity, by convenience, statistical software packages compute the log likelihood multiplied by -2 (-2LL). The higher is the value of this statistics, the worse the prediction on the dependent variable is. The statistics (-2LL) is computed as in the following equation:

$$-2LL = -2\{\eta_{Y=1} \ln[P(Y = 1)] + (N - \eta_{Y=1}) \ln[1 - P(Y = 1)]\}$$

Where N is the total number of observations, $\eta_{Y=1}$ is the number of cases where Y=1 and P(Y=1) is the probability that the variable Y takes value 1.

This statistic can be computed both on the model involving just the constant ad on the complete model, including all the regressors. The difference between the two is equivalent to the Sum of Squares statistic in the linear regression.

The log likelihood computed on the complete model has historically been used to test the goodness of fit, meaning the statistical significance of the information not explained by the predicted logistic model, parallel to testing statistical significance of unexplained variance in linear regressions.

A second useful tool is the likelihood ratio, which allows to compare two models that differs for the presence of one or more parameters. The null hypothesis is that the parameters in the complete model are equal to zero. In the simplest case, the one with one parameter, we confront the model $logit(Y) = \alpha + \beta\text{X}$ and the model $logit(Y) = \alpha$. The test is once again based on the likelihood function. The ratio is given by the proportion between the maximum likelihood when the null hypothesis is real (l_0), and the same quantity under the alternative hypothesis (l_1).

$$LR = -2\ln\left(\frac{l_0}{l_1}\right) = -2[\ln(l_0) - \ln(l_1)]$$

The statistic test distributes as a χ^2 with degrees of freedom equal to the number of parameters that the complete model has in addition to the base one.

Evaluating the statistical significance of an independent variable in the prediction of the dependent variable is one of the most important steps in a regression model. Specifically, in a logistic regression, one possible way to accomplish that is to use the likelihood ratio test, as explained before, with and without the variable being tested. The only disadvantage of this procedure is that it is computationally demanding and requires more time with respect to other statistical tests. The simpler alternative is the Wald test, similar to the *t* test in linear regressions: its null hypothesis examines the effect given by the absence of one independent variable on the regression, by putting the relative β equal to zero. In this case, the Wald statistic is computed as the ratio between the estimated coefficient and its standard error. The square of this test is asymptotically distributed as a χ^2. Even the Wald test has drawbacks: for large values of β the standard errors is inflated, which leads to accept the null hypothesis, when it is, indeed, false.

3.3. Interpretation of logistic regression coefficients

Interpreting β in a logistic regression is not simple. We know that a positive value indicates an increase in the probability of observing $Y = 1$, when the regressor to which it is referred increases, while the contrary is true for negative values. Since we are referring to an S-shaped curve, the slope is not constant throughout the curve, which links the P(Y=1) to the values taken by the independent variables in a given point. It is possible to compute the slope of the curve with reference to two distinct points, examining the change in P(Y=1).

Another method for the interpretation of β includes the use of the odds ratio: if we look at the expression of the odds ratio (probability of having Y=1 divided by its complementary), we can extrapolate e^{β_k}, which is the effect of an increase in the independent variable X_k on the ratio between the probabilities. An odds ratio smaller than one indicates that an increase in the independent variable to which we are referring to causes a decrease in the probability, and vice versa.

The odds ratio is not a separate measure of the relationship between dependent and independent variables, but rather a different way to communicate the information logistic regression coefficients supply.

3.4. Multinomial logistic regression

So far, the chapter has served as an introduction to binary logistic regression, its use and how to interpret the outcome. In this section we will expand the model to response variables with more than two outcome categories. In this case, the model is called multinomial logistic regression.

Assume a response variable can take three values, -1, 0 and 1. Since the regression used for binary response variables is parametrized looking at the logit of Y=1 and Y=0, when a three outcome variable is included, two logit functions are needed. One must choose an outcome category as a reference (typically the case in which Y=0), then compare the other two to it. Therefore, indicating with $g_1(x)$ the logit transformation, the model will be outlined by the following equations:

$$g_1(x) = \ln\left[\frac{P(Y=1|x)}{P(Y=0|x)}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{1p}x_p$$

And

$$g_{-1}(x) = \ln\left[\frac{P(Y=-1|x)}{P(Y=0|x)}\right] = \beta_{-10} + \beta_{-11}x_1 + \beta_{-12}x_2 + \cdots + \beta_{-1p}x_p$$

The likelihood function of this model is

$$l(\beta) = \prod_{i=1}^{n}[\pi_0(x_i)^{y_{0i}}\,\pi_1(x_i)^{y_{1i}}\pi_{-1}(x_i)^{y_{-1i}}]$$

Where $\pi_1(x_i) = E(Y=1|x)$ is the conditional probability of Y being 1, given the predictors, and the same for the categories 0 and (-1). Instead, the log-likelihood equation takes the form

$$L(\beta) = \ln\left[l(\beta)\right] = \sum_{i=1}^{n}\{y_{1i}g_1(x_i) + y_{-1i}g_{-1}(x_i) - \ln\left(1 + e^{g_1(x_i)} + e^{g_{-1}(x_i)}\right)\}$$

The likelihood equations are found by partially differentiating $L(\beta)$ with respect to each unknown parameter, then finding the values of the parameters that set the equations to zero. Those values will become the estimate of the regressions.

3.5. Penalised logistic regression:

When the predictors included in an analysis are large numbers, a powerful tool is represented by penalised logistic regression. This model allows the user to identify the most important predictors, that better explain the response variable.

The most implemented specification used in practice is the Elastic Net, which is a composition of two terms, the Lasso penalisation and the Ridge penalisation. The first includes a penalty term constraining the size of estimated coefficient, so it allows to set some parameters to zero, while the second allows to select even high correlated regressors or improve the performance. Ridge performs best in scenarios without a large number of noisy predictors. Lasso works poorly when correlated regressors are present. On the contrary, Elastic Net is proved to work well in every scenario.

The Elastic Net technique, applied to the logistic regression, minimizes the following objective function:

$$min_{\beta_0,\beta}[\frac{1}{T}\sum_{t=1}^{T} y_t(\beta_0 + \beta' x_t) - \log\left(1 + e^{\beta_0 + \beta' x_t}\right)] + \lambda[\frac{(1-\alpha)}{2}\left||\beta|\right|_2^2 + \alpha\left|\left|\beta\right|\right|_1]$$

Where β_0 is the intercept, β is the vector of estimated coefficients for the regressors, $\lambda \geq 0$ is a complexity parameter and α is in between Ridge and Lasso, that have respectively $(\alpha = 0)$ and $(\alpha = 1)$, so it will assume values comprised between zero and one. The λ selected will be the one that, among the different values used in the estimation, will guarantee the maximum area under the ROC curve. The lower α is, the better the algorithm should perform.

A second alternative to implement a penalised regression is using the Adaptive Lasso, as was done in Caporin and Poli (2018). Adaptive Lasso is a two-stage procedure able to decrement the number of false positive predictions in a regression, by minimizing an objective function similar to the one of Elastic Net:

$$min_{\beta_0,\beta}[\frac{1}{T}\sum_{t=1}^{T} y_t(\beta_0 + \beta' x_t) - \log{(1 + e^{\beta_0+\beta' x_t})}] + \lambda \sum_{i=1}^{p} \left| \frac{\beta_i}{\hat{\beta}_{i,initial}} \right|$$

In the previous equation $\hat{\beta}_{i,initial}$ is the coefficient estimated in the first-stage in the equation of Elastic Net, while also λ and α assume the same values of the method indicated before. The implementation of this second technique should grant a lower number of false positives.

An arising issue is the one of class imbalance, which occurs whenever one class is outnumbered by another: as Caporin and Poli (2018) explain, this problem is born from the rarity of jumps, which worsens the performance of standard classification systems. To deal with the problem, the authors apply four machine learning techniques, namely cost-sensitive learning, oversampling, under-sampling and synthetic sampling. In under-sampling, observations from the major and minor classes are randomly picked in order to have a balanced sample in which both classes have equal distributions. In oversampling, the minority class is replicated, allowing to achieve more equal distributions. Both these methodologies, as a result of picking observations randomly, could discard important information in one of the two classes. Synthetic sampling permits to create a new minority sample, as a result of an interpolation between two minority class samples chosen randomly. Cost-sensitive learning introduces costs into the classifier, which has a higher value when referring to the minor class. This leads the classifier to weight more the imbalanced class.

Adaptive Lasso applied to the methodologies described is estimated using cross-validation with blocks of contiguous time, to achieve more reliable estimations. Caporin and Poli (2018) also remove first observations for each test, thus reducing the dependence between training sets and the test sets.

In the context of this elaborate, the problem of class imbalance will not be addressed.

4. Dataset

The dataset used in this elaborate comprises data going from February 4, 2005 to February 25, 2015 about 5-minute stock prices of 88 of the constituents of S&P 100, EPS and firm-specific news stories for each one, and 23 macro-economic announcements. The observation period covers also the years of the 2008 financial crises.

The news data were retrieved from StreetAccount from Factset and Thompson One form Thompson Reuters, while prices were taken by Kibot.com, a less known but still reliable provider, as explained in Caporin and Poli (2018). These two news providers have been chosen specifically because they are professional providers, and they are supposed to publish only relevant firm-specific news, disentangling them from irrelevant news that could create noise.

The list of stocks is reported in Table 1. From the dataset where excluded stocks that either didn't have news stories available on the selected providers, or where included in the S&P 100 in the middle of the observation period.

Table 1 – List of stocks, ticker and sector to which they belong

Ticker	Name	Sector
AAPL	Apple	Consumer goods
ABT	Abbot laboratories	Healthcare
ACN	Accenture Plc	Technology
AEP	American Electric Power Co., Inc	Utilities
AIG	American International Group, Inc	Financial
ALL	The Allstate Corporation	Financial
AMGN	Amgen Inc	Healthcare
AMZN	Amazon.com, Inc	Service
APA	Apache Corp	Basic Materials
APC	Anadarko Petroleum Corporation	Basic Materials
AXP	American Express Company	Financial
BA	The Boeing Company	Industrial Goods
BAX	Baxter International Inc	Healthcare
BHI	Baker Hughes Incorporated	Basic Materials
BIIB	Biogen Inc	Healthcare
BK	The Bank of New York Mellon Corporation	Financial

BMY	Bristol Myers Squibb Company	Healthcare
BRK.B	Berkshire Hathaway Inc	Financial
C	Citigroup Inc	Financial
CAT	Caterpillar Inc	Industrial Goods
CELG	Celgene Corporation	Healthcare
CL	Colgate-Palmolive Co	Consumer Goods
CMCSA	Comcast Corporation	Service
COF	Capital One Financial Corporation	Financial
COP	ConocoPhillips	Basic Materials
COST	Costco Wholesale Corporation	Service
CSCO	Cisco Systems, Inc	Technology
CVS	CVS Health Corporation	Healthcare
CVX	Chevron Corporation	Basic Materials
DD	E. I. du Pont de Nemours and Company	Basic Materials
DIS	The Walt Disney Company	Services
DOW	The Dow Chemicals Company	Basic Materials
EBAY	eBay Inc	Services
EMC	EMC Corporation	Technology
EMR	Emerson Electric Co	Industrial Goods
EXC	Exelon Corporation	Utilities
FCX	Freeport-McMoRan Inc	Basic Materials
FDX	FedEx Corporation	Services
GD	General Dynamics Corporation	Industrial Goods
GE	General Electric Company	Industrial Goods
GILD	Gilead Science Inc	Healthcare
GS	The Goldman Sachs Group, Inc	Financial
HAL	Halliburton Company	Basic Materials
HD	The Home Depot, Inc	Services
HON	Honeywell International Inc	Industrial Goods
HPQ	HP Inc	Technology
IBM	International Business Machine Corporation	Technology
INTC	Intel Corporation	Technology
JNJ	Johnson & Johnson	Healthcare
JPM	JPMorgan Chase & Co	Financial

KO	The Coca-Cola Company	Consumer Goods
LLY	Eli Lilly and Company	Healthcare
LMT	Lockheed Martin Corporation	Industrial Goods
LOW	Lowe’s Company, Inc	Services
MCD	McDonald’s Corp	Services
MDT	Medtronic plc	Healthcare
MET	MetLife, Inc	Financial
MMM	3M Company	Industrial Goods
MO	Altria Group, Inc	Consumer Goods
MON	Monsanto Company	Basic Material
MRK	Merck & Co, Inc	Healthcare
MSFT	Microsoft Corporation	Technology
NKE	NIKE, Inc	Consumer Goods
NSC	Norfolk Southern Corporation	Services
ORCL	Oracle Corporation	Technology
OXY	Occidental Petroleum Corporation	Basic Materials
PEP	Pepsico, inc	Consumer Goods
PFE	Pfizer Inc	Healthcare
PG	The Procter & Gamble Company	Consumer Goods
QCOM	QUALCOMM Incorporated	Technology
RTN	Raytheon Company	Industrial Goods
SBUX	Starbucks Corporation	Services
SLB	Schlumberger Limited	Basic Materials
SO	Southern Company	Utilities
SPG	Simon Property Group In	Financial
T	AT&T, Inc	Technology
TGT	Target Corp	Services
TXN	Texas Instruments Inc	Technology
UNH	UnitedHealth Group Incorporated	Healthcare
UNP	Union Pacific Corporation	Services
UPS	United Parcel Service	Services
USB	U.S. Bancorp	Financial
UTX	United Technologies Corporation	Industrial Goods
WBA	Walgreen Boots Alliance, Inc	Services

WFC	Wells Fargo & Company	Financial
WMB	Williams Companies, Inc	Basic Materials
WMT	Wal-Mart Stores Inc	Services
XOM	Exxon Mobil Corporation	Basic Materials

In the dataset, jumps for each stock are included, computed from 5-minute returns. The returns were obtained as $100\log\,(\frac{p_j}{p_{j-1}})$, with p_j being the price at the end of the 5-minute interval. To identify the exact timing of the jump, the methodology described by Andersen et al. (2007b) was applied, but adopting the corrected threshold bi-power variation from Corsi et al (2010).

The approach developed by Corsi et al. (2010) to identify jumps starts form the assumption that stock prices follow a Brownian process with a jump term.

$$dX_t = \mu_t dt + \sigma_t dW_t + dJ_t$$

Where X_t is the price process at time *t*, μ_t is a drift term, $\sigma_t dW_t$ is the continuous volatility component and dJ_t is the jump part.

Fixing a time T, the equation above can be rewritten in discrete time. In particular, the quadratic variation of the previous process takes the form of

$$[X]_t^{t+T} := X_{t+T}^2 - X_t^2 - 2\int_t^{t+T} X_{s-} dX_s$$

With *t* being the day. This form allows to distinguish continuous and discontinuous part of the process.

$$[X]_t^{t+T} = [X^c]_t^{t+T} + [X^d]_t^{t+T}$$

The continuous part is given by $[X^c]_t^{t+T} = \int_t^{t+T} \sigma_s^2 ds$, while the discontinuous one is $[X^d]_t^{t+T} = \sum_{j=N}^{N_{t+T}} c_j^2$. The most common estimator of the process $[X]_t^{t+T}$ is realized variance, namely the sum of squared realized returns $(\Delta_j X)$. If the intervals on which returns are computed are small and approach zero, realized volatility converges in probability to $[X]_t^{t+T}$.

$$RV_\Delta(X)_t = \sum_{j=1}^{N} (\Delta_j X)^2$$

An estimator of the continuous part of realized volatility is bi-power variation, since it converges in probability to $\int_t^{t+T} \sigma_s^2 ds$.

$$BPV_\delta(X)_t = {\mu_1}^{-2} \sum_{j=M}^{[T/\delta]} |\Delta_{j-1}X| \cdot |\Delta_j X|$$

From the equation above, it can be seen that bi-power variation is the sum of the product of two adjacent returns, multiplied by the square root of μ_1, which is approximately equal to 0,7979. This estimator has a drawback: bi-power variation asymptotically converges to the integrated continuous volatility. When δ is finite, a jump occurring in $|\Delta_j X|$ will not vanish, instead increasing, the greater the observed return is.

Corsi et al. (2010) introduce a new estimator, **threshold multipower variation**, which is defined as

$$TMPV_\delta(X)_t^{[\gamma_1,\dots,\gamma_m]} = \delta^{1-\frac{1}{2}(\gamma_1+\dots+\gamma_m)} \sum_{j=M}^{[T/\delta]} \prod_{k=1}^{M} |\Delta_{j-k+1}X|^{\gamma k} I_{\{|\Delta_{j-k+1}X|^2 \leq \vartheta_{j-k+1}\}}$$

ϑ is a strictly positive threshold function. Threshold multipower variation avoids the problem presented by bi-power variation, since the indicator function assumes value zero and vanishes if the jump in return is higher than the threshold ϑ.

To build the dataset on which the Thesis is based, Caporin and Poli (2018) apply a corrected version of threshold multipower variation, defined as:

$$C - TMPV_\Delta(X)_t^{[\gamma_1,\dots,\gamma_m]} = \delta^{1-\frac{1}{2}(\gamma_1+\dots+\gamma_m)} \sum_{j=M}^{[1/\delta]} \prod_{k=1}^{M} Z_{\gamma k}(|\Delta_{j-k+1}X|, \vartheta_{j-k+1})$$

$Z_\gamma(x, y)$ is a function taking the form

$$Z_\gamma(x,y) = \begin{cases} |x|^\gamma & if\ x^2 \leq y \\ \dfrac{1}{2N(-c_\vartheta)\sqrt{\pi}} \left(\dfrac{2}{c_\vartheta^2} y\right)^{\frac{\gamma}{2}} \Gamma\left(\dfrac{\gamma+1}{2}, \dfrac{c_\vartheta^2}{2}\right) & if\ x^2 > y \end{cases}$$

$N(x)$ is the standard normal cumulative function, $\Gamma(\cdot)$ is the upper incomplete gamma function, ϑ is set equal to $c_\vartheta^2 \sigma^2$, σ^2 being the variance of the return on interval j, assuming it is distributed as a $N(0, \sigma^2)$. As in the study of Corsi et. Al (2010), c_ϑ was set equal to 3.

Just like threshold multipower variation, the corrected version converges in probability to the continuous part of a price process. This feature allows to compute the discontinuous part as the difference between the realized volatility and the corrected threshold multipower variation.

Jumps are detected implementing the Barndorff-Nielsen and Shepard (2006) test, where the estimator illustrated replaces the ones based on multipower variation.

For what concerns news stories and EPS, each of them reports the precise time of announcement. Both StreetAccount and Thompson One filter announcements based on topic and relevance, while the latter assigns also an importance level. As in Caporin and Poli (2018) the news included in this study are categorized as:

- Unscheduled News Stories, comprising seven topics from each of the providers. The topics' list is reported in Table 2.
- Prescheduled Earnings Announcements. Retrieved from StreetAccounts, this category reports the EPS announced by the companies along with the public consensus at the time of announcement.
- Prescheduled Macroeconomic Announcements. This category groups 23 macroeconomic announcements released in trading hours, which were recovered from Thompson Reuters. Like the previous category, figures and consensus forecast are included. The list of macroeconomic topics can be found in Table 3, along with their usual announcement time.

Table 2 – List of topics from both news providers

Thompson Reuters	**StreetAccount**
All	All
Earnings pre-announcements	Earnings related
Dividends	M&A
Financial	Litigations
Medium	Regulatory
High	Newspapers
Top	Up/downgrades

For both providers, the "*all*" category covers all the firm-specific news collected. The news with the lowest relevance were disregarded, since they coincided with "*all*".

As in Caporin and Poli (2007), a sentiment indicator has been attached to each news story, basing the procedure on the one introduced by Loughran and McDonald (2011). The method used extracts the sentiment, in the form of a variable assuming values -1, 0, or 1, both from the headlines and the body.

Table 3 – Macroeconomic news and announcement time.

Announcement	**Release Time**
Business inventories	10.00
Chicago PMI	9.45/10.00
Construction Spending	10.00
Consumer Confidence	10.00
Consumer Credit	15.00
Michigan Consumer Sentiment Index	9.45/9.55/10.00
EIA Crude Oil Stocks	10.30
ECRI Weekly	10.30
IBD Economic Optimism	10.00
Employment Trends Index	10.00
Existing Home Sales	10.00
Factory Orders	10.00
Federal Budget	14.00
FOMC Rate Decision	12.30/14.00/14.15
NAHB Housing Market	10.00/13.00
Leading Index	10.00
ISM Manufacturing Index	10.00
EIA Natural Gas Stocks	10.30
New Home Sales	10.00
New York NAPM Index	9.45
Pending home Sales	10.00
Philadelphia Fed business Index	10.00/12.00
Wholesale Inventories	10.00

The dataset contains news indicators built at time *t*, based on news released in the 30 minutes before *t*, in the same interval, and in the 10 minutes after *t*, to gauge the reaction the release causes. To assign a value to a text, the methodology of Caporin and Poli (2017) was implemented. This procedure focuses on concepts that could cause different feedbacks on the market. All of them relate to a reference period and to previous ones. The introduction of the concepts was intended as a mean to identify the portion of news on which investors based their decisions. The concepts included on the building of the dataset are:

- Standard measures, like the presence of news or sentiment.
- Abnormal quantity, consisting in quantity of news being above a threshold. Unseen amounts of information could cause investors to react, which, in turns, could lead to jumps.
- Uncertainty, identifying the presence of news whose sentiment is opposite with respect to the one expected.
- News persistence. Every time the quantity of news exceeds the threshold for two consecutive periods, the news presents this characteristic. It is important to take into account news persistence, since professional providers do not report redundant news. This means that the presence of a higher quantity of news is linked to something else.

Caporin and Poli (2017) also include quantity variation, sentiment inversion, quantity variation conditional on sentiment, sentiment conditional on quantity and a news burst index.

Two ulterior measures are built on news announcements: Standardized Unexpected Earnings (*SUE*) and a standardized indicator of surprise for macro announcements (*Std_Macro*).

$$SUE_t = \frac{EPS_t^{actual} - EPS_t^{forecast}}{\hat{\sigma}_{surp,EPS}}$$

In the equation, $\hat{\sigma}_{surp,EPS}$ is the standard deviation of the quantity at the numerator. SUE_t is a score measuring the number of standard deviations between the actual earnings per share and the ones forecasted by the market.

$$Std\_macro_t = \frac{Macro_{k,t}^{actual} - Macro_{k,t}^{forecast}}{\hat{\sigma}_{surp,Macro}}$$

Even in this case, $\hat{\sigma}_{surp,Macro}$ is the computed standard deviation of the quantity at the numerator. The standard surprise is calculated for each of the macroeconomic variables reported in Table 3, except for ECRI Weekly, Employment Trends and New York NAPM Index, for which market consensus was not available. In those cases, the dataset includes the standardized change from previously released values.

The dataset contains 624 news measures, 160 based on news coming from Thompson Reuters, 188 use news from StreetAccount (12 of which are Standardize Unexpected Earnings indicators) and the remaining 276 created on macroeconomic news. To the previous news measures were also added the time series of jumps, in order to see if jumps observed in other assets have a significant effect on the probability of observing jumps in the future, and also look for a dynamic effect that past jumps of the variable in analysis may have. The first could be considered like a sort of spillover effects among stocks. Including stocks' time series increases the number of regressors, taking them up to 712 variables.

4.1. Preliminary analysis.

For each stock, Figure 2 reports the number of jumps registered, following the method previously described. It is interesting to notice that the highest number of jumps is shown by financial corporations and companies working in the technological industry, or closely related to it. The first is probably related to the effect that 2008 crisis had on the financial industry, which was among the most affected by it. It is worth noticing that none of the assets reported has a number of jumps smaller than twenty.

In Figure 3, the total number of jumps are represented. The number of jumps was obtained by aggregating the number of jumps per day in every single stock (the same has been done in Caporin and Poli (2018)). It can be noticed that the number of jumps spikes in the period of the 2008 crisis, and also in the years 2012-2013, during which the sovereign debt crisis raged in Europe.

