Master Thesis

Swiss Federal Institute of Technology

Forecasting Intensity of Mid-Quote Price Changes Via the Hawkes Model

Author: Joel Bloch
Examinor: Prof. Dr. Didier Sornette
Supervisor: Dr. Vladimir Filimonov

Zurich, Switzerland
August 2014

Contents

1 Abstract

Knowledge of the expected number of short-term trades or short-term price changes impacts trading performance in financial markets by influencing the order size process. The Hawkes Poisson process is a known model to predict these changes in liquid financial markets. The quality of the algorithm depends on the calibration of its parameters. This work performs the calibration of these parameters by testing real data from the E-Mini market 2011. The analyzed calibration parameters are "number of simulations", "sample time frame", "kernel choice", and "detrending method".

The calculations reveal that a larger number of simulations and a longer sample time frame improves the results. These findings may be intuitive; however, the results have also show that there is no substantial increase in prediction-quality above the threshold of 300 simulations and a 30 minute sample time frame.

The basic concept of the Hawkes Poisson process stipulates the distinction between the arrival of exogenous events - such as external market news - and the self-excited arrival of endogenous events, i.e. trades triggered by past trades or price changes. In the Hawkes Poisson process, the probability drop-off of an event triggered by a past event is called a "kernel". Whether this kernel has an exponential or power law form is a discussion within the literature. The application of this forcasting algorithm with real market data has shown that the forecasting performance is higher with an exponential kernel than with a power law form. The forecasting algorithm assumes that the exogenous intensity is not time-dependent. As in real markets, the exogenous intensity may be time-dependent thus it is necessary to detrend in-frame-data as a first step before applying the algorithm. The application of the forecasting algorithm has shown that detrending of historical data decreases the forecasting performance for the trend chosen in this thesis. This could be due to a wrong "bias" added by choosing a wrong trend. Further trend determination methods are proposed.

The fully calibrated model overestimated the number of mid-price changes by 21.3%. The false positive rate of the prediction of extreme events, which are defined as number of minutes with 95 or more mid-price changes, was 1.2%; and the true positive rate of this parameter was 59.4%. The positive predictive value lies at 85.9%, meaning, that if the algorithm predicts an extreme event for the next minute, it is correct in 85.9% of the time. Equally, if the algorithm predicts that no extreme event will occur in the following

minute, it will be correct in 95.3% of the times (negative predictive value)

An application of the Hawkes Poisson process for a bi-variate setup (mid-price up, mid-price down) is proposed.

2 Introduction

Knowledge of the expected number of short-term trades or short-term price changes impacts the trading performance in financial markets by influencing the order size process.

In order to define forecast models for market parameters, it is necessary to understand the dynamics of the underlying markets. Often in economic theory, these models are based on the Efficient Market Hypothesis (EMH), which stipulates that the markets are fully absorbing all available information instantaneously and this information is fully reflected in the market prices. [19][20][21][22]

A more sophisticated nonemodelling approach is based on the observation that real markets do not behave entirely upon EMH. According to the EMH, large changes in market prices can only occur due to "big news," which influences the market from the outside. The analysis of extreme market changes in the past has proven this assumption wrong. The number of short-term trades or short-term price changes are not only based on news impacting the market from the outside but also on information from the market itself that influenced its participants, e.g. the number of trades or the market price changes in the recent past. The Hawkes Poisson process is a general model describing such systems.

In fact, Hardiman et al. [24] and Filimonov and Sornette [3] demonstrated that the Hawkes Poisson process is a more valuable tool to predict parameters of such markets than models based on the EMH.

This thesis exploits the Hawkes Poisson model for the prediction of the number of short-term trades and/or short-term price changes. It has two main goals:

1. Defining the calibration parameters of the model algorithm in order to increase its prediction quality for the number of mid-price changes; and
2. Understanding the stability of the prediction algorithm based on the impact of different circumstances on the optimal calibration parameters.

As a first step, the theory of the Hawkes Poisson process is described. In the second step, each Hawkes Poisson parameter is calibrated by applying real data from the E-Mini market 2011. The calibration process is then evaluated based on performance parameters that reflect the forecasting quality of the algorithm. As a summary, in the third step, the stability of the prediction quality with the optimal set of calibration parameters is tested under different circumstances.

3 Theoretical Background

3.1 Model: Hawkes Poisson Process

If we look at a (stock) market as a series of mid-price change events, we can describe it as a discontinuous point process with a specific event arrival rate. In a financial market, mid-price is defined as the price between the market bid and the market ask price. This arrival rate is named λ. The simple case is called a Poisson Process. In a Poisson Process, the individual events do not depend on each other. The Hawkes Poisson process is a further complication of the general Poisson process that assumes that the intensity is time-dependent as well as dependent on the arrival of past events.