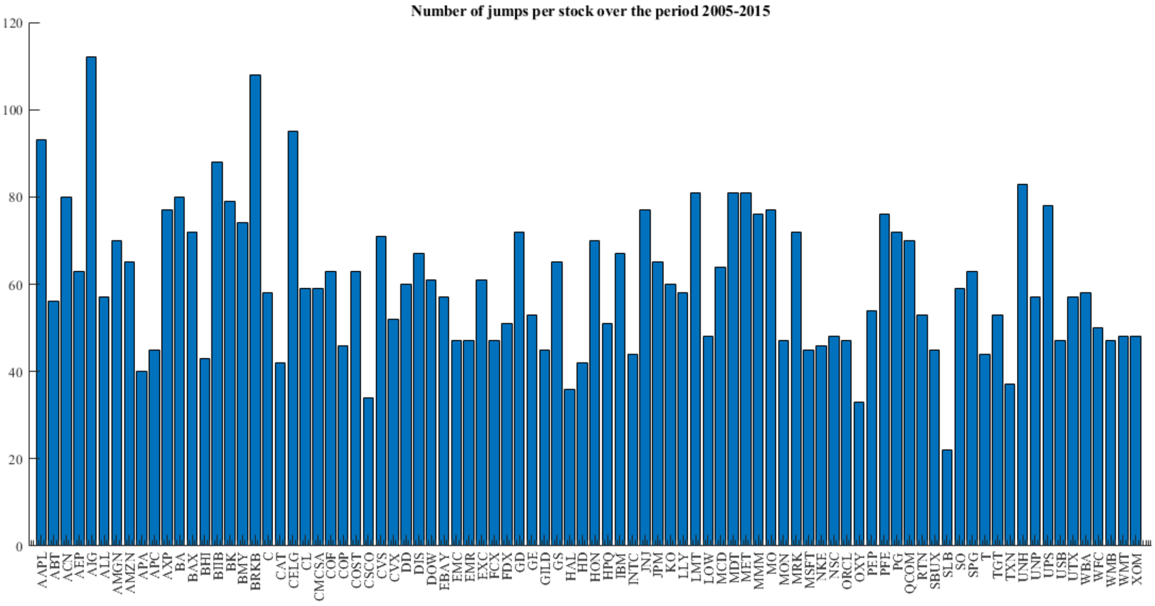

*Figure 2 – Number of jumps per stock over the observed period. **Source**: own elaboration.*

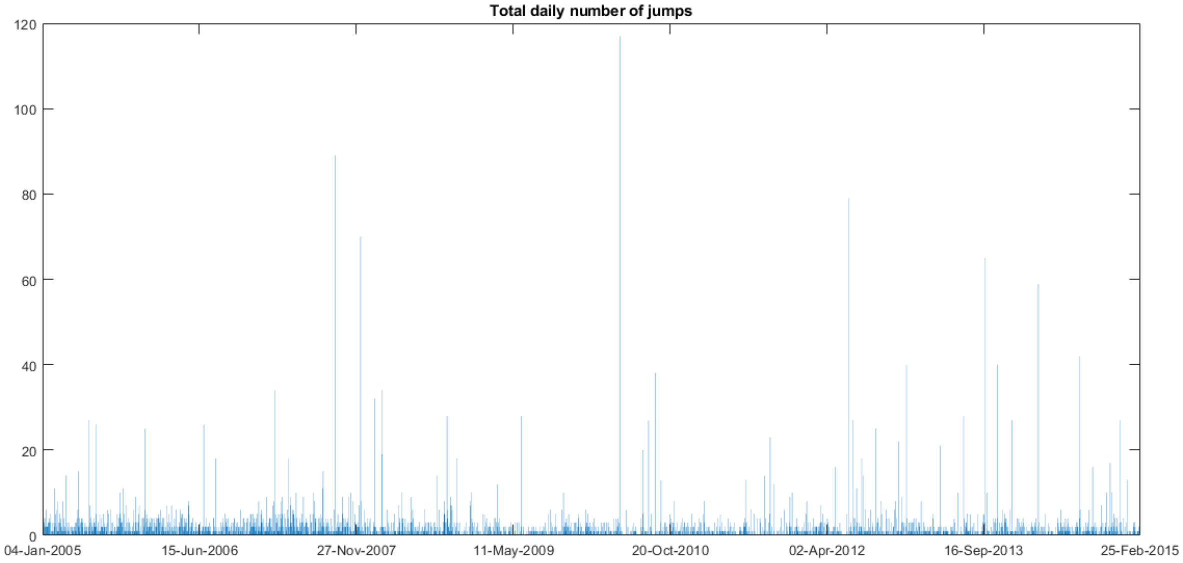

Figure 3 – Number of daily jumps, summed over all stocks. ***Source****: own elaboration.*

In addition to the time series of the total number of jumps, Caporin and Poli (2018) compute the median of absolute jumps' size, and the number of news stories from Thompson One and StreetAccount, which are reported in Figure 4. As a further confirmation of the impact the financial crisis had on returns and, consequently, jumps, their absolute size shows higher values from the end of 2008 and throughout 2009.

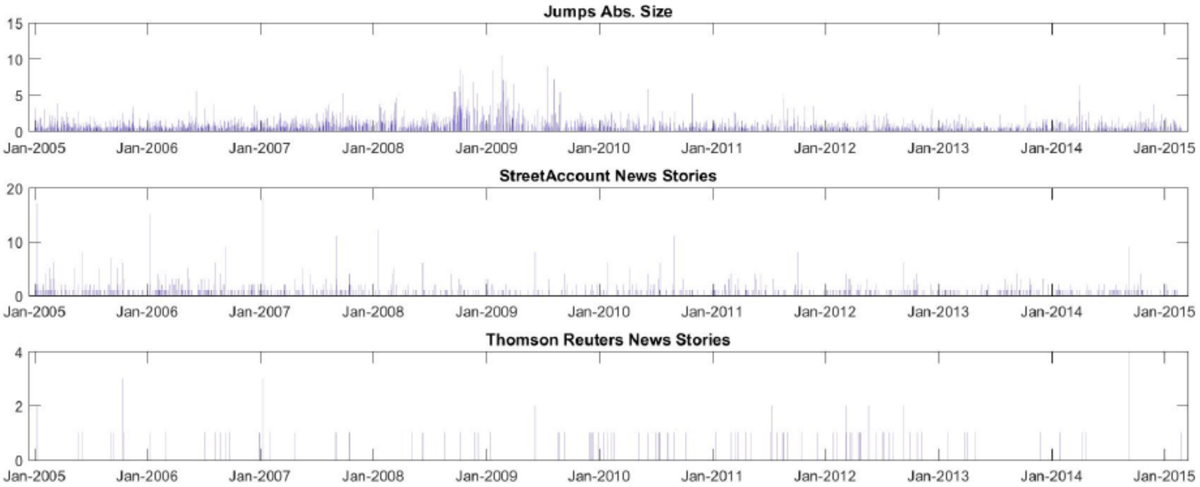

Figure 4 – From top to bottom, median of jumps' absolute size, number of news stories released by StreetAccount and by Thompson Reuters, respectively. ***Source:*** *Caporin and Poli (2018) – News and intraday jumps: variable selection, regularization and the economic impact of rare events.*

Regarding the news stories, it is evident that StreetAccount's releases are more frequent than the news released by Thompson Reuters. It can also be seen that it seems to be present a form of correlation between the release of news stories and jumps in returns: in the periods in which there is an unusually high number of stories, the number of jumps increases. There is a decreasing number of news stories released over time, which Caporin and Poli (2018) attribute to a more accurate selection of relevant news. The number of news stories has been obtained summing across the assets.

In Figure 5, are included 5 graphs: frequency of jumps, median of jumps absolute size and the frequency of news divided by type, all distributed over the intraday interval. The highest presence of jumps is at opening hours and closing hours, showing peaks at 10:00 AM, and from 14:00 to 15:00 PM. A similar case can be found in the absolute size graph, which shows also a peculiar U-shape. Looking at the frequency of news stories, it seems to be higher at the top of each hour. Furthermore, it is confirmed the fact that StreetAccount releases news with a higher regularity than Thompson Reuters. EPS are usually released between 15:00 and 15:30, as evidenced by the peaks in the frequency. Once again, it seems to exist a correlation between the frequency of news in certain hours and the frequency of jumps, in particular, around 10 AM, 14 and 15 PM. All data are presented in percentage, except for the median of jumps' absolute size.

As a part of their work, Caporin and Poli (2018) present a matching analysis, where they compute three indicators to link a jump occurrence with news released up to 30 minute earlier. The three indicators are:

- $P(J|N)$, namely the number of jump-news matches as a percentage of the total number of news released.
- $median(J|N)$, i.e. the median absolute size of jumps coinciding with news.
- $P(N|J)$, that is the number of jump-news matches divided by the number of total jumps.

$P(J|N)$ indicates the likelihood that news cause a jump, while $P(N|J)$ accounts for the portion of jumps linked to a certain type of news. Table 4 contains the results of the indicators applied to the matching analysis, on the basis of the type of news considered. "Other Sources" is used to indicate the absence of news in the dataset, while the results for StreetAccount and Thompson Reuters refer to topic *all*.

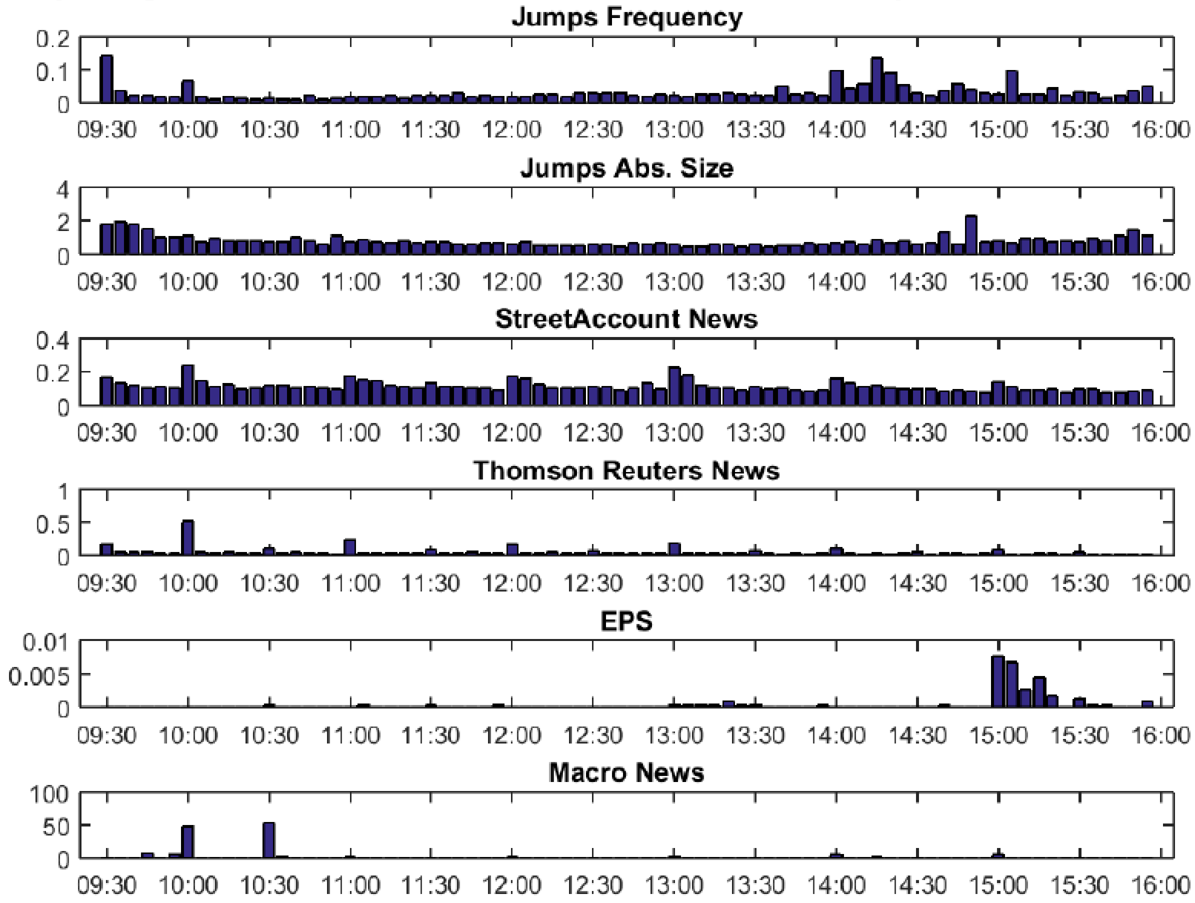

Figure 5 – Frequency of jumps (in %), median of jumps absolute size, frequency of news measures, divided by source (in %), all over intraday intervals. ***Source:*** *Caporin and Poli (2018) – News and intraday jumps: variable selection, regularization and the economic impact of rare events.*

It can be observed that the type of news causing more frequently a jump is EPS, with a $P(J|N)$ equal to 5.09%. Instead, the number of jumps associated with EPS is the lowest, with a $P(N|J)$ of 0.22%, meaning that only a 0.22% out of total jumps are anticipated by an EPS announcement. The second highest $P(J|N)$ is the one referring to StreetAccounts News, followed by Thompson Reuters News, Macro News and Other Sources. Higher StreetAccount $P(J|N)$ and $P(N|J)$, with respect to the same measures on Thompson Reuters, seem to indicate a greater impact of news released by the first provider in the determination of jumps. More in general, negative news present higher values both for $P(J||JN)$ and $P(N|J)$, meaning negative news are followed by jumps, and jumps are associated more often to negative news, respectively.

The fact that $P(N|J)$ for Other Sources is considerable (84.50%) suggests that the majority of jumps is not associate to news releases present in the dataset.

Going more in detail in each category, the topic with the largest $P(J|N)$ in StreetAccount news stories is *newspapers*, followed by *M&A*. In Thompson Reuters news stories, the topic with the highest number of jumps observed after the release is *top*. All other Thompson Reuters topics present greater values of the indicator, with respect to the topic *all*, indicating the ability of the provider to classify the news. Regarding Macro announcements, the one that plays the most important role in causing jumps is FOMC rate decision, with a $P(J|N)$ equal to 0.72%.

The authors conclude that, even though most of the jumps do not occur after the release of news, as evidenced by the outstanding $P(J|N)$ of Other Sources, some types of news in the dataset could still be significant in determining jumps.

Table 4 – Results of matching analysis. ***Source:*** *Caporin and Poli (2018) – News and intraday jumps: variable selection, regularization and the economic impact of rare events.*

News		**All**	**Positive**	**Negative**
EPS	$P(J\|N)$	5.09	1.22	3.87
	$median(J\|N)$	1.64	1.78	1.74
	$P(N\|J)$	0.22	0.16	0.28
StreetAccount	$P(J\|N)$	0.20	0.11	0.09
News Stories	$median(J\|N)$	1.49	1.58	1.87
	$P(N\|J)$	4.04	3.94	4.12
Thompson Reuters	$P(J\|N)$	0.15	0.08	0.07
News Stories	$median(J\|N)$	1.34	1.29	1.63
	$P(N\|J)$	1.55	1.69	1.37
Macro	$P(J\|N)$	0.03	0.01	0.02
Annoucements	$median(J\|N)$	0.81	0.90	0.80
	$P(N\|J)$	11.22	9.39	13.70
Other Sources	$P(J\|N)$	0.03	0.02	0.01
	$median(J\|N)$	0.76	0.80	0.73
	$P(N\|J)$	84.50	86.35	81.99

As further preliminary analysis, the regressors that presented no jump in their time series were studied. Logically, these variables should not have explanatory power and their relating coefficient should take value 0 in the following regressions. To find which of the independent variables contain no useful data, a matrix has been created, with the 88 assets on rows, and 712 measures on columns. Each cell takes the value 1, if the corresponding measure in the specific asset does not show jumps, or value 0 if at least one jump was detected. In Table 5 are reported the measures that, most frequently, do not contain information. On the contrary, non-zero coefficients should be expected for the measures that have at least one observation in the time series, depending on their significance.

Table 5 – Results of preliminary analysis on measures, divided by the number of stocks in which they present no information. ***Source****: own elaboration.*

Measures	**Stocks in which they are absent**
98 measures, subdivided as follows: - <u>StreetAccount:</u> 4 measures on Litigation (two with uncertain sentiment and two with positive sentiment), 3 measures on M&A (all with uncertain sentiment), one Newspapers measures with uncertain sentiment. - <u>T. Reuters:</u> 7 measures on Dividends (all sentiment types), 7 measures on Earnings (all sentiment types), 9 measures on Financial (all sentiment types), 3 measures of Medium relevance (all with uncertain sentiment), 3 measures of High relevance (all with uncertain sentiment), 7 measures of Top relevance (all sentiment types). - <u>Macro:</u> 4 measures on Inventories, 4 measures based on Chicago PMI, 4 measures on Cons. Confidence, 4 measures on Constr. Spending, 4 measures on Employment Trends, 4 measures on Home Sales, 4 measures on Factory Orders, one measures on Economic Optimism, 4 measures on Manufacturing PMI, 4 measures on Leading Index, one measure on Michigan Sentiment, 4 measures on New Home Sales, 8 measures on NY NAPM, 4 measures on Wholesale Inventories.	No observations in all 88 stocks
102 measures, of which 57 are measures based on StreetAccount news and 45 are measures computed on T. Reuters news and none of it are Macro measures.	No observations in almost all stocks (60-87)

55 measures, of which 25 are measures based on StreetAccount news and 45 are measures computed on T. Reuters news and none of it are Macro measures.	No observations in the majority of stocks (30-59)
114 measures, of which 67 are measures based on StreetAccount news and 47 are measures computed on T. Reuters news and none of it are Macro measures.	Observations presents in almost all stocks (1-29)
343 measures, divided as follow: - All 88 stocks time series - StreetAccount: 15 measure with topic "All" (all sentiment types), 8 measure with topic Upgrade/Downgrade (all sentiment types) - T. Reuters: 5 measures with topic "All" (all sentiment types) and 5 measures on Medium relevance (all sentiment types) - Macro: remaining 222 measures that were not included in previous descriptions.	At least one observation is present in all stocks

Furthermore, correlation between variables has been analysed, in order to eventually exclude from regressors one or more highly correlated variables. All values of the correlation computed were hugely below the value of 0,9, which was reassuring but also to be expected, since in the majority of cases news measures present a low number of observations.

5. Empirical analysis

The empirical analysis conducted aims to find the main determinants of the jumps in the 88 stocks of S&P 100, listed in Chapter 4, with the implementation of Elastic Net and Adaptive Lasso. As in previous literature, the variables that could cause significant impact on the jump process are macroeconomic measures and firm-specific news stories. Unlike previously revised literature, jumps in other stocks' prices are simultaneously included among possible determinants. Thus, the specification of the model contains 712 variables, among which are present both company-specific news, macroeconomic news, and jumps series of other assets. This means that for each stock estimated coefficients have been retrieved for the most important time series, through the two mentioned regularization procedures. Once the most significant regressors and relative coefficients have been found, the probability of observing a jump has been computed through logistic regression.

The first step is the implementation of the Elastic Net technique on logistic regression, with which coefficients for the most significant covariates are retrieved. The value of α applied is 0.1, since the lower its value, different from zero, the best the two techniques for Penalised Logistic Regression perform. Also, Caporin and Poli (2018) demonstrate that $\alpha = 0.1$, without the addition of any machine learning technique, grants the highest area under the ROC curve, indicating the best performing model. Instead, the value of λ, the second penalisation parameter, is chosen as the one that grants the highest area under the ROC curve (AUC) among ten different values, ranging from 0,00001 to 10 million. The extremities of the interval for λ are the same used in Caporin and Poli (2018), in order to guarantee a fair comparison between the results obtained in the two empirical analysis. Values of λ higher than one are never chosen as the best penalty factor in the regressions.

The area under the ROC (Receiver Operating Characteristic) curve measures the accuracy of the model in terms of classification. The area can assume values between zero and one, and plots the probability of observing a true signal, or a false one, depending on a range of cutpoints. It could be interpreted as the ability of the model to distinguish, among the subjects analysed, which shows the desired outcome and which do not. The closer to one the area is, the greater is the ability of the model to discriminate.

The second step consists in the application of Adaptive Lasso, to either confirm or adjust the coefficients resulting from Elastic Net. λ and α will be the same as the previous step. The same analysis will then be repeated just for positive and negative jumps.

5.1. Results of Penalised Logistic Regression

To perform the Elastic Net regularization, the package *glmnet* on R has been used. This package allows to perform regularization with different distributions and parameters. The λ that guarantees the highest area under the ROC has been picked, and then used also in the following step, in Adaptive Lasso. The coefficients have been retrieved from the regularization based on half the sample, while the performance of the model was tested on the other half of the sample, the most recent one.

Table 6 – Values of λ, chosen among ten values in the range between 0,00001 and 10 million, that grant the highest area under the ROC curve for each stock. ***Source****: own elaboration.*

Asset	λ	Asset	λ	Asset	λ	Asset	λ
AAPL	0,02154	CMCSA	0,02154	HON	0,00001	PEP	0,00001
ABT	0,00046	COF	0,00001	HPQ	1,00000	PFE	0,00046
ACN	0,00001	COP	1,00000	IBM	0,00046	PG	0,00046
AEP	1,00000	COST	0,02154	INTC	0,00001	QCOM	0,00046
AIG	0,02154	CSCO	0,02154	JNJ	0,00001	RTN	0,02154
ALL	1,00000	CVS	1,00000	JPM	0,00001	SBUX	0,00046
AMGN	0,02154	CVX	0,00001	KO	0,00046	SLB	0,02154
AMZN	0,00046	DD	0,00001	LLY	0,00001	SO	1,00000
APA	0,02154	DIS	1,00000	LMT	0,00046	SPG	1,00000
APC	0,00046	DOW	0,02154	LOW	0,00001	T	0,02154
AXP	1,00000	EBAY	1,00000	MCD	0,02154	TGT	0,00046
BA	0,00001	EMC	1,00000	MDT	0,00046	TXN	0,00001
BAX	0,00001	EMR	1,00000	MET	0,02154	UNH	1,00000
BHI	0,02154	EXC	1,00000	MMM	0,00046	UNP	1,00000
BIIB	0,02154	FCX	1,00000	MO	1,00000	UPS	0,00001
BK	1,00000	FDX	0,00001	MON	1,00000	USB	0,02154
BMY	1,00000	GD	0,00001	MRK	0,00046	UTX	0,02154
BRK.B	0,00001	GE	1,00000	MSFT	0,00046	WBA	0,02154
C	0,02154	GILD	0,02154	NKE	0,00001	WFC	1,00000
CAT	1,00000	GS	0,00046	NSC	0,00001	WMB	0,00046
CELG	0,00001	HAL	0,02154	ORCL	0,02154	WMT	1,00000
CL	0,00001	HD	0,02154	OXY	0,00001	XOM	0,02154

Table 6 contains the best values of λ for each of the 88 regressions, namely the ones that guarantee the highest area under the ROC curve. Not one of the values assumed by λ exceeds one. In fact, the majority of them is close to zero, which signifies a low level of penalisation. This could be due to the fact that the variables used as regressors often assume the value zero (some of them never show a non-zero observation, as highlighted before). A low penalisation parameter could be set in order not to exclude information, a priori.

In Table 7 are shown the 30 variables that are included in the results of the penalised logistic regressions, while in Appendix A, a graphical representation of the coefficients resulting from the regularization procedure can be find in Figure A.1 to A.6.

As could be expected, out of the 712 measures used as regressors, the majority of them has been discarded. Even from a graphical standpoint, it is evident that there is no general behaviour across time series: while some stocks seem to react to a greater number of regressors, with relatively small coefficients, others present a lower number of significant regressors, whose related coefficients are high in absolute value. It is worth observing that most Macro measures exhibit a coefficient in almost all regressions, meaning that jumps observed in different stocks, collocated in different sectors are influenced by macroeconomic news. One striking feature is the sign of the coefficients: in fact, almost the totality of coefficients related to macroeconomic news measures are negative. This is implying that macro news measures influence negatively the probability of observing a jump in the future. The most recurrent among Macroeconomic news appears to be the announcement of the FOMC, as could be expected, since the interest rates decided by the Federal Reserve impact all sectors. It should be highlighted that the news more often selected are often the ones with sentiment -1, which is to say the ones showing a negative sentiment in the text of the article released.

Table 7 – Most significant variables, as resulting from Elastic Net Regularization. ***Source****: own elaboration.*

Most significant variables, resulting from Elastic Net Regularization
Macro FOMC Rate Dec. -1 announcement
Macro FOMC Rate Dec. 0 announcement
StreetAcc. All 0 and +1 persistence
Macro Cons. Credit -1 announcement

Macro ECRI -1 announcement

Macro Nat. Gas Stocks -1 announcement

Macro Oil Stocks -1 announcement

Macro Business Inv. -1 announcement

Macro Chicago PMI -1 announcement

Macro Cons. Confidence -1 announcement

Macro ECRI 0 announcement

Macro ECRI +1 announcement

Macro Ex. Home Sales -1 announcement

Macro Ex. Home Sales -1 abs(surprise)

Macro Factory Orders -1 announcement

Macro ISM Man. PMI -1 announcement

Macro Leading Index -1 announcement

Macro Michigan Sent. -1 announcement

Macro Nat. Gas Stocks +1 announcement

Macro Oil Stocks 0 announcement

Macro Pend. Home Sales -1 announcement

Macro Phil. Fed -1 announcement

Macro Wholesale Inv. -1 announcement

Macro Business Inv. 0 announcement

Macro Cons. Confidence -1 abs(surprise)

Macro Cons. Confidence 0 announcement

Macro Constr. Spending -1 announcement

Macro Cons. Credit +1 announcement

Macro Empl. Trends -1 announcement

Macro Factory Orders 0 announcement

For what concerns company specific news measures, the most relevant are the ones covering all the topics included in the analysis (*"all"*) from both news providers, even though measures based on the news retrieved by StreetAccount are more frequent with respect to the same type of variables from Thompson Reuters. This fact confirms that news regarding a company are weighted by investors and are a determinant in decision making. This seems to confirm what already stands in the literature. Apart from the most comprehensive topic, the ones that get most frequently a non-zero coefficient are Thompson Reuters measures computed on dividends and with high and medium relevance, and StreetAccount measures based on up and downgrades, earnings, litigations and M&A. Still, StreetAccount news measure are more recurrent than Thompson Reuters one, as highlighted in Caporin and Poli (2018).