$$\lambda(t|\mathcal{F}_{t-}) = \mu(t) + \int_{-\infty}^{t} \varphi(t-s)dN(s) \tag{1}$$

where $\lambda(t|\mathcal{F}_{t-}) = lim_{h\downarrow 0}\frac{1}{h}Pr[N(t+h)-N(t)>0|\mathcal{F}_{t-}]$. $N(t) = max(i : t_i \leq t)$ refers to the number of events which happened in the past and shall be called the *counting process*. The *filtration* is defined as $\mathcal{F}_{t-} = t_1, ..., t_i; \forall i<N(t)$. We define the exogenous part of the intensity μ as the *background intensity*. The endogenous feedback in the system is described by the *kernel function* $\varphi(t)$. This kernel function denotes the probability of new events to occur based on past events. We call an original (e.g. exogenous) event a *mother event*. The number of events that follow a mother event are logically named *daughter events*.
The integral of the kernel function is called a *branching ratio*.

$$n := \int_{0}^{\infty} \varphi(t)dt>0 \tag{2}$$

The branching ratio defines the average number of children a mother event creates. Intuitively one understands that if a mother event has, on average, more than one daughter event, the system is super-critical and grows uncontrolled. Stationarity of the Hawkes Poisson process requires a branching ratio $n < 1$.

We can rewrite the conditional intensity of the Hawkes Poisson process as:

$$\lambda(t|\mathcal{F}_{t-}) = \mu(t) + n\sum_{t_i<t} h(t-t_i) \tag{3}$$

Here we used the normalized kernel function

$$h(t) = \varphi(t)/n \tag{4}$$

with

$$\int_0^\infty h(t)\mathrm{d}t = 1 \tag{5}$$

The Hawkes self-excited Poisson process is a generalization to the Poisson Point Process, where past events can influence the current intensity. This dependency is described by the so called "kernel function."

The Hawkes Poisson process describes a chain of individual events that can either happen exogenously (*immigrant* events) or endogenously due to a past event (*descendant*). The occurrence of exogenous events is described by μ. Every exogenous event can create multiple following events in multiple generations. We can group all these subsequent events into clusters. Every mother event can statistically create a defined number of daughter events. These daughter events can, in turn, create further generations of events. We define the *branching ratio* n as the average number of daughter events triggered by a mother event. A system with a branching ratio smaller than one is called *subcritical*. Analogously a system can be *critical* ($n = 1$) or *super-critical* ($n > 1$). In a subcritical system with one immigrant (mother) event, the generations of daughters are limited and the system will die out eventually. On the other hand, a single immigrant can lead to an explosion (infinite events) of the system if every event statistically creates multiple daughter events.
In a subcritical system with a constant (non time-dependent) background intensity μ, the branching ratio equals the average fraction of the number of descendants in the whole population of events [3].

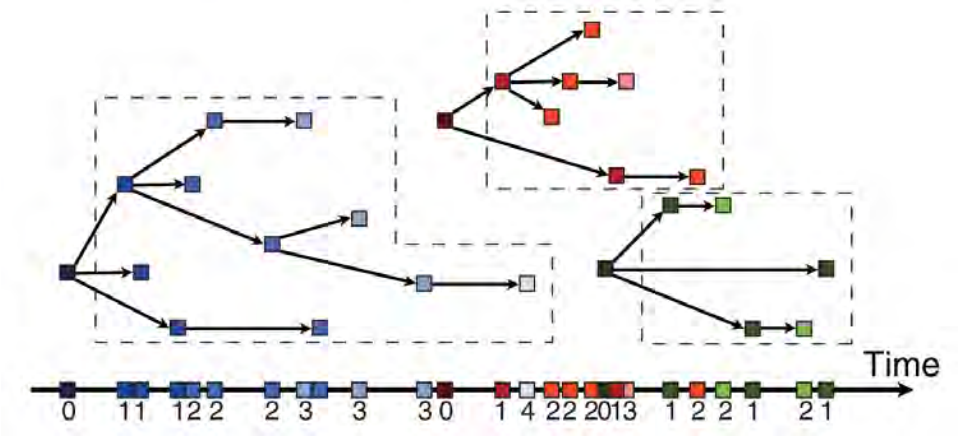

Figure 1: Source: [1], Illustration of the branching structure of the Hawkes Poisson process (top) and events on the time axis (bottom). Different coloured markers correspond to different clusters; the dashed lines denote descendants of the same cluster, and the number next to each event denotes its order within the cluster. This picture corresponds to a branching ratio equal to n = 0.88.

3.2 Kernels: Exponential vs. Power Law

The most commonly used kernel in finance is the exponential kernel, with memoryless properties in such a stochastic process. [3],[1],[7],[8],[9]

$$h(t) = \frac{1}{\tau} \exp(-\frac{t}{\tau})\chi(t) \tag{6}$$

This exponential kernel states that the probability for a daughter event to be created after a mother event decreases exponentially with time.

However, if we assume the stock market behaves similarly to the energy build-up and release of tectonic plates (resulting in earthquakes) we would work with a kernel in the form of power-law (as done in [25]). This kernel is often used for geophysical simulations. Compared to the exponential kernel (formula 6), this kernel states that events can cause new events at a much later time. In this context we can call a kernel which describes a situation where mother events can still trigger daughter events even after a very long

time a "long memory kernel". This long memory power law kernel looks as follows:

$$h(t) = \frac{\theta c^{\theta}}{(t+c)^{1+\theta}} \chi(t) \tag{7}$$

The kernel for this Epidemic-Type Aftershock sequence [10],[11],[12],[13] has a power law time-dependance and describes the modified Omori-Utsu law of aftershock rates [3],[14],[15] Similarly to the exponential kernel, the *heavy-side function* χ ensures the causality principle [3].

3.3 Estimation of kernel parameters

Confronted with real life market data, if we want to forecast the number of mid-quote price changes, we need to estimate the kernel parameters (μ, n, τ for exponential kernel or μ, n, c, θ for power law kernel). This can be accomplished in multiple ways. The most obvious method is to reverse engineer which events belong to which cluster. (The idea behind clusters is illustrated in figure 1) Knowing the cluster, we can easily calculate the average number of endogenous daughter events based on an immigration event. However, Sornette and Utkin proved in 2009 [26] that this approach has many drawbacks in the case of long-memory kernels.

Instead, we used the Maximum Likelihood Estimation method. The log-likelihood function is already known in closed form for Hawkes Poisson processes [23],[12]

The parameters for the Hawkes Poisson arrival rate in the sum-form (Formula 3) can be calculated by maximising the following log-likelihood function

$$\log L(t_1, ..., t_N) = -\int_0^T \lambda(t \mid \mathscr{F}_{t-}) dt + \sum \log \lambda(t_i \mid \mathscr{F}_{t_i-}) \tag{8}$$

where t_i are the observed times of the events. In this specific case, where we know the kernel to be of exponential form (Formula 6) and the background intensity is not dependent on time, the log-likelihood function to be maximised becomes:

$$\log L(\mu, n, \tau \mid t_1, ..., t_N) =$$
$$-\mu T - n \sum_{t_i<T} \left(1 - e^{-(T-t_i)/\tau}\right) + \sum_{t_i<T} \log \left(\mu + \frac{n}{\tau} \sum_{t_j<t_i} e^{-(t_i-t)/\tau}\right) \quad (9)$$

Through maximisation we find the estimated parameters for the exponential kernel

$$\{\widehat{\mu}, \widehat{n}, \widehat{\tau}\} = \arg\min_{\mu,n,\tau}[-\log L(\mu, n, \tau \mid t_1, ..., t_n)] \quad (10)$$

This method does not take events into account which occurred before the sample started. This, however, could explain some of the first events in the sample. We thus believe those events are immigrants and are being described with the background intensity μ, when in fact they are children events of a past generation. This leads us to underestimate n and overestimate μ. Similarly we have a distortion effect at the cut-off, after the sample ends. We do not see all the daughter events of the events in our sample size as these would lie outside our sample. This is especially true for events on the far right side of our sample. This effect would lead to an underestimation of n, which gets balanced by an overestimation of τ. The longer a memory of a kernel is, the stronger this effect becomes.

Once the parameters μ, n and τ are estimated, they can be validated with a goodness-of-fit between the real data and the model. A particularly useful goodness-of-fit test, according to Ogata [12], is the residual analysis. This is done by analyzing the residual process, defined as the nonparametric transformation of the initial series of the event time stamps t_i into [3]

$$\xi_i = \int_0^{t_i} \widehat{\lambda}_t(t)dt \quad (11)$$

where $\hat{\lambda}_t(t)$ is the conditional intensity of the Hawkes Poisson process estimated with the maximum likelihood method. We can test the statement of [18], that if the data set of events has been created using a Hawkes Poisson process with kernel $h(T)$, the residual process ξ_i must be Poisson with unit intensity. This goodness-of-fit can be tested in multiple ways:

- visual cusum plot or Q-Q plot analysis
- statistical tests. e.g.
  - independence tests applied to the sequence of ξ_i
  - tests of the exponential distribution of the transformed inter-event times $\xi_i - \xi_{i-1}$. This is equivalent to testing the uniformity of distribution of the variable $U_i = 1 - \exp(\xi_i - \xi_{i-1})$ between $[0, 1]$.

The calibration process could be used for real time analysis in, for example, a high frequency trading environment. As such, it is important to optimize the speed of such a calibration. Even though the log-likelihood function uses $\mathcal{O}(N^2)$ complexity if the kernel is of exponential or an approximate power law form, it can be reduced to $\mathcal{O}(N)$ using recursive relation [23].

A further simplification of the computational process is the method devised by Filimonov and Sornette in 2013 [3]. The following method not only decreases the number of local minimas and lets us present visually the search space, it also decreases the computational complexity using dynamic programming [3]. The trick is to split the set of parameters into two groups, e.g. for the case of the exponential kernel. We split the parameter search for μ, n, τ into two separate searches for μ, n and subsequently τ. The original optimization problem

$$\{\widehat{\mu}, \widehat{n}, \widehat{\tau}\} = \arg\min_{\mu,n,\tau}[-\log L(\mu, n, \tau \mid t_1, ..., t_n)] \tag{12}$$

becomes the optimization problem

$$\{\widehat{\tau}\} = \arg\min_{\tau} S(\tau \mid t_1, ..., t_N) \tag{13}$$

with

$$S(\tau \mid t_1, ..., t_N) = \min_{\mu,n} [-\log L(\mu, n, \tau \mid t_1, ...t_N)] \tag{14}$$

3.4 Detrending

The goal of this work is to find a way to predict the rate of arrival of mid-price changes in a market. To do this, assuming the market follows Hawkes Poisson with an exponential kernel, we need to estimate the parameters for the exponential kernel n and τ. If the exogenous intensity μ is constant, this estimation can be done using the methods described in chapter 3.3 using the maximum likelihood function.