Looking at the spillover effect, it is meaningful to mention that stock time series get a coefficient from the Elastic Net regularization in the majority of the regressions performed. Also in this case, the coefficients are both negative and positive. Since one should rationally expect a greater spillover effect from stocks belonging to the same sector, to further investigate the matter, the assets were divided into groups based on the sector and the coefficients were analysed. Neglecting momentarily company-specific and macroeconomic news, the mean of the coefficients has been computed, first including all the assets used as regressors, and secondly considering sectors only. Table 8 summarizes the results. The coefficients mean, when taken including only the asset in the sector considered, is higher than the one taken considering all assets, in the great majority of cases. This feature seems to point at a larger importance of jumps in similar firms in predicting jumps, corroborating the hypothesis of the presence of spillover effect, which is also proved extensively in the literature. Some exceptions exist: for some assets, the mean considering the sector is equal to zero, while the most comprehensive one is different from zero. This implies that in some cases, stocks in the sector are not chosen from the regularization. Such a behaviour could be due to time series with fewer observations, that are neglected from Elastic Net and Adaptive Lasso.

Moreover, looking at the dynamic component of the model (jumps of the analysed variables have been included among the regressors, up to 5 minutes before the time of analysis), it shows a coefficient for all regularizations performed, in many cases even higher than the coefficients assigned to other assets. This points to a great importance of previous jumps, in determining the probability of observing another, in the same asset.

*Table 8 – Coefficient mean, including all assets (first row) and those belonging to the sector only (second row). **Source:** own elaboration.*

Sector: Consumer Goods

	AAPL	CL	KO	MO	NKE	PEP	PG
All	-0,0181	0,0317	-0,0499	0,0884	0,0721	0,1227	0,0945
Sector	0,0624	-0,8716	0,2842	0,3951	0,1500	0,1273	-0,6641

Sector: Basic Materials

	APA	APC	BHI	COP	CVX	DD	DOW	FCX

All	-0,0497	-0,0547	-0,0129	0,1205	0,0034	0,0814	0,0000	0,0000
Sector	0,0346	0,1723	-0,2072	0,6406	0,0000	0,2920	0,0000	0,0000
	HAL	**MON**	**OXY**	**SLB**	**WMB**	**XOM**		
All	-0,0767	0,0055	-0,1193	-0,1105	-0,0134	0,1005		
Sector	-0,2402	0,0000	-0,5003	-0,5420	-0,1267	-0,0087		
				Sector: Financial				
	AIG	**ALL**	**AXP**	**BK**	**BRKB**	**C**	**COF**	**GS**
All	0,0281	0,0013	0,0003	0,0003	0,0100	0,0807	0,0248	0,0704
Sector	0,2479	0,0000	0,0018	0,0005	-0,2255	0,0785	-0,0077	-0,0907
	JPM	**MET**	**SPG**	**USB**	**WFC**			
All	0,2386	0,0924	0,0001	-0,0093	0,0028			
Sector	0,7766	0,5283	0,0005	-0,0782	0,0026			
				Sector: Healthcare				
	ABT	**AMGN**	**BAX**	**BIIB**	**BMY**	**CELG**	**CVS**	**GILD**
All	0,0850	0,0467	0,1648	0,0086	0,0021	0,1870	0,0391	-0,0005
Sector	0,5474	0,4597	-0,0170	0,0402	0,0015	0,2945	0,0267	-0,0029
	JNJ	**LLY**	**MDT**	**MRK**	**PFE**	**UNH**		
All	-0,0171	0,2505	-0,0794	0,0000	0,1839	0,0017		
Sector	0,2135	0,2542	0,4232	0,0000	0,4126	0,0016		
				Sector: Industrial Goods				
	BA	**CAT**	**EMR**	**GD**	**GE**	**HON**	**LMT**	**MMM**
All	0,1225	0,0006	0,0053	-0,1808	-0,0012	0,0469	-0,0682	-0,0193
Sector	0,5086	0,0000	0,0000	0,2289	0,2426	0,4001	-0,9246	-0,4854
	RTN	**UTX**						
All	0,1035	0,0734						
Sector	0,0525	0,2541						
				Sector: Services				
	AMZN	**CMCSA**	**COST**	**DIS**	**EBAY**	**FDX**	**HD**	**LOW**
All	-0,0529	0,0719	0,0003	0,1540	-0,0138	0,0023	0,0222	0,1338
Sector	-0,0051	-0,1749	0,0000	-0,3979	-0,0367	0,0012	0,3501	0,4846
	MCD	**NSC**	**SBUX**	**TGT**	**UNP**	**UPS**	**WBA**	**WMT**
All	-0,1759	0,0864	0,0431	0,0869	0,0006	0,2574	-0,0510	0,0027
Sector	-0,2829	-0,5146	0,6395	-0,5132	0,0031	0,0196	0,1938	0,0091
				Sector: Technologies				

	ACN	**CSCO**	**EMC**	**HPQ**	**IBM**	**INTC**	**MSFT**	**ORCL**
All	0,1404	0,0886	0,0007	0,0922	0,0000	0,1405	0,1876	0,1126
Sector	-0,4077	-0,0420	-0,0060	-0,3640	0,0000	0,1465	-0,3651	0,2393
	QCOM	**T**	**TXN**					
All	-0,0729	0,0522	-0,0422					
Sector	-0,1295	-0,4080	0,7659					
Sector: Utilities								
	AEP	**EXC**	**SO**					
All	0,0038	0,0010	0,0041					
Sector	0,0000	0,0000	0,0091					

To evaluate how the model is capable of predicting the presence of jumps, one can refer to the area under the ROC curve. In Table 9 are summarized the values of the AUC for each asset.

As Table 9 shows, the area under the ROC curve varies, depending on the asset considered. The maximum value reached is 0,791 in the model performed on the asset TGT (Target Corp). Looking at the AUC per sector, the ones showing the highest values are the Utilities sector, followed by the sectors of Financial and Services, meaning that the model performs better in those sectors, predicting the exact value of the dependent variable more often.

Overall, the mean area under the ROC curve is 0,5238, while in Caporin and Poli (2018), which was the starting point of this analysis, it was 0,5943 (for the same level of α). This could mean that the addition of many regressors is not the best policy, and the model could be improved.

Table 9 – Area under the ROC curve, resulting from the regularization, for each asset. ***Source:*** *own elaboration.*

Asset	**AUC**	**Asset**	**AUC**	**Asset**	**AUC**	**Asset**	**AUC**
AAPL	0,722	CMCSA	0,613	HON	0,677	PEP	0,393
ABT	0,646	COF	0,618	HPQ	0,317	PFE	0,583
ACN	0,580	COP	0,278	IBM	0,619	PG	0,588
AEP	0,585	COST	0,638	INTC	0,677	QCOM	0,436
AIG	0,655	CSCO	0,629	JNJ	0,403	RTN	0,663
ALL	0,647	CVS	0,658	JPM	0,664	SBUX	0,657
AMGN	0,630	CVX	0,620	KO	0,287	SLB	0,333

AMZN	0,613	DD	0,403	LLY	0,385	SO	0,676
APA	0,392	DIS	0,483	LMT	0,390	SPG	0,705
APC	0,364	DOW	0,551	LOW	0,380	T	0,675
AXP	0,575	EBAY	0,610	MCD	0,354	TGT	0,791
BA	0,350	EMC	0,455	MDT	0,648	TXN	0,124
BAX	0,340	EMR	0,642	MET	0,314	UNH	0,571
BHI	0,293	EXC	0,593	MMM	0,372	UNP	0,700
BIIB	0,579	FCX	0,578	MO	0,324	UPS	0,379
BK	0,682	FDX	0,608	MON	0,570	USB	0,699
BMY	0,657	GD	0,333	MRK	0,557	UTX	0,674
BRK.B	0,416	GE	0,461	MSFT	0,430	WBA	0,612
C	0,606	GILD	0,438	NKE	0,363	WFC	0,592
CAT	0,732	GS	0,525	NSC	0,387	WMB	0,379
CELG	0,479	HAL	0,284	ORCL	0,330	WMT	0,734
CL	0,362	HD	0,703	OXY	0,428	XOM	0,631

5.2. Result of Penalised Logistic Regression on Positive and Negative jumps

The same analysis illustrated in the previous section has been implemented considering positive and negative jumps only, separately. First of all, two different time series were created for each asset, one containing positive jumps only, while the other containing negative jumps. Once again, Elastic Net regularization has been implemented using the package *glmnet* in R, with the same methodology to establish the penalisation parameters seen in the previous section. As in the case of the analysis executed on the whole time series, the resulting coefficients are displayed in Appendix C, dividing the results following the category of the regressors, in comparison with the previous ones.

Table 10 reports the 30 most selected variables from Elastic Net and, subsequently, Adaptive Lasso, as done for the entire time series. It should be underlined that the FOMC Rate appears only in the time series of negative jumps in the measure based on the announcement of December. This fact seems to suggest that Federal Reserve policies influence stock jumps in the negative sense: given that the relative coefficients often have a positive sign, FOMC rates measures increase the probability of observing a negative jump. Looking at the most selected variables in the positive time series, it catches the attention the fact that company specific news measures do not have an outstanding role. On the contrary, when the analysis is performed only on negative jumps, their role is greater, since they receive a coefficient from the regularization process more often. Even in this case, this type of variables increases the probability of observing a negative jump, but not a positive one. Among the macroeconomic news measure, it is worth mentioning the great impact news based on oil and natural gas have on the positive time series. This feature accentuates the importance of the sector, and its shocks, on financial markets. A striking feature that needs to be addressed is the fact that other stocks' jumps play a great role in negative time series: even though in the analysis performed on negative jumps the variable selected often are less in general, among the most selected ones appear many of the stocks present in the dataset, almost all of them receiving a negative coefficient from the regularization. So, observing jumps in other assets appears to be important, even more so if we expect a negative jump.

Once again, the sentiment -1 (negative sentiment emerging from title and text of news articles) is more recurrent, indicating that jumps are more sensible to occur after the negative news are released.

*Table 10 – Most selected variables, as resulting from Elastic Net Regularization, for positive jumps and negative jumps time series. **Source**: own elaboration.*

Most selected variables for positive jumps	Most selected variables for negative jumps
Macro Nat. Gas Stocks -1 announcement	IBM
Macro Business Inv. -1 announcement	Macro FOMC Rate Dec. -1 announcement
Macro Chicago PMI -1 announcement	COF
Macro Cons. Confidence -1 announcement	SPG
Macro Cons. Credit -1 announcement	EBAY
Macro Cons. Credit +1 announcement	GD
Macro ECRI -1 announcement	BK
Macro ECRI 0 announcement	SBUX
Macro ECRI +1 announcement	LMT
Macro Empl. Trends -1 announcement	ALL
Macro Ex. Home Sales -1 announcement	GS
Macro Factory Orders -1 announcement	MMM
Macro Federal Budget -1 announcement	MSFT
Macro Federal Budget +1 announcement	HON
Macro Leading Index -1 announcement	USB
Macro Michigan Sent. -1 announcement	AIG

Macro NAHB +1 announcement	StreetAcc. All -1 and 0 persistence
Macro Nat. Gas Stocks 0 announcement	CVX
Macro Oil Stocks 0 announcement	NKE
Macro Pend. Home Sales -1 announcement	QCOM
Macro Phil. Fed -1 announcement	StreetAcc. All 0 and +1 persistence
Macro Wholesale Inv. -1 announcement	COST
Macro Constr. Spending -1 announcement	DD
Macro IBD Ec. Opt. -1 announcement	JNJ
Macro ISM Man. PMI -1announcement	C
Macro NAHB -1 announcement	DIS
Macro Nat. Gas Stocks +1 announcement	WFC
Macro New Home Sales -1 announcement	StreetAcc. All +1 announcement
Macro Oil Stocks -1 announcement	EXC
Macro Oil Stocks -1 abs(surprise)	HPQ

Regarding firm-specific news, as in the case of the entire time series, StreetAccount variables are predominant with respect to Thompson Reuters ones, since they acquire a coefficient from the penalised regression in a higher number of cases. The topics chosen in the analysis performed on the whole series are confirmed to be important even when positive and negative jumps only are investigated. For StreetAccount measures, the more frequent topic is *"all"*, with all three different sentiments (-1, 0 and 1), both in the positive series and in the negative one. This does not necessarily mean that all three sentiment are chosen in the regression of on asset. Following those variables, the ones getting a coefficient in most regressions are those based on Earnings and Upgrade and Downgrade news, confirming the results obtained in the previous section. Looking at Thompson Reuters measures in both series, the topic *"all"* has the greatest importance, immediately followed by the measures built on news of High and Medium relevance. In the analysis performed on the positive series stands out the fact that Thomson Reuters measures on Financial announcements with sentiment -1 are among the most significant in the prediction of positive jumps. Since on average their coefficient has a positive sign, it is stressed the relevance that financial statements have on creating positive jumps, and on investment decisions for investors.

Studying the presence of other assets among the determinants of positive and negative jumps, they play a larger role in the latter case. Still, it is not clear the predominance of one sector on all the others: in fact, assets with a coefficient resulting in the majority of regressions belong to different sectors.

As before, to study the impact of sectors on other stocks belonging to the same industry, means of coefficients where taken, disregarding momentarily all other variables (Appendix B). It is confirmed that when taken only on the companies of the sector, the mean of the coefficients is larger than the one including the betas of all assets, in every case. This is to say that also when studying positive and negative jumps only, a jump occurring in the price of firms belonging to a sector has a higher influence on the probability of observing a positive or negative jump in a company of the considered industry.

At last, a comparison of the betas resulting from all three regularizations are displayed. In Figures from 6 to 9 are reported the coefficients, divided on the basis of measure categories of one stock for seven stocks, belonging to different industries included in the analysis.

The figures are built in the following way: for each category of news measures (StreetAccount, T. Reuters, Macroeconomic, other assets) are displayed two scatter plots, one having on the x axis the values of the coefficients from Adaptive Lasso on the entire time series, while on the y axis coefficients of penalised regression on positive jumps can be found; the second has the same structure, with the only difference being that values on the y axis, which are the betas resulting from penalisation on negative jumps only. In all four figures, it is evident that, when performing the regularization on positive or negative jumps only, the variables determining the probability of observing a jump are less, even more so in the second case. In fact, more coefficients are set to zero, with respect to the regression on the entire series. As highlighted before, firm-specific news measures seem to have a lower impact on the determination of the probability of observing a positive or negative jump, while macroeconomic news remain central, but only for the positive series.

Also, other assets show coefficients almost always set to zero, which is to say they contribute less to the probability of observing positive or negative jumps, contrary to the case of the whole series. The only exception is often the case of past jumps of the variable taken into consideration: this dynamic component of the model shows, in most case, the only non-zero coefficient. This is to say that a jump occurred up to 5 minutes before the time of analysis influences the probability to observe another one.

Another striking feature emerging is the fact that the fewer variable chosen from the regressions on positive or negative jumps maintain the same sign they displayed from the broader regressions, even though the magnitude is smaller. In fact, variables that in some cases showed a coefficient even up to 20 or more are often scaled back to inferior figures. This happens not only on such high scales, but also on reduced scales, and for both positive and negative coefficients. In general, looking at the values of the coefficients, when the same variables are chosen for both positive and negative series, their coefficient is of the opposite sign.

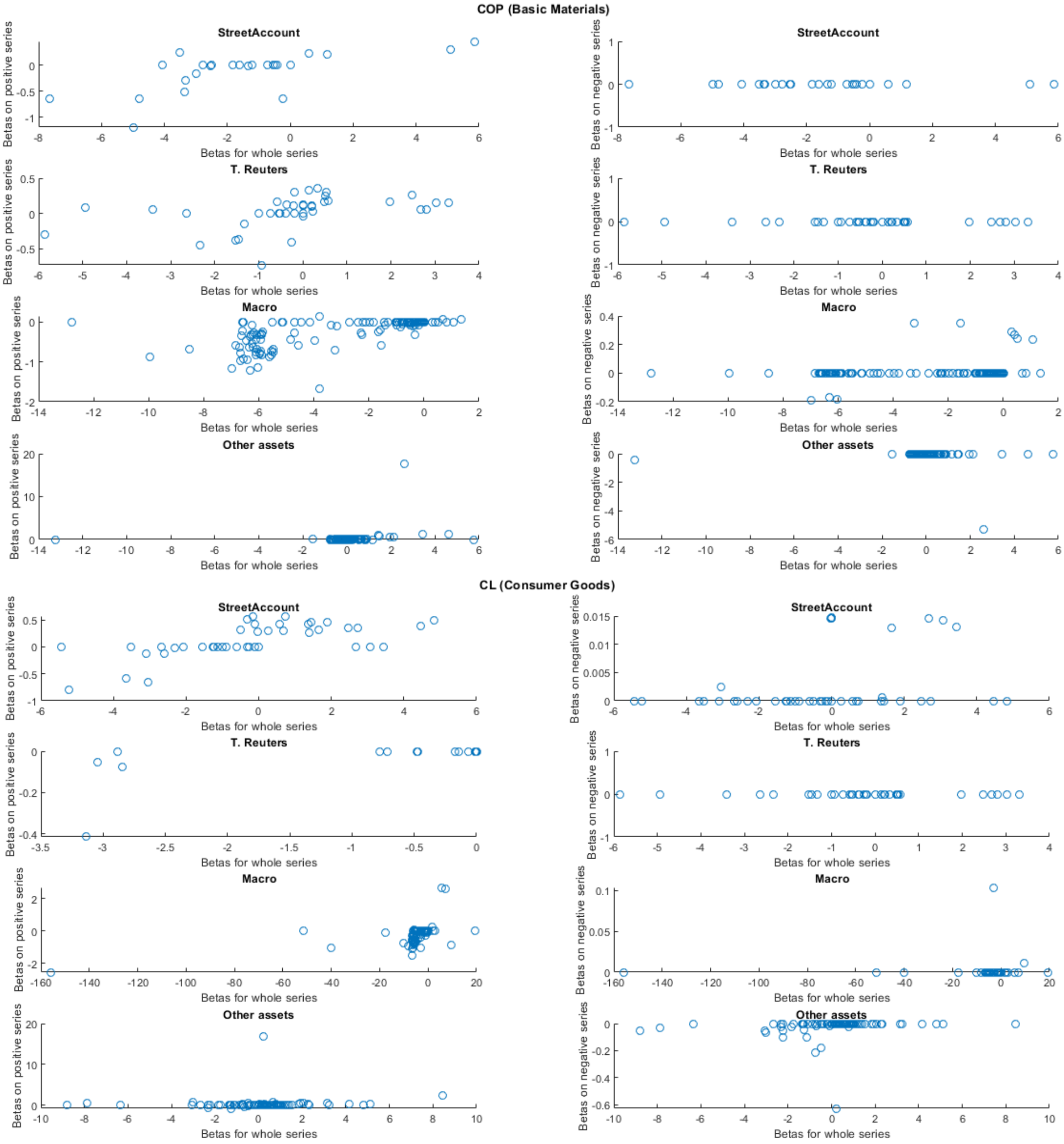

Figure 6 – Comparison between the coefficients resulting from Adaptive Lasso regularization, on entire series, positive jumps only and negative jumps only, for the first asset of Basic Materials and Consumer Goods sectors. ***Source****: own elaboration.*

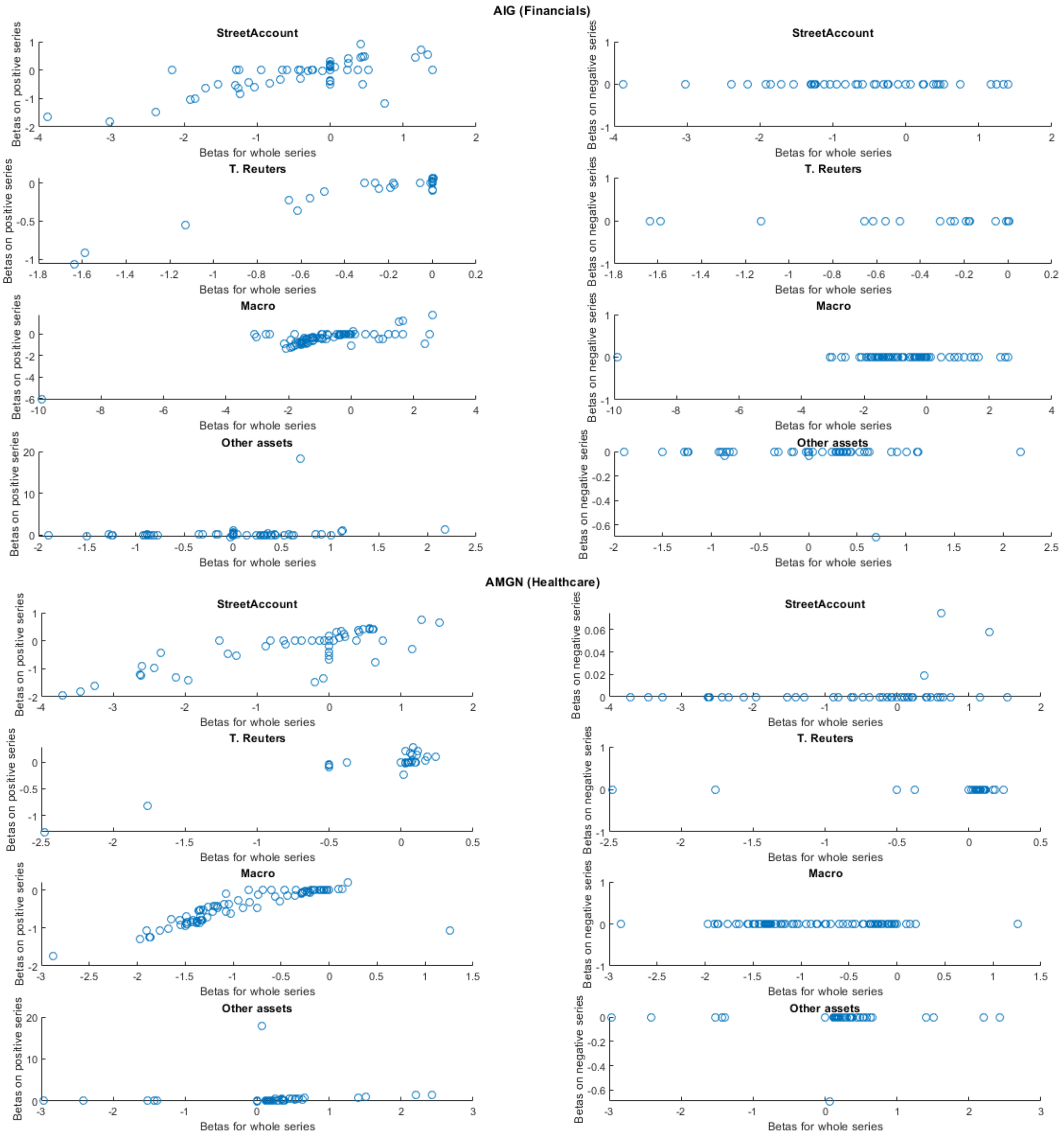

Figure 7 - Comparison between the coefficients resulting from Adaptive Lasso regularization, on entire series, positive jumps only and negative jumps only, for the first asset of Financial and Healthcare sectors. ***Source****: own elaboration.*

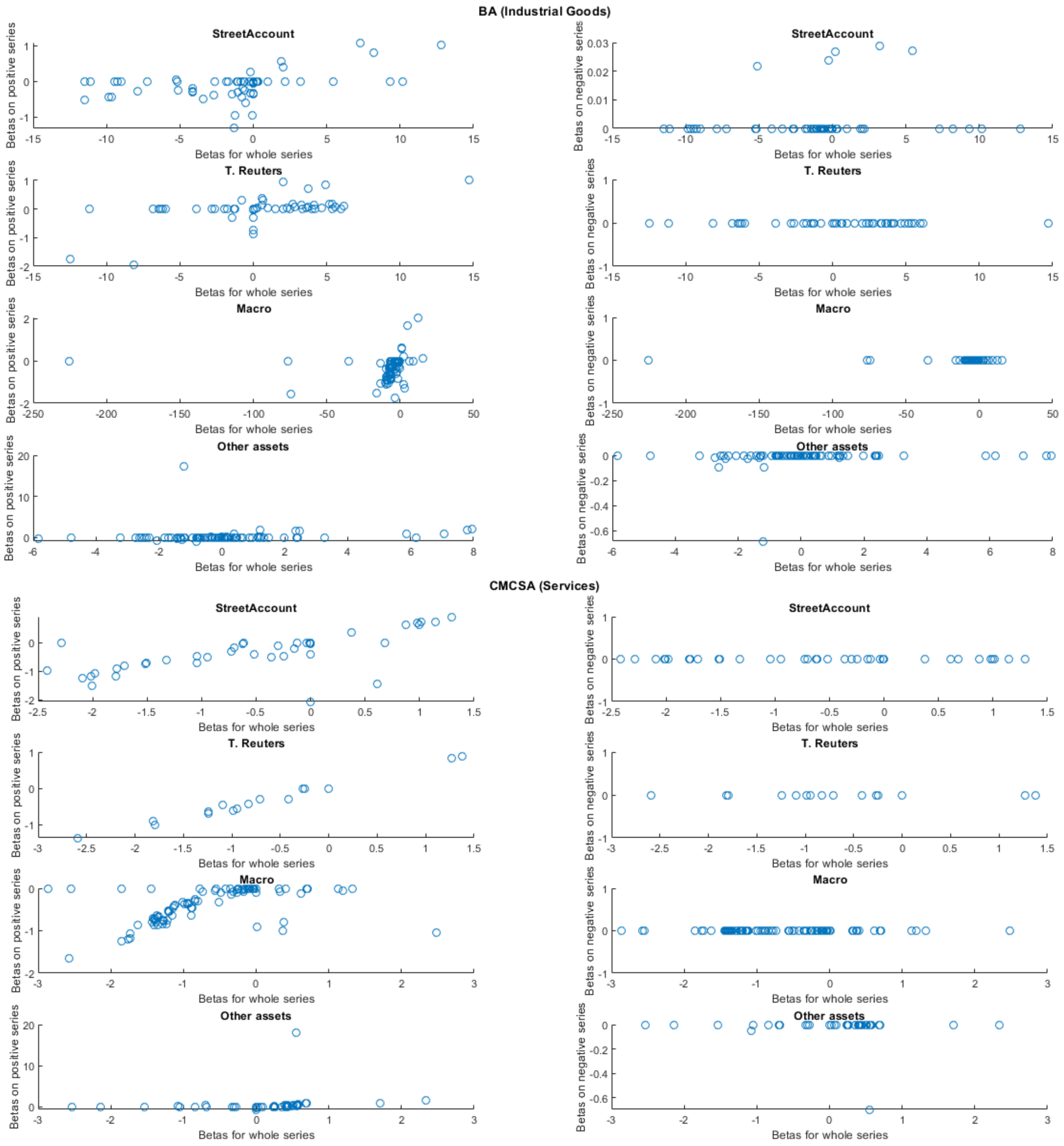

Figure 8 - Comparison between the coefficients resulting from Adaptive Lasso regularization, on entire series, positive jumps only and negative jumps only, for the first asset of Industrial Goods and Services sectors. ***Source****: own elaboration.*