However, if the exogenous intensity μ is time dependent, a good estimation of n and τ (the endogenous part, the kernel) is not possible as the model assumes a constant exogenous intensity. If, for example, traders tend to trade more frequently during early morning (many exo events), the algorithm would mistakenly interpret exo events as endo events. To solve this issue, if the underlying time dependent exogenous intensity is known, the market data can and should be de-trended.

Detrending transforms the timestamps of our data. It is a sound assumption that the original data set of timestamps follows a time dependent trend line. The modified time stamps of the same data set follow a constant trend line. The intensity deviations from this trend line will then be due to endogenous processes. [6]

The original timestamps shall be called t_i, and the detrended timestamps t_i^*. Detrending of t_i works by integrating the trend function, here the exogenous intensity $\mu(t)$, up to t_i.

$$t_i^{**} = \int_0^{t_i} \mu(t)dt \tag{15}$$

The logical interpretation here would suggest that t_i^{**} is the number of events which would occur in the trend until the timestamp t_i. This is directly proportional to the detrended timestamps. We then need to rescale so that the transformed time stamps are matching the original time stamps.
$t_0^* \stackrel{!}{=} t_0 = 0$ and $t_{\max}^* \stackrel{!}{=} t_{\max}$ with $K = \frac{t_{\max}^{**}}{t_{\max}}$

$$t_i^* = \frac{1}{K} \int_0^{t_i} \mu(t)dt \tag{16}$$

4 Calibration of Prediction Algorithm

We want our algorithm to predict the number of events in the out-frame based on input from the in-frame. The parameters of the algorithm need to be calibrated in order to optimize its prediction performance which will be measured by the performance indicators. A set of calibrated parameters defines the setting of an algorithm. The calibration parameters and performance indicators will be described shortly.

The algorithm to predict the number of mid-price changes works as follows:

1. For a given day, a trading activity trend is determined based on the trading activity within an interval of certain trading days before the given day. This time frame shall be called the "trend determination timeframe." (See section 4.1.3)

2. The data of the day under examination is detrended using the trend found in step 1.

3. A "sample timeframe" - on which the prediction itself is based - is selected. The Hawkes Poisson model is calibrated on this time frame by using the parameter estimation method for the two kernels (see chapter 3.3).

4. Combining the knowledge of these parameters and the mid-price changes in the sample timeframe (in-sample), we can calculate the event (mid-price change) arrival rate, or "intensity" using the Hawkes Poisson model (eq. 3). For every mid-price change in the in-sample, the probability for descendant events, i.e. triggered mid-price changes, is calculated for the out-sample (also called prediction time frame. In this work the prediction time frame is one minute). The implementation of this stochastic process is done using the thinning method described in [5]. Because this is a stochastic process, we run the simulation to estimate the prediction multiple times and calculate averages.

4.1 Introducing Calibration Parameters

The algorithm can be calibrated by different parameters and settings. This thesis conducted performance tests with different settings in order to find an optimal prediction algorithm.

The settings to be tested against each other are as follows:

- Number of simulations
- Choice of kernel
- Detrending method
- Length of sample time frame

4.1.1 Number of Simulations

The Hawkes Poisson process (3) describes the intensity of occurring events. Since the memory kernel is the probability for a descendant event to be triggered, the entire Hawkes Poisson process is a stochastic process. To obtain an adequate prediction for the number of mid-price changes, we have to simulate the process multiple times and calculate averages.

4.1.2 Choice of Kernel

As discussed in chapter 3.2 different kernels can be used in the Hawkes Poisson formalism and algorithm. We decided to compare the prediction power of mid-price changes for the exponential short memory kernel (6) as well as for the power law long memory kernel (7).

4.1.3 Detrending Method

Detrending of our data was done the following way:

1. A number of days leading up to the day we would want to detrend is chosen. This time frame shall be called trend determination time frame
2. The chosen timeframe is cleaned of days with different behaviour, as using these days would skew our final trend as shown in [3]
3. The time series of events for the remaining days in the trend determination time frame are averaged. This averaged time series shall be used as our trend. Why we can use this averaged time series as our trend will shortly be described.
4. Using formula 16 the current day is detrended.

The trend represents the average behavior of the exogenous intensity μ. We found the trend however averaging the total intensity λ. We could do this using the following following working assumption, which is illustrated in figure 2:

- We want to find the trend for $\mu(t)$
- As the intensity for the trend is normalized via the equation 16, the absolute value is less important than simply the behavior over time.
- A trend is usually found by averaging time stamp data over multiple days in the past. However, we can only find a trend for $\lambda(t)$, not directly for $\mu(t)$.
- Working assumption: The added average endogenous activity is constant (Explanatory illustration in figure 2)
- If this working assumption holds true then the form of $\lambda(t)$ = form of $\mu(t)$ and we can use the total averaged intensity for detrending.