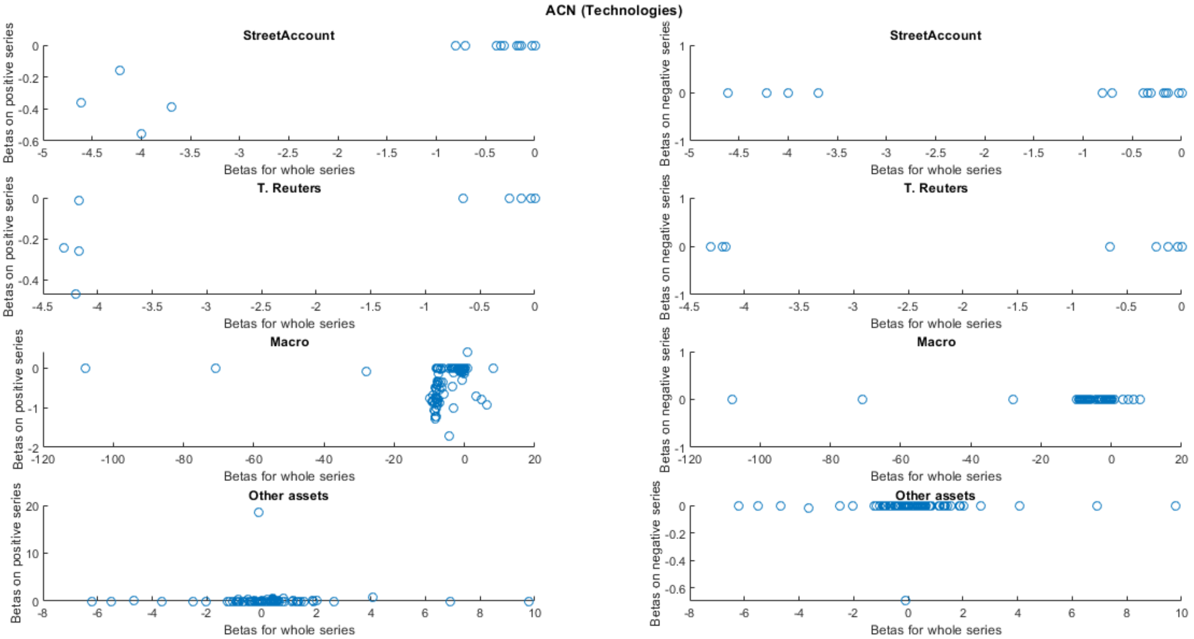

Figure 9 - Comparison between the coefficients resulting from Adaptive Lasso regularization, on entire series (on x axis, on both graphs), positive jumps only and negative jumps only (on y axis, in the right and left graph, respectively), for the first asset of Technologies and Utilities sectors. ***Source****: own elaboration.*

Both these Figures and the other comparisons are also reported in Appendix C (Figures C.1-C.24).

6. Conclusions

The prediction of jumps in stock price is a matter of the highest importance, since an unpredicted movement in assets' prices could impact on the performance of portfolios and, indirectly, on people's savings and living. This is the main reason why the topic has been so widely investigated through the years.

As exposed in the chapter on literature review, the approaches to this problem have been several, both looking at the models implemented, and to the variables used as predictors. Some published papers include among regressors only firm-specific news (Ryan and Taffler (2004), Lee and Mykland (2008), Evans (2011)); others include both company-specific news and macroeconomic news (Kanneiainen and Yue (2019)). Most recent papers are able to use also the sentiment that an article conveys, along with the canonical types of news measures (firm-specific and macroeconomic). Among those, we can comprehend Tetlock (2007), Loughran and McDonald (2011), Garcia (2013), Caporin and Poli (2017). The contribution of this work is the extension of regressors to other assets' jumps time series.

This work starts from a dataset containing jumps for 88 stocks of S&P 100 over a 10-year period and 624 news measures; jumps were detected implementing the Barndorff-Nielsen and Shepard (2006) test, with a corrected threshold bi-power variation estimator. Penalised logistic regression was performed for each asset, using Elastic Net and Adaptive Lasso, in order to diminish the number of variables contributing to the prediction. Among the regressors were comprehended four categories of measure: company-specific news measures built on data coming from the provider StreetAccount, company-specific measures based on Thompson Reuters news, macroeconomic news and other assets' jumps time series.

After finding the best penalisation parameters, coefficient for the whole jump series were retrieved. It is found that macroeconomic news are the most commonly selected as determinants, one feature that confirms what already stands in the literature. Furthermore, StreetAccount measures are chosen more frequently than Thompson Reuters ones, which could be due to the fact that in the first provider news are released more often with respect to the second. Jumps in other assets appear to have an impact on prediction of jumps, even more so if they belong to the same sector of the stock considered. Also, there seems to be a dynamic effect, since the time series containing the jumps observed up to 5 minutes before time of analysis of the asset considered, often shows a higher coefficient with respect to the other stocks included.

To assess the goodness of the model, the area under the ROC was computed: on average the resulting AUC was 0,5238, which was smaller than the same measure obtained in Caporin and

Poli (2018), the main comparison for this analysis. This could mean that adding all the assets may not be the right choice, a topic that be investigated with further work.

As a further analysis, time series containing only positive and only negative jumps were considered, and the same regularization was applied. Macroeconomic news still result as the most determinant in predicting the occurrence of a jump in the positive series, while the same role belongs to other assets in the negative ones. StreetAccount measures still prevail on Thompson Reuters one, since they were chosen more often. In both series, the number of news measures to which the model assigns a coefficient was lower with respect to the same in the model performed on the whole time series. Moreover, the coefficients themselves were smaller. Firm-specific news measures and other assets' jumps often present a neglectable influence when used in this second instance. However, even in this case, the sector to which a stock belongs to impacts more than the others.

The analysis performed in this thesis also has some drawbacks: even though the conjunction of news measures and jumps in other assets is novel, the model does not seem to perform better with respect to the starting point to which it compares to, suggesting that including all assets could not be the best choice. Moreover, machine learning techniques applied in Caporin and Poli (2018) were not performed in the context of this work, which could be an additional extension.

Further possible research could be implemented with respect to the hypothesis of including among the regressors only the asset belonging to the sector in which the considered company works, while simultaneously adding machine learning techniques, to find the combination that grants the most performing model.

7. References and sitography

AIT-SAHALIA, Y., CACHO-DIAZ, J., LEAVEN, R.J.A. (2013). *Modelling financial contagion using mutually exciting jumps processes.* Journal of Financial Economics, Elsevier, vol. 117(3), pages 585-606.

ANDERSEN, T.G. et al. (2003). *Modelling and forecasting realized volatility*. Econometrica 71, pp. 579–625.

ASGHARIAN, H., BENGTSSON, C. (2006). *Jumps spillover in international equity markets.* Journal of Financial Econometrics, Vol. 4, No. 2, pp. 167-203

ASGHARIAN, H., NOSSMAN, M. (2011). *Risk contagion among international stock markets.* Journal of International Money and Finance, Vol. 30, Issue 1, pp. 22-38.

AUDRINO, F., TETEREVA, A., (2019). *Sentiment spillover effects for US and European companies.* Journal of Banking and Finance, 106, pp. 542-567.

BAKER, S.R. et al. (2019). *What triggers stock market jumps?.* Working Paper, Northwestern University.

BAJGROWICZ, P., SCAILLET, O., TRECCANI, A. (2016). *Jumps in High-frequency Data: Spurious Detections, Dynamics and News*. Management Science, 62, pp 2198-2217.

BERRY, W.D., and FELDAMN, S. (1985). *Multiple regression in practice*. Sage University Paper, Series on Quantitative Applications in the Social Sciences, 07-050. Beverly Hill, CA: Sage.

CAPORIN, M., ROSSI, E., SANTUCCI DE MAGISTRIS, P. (2011). *Conditional jumps in volatility and their economic determinants*. Available at SSRN 1924812. September 2011.

CAPORIN, M., POLI, F. (2017). *Building news measures from textual data and an application to volatility forecasting*. Econometrics, 5, 1-46.

CAPORIN, M., POLI, F. (2018). *News and Intraday Jumps: variable selection, regularization and the economic impact of rare events.*

CHAN, W.S. (2003). *Stock price reaction to news and no-news: drift and reversal after headlines.* Journal of Financial Economics, Vol. 70, Issue 2, pp. 223-260.

CORSI, F., PIRINI, D., RENÒ, R. (2010). *Threshold Bipower Variation and the Impact of Jumps on Volatility Forecastin*g. Journal of Econometrics, 114 (3), pp 276-288.

DUMITRU, A-M. AND URGA, G. (2012). *Identifying jumps in financial assets: A comparison between nonparametric jump tests*. Journal of Business and Economic Statistics, 30(2), pp. 242-255.

EVANS, K.P. (2011). *Intraday Jumps and US Macroeconomic News Announcements*. Journal of Banking and Finance 35, pp 2511-2527.

GARCIA, D. (2013). *Sentiment during recessions*. Journal of Finance, Vol. 68(3), pp. 1267-1300.

GLONEK G.F.V., MCCULLAGH, P. (1995). *Multivariate Logistic Model*. Journal of the Royal Statistical Society, Series B (Methodological). Vol. 57, No. 3, pp. 533-546.

HASTIE, T., QIAN, J. (2016). *An introduction to glmnet*. Available online: https://cloud.r-project.org/web/packages/ glmnet/vignettes/glmnet.pdf

HESTON, S.L, SINHA, N.R. (2016). *News versus Sentiment: Predicting Stock Returns from News Stories*. Finance and Economics Discussion Series, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.

HOSMER, D.W., LEMESHOW, S. (2004). *Applied Logistic Regression*, Second Edition. John Wiley and Sons, 2004.

HUANG, X. (2015). *Macroeconomic News Announcements, Systemic Risk, Financial Market Volatility and Jumps*. Economics Discussion Series 2015-097. Washington: Board of Governors of the Federal Reserve System.

JAWADI, F., LOUHICHI, W., CHEFFOU, A.I., (2015). *Testing and modelling jump contagion across international stock markets: A nonparametric intraday approach.* Journal of Financial Markets, 26, pp. 64-84.

JIANG, G.J., LO, I., VERDELHAN, A. (2011). *Information shocks, liquidity, and price discovery: Evidence from the US treasury market.* Journal of Financial and Quantitative Analysis, Vol. 46, pp. 527-551.

KANNIAINEN, J., YUE, Y. (2019). *The arrival of news and return jumps in stock markets: a nonparametric approach*. arXiv, Quantitive Finance.

KAPETANIOS, G. et al. (2019). *Jumps in option prices and their determinants: Real-time evidence from the E-mini S&P 500 option market.* Journal of Financial Markets, November 2019, 100506.

LAHAYE, J., LAURENT, S., NEELY, C.J. (2011). *Jumps, Cojumps and Macro Announcement*. Journal of Applied Econometrics, 26, pp 893-921.

LEE, S.S. (2012). *Jumps and Information Flow in Financial Markets*. Review of Financial Studies, 25, pp 439-479.

LEE, S.S., MYKLAND, J. (2008). *Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics*. Review of Financial Studies, 21, pp 2535-2563.

LOUGHRAN, T., MCDONALD, B. (2011). *When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks*. Journal of Finance, 66, pp 35-65.

MAHEU, J.M., MCCURDY, T.H. (2004). *News arrival, jump dynamics and volatility components for individual stock returns.* The Journal of Finance, Vol. 59, No. 2, April 2004.

MENARD, S. (2002). *Applied logistic regression analysis*. Second edition. Sage University Paper, Series on Quantitative Applications in the Social Sciences, 07-106. Thousand Oaks, CA: Sage.

MERTON, R.C. (1976). *Option pricing when underlying stock returns are discontinuous*. Journal of Financial Economics. Vol. 3 Issue 1-2, pp 125-144.

O'BRIEN, S.M., DUNSON, D.B. (2004). *Bayesian Multivariate Logistic Regression*. Biometrics, 60(3), pp. 739-746.

PENNY, W.D, ROBERTS, S.J. (1999). *Dynamic Logistic Regression*. International Joint Conference of Neural Network.

RANGEL, J.G. (2011). *Macroeconomic news, announcements, and stock market jump intensity dynamics.* Journal of Banking and Finance, 35, 1263-1276.

RYAN, P., TAFFLER, R.J. (2004). *Are economically significant stock returns and trading volumes driven by firm-specific news releases?*. Journal of Business Finance & Accounting, Vol. 31, No. 1-2, pp 49-82, January 2004.

SIDOROV, S.P. et al. (2014). *GARCH Model with Jumps: Testing the Impact of News Intensity on Stock Volatility*. World Congress on Engineering, WCE, C. 110-115.

TETLOCK, P.C. (2007), *Giving content to investor sentiment: The role of media in the stock market.* Journal of Finance, Vol. 62, pp. 1139-1168.

XIAO, Y., YIN, X., ZHAO, J. (2019). *Price dynamics of individual stocks: jumps and information.* Finance Research Letters. Available online December 2019, 101404.

Sitography

The Comprehensive R Archive Network: https://cran.r-project.org/

Matlab Central: https://it.mathworks.com/matlabcentral/?s_tid=srchtitle

8. Appendixes

8.1. Appendix A: Graphical representation of results from penalised regressions

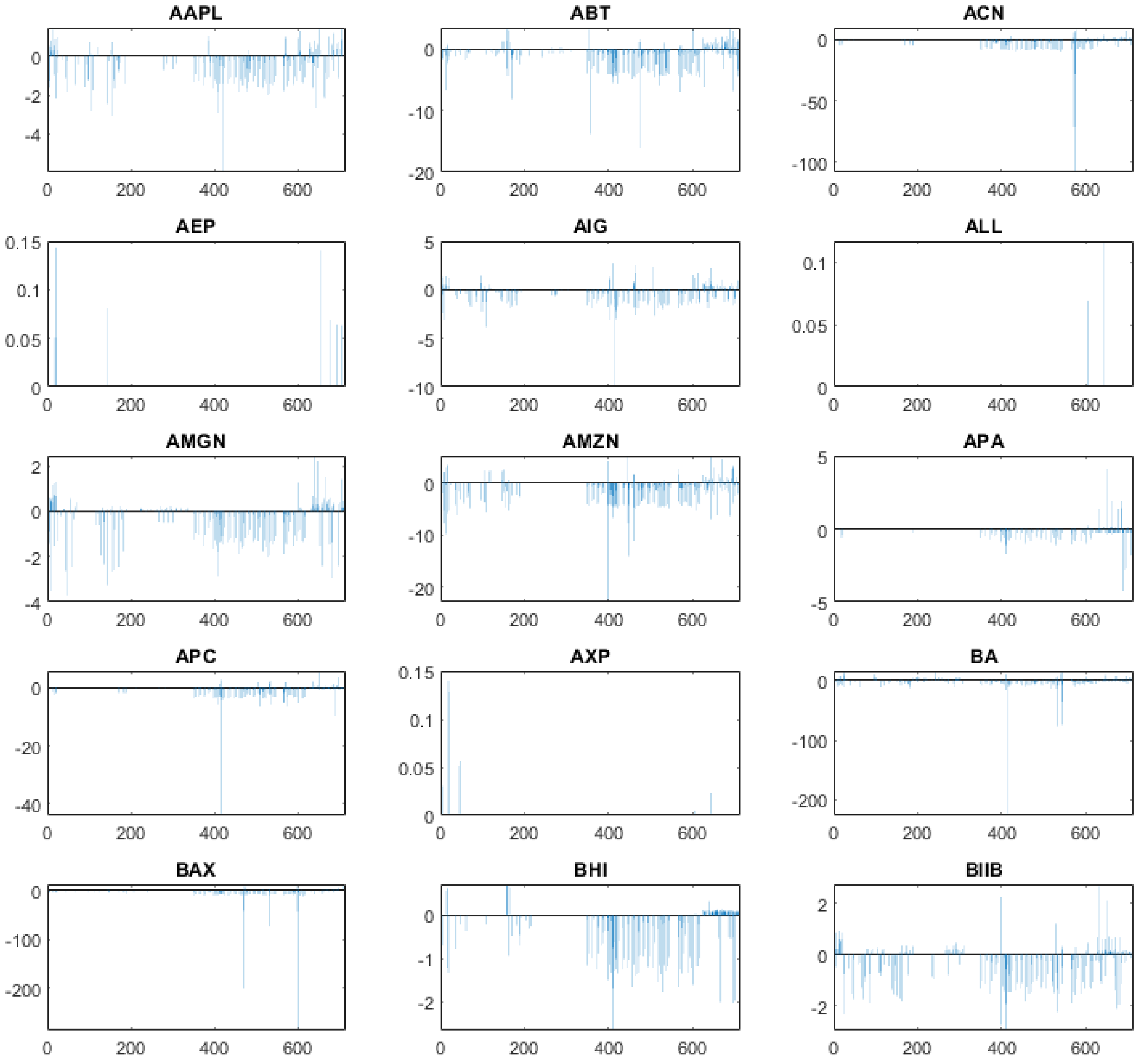

Figure A.1 – Plots displaying the representation of the coefficients resulting from Elastic Net Regularization, then confirmed with Adaptive Lasso, for the first fifteen stocks. ***Source****: own elaboration.*

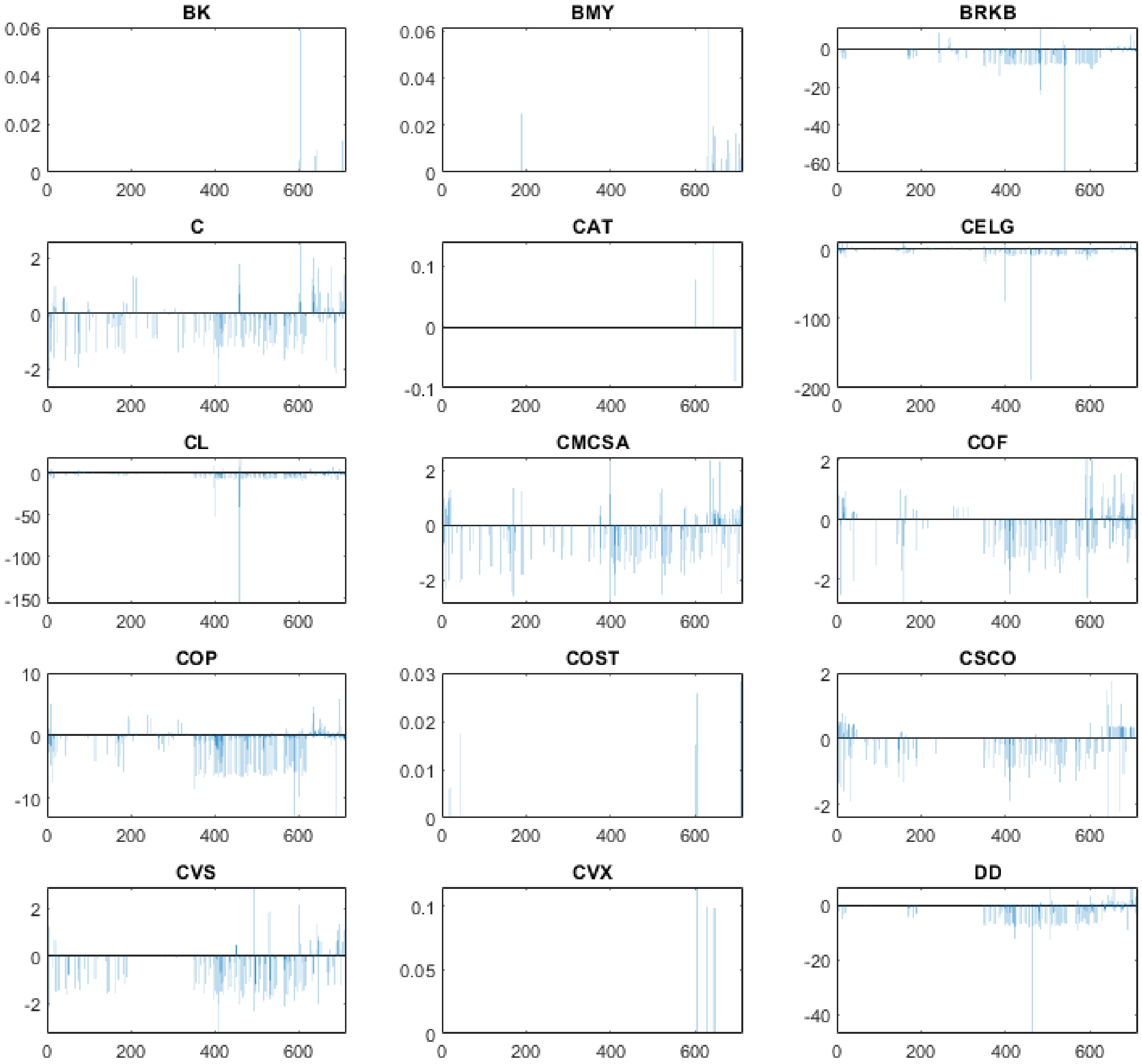

Figure A.2 - Plots displaying the representation of the coefficients resulting from Elastic Net Regularization, then confirmed with Adaptive Lasso, for stocks going from BK to DD. ***Source****: own elaboration.*

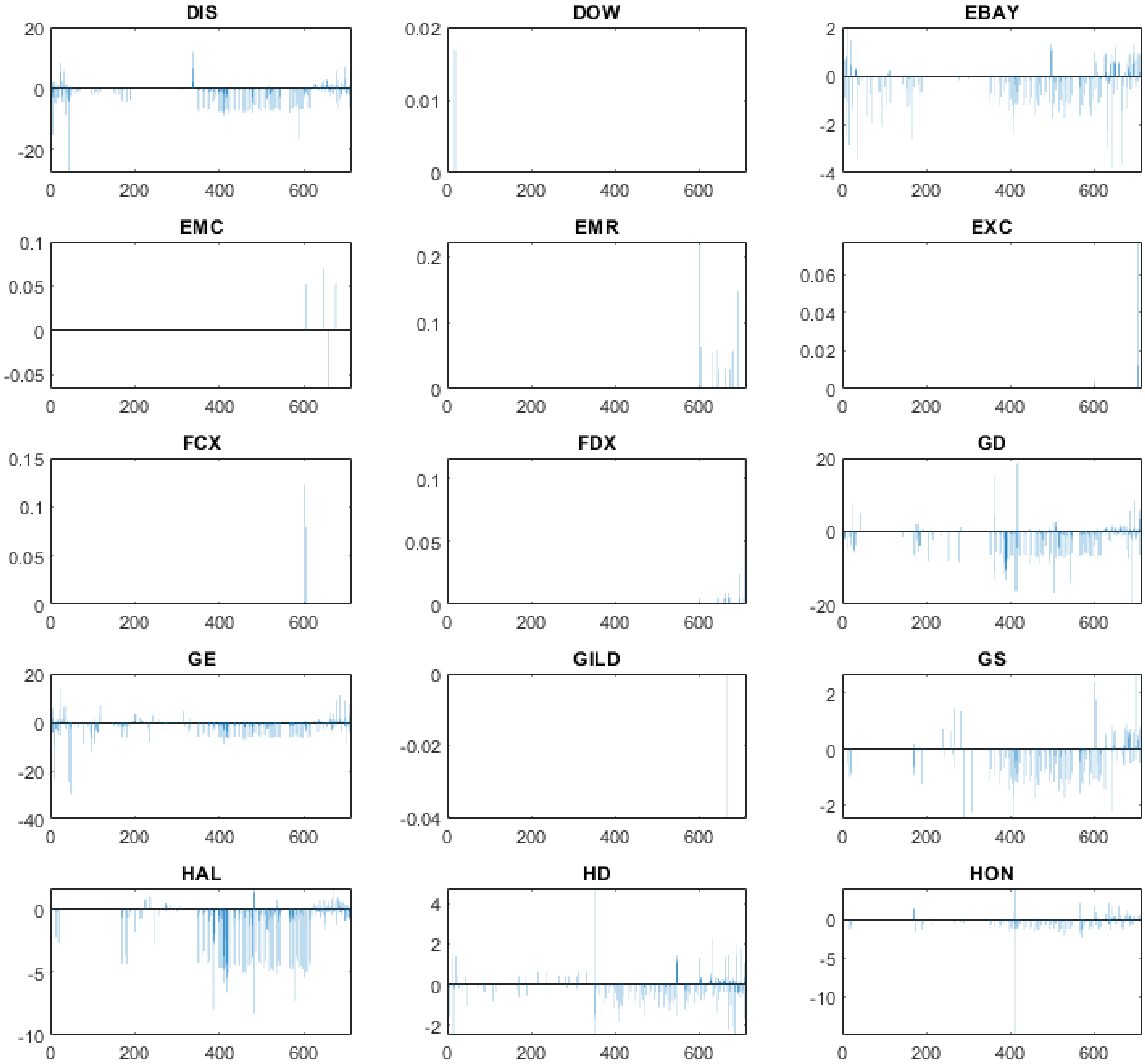

Figure A.3 - Plots displaying the representation of the coefficients resulting from Elastic Net Regularization, then confirmed with Adaptive Lasso, for stocks going from BK to DD. ***Source****: own elaboration.*

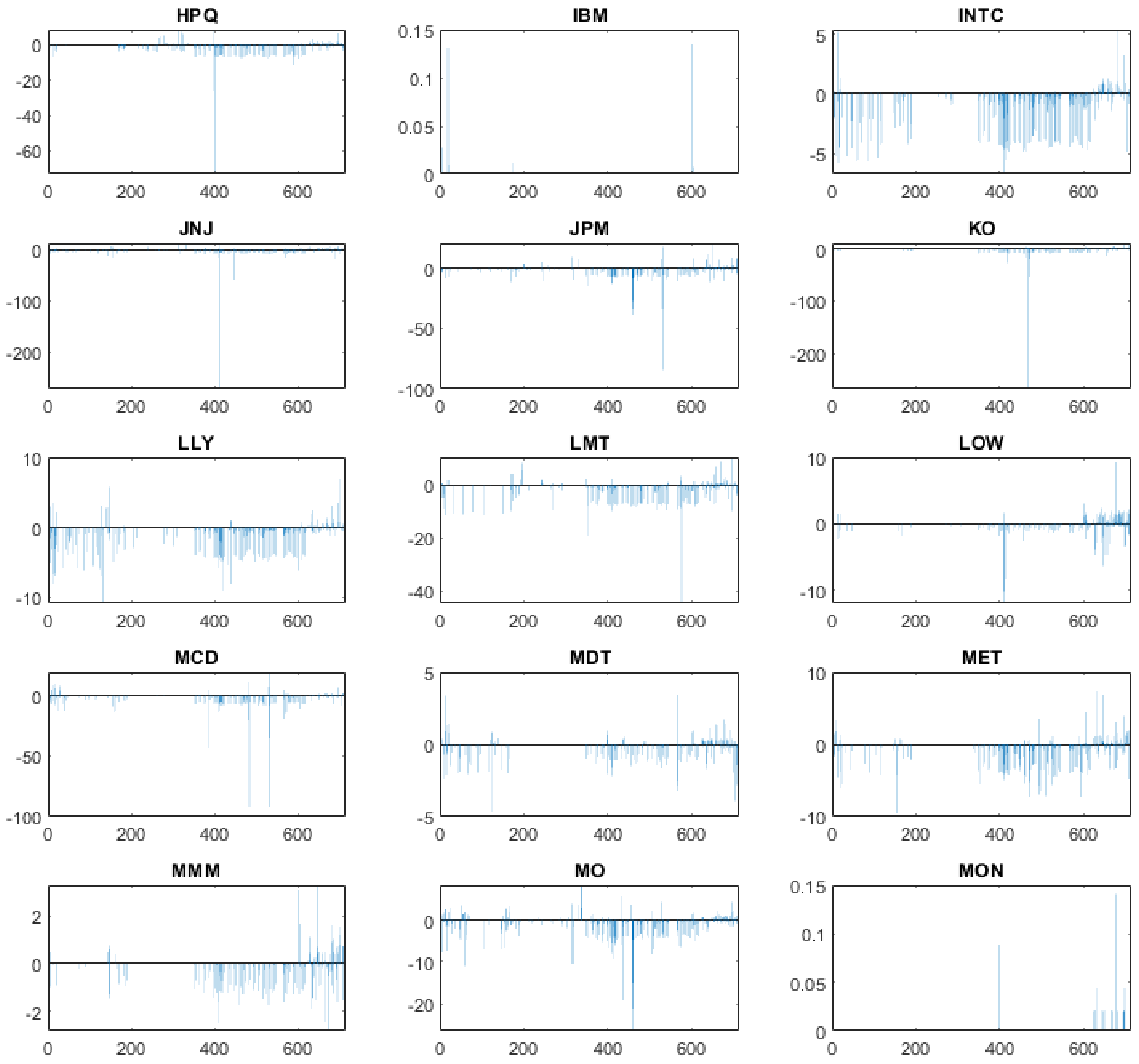

Figure A.4 - Plots displaying the representation of the coefficients resulting from Elastic Net Regularization, then confirmed with Adaptive Lasso, for stocks going from HPQ to MON. ***Source****: own elaboration.*

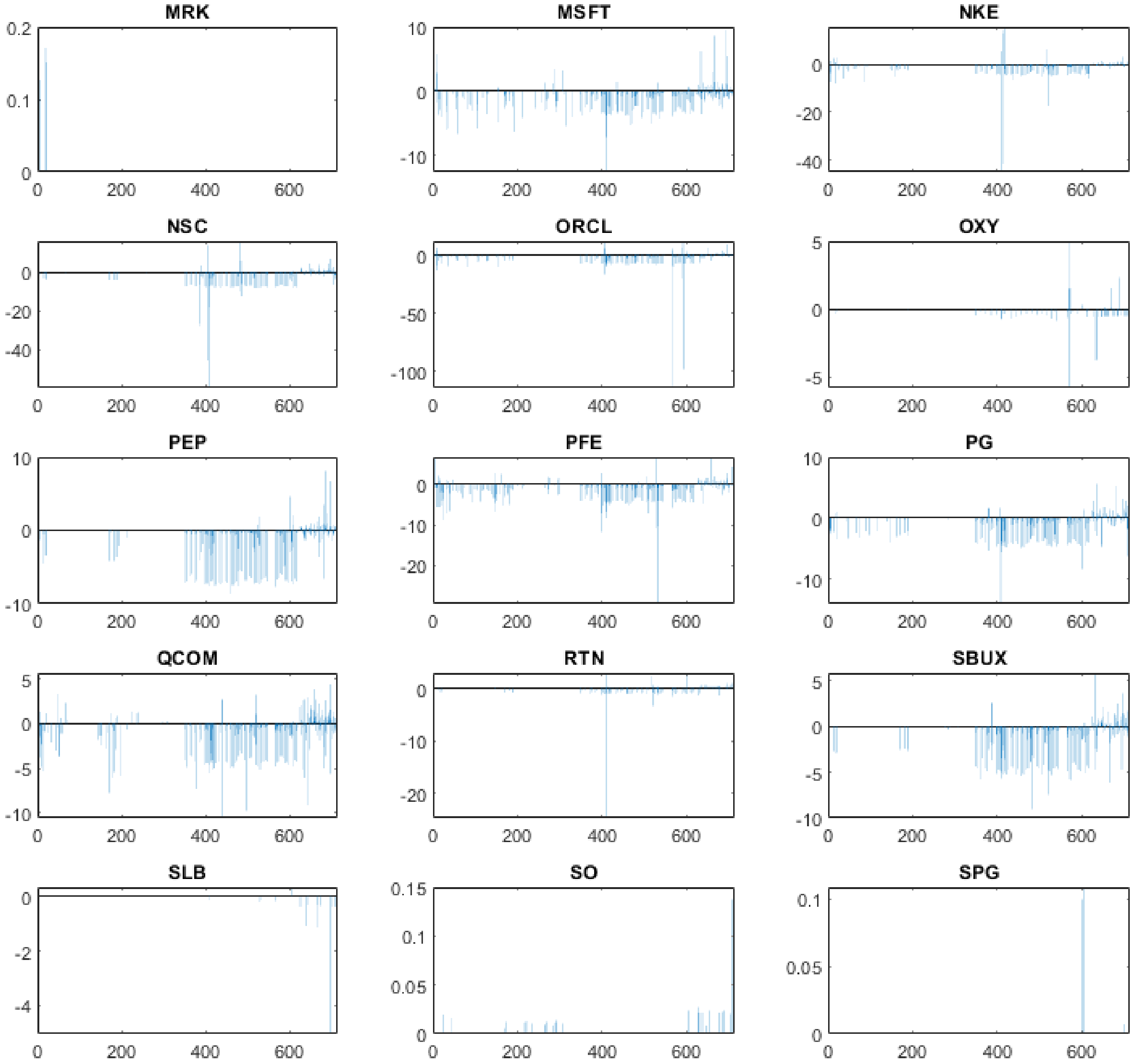

Figure A.5 - Plots displaying the representation of the coefficients resulting from Elastic Net Regularization, then confirmed with Adaptive Lasso, for stocks going from MRK to SPG. ***Source****: own elaboration.*

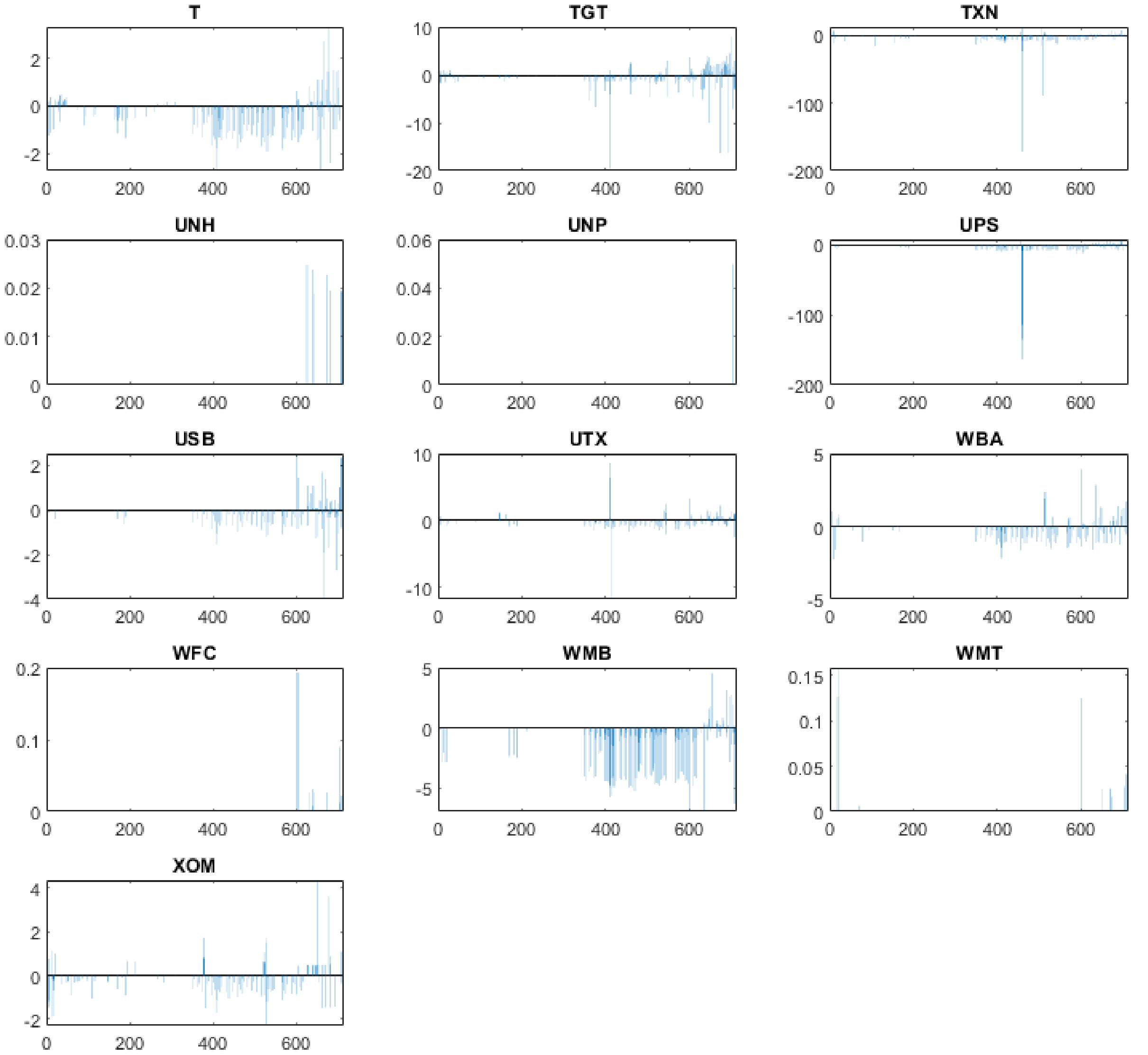

Figure A.6 - Plots displaying the representation of the coefficients resulting from Elastic Net Regularization, then confirmed with Adaptive Lasso, for stocks going from T to the last asset. ***Source****: own elaboration.*

8.2. Appendix B: Comparison between means of coefficient, resulting from penalised regression on positive and negative jumps only.

Table B.1 - Comparison between coefficients means resulting from Adaptive Lasso regularization performed on positive and negative time series. For each asset, on the first two rows are reported the mean for coefficients for positive jumps only (respectively, mean over all assets and over the ones belonging to the same sector), while on last two rows are displayed the mean for negative jumps only. ***Source:*** *own elaboration.*

Sector: Consumer Goods

	AAPL	**CL**	**KO**	**MO**	**NKE**	**PEP**	**PG**	
All (pos)	0,2882	0,2837	0,2541	0,2446	0,2633	0,2575	0,2742	0,2882
Sector	2,5841	2,5047	2,6640	2,5650	2,5507	2,6621	2,5833	2,5841
All (neg)	-0,0078	-0,0186	-0,0119	-0,0079	-0,0158	-0,0108	-0,0099	-0,0078
Sector	-0,0978	-0,0944	-0,0970	-0,0984	-0,0943	-0,0983	-0,1015	-0,0978

Sector: Basic Materials

	APA	**APC**	**BHI**	**COP**	**CVX**	**DD**	**DOW**	**FCX**
All (pos)	0,3076	0,2411	0,2396	0,2779	0,2947	0,3158	0,2567	0,2622
Sector	1,5122	1,5152	1,3762	1,5575	1,3886	1,4629	1,4016	1,3933
All (neg)	-0,1191	-0,1867	-0,0158	-0,0651	-0,0164	-0,0131	-0,0086	-0,0086
Sector	-0,4958	-0,9806	-0,0676	-0,4092	-0,0560	-0,0521	-0,0497	-0,0495

	HAL	**MON**	**OXY**	**SLB**	**WMB**	**XOM**
All (pos)	0,2129	0,2644	0,2482	0,1298	0,2642	0,3081
Sector	1,3380	1,2877	1,3808	1,4384	1,3070	1,2701
All (neg)	-0,0080	-0,0119	-0,2575	-0,0206	-0,0133	-0,0128
Sector	-0,0505	-0,0538	-1,7699	-0,0837	-0,0514	-0,0467

Sector: Financial

	AIG	**ALL**	**AXP**	**BK**	**BRKB**	**C**	**COF**	**GS**
All (pos)	0,2833	0,2987	0,2776	0,2995	0,2310	0,3347	0,3080	0,2725
Sector	1,6193	1,5863	1,6448	1,6245	1,4997	1,6086	1,6092	1,5904
All (neg)	-0,0086	-0,0111	-0,0092	-0,0118	-0,0079	-0,0114	-0,0115	-0,0189
Sector	-0,0559	-0,0645	-0,0614	-0,0632	-0,0536	-0,0747	-0,0710	-0,0947

	JPM	**MET**	**SPG**	**USB**	**WFC**
All (pos)	0,2944	0,2928	0,2911	0,2830	0,3230
Sector	1,8007	1,6896	1,6054	1,6064	1,5649
All (neg)	-0,0157	-0,0079	-0,0159	-0,0284	-0,0136

Sector	-0,0866	-0,0537	-0,0812	-0,1269	-0,0831			
Sector: Healthcare								
	ABT	**AMGN**	**BAX**	**BIIB**	**BMY**	**CELG**	**CVS**	**GILD**
All (pos)	0,2873	0,3200	0,2608	0,2685	0,3325	0,2644	0,2917	0,2626
Sector	1,3682	1,6210	1,3178	1,4396	1,4391	1,4871	1,3909	1,4070
All (neg)	-0,0126	-0,0078	-0,0087	-0,0078	-0,0079	-0,0079	-0,0083	-0,0091
Sector	-0,0507	-0,0493	-0,0496	-0,0490	-0,0498	-0,0498	-0,0497	-0,0495
	JNJ	**LLY**	**MDT**	**MRK**	**PFE**	**UNH**		
All (pos)	0,3375	0,3078	0,2799	0,2549	0,2878	0,2982		
Sector	1,5070	1,4260	1,4165	1,4301	1,5181	1,3901		
All (neg)	-0,0115	-0,0079	-0,0119	-0,0078	-0,0106	-0,0079		
Sector	-0,0489	-0,0498	-0,0499	-0,0490	-0,0523	-0,0498		
Sector: Industrial Goods								
	BA	**CAT**	**EMR**	**GD**	**GE**	**HON**	**LMT**	**MMM**
All (pos)	0,2962	0,2656	0,3226	0,3066	0,2818	0,2897	0,2550	0,2629
Sector	2,0054	2,0066	1,9097	2,0180	1,8543	2,1473	1,9443	1,7904
All (neg)	-0,0114	-0,0121	-0,0172	-0,0182	-0,0150	-0,0116	-0,0153	-0,0091
Sector	-0,0710	-0,0750	-0,0836	-0,0817	-0,0905	-0,0764	-0,0687	-0,0763
	RTN	**UTX**						
All (pos)	0,2735	0,3338						
Sector	1,7579	1,9987						
All (neg)	-0,0109	-0,0123						
Sector	-0,0777	-0,0813						
Sector: Services								
	AMZN	**CMCSA**	**COST**	**DIS**	**EBAY**	**FDX**	**HD**	**LOW**
All (pos)	0,2768	0,3048	0,3358	0,3120	0,2715	0,2713	0,3402	0,2756
Sector	1,3132	1,2050	1,3673	1,2689	1,1968	1,2672	1,4911	1,2471
All (neg)	-0,0091	-0,0085	-0,0095	-0,0139	-0,0134	-0,0107	-0,2202	-0,0325
Sector	-0,0488	-0,0434	-0,0464	-0,0487	-0,0565	-0,0506	-0,7623	-0,0621
	MCD	**NSC**	**SBUX**	**TGT**	**UNP**	**UPS**	**WBA**	**WMT**
All (pos)	0,2553	0,2606	0,2601	0,2803	0,2507	0,3729	0,2657	0,2931
Sector	1,1440	1,2509	1,2283	1,3268	1,2102	1,4240	1,2560	1,1961
All (neg)	-0,0081	-0,0079	-0,0105	-0,0155	-0,0079	-0,0126	-0,0122	-0,0099
Sector	-0,0434	-0,0436	-0,0495	-0,0600	-0,0436	-0,0476	-0,0419	-0,0452

Sector: Technologies

	ACN	**CSCO**	**EMC**	**HPQ**	**IBM**	**INTC**	**MSFT**	**ORCL**
All (pos)	0,2607	0,2990	0,2434	0,2695	0,2936	0,2460	0,2961	0,2422
Sector	1,6752	1,7285	1,6584	1,6692	1,6381	1,7061	1,5543	1,6821
All (neg)	-0,0082	-0,0933	-0,0086	-0,0105	-0,0170	-0,0078	-0,0201	-0,0110
Sector	-0,0654	-0,5665	-0,0632	-0,0708	-0,0741	-0,0627	-0,0812	-0,0624
	QCOM	**T**	**TXN**					
All (pos)	0,2625	0,2930	0,2503					
Sector	1,6978	1,5565	1,7900					
All (neg)	-0,0101	-0,0096	-0,0103					
Sector	-0,0655	-0,0623	-0,0625					

Sector: Utilities

	AEP	**EXC**	**SO**
All (pos)	0,2760	0,3258	0,3028
Sector	6,6797	6,5818	6,6188
All (neg)	-0,0078	-0,0102	-0,1935
Sector	-0,2276	-0,2291	-3,5916

8.3. Appendix C: Comparison between results from penalised regressions, divided for category of covariates.

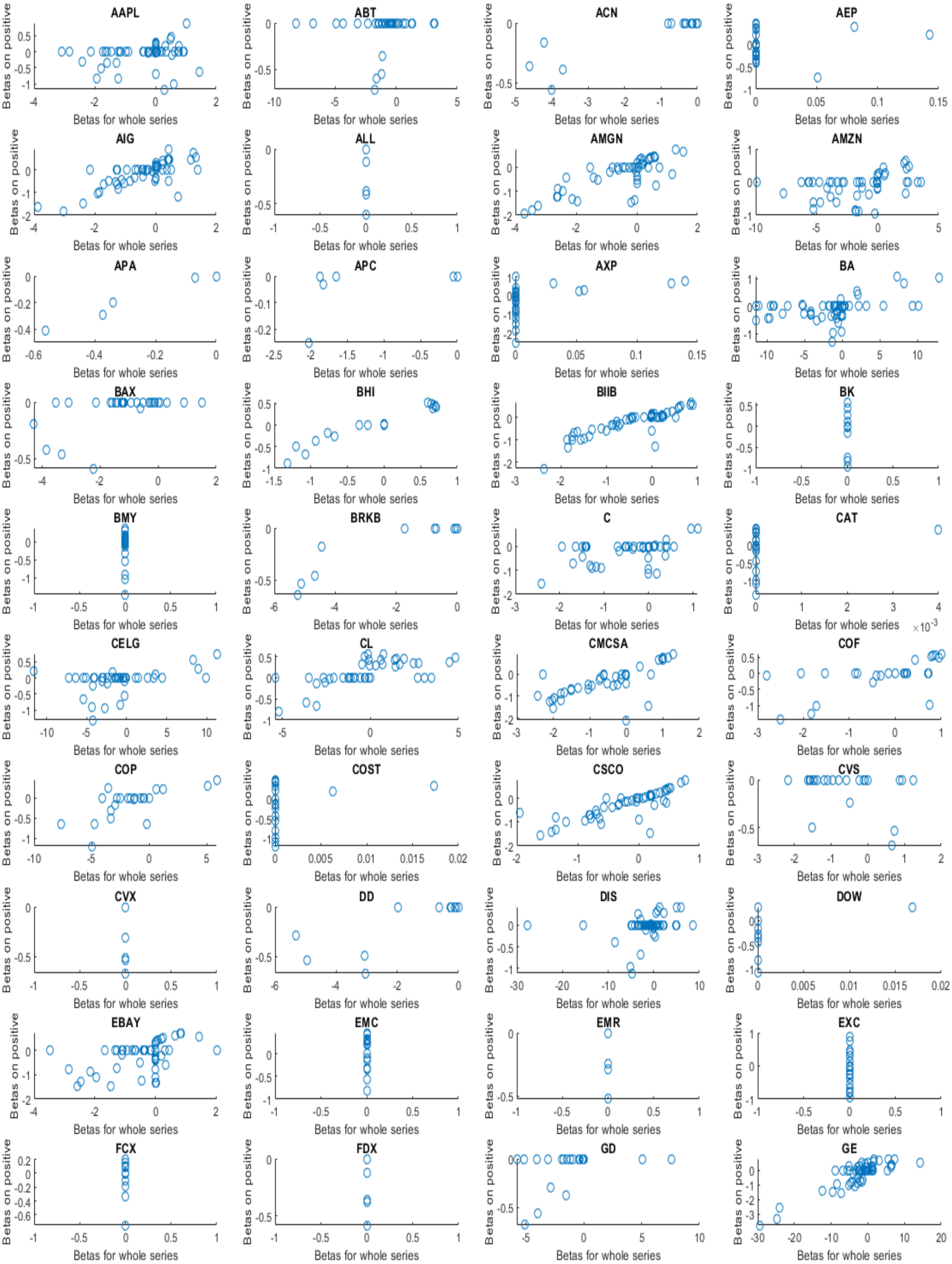

Figure C.1 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from AAPL to GE. The coefficients refer to StreetAccount measures only. ***Source:*** *own elaboration.*