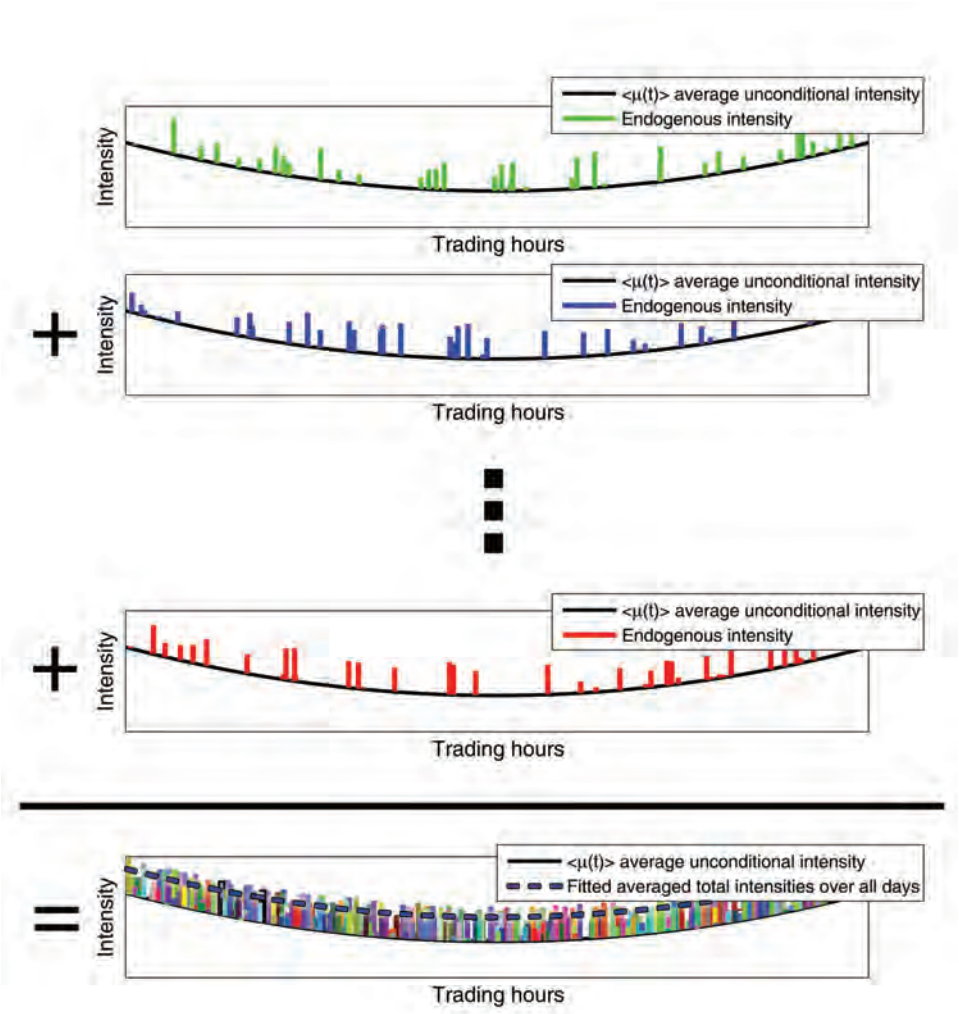

Figure 2: We use the averaged total intensity of multiple days in the past as the trend for unconditional intensity $\mu(t)$. This figure explains why we can do this, even though the total intensity includes the endogenous intensity. The time stamp series (e.g. of mid-price changes) for multiple days in the past are added. The time stamps are either exogenous (black line) or endogenous (green, blue, red line for different days respectively). As the endogenous time stamps added from multiple days appear to occur evenly distributed, the total measured intensity has the similar form as the average unconditional intensity. This can be seen by the fitted averaged total intensity line (dotted)).

Next we had to decide which days we use for detrending and how best to find the average of intensity. In this chapter we will illustrate the rationale behind those choices. In a later stage we will use the different methods on real trading data to see which performs best.

a. Choosing a time frame over which the trend is determined: The more days we use for detrending the less noise we have in our final averaged trend. However the daily trends tend to change exceedingly over long periods of time, as mentioned above. It has been shown, for example, that the exogenous part has decreased relatively to the endogenous part over the last 10 years [3]. Thus taking an average over too long of a period does not yield better results. We tested detrending on real data for one month and three month periods.

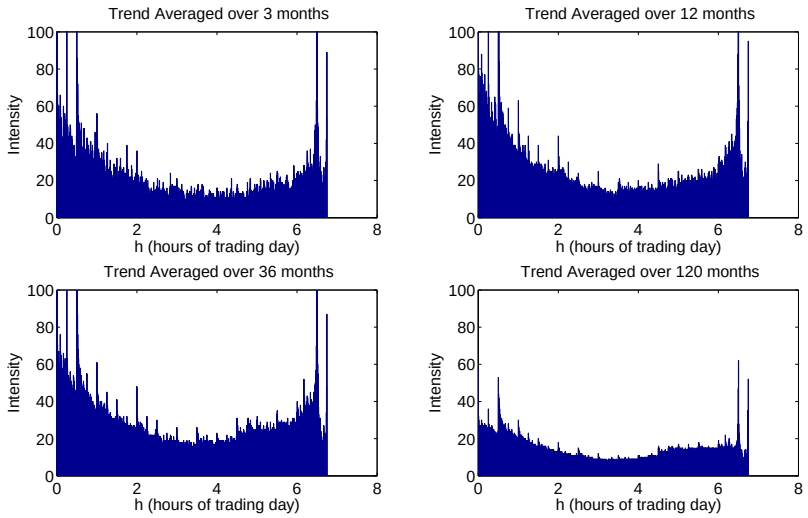

Figure 3: Average intensity over different time frames, all ending on the date 04-01-2010. Resolution over 1000 packages. One clearly sees the typical U-shape attributed to a lower activity during lunch time.

b. Cleaning of the timeframe on days of extreme trades and different behavior: We clean the days of the pre-defined timeframe by discarding non-trading days as well as days with extreme trading behavior, i.e. extreme peaks in the trading activity. Furthermore, time-shifts, such as summer or

wintertime, were accounted for. As Filimonov and Sornette have shown in [3], using days with extreme trading behavior to find trends will significantly skew the trend. One way to find these days is to manually go through the time period and flag days which exhibit this extreme behavior. However, since part of the goal of this paper is to create an algorithm, which could, in theory, be automatically used as a tool on the trading floor, this is not an option. Days with *different* trading behavior often show an increase in the number of trades. Big peaks of trading activity are usually so extreme such that the total number of trades on those days is significantly larger than on days without those big peaks. We thus use the extreme number of trades per days as a proxy for extreme trading behavior on those days. In practice, we consider a quantile interval of days cut off from the extreme days in terms of trading activity. In concrete, we tested both for (15% - 85%) as well as (25% to 75%) quantiles. Furthermore days with big news, such as fed rate announcements have a very different dynamics. We excluded those days from the trend averaging.

c. Averaging: Every event on the days included in the trend analysis timeframe were aggregated to one fictional day. As detrending works with the rescaled (by K, see eq. 16) absolute averaged intensity, we can use this as a trend. Since the aggregated day has many more time stamps, we can coarse-grain the time stamps by combining N consecutive timestamps on the aggregated day and average the time for an averaged time stamp. In this way the total number of events for our aggregated day would more closely resemble that of an average day. Hence N would approximately be the number of days over which the average was performed. This does not in any way increase the quality of our detrending, but it is, however, a good method for illustration.

The shape of the trend changes over time. It is natural to assume that the days closer to the day we want to predict the number of trades for are more influential in finding the correct trend of the current day than days further in the past. To include this assumption into the trend determination, one could use a **E**xponentially-**W**eighted-**M**oving-**A**verage EWMA to give days closer to the target day more importance. For every day we use for averaging, we include every time stamp of the mid-price change only with a probability of $p_i = e^{-\alpha \cdot \Delta t_i}$ where Δt is the difference in days to the day we are trying to predict. For the forecasting algorithm we decided to test EWMA with $\alpha = 0.1$, which equals an exponential drop off after two months.

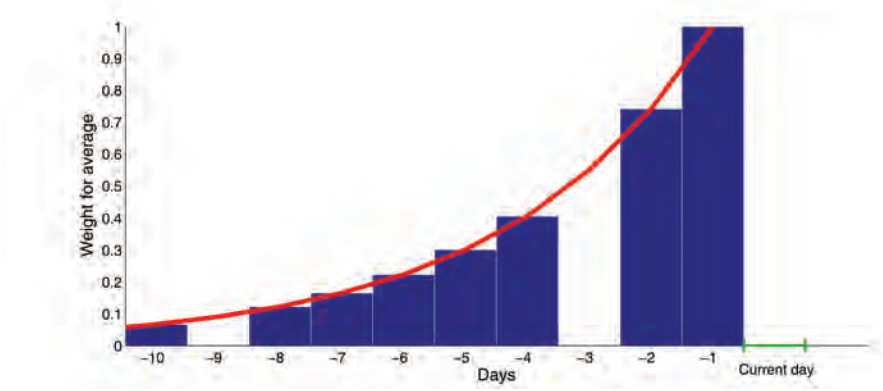

Figure 4: Illustration for **E**xponentially **W**eighted **M**oving **A**verage. After having decided which days should be used for the in-frame, the remaining days in the in-frame are given a weight each. This weight is calculated based on an **E**xponential drop off (red line). The further the day lies in the past, the less it is **W**eighted for the **A**verage (blue bars). This process gives us the trend used for detrending the current day. For the consecutive days the process has to be **M**oved accordingly.