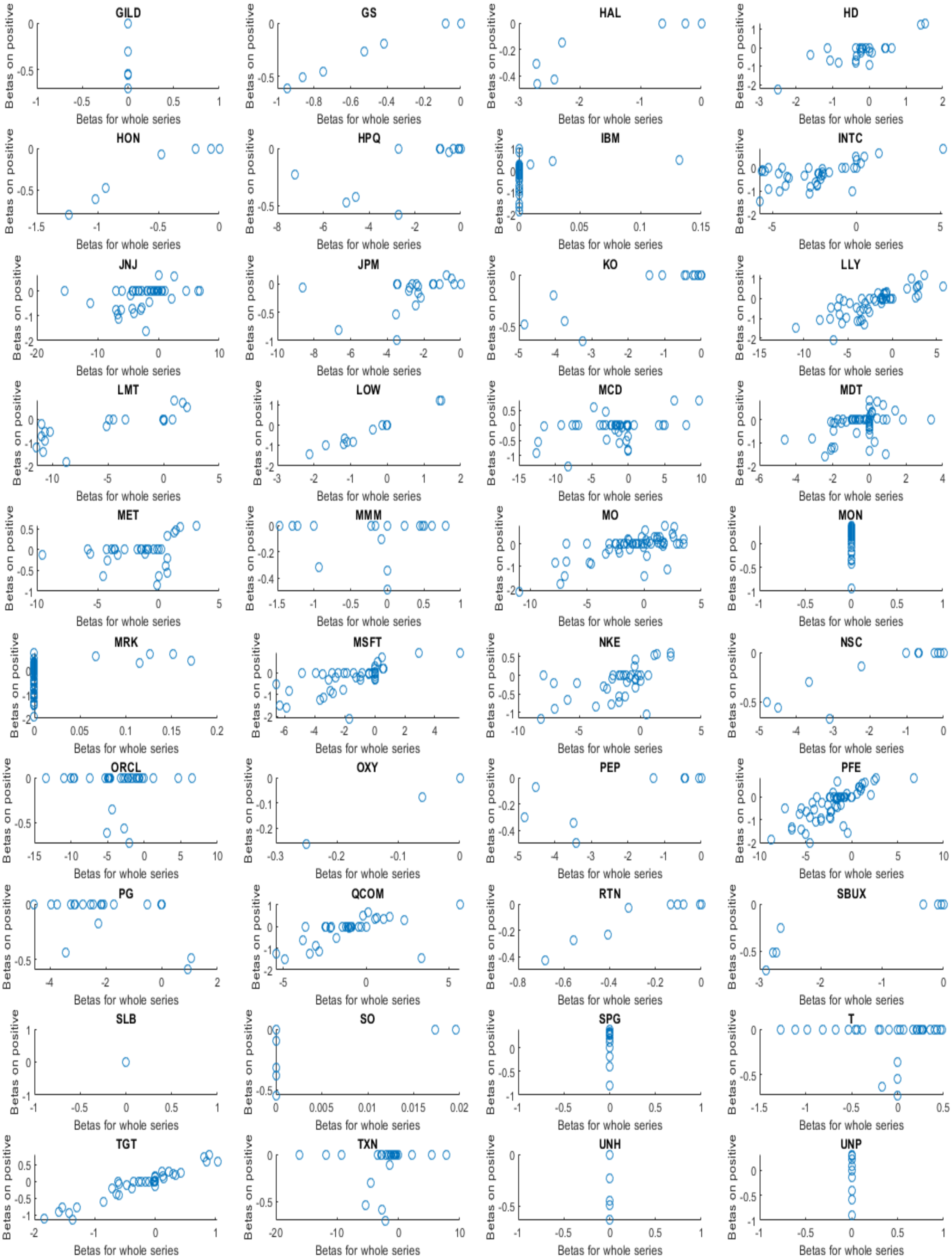

Figure C.2 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from GILD to UNP. The coefficients refer to StreetAccount measures only. ***Source:*** *own elaboration.*

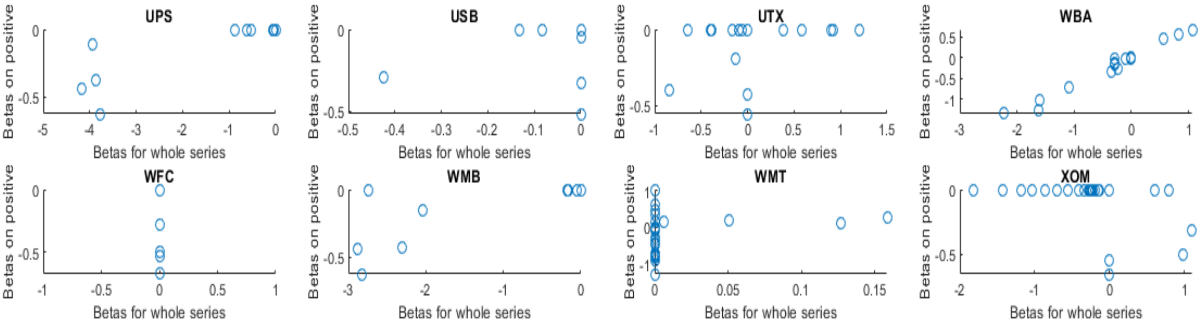

Figure C.3 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from UPS to XOM. The coefficients refer to StreetAccount measures only. ***Source:*** *own elaboration.*

.

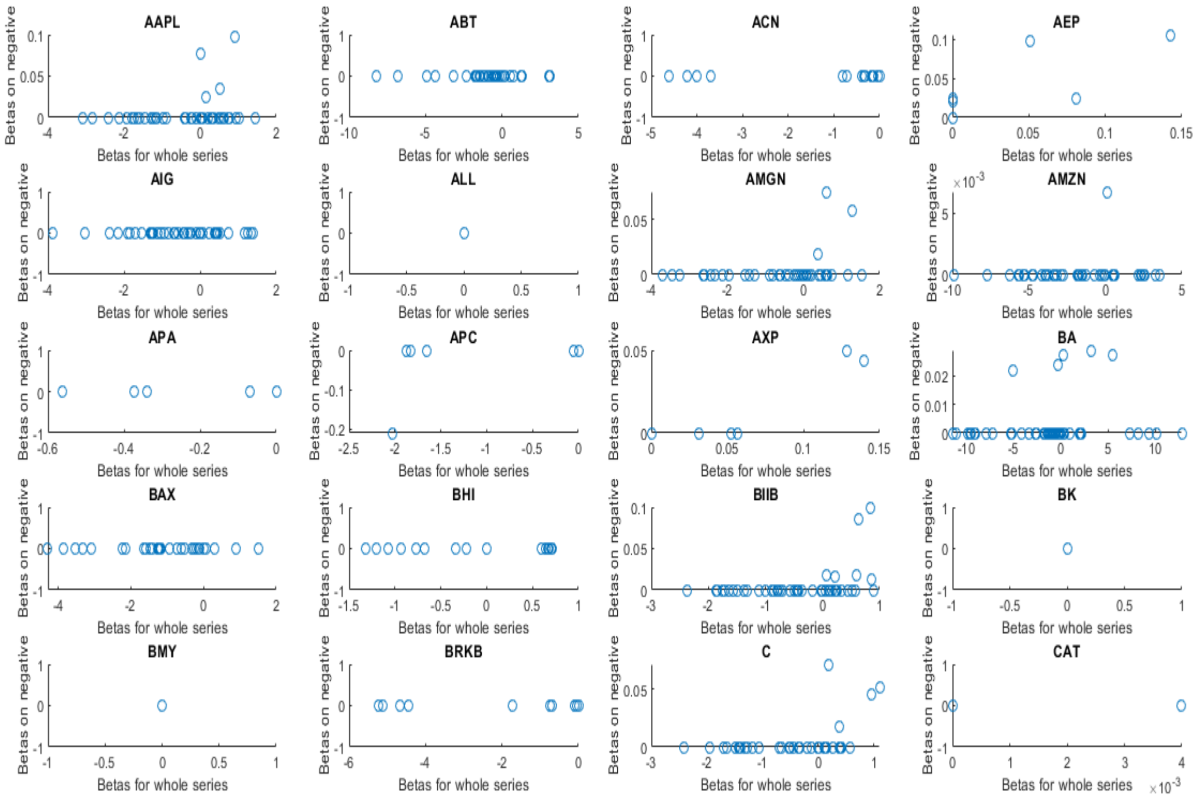

Figure C.4 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from AAPL to CAT. The coefficients refer to StreetAccount measures only. ***Source:*** *own elaboration.*

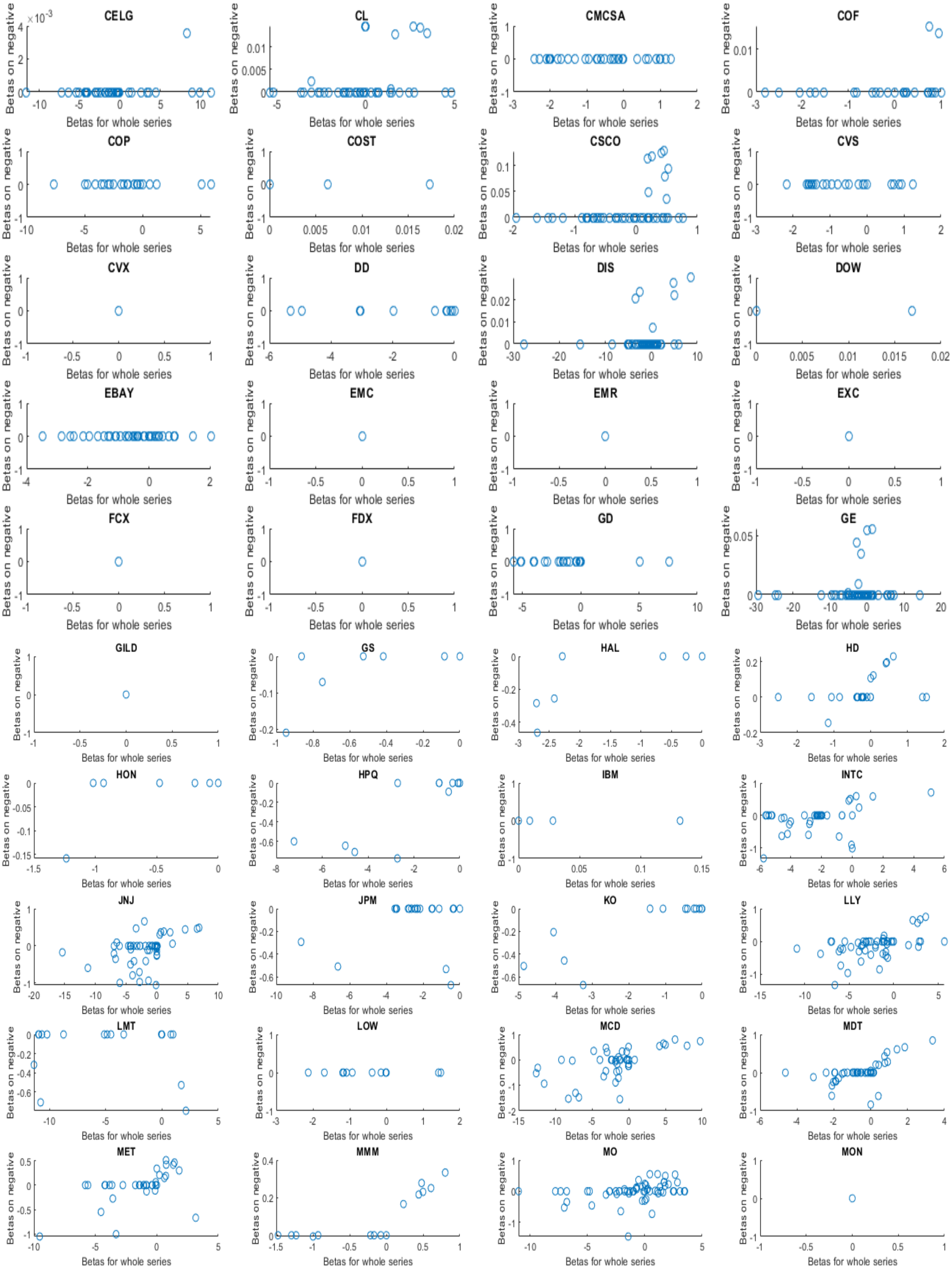

Figure C.5 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from CELG to MON. The coefficients refer to StreetAccount measures only. ***Source:*** *own elaboration.*

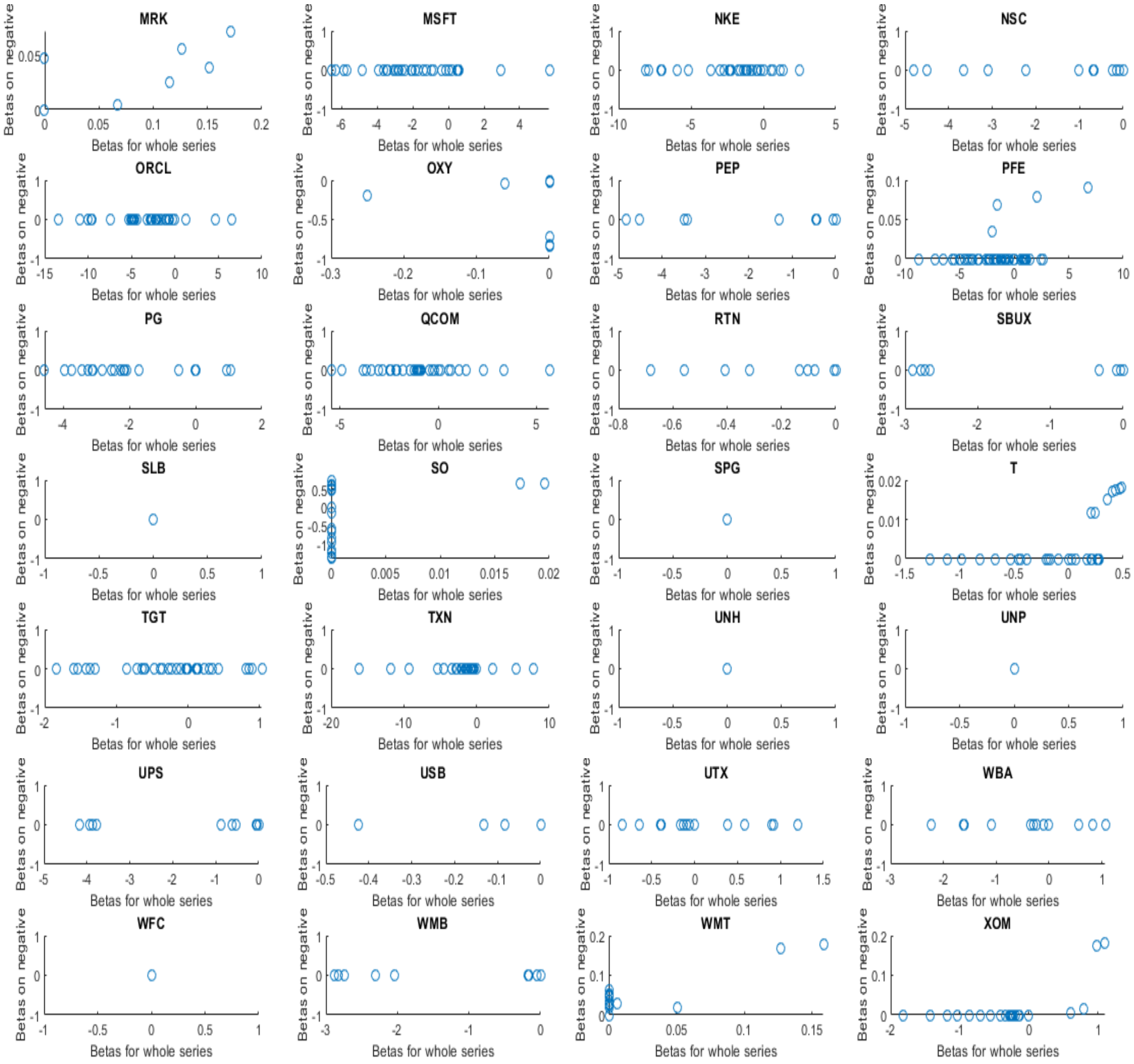

Figure C.6 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from MRK to XOM. The coefficients refer to StreetAccount measures only. ***Source:*** *own elaboration.*

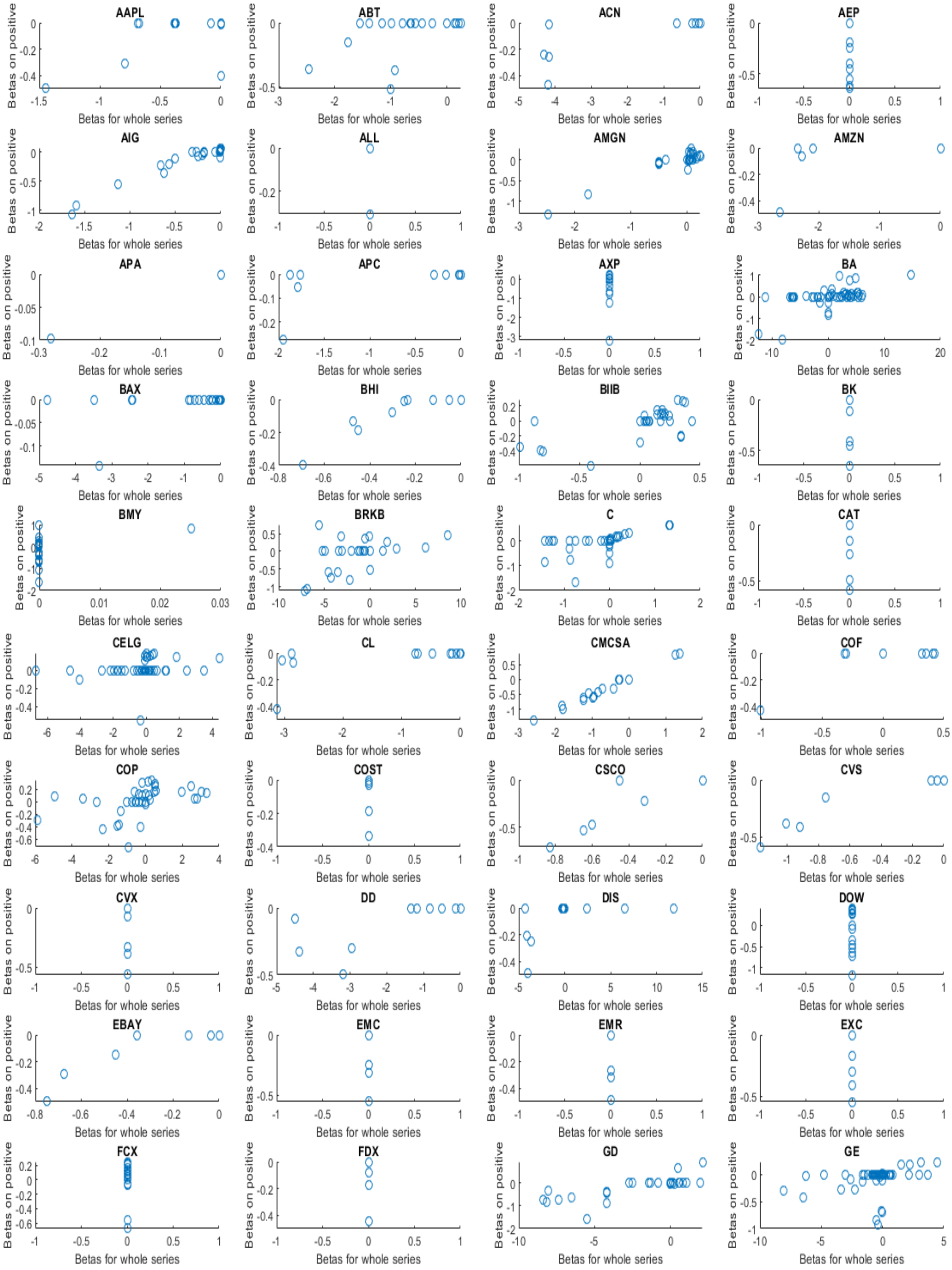

Figure C.7 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from AAPL to GE. The coefficients refer to T. Reuters measures only. ***Source:*** *own elaboration.*

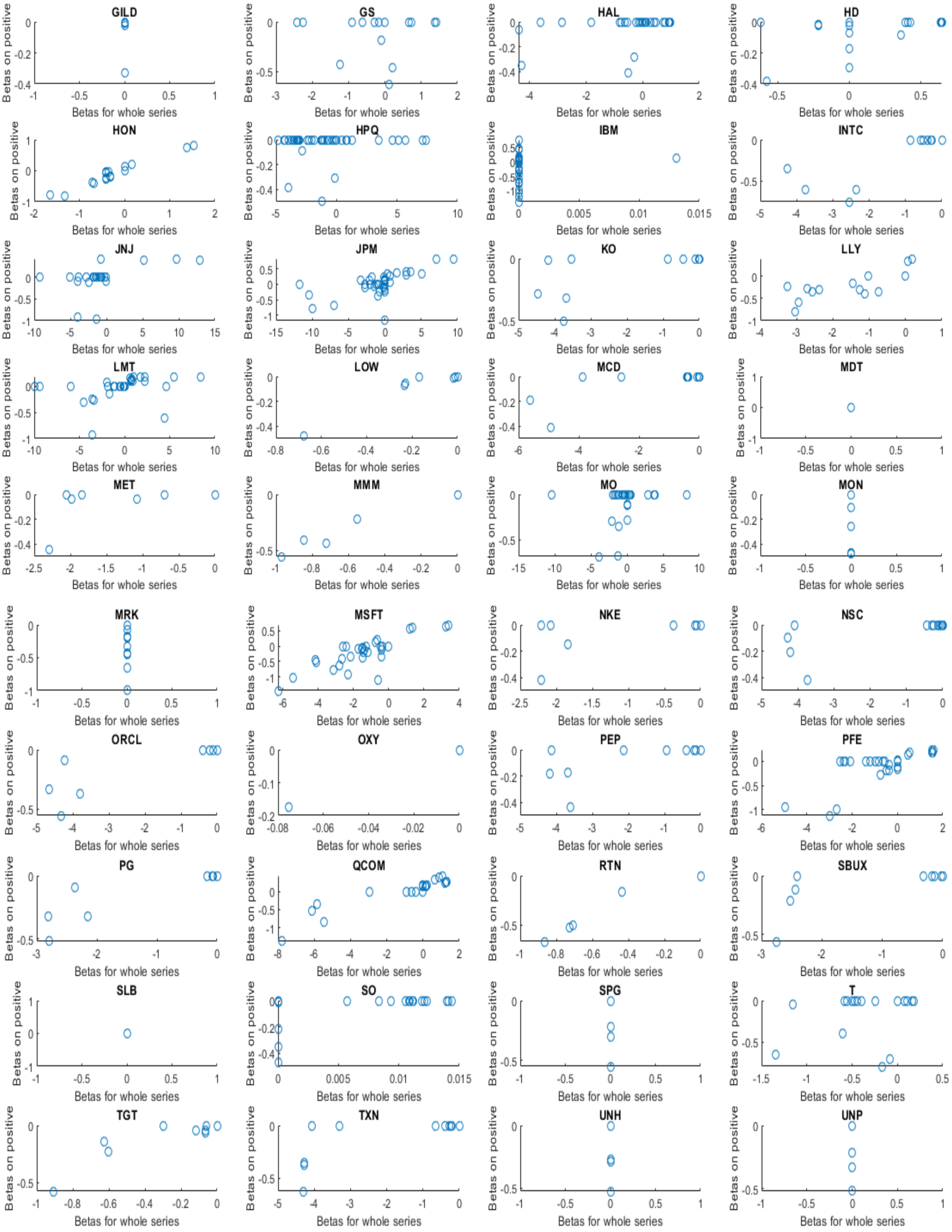

Figure C.8 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from GILD to UNP. The coefficients refer to T. Reuters measures only. ***Source:*** *own elaboration.*

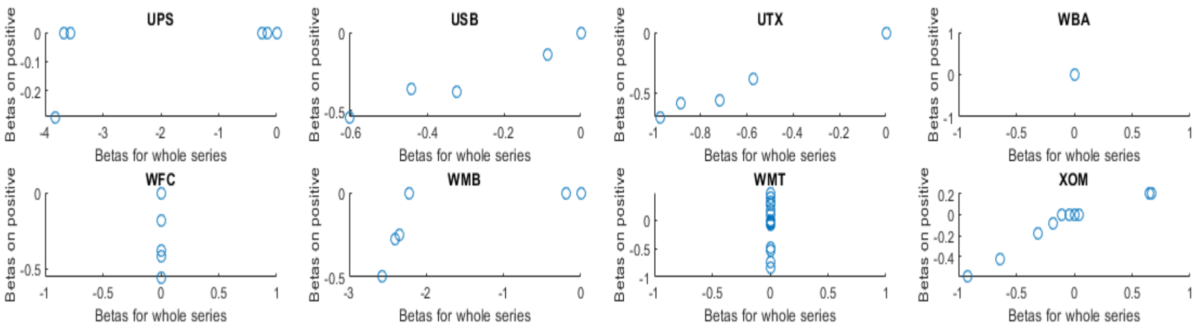

Figure C.9 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from UPS to XOM. The coefficients refer to T. Reuters measures only. ***Source:*** *own elaboration.*

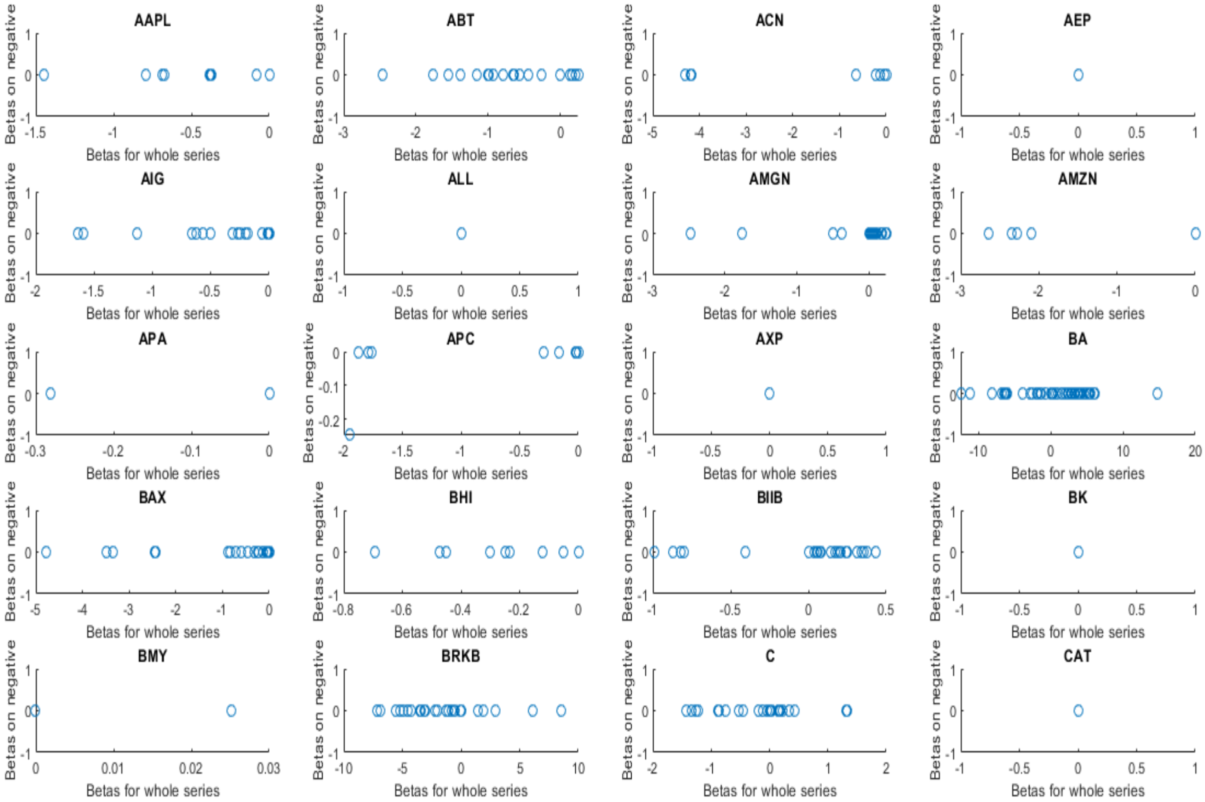

Figure C.10 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from AAPL to CAT. The coefficients refer to T. Reuters measures only. ***Source:*** *own elaboration.*

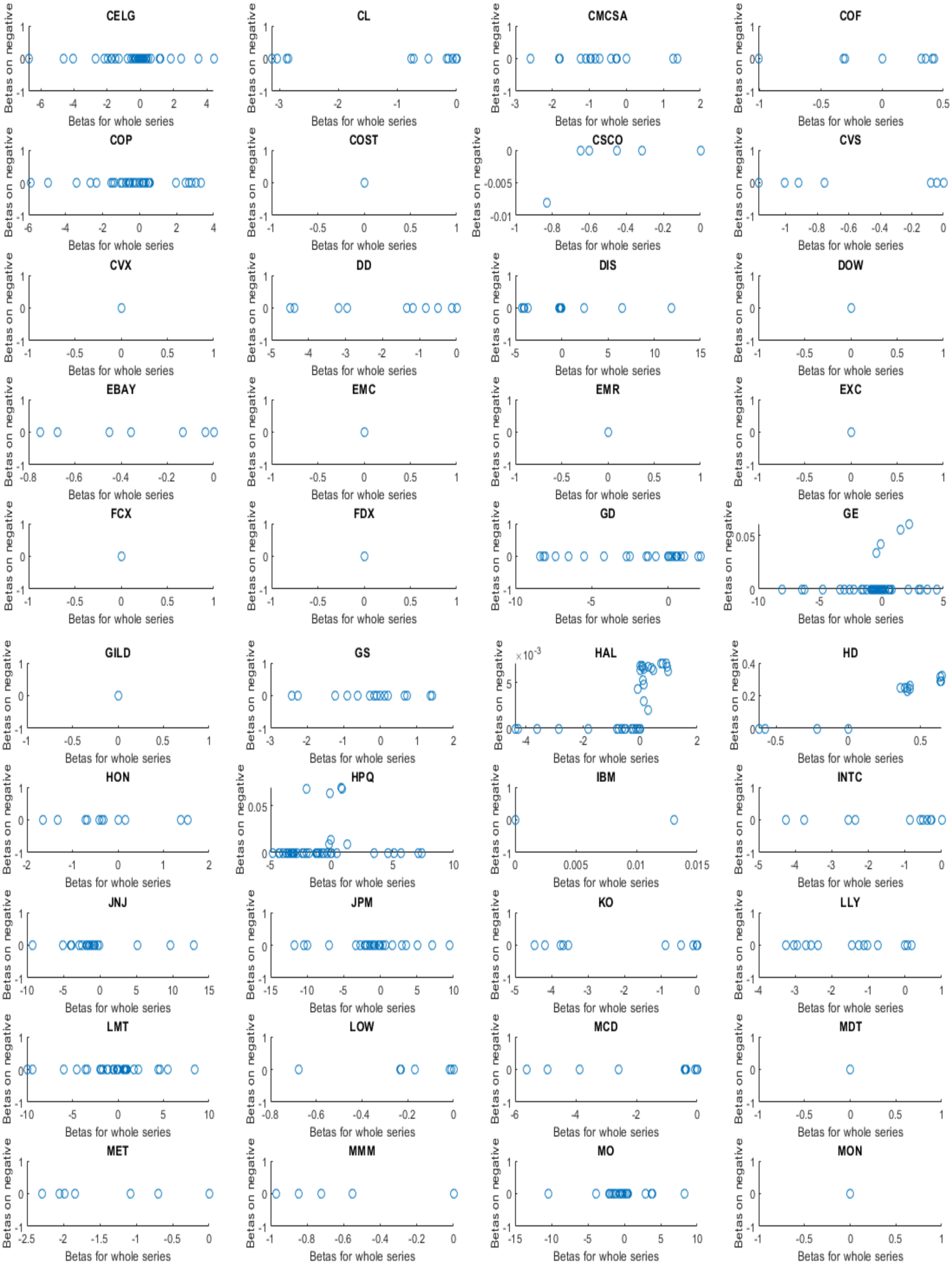

Figure C.11 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from CELG to MON. The coefficients refer to T. Reuters measures only. ***Source:*** *own elaboration.*

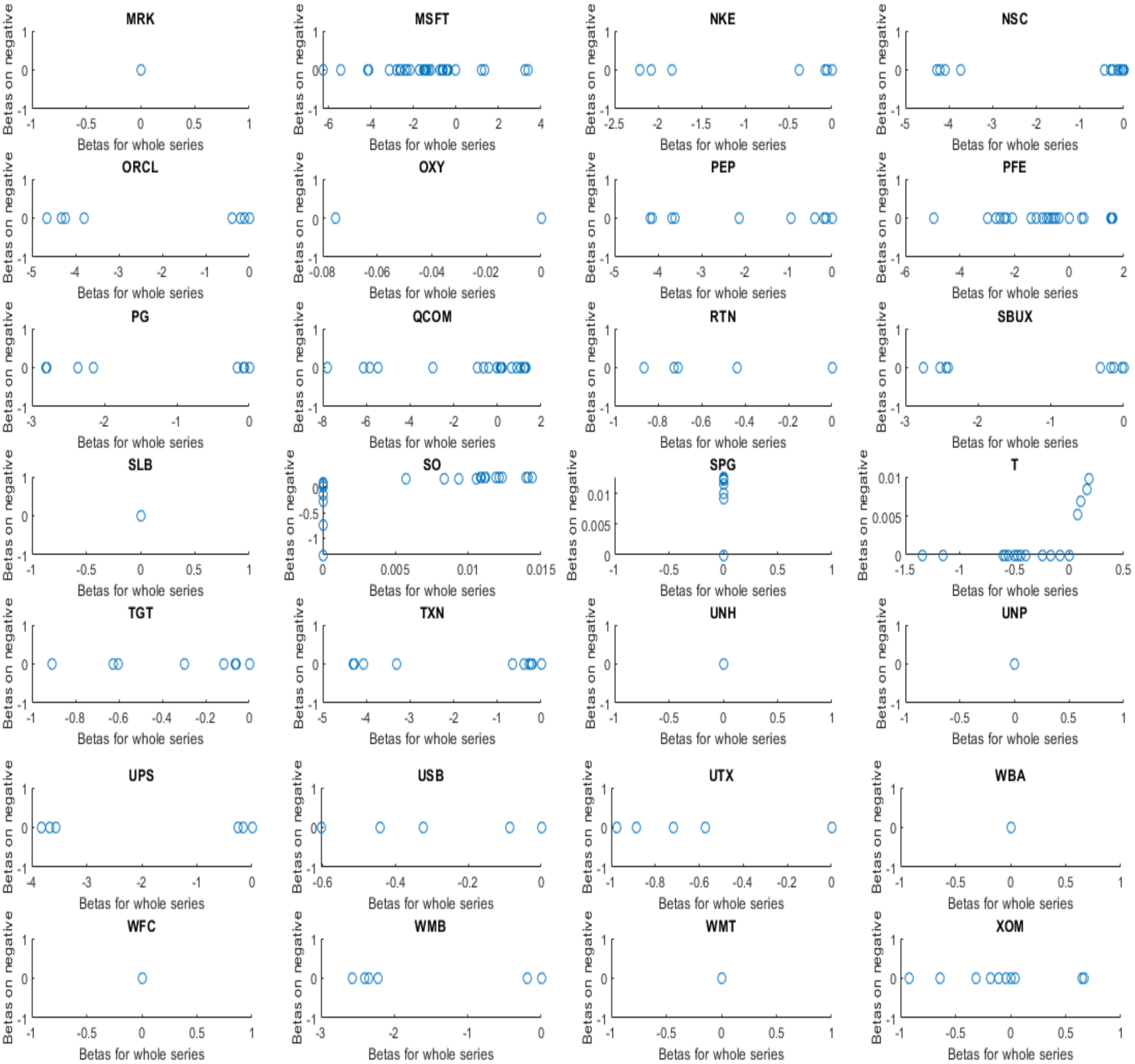

Figure C.12 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from MRK to XOM. The coefficients refer to T. Reuters measures only. ***Source:*** *own elaboration.*

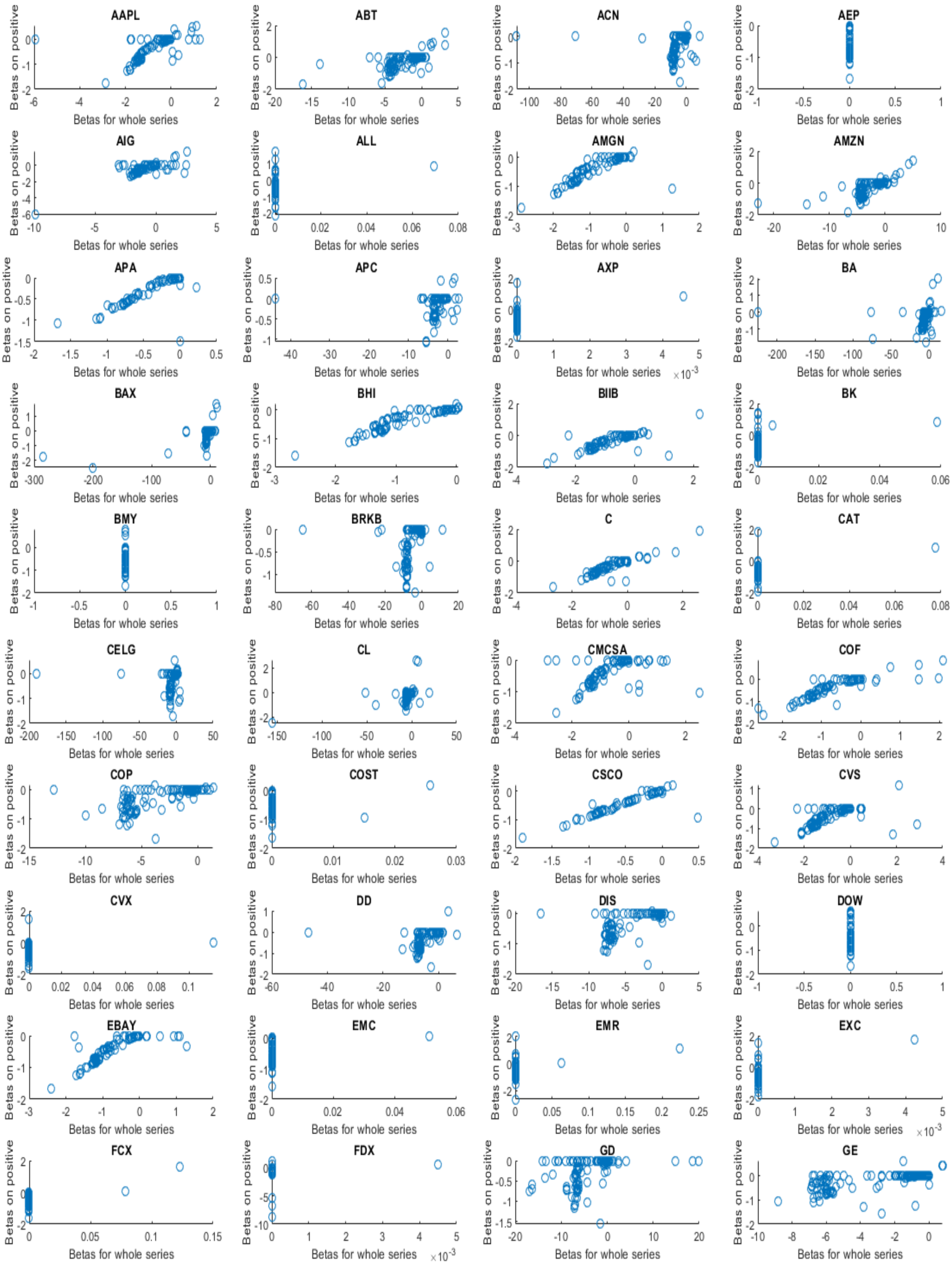

Figure C.13 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from AAPL to GE. The coefficients refer to Macroeconomic news measures only. ***Source:*** *own elaboration.*

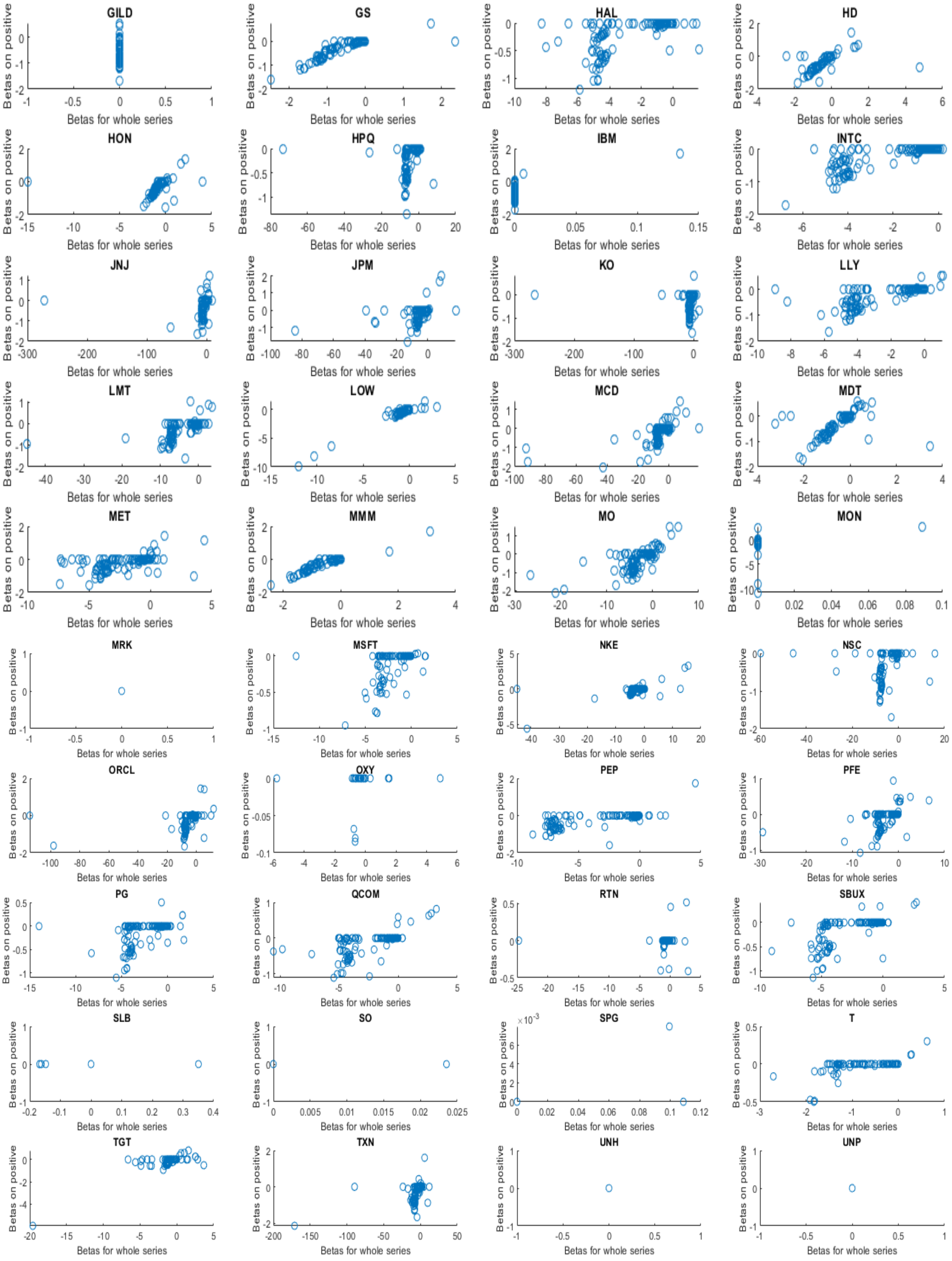

Figure C.14 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from GILD to UNP. The coefficients refer to Macroeconomic news measures only. ***Source:*** *own elaboration.*

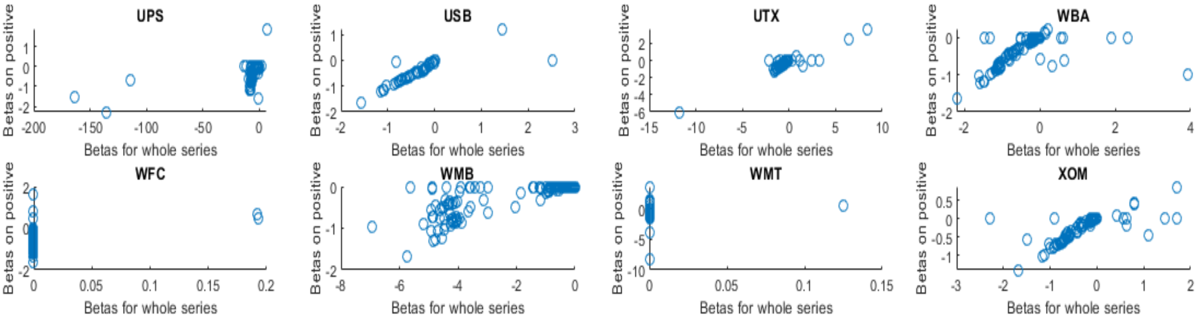

Figure C.15 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from UPS to XOM. The coefficients refer to Macroeconomic news measures only. ***Source:*** *own elaboration.*

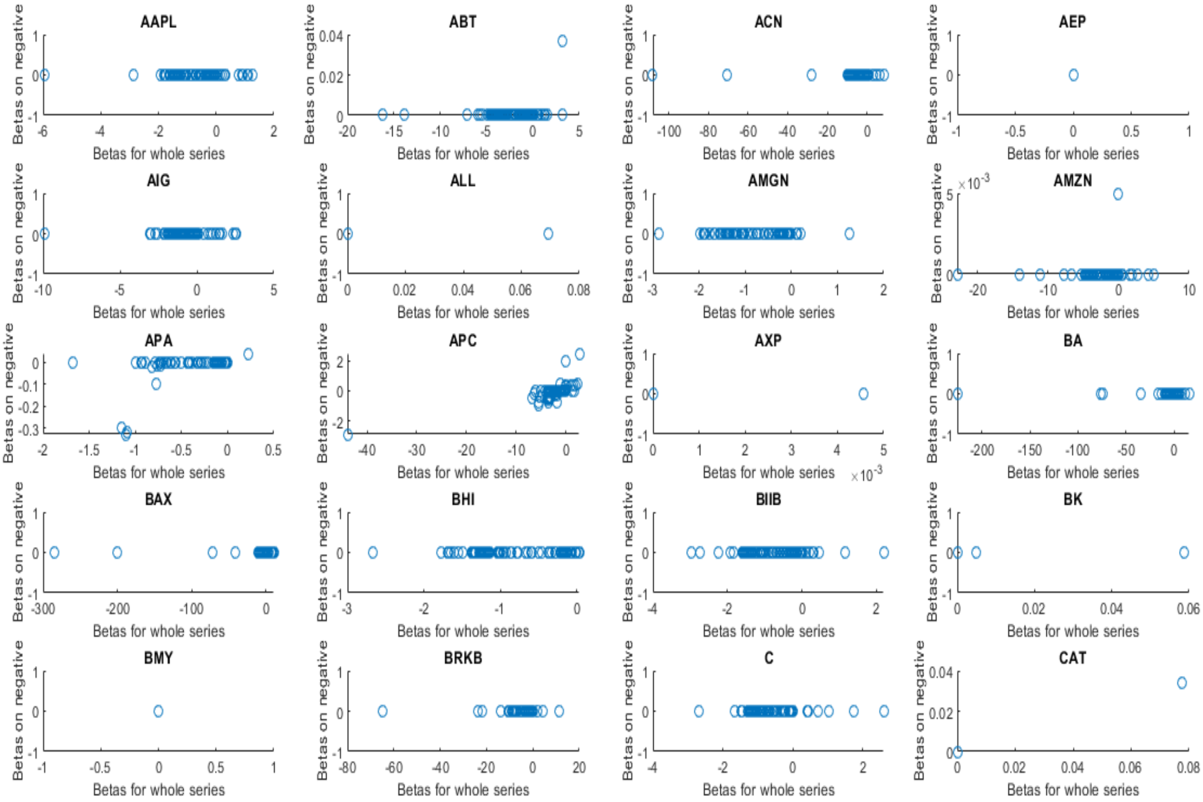

Figure C.16 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from AAPL to CAT. The coefficients refer to Macroeconomic news measures only. ***Source:*** *own elaboration.*

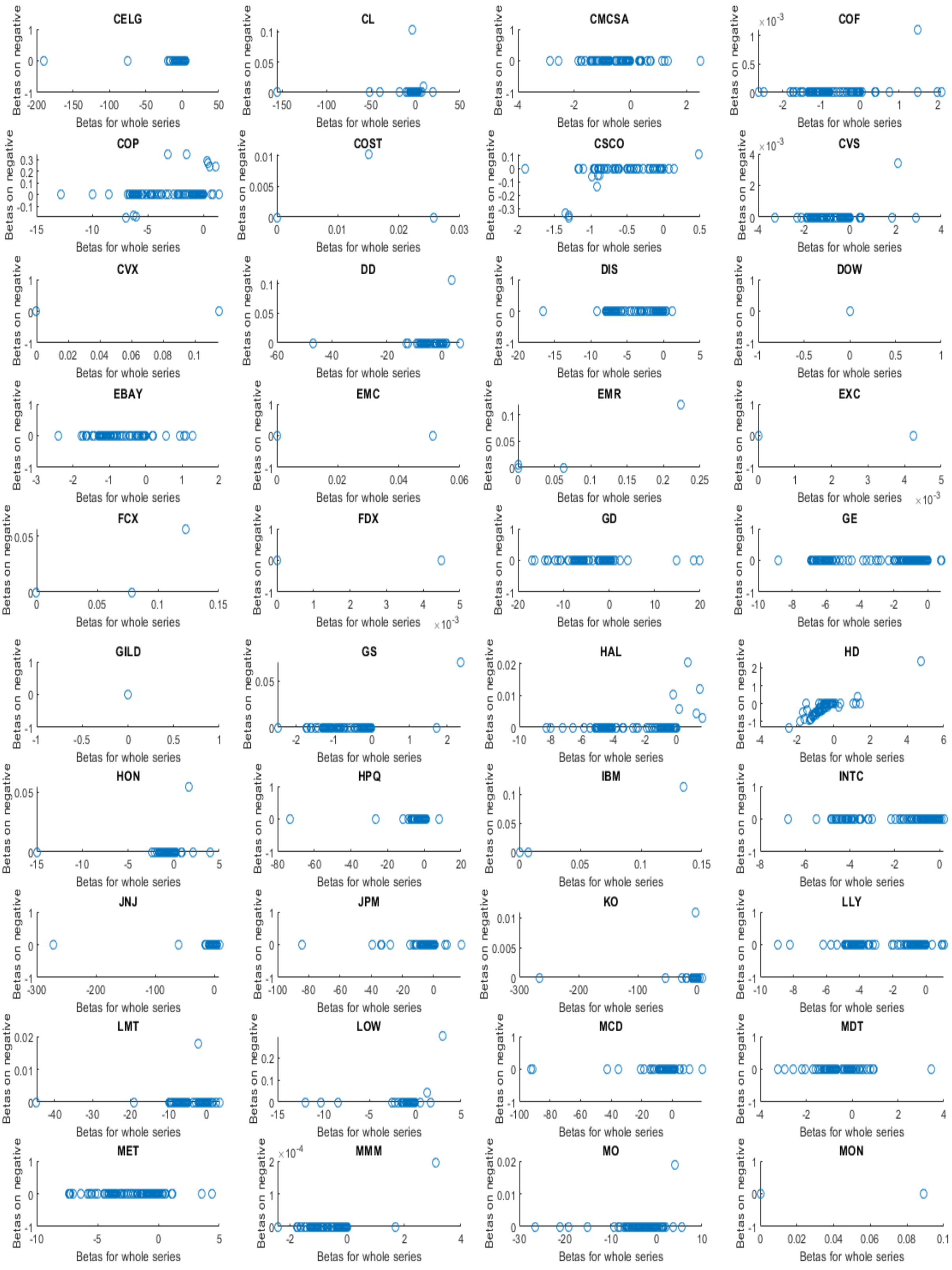

Figure C.17 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from CELG to MON. The coefficients refer to Macroeconomic news measures only. ***Source:*** *own elaboration.*