If we look at figure 3 we may be tempted to use a polynomially fit to smooth the averaged trend. To analyse the difference between the *raw* - i.e. un-smoothed - trend as well as the polynomially-fitted trend we plot these two trends and their respective detrended day in the following figure 5

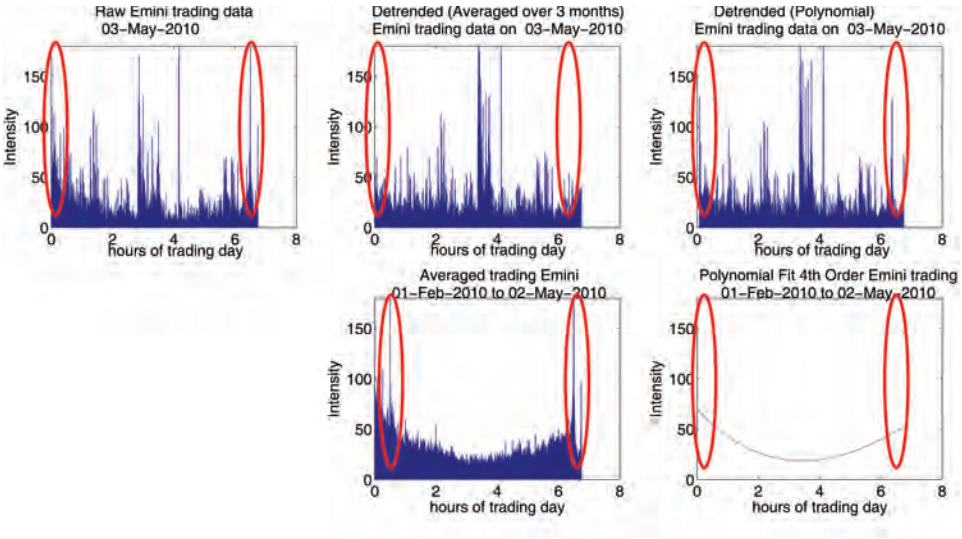

Figure 5: From left to right. First column: The raw un-detrended trading data of an exemplary day (here 03-May-2010). Second column, bottom row: the trend using normal averaging over a time window of three months. Second column, top row: Detrended day using the trend of the bottom row. Third column, bottom row. Polynomial fit to the raw trend. Third column, top row: Detrended day using the polynomial trend of the bottom row. The peaks early in the morning and late in the evening are marked with red circles. One can clearly see in the bottom right figure, that the polynomial fit does not include those peaks. Consequently the detrended data with the polynomial fit (top row, right figure) still show these peaks even though these peaks are part of the unconditional time dependent intensity $\mu(t)$ we would like to detrend.

The trend for the unconditional intensity shows the expected U-shape ([3],[24]). Over lunchtime fewer trades occur than early in the morning and later in the evening. Just at the time of opening and at the time of closing, two peaks can be seen – these are the first and last auctions of the day.

We can clearly see that using polynomial fitting would lead to counting the opening and closing peak as an endogenous events. Using the *raw* average detrended these peaks as well. This effect can clearly be seen in figure 5. The top left graph is the un-detrended, original trading data for a given day. Using the trends (bottom row) once with the polynomial assumption and

once with the *raw* trend we see the U-shape disappear. However, only the raw trend is able to detrend the peaks early in the morning and late in the evening.

For the experiment on the prediction power of different models, we used the non-fitted trend both for EWMA and for normal averaging.

Furthermore, the whole set-up will be tested without performing the detrending.

4.1.4 Sample Timeframe

The sample timeframe is the timeframe on which the estimation of parameters of the Hawkes Poisson model is based. The longer the sample timeframe, the better the algorithm results, but also the higher the computable requirements to perform the calculation. As the dynamics of the market changes substantially from trading hours to non-trading hours the sample timeframe must fully lie in the trading hours. This small caveat means that predictions cannot be made for the first few minutes of the trading day as a sufficiently large sample timeframe has not been reached yet. We cannot use data from the end of the previous trading day as these events could have offspring events after closing hour and do not influence new events over 12 hours later at the opening of the new trading day. Nevertheless, one could employ smaller sample time frames in the morning but those predictions would be less reliable.
The sample frame does not include any data from the future in the experimental setup. There is no look ahead bias.

4.2 Calibration: Number of Simulations

As discussed previously the stochastic nature of the multiple Hawkes Poisson model simulations have to be performed in order to average and improve the results. To find the number of simulations needed for the prediction of the relative deviation to no longer fluctuate, we performed the following test.

4.2.1 Test Setup

We try to find the number of simulations needed for the averaged relative deviation between the algorithmically predicted number of events and the

actual number of events in the prediction timeframe. Once the relative deviation is sufficiently small and stops fluctuating, a sufficient number of simulations has been reached.

We set up a test with a synthetic time series. When we talk of synthetic time series we mean a time series we created using the Hawkes Poisson algorithm 3 with a kernel and its corresponding set of parameters that we choose. The advantage of the synthetic time series are two-fold.

1. The actual parameters of the Hawkes Poisson process (μ, n and τ for exponential kernel) are known and thus the deviation we measure comes only from the simulation, not from the estimation of the parameters.

2. The prediction algorithm assumes that the Hawkes Poisson process governs the dynamics. However, other dynamics can also play a role in real life trading data. With a synthetic time series we can estimate a good number of simulation for a purely Hawkes Poisson "reality".