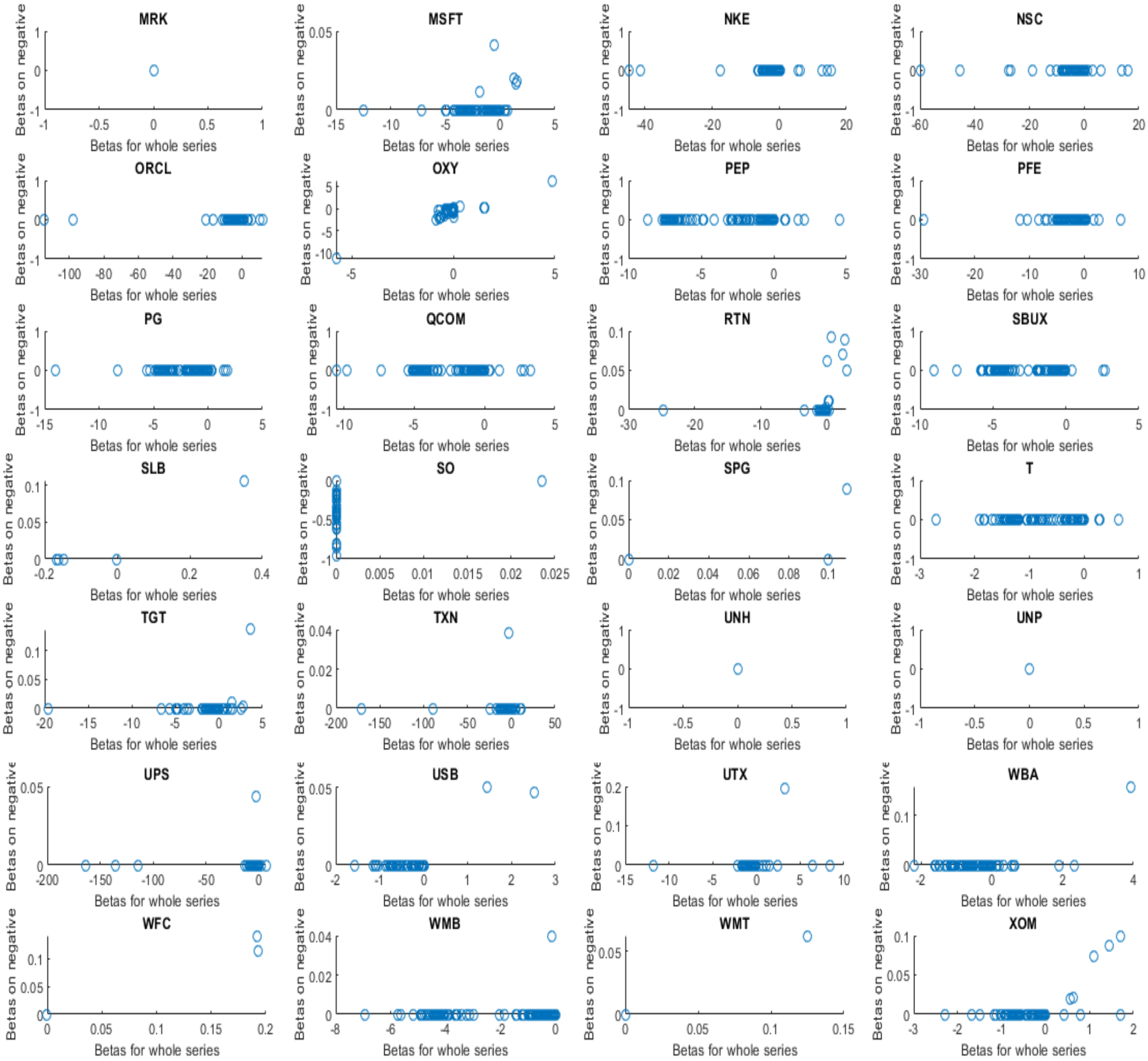

Figure C.18 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from MRK to XOM. The coefficients refer to Macroeconomic news measures only. ***Source:*** *own elaboration.*

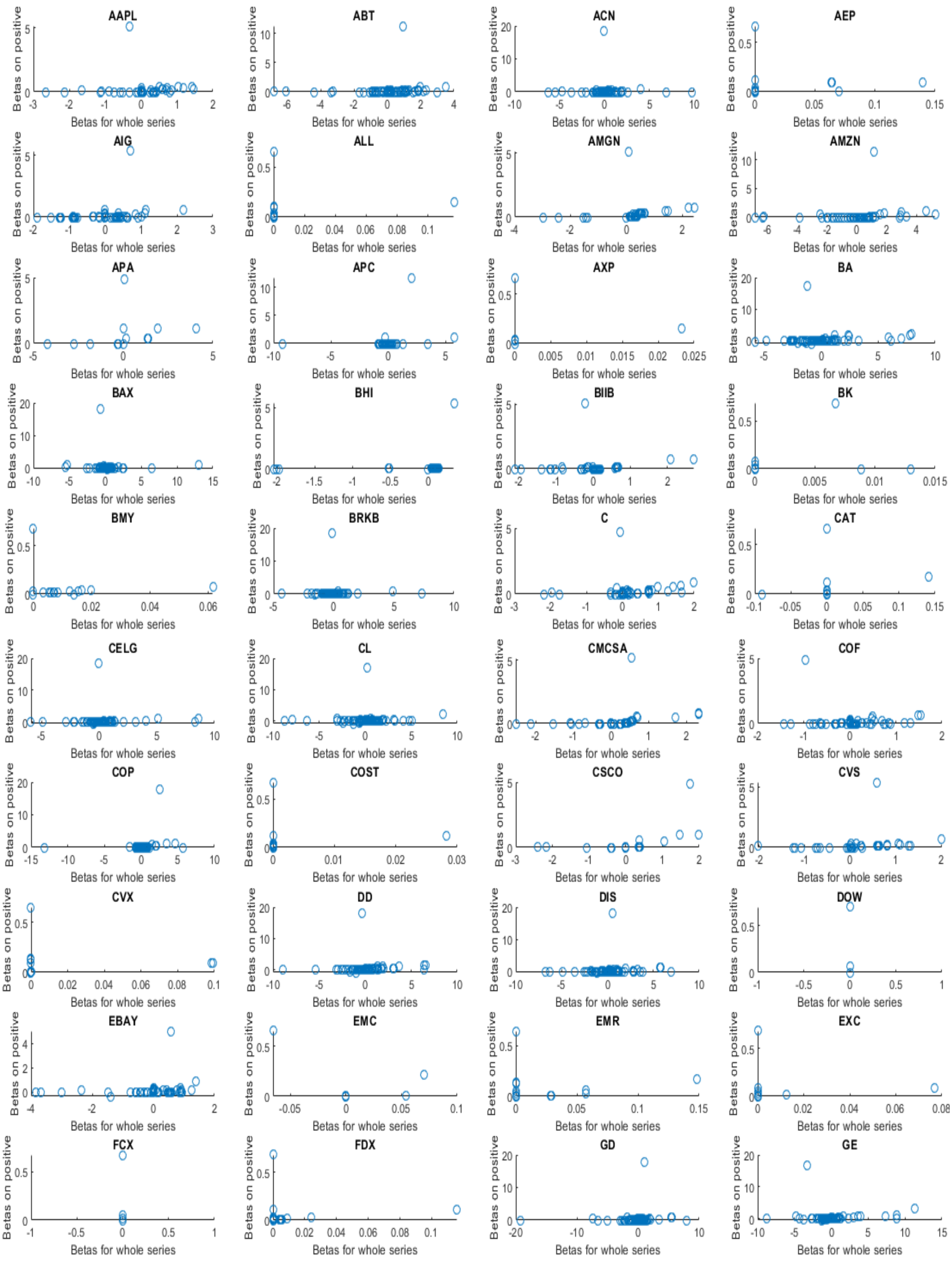

Figure C.19 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from AAPL to GE. The coefficients refer to other stocks included as measure. ***Source:*** *own elaboration.*

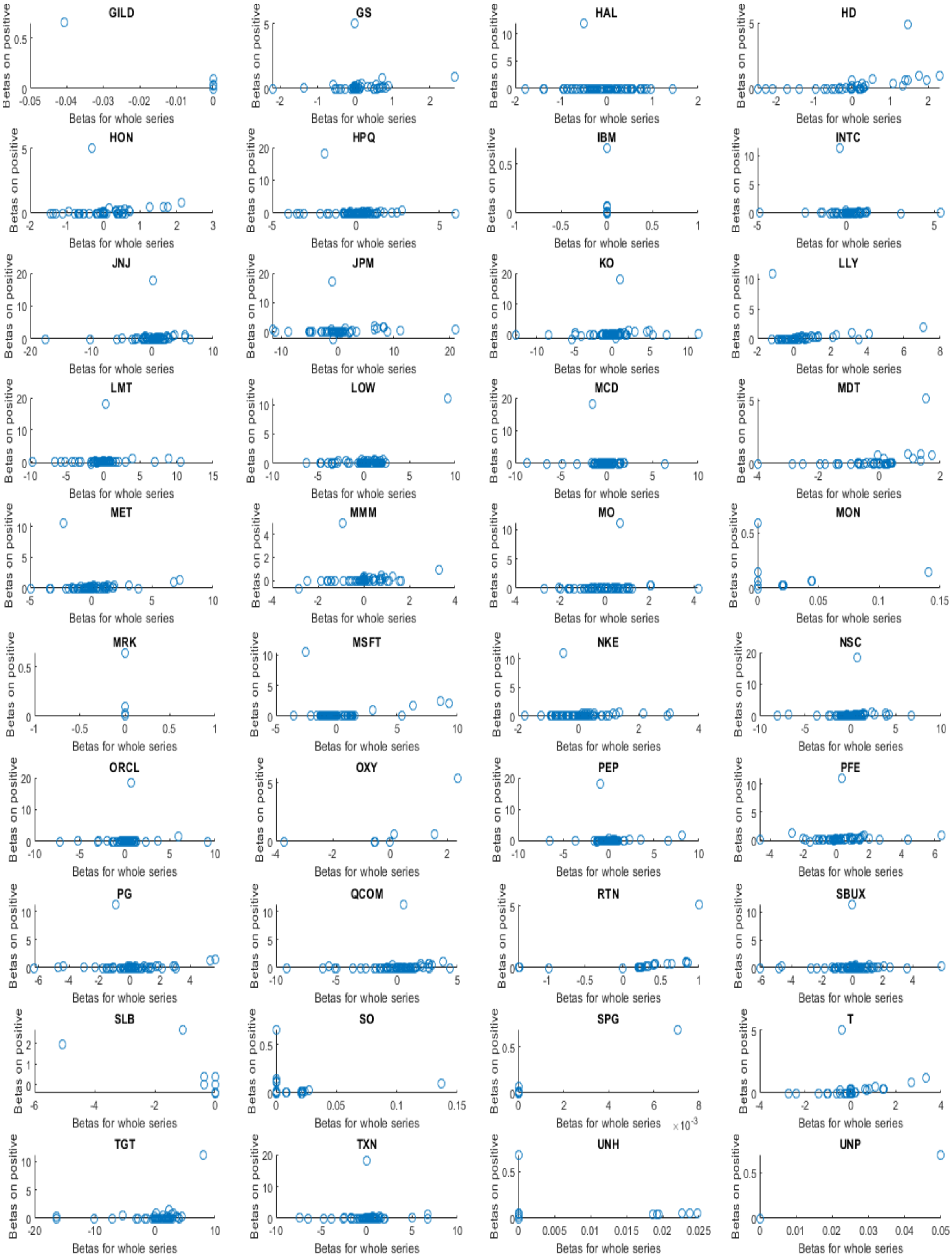

Figure C.20 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from GILD to UNP. The coefficients refer to other stocks included as measure. ***Source:*** *own elaboration.*

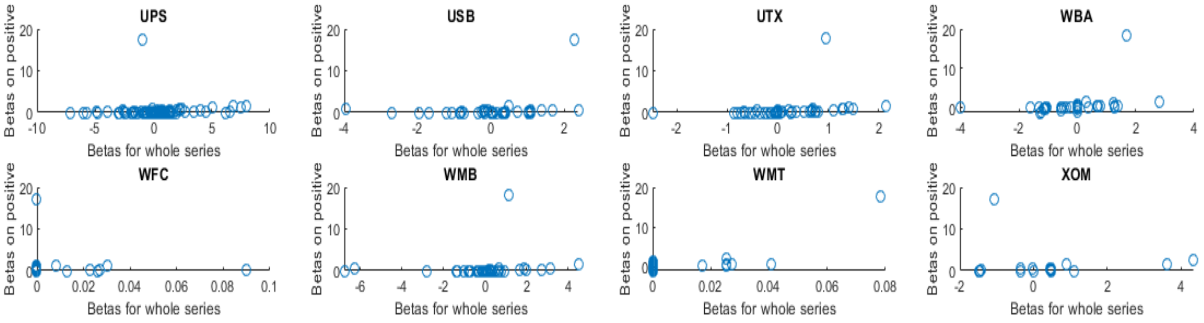

Figure C.21 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on positive jumps only (on y axis), for assets going from UPS to XOM. The coefficients refer to other stocks included as measure. ***Source:*** *own elaboration.*

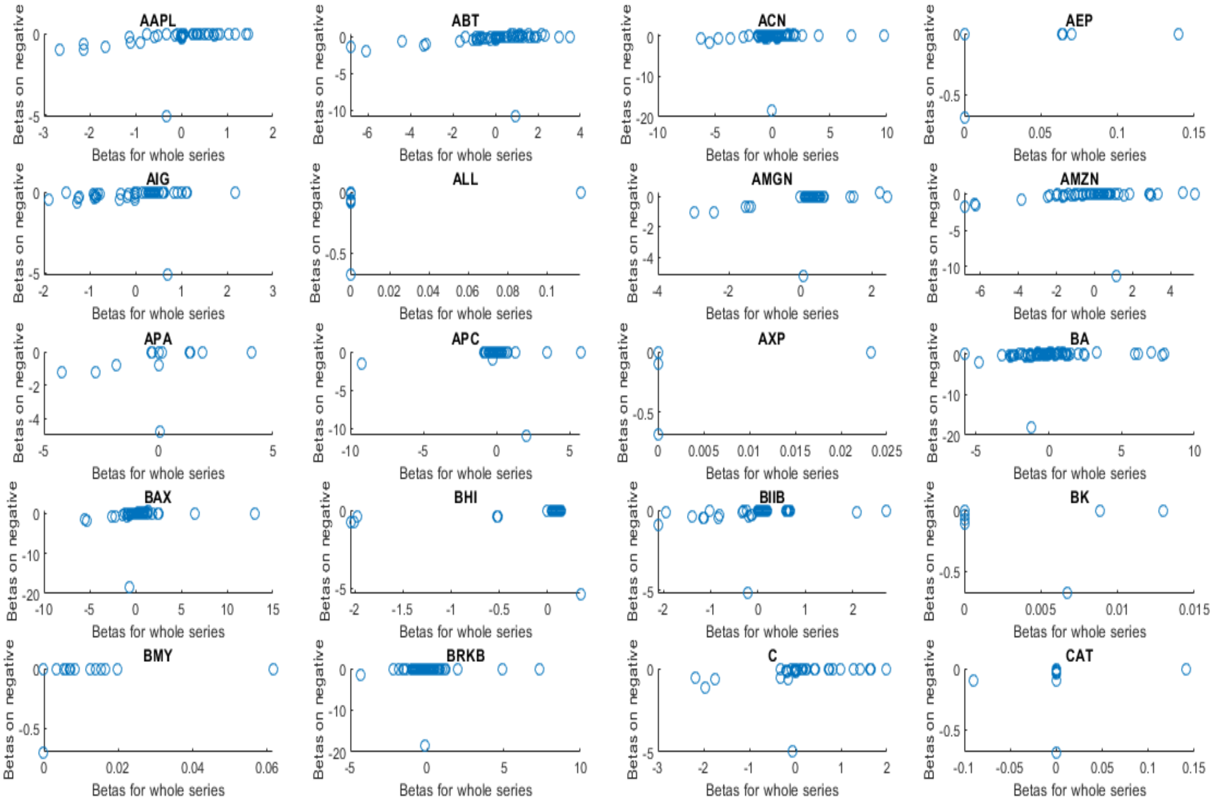

Figure C.22 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from AAPL to CAT. The coefficients refer to other stocks included as measure. ***Source:*** *own elaboration.*

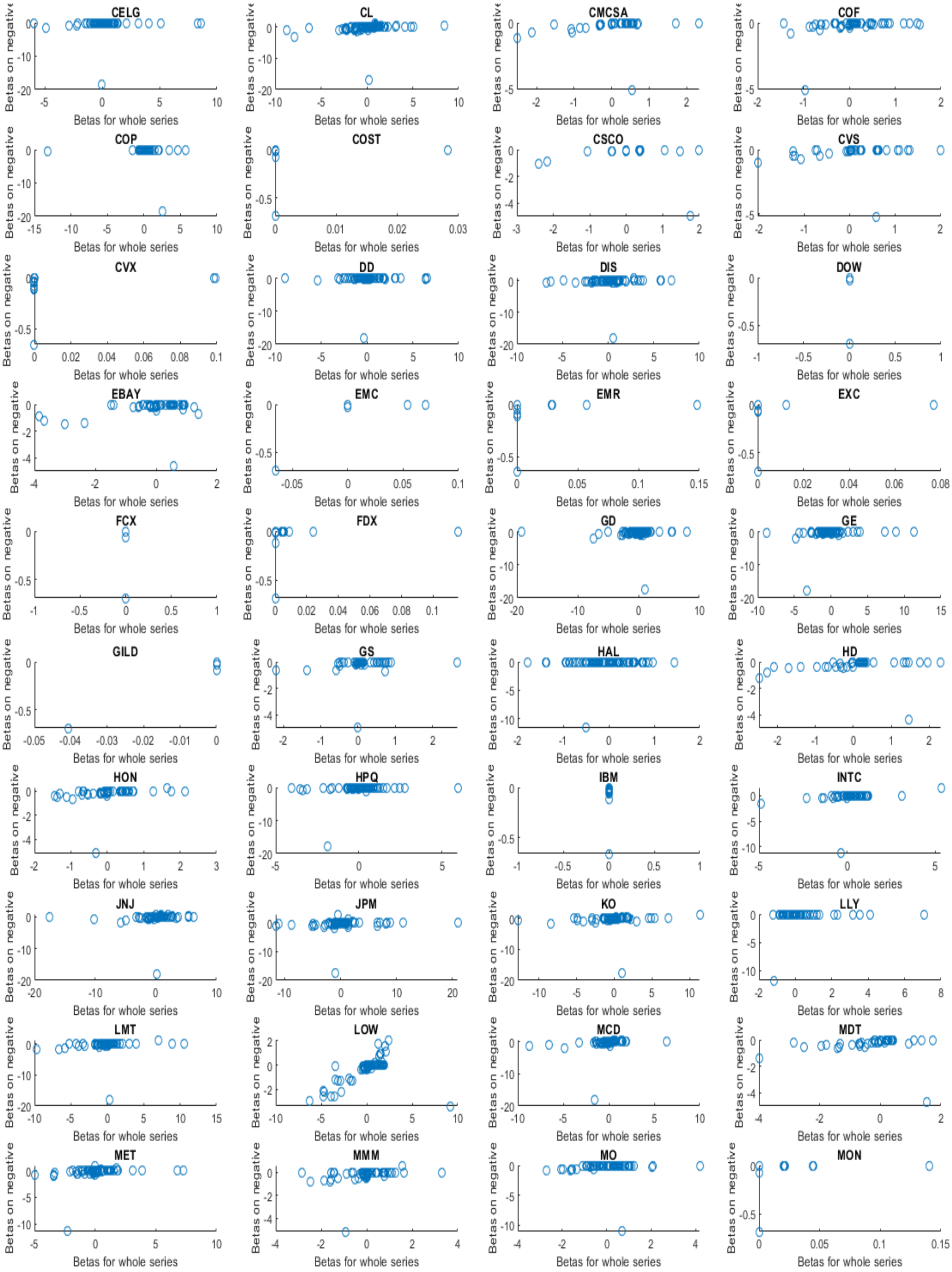

Figure C.23 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from CELG to MON. The coefficients refer to other stocks included as measure. ***Source:*** *own elaboration.*

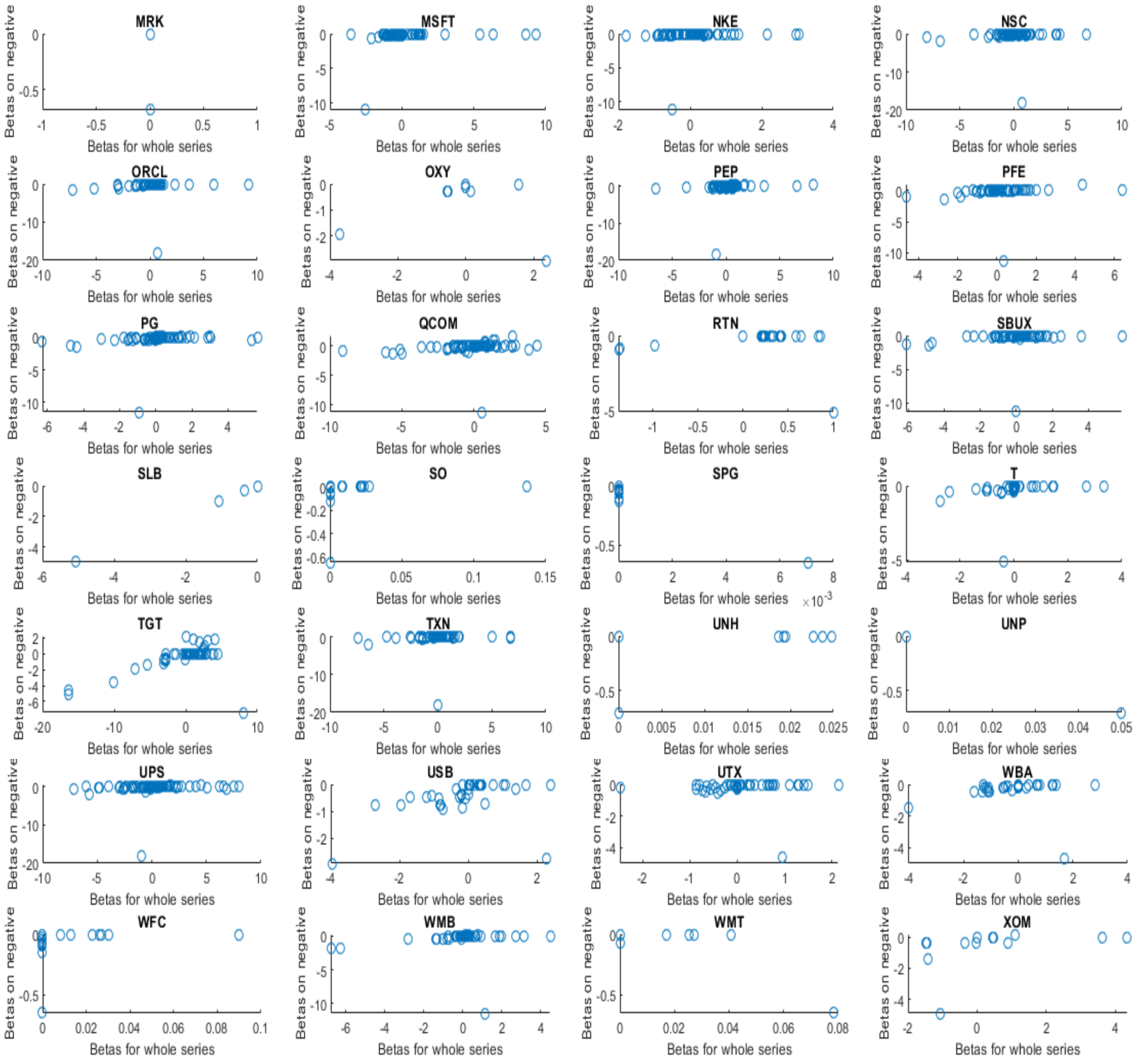

Figure C.24 – Comparison between coefficients obtained by penalised regressions performed on whole time series (on x axis) and on negative jumps only (on y axis), for assets going from MKR to XOM. The coefficients refer to other stocks included as measure. ***Source:*** *own elaboration.*