Our test setup is as follows:

1. A kernel and its corresponding parameters are chosen.

2. The length of the synthetic time series is defined from [0, t_{end}].

3. We define a time t_0 with $0 < t_0 < t_{end}$ from which we would like to predict and the length of the time frame which ends at t_0 and will be the input to the algorithm described in chapter 3.3 to predict the parameters (μ, n, τ for short memory exponential Kernel and μ, n, c, θ for the long memory power law kernel.) (Sample frame)

4. A synthetic time series for the time window [0, t_{end}] length is simulated applying the intensity λ of the Hawkes Poisson model (3) for every event in the sample time window. Thinning method as described in [5] was used.

5. Simulation for a prediction time frame starting at time t_0 is done using the kernel and parameters we used to create the synthetic reality with the thinning method from [[5]]. We measure the discrepancy between the number of events our algorithm predicted in the timeframe [t_0 t_{end}] and the actual number of events in the (synthetic) reality. This step is repeated N_{sim} times and the results are averaged.

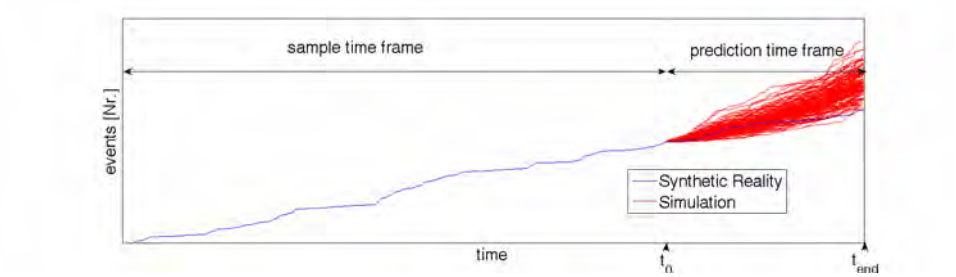

Figure 6: A visualisation of the synthetic simulation. The total averaged number of simulated events (red) between t_0 and t_{end} are compared with the synthetic reality (blue).

The performance indicator of the test was the decrease of fluctuation in the relative deviation of predicted number of events versus number of events in the synthetic reality for the same prediction timeframe.

The estimation value for the relative deviation was the following:

$$\mathrm{E}[\text{Relative Deviation}] = \frac{\mathrm{E}[N_e - N_r]}{\mathrm{E}[N_e]} \tag{17}$$

where N_e is the number of estimated events in the prediction time frame and N_r the actual number of events in the prediction time frame. Here. N_r is from the synthetic time series. For later experiments we used N_r of actual trading data.

4.2.2 Test Results

The synthetic reality test was performed for different parameter values of the Hawkes Poisson equation. While the relative deviation for different parameter value combinations was different, the qualitative behavior of fluctuation was similar across all different parameter realities. The following figure therefore serves as an example of this qualitative trend.

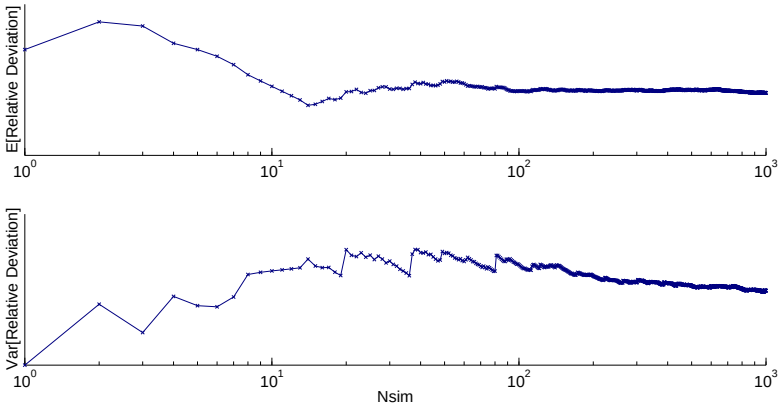

Figure 7: Expectation value and variance of relative deviation averaged over different number of simulation. One clearly sees the fluctuation when we average over too few simulations. After ca. 100 simulations the fluctuation gets smaller. To be certain to avoid overly large deviations in the prediction algorithm due to averaging over too few simulations, we chose to average over 300 simulations.

Result: We choose to average over **300 Simulations**. This is a good compromise between short computational time and sufficiently large sample size for averaging in order to decrease the bias to do being a stochastic process.

4.3 Calibration: Multiple Trend-Determination Methods

As described in chapter 3.4, determining the correct way to find a trend is not trivial thus we decided to test several trend-detection methods on real data.

4.3.1 Test Setup

We want to know which trend-detection method (as well as which sample frame length and which choice of kernel in the consecutive chapters) yields the best prediction strength for the number of mid price changes. To do this, we tested the prediction algorithm with the different parameter values

on real data and compared the prediction strength through various performance indicators.

Data Experiments were performed on data from the E-mini S&P500 future contracts, which are being traded in the Chicago Mercantile Exchange (CME). The CME started selling E-minis in 1997 as a smaller contract size alternative to pre-existing S&P500 future contracts. The calibration method developed in this work is based on the data of the trades in the year 2011. All time stamps were rounded to the nearest milli-second. The data set was cleaned of gaps, non-trading days and non-trading hours. Further, rollover-weeks were not used for the experiment. A rollover week is based on the following situation. Every third Friday of March, June, September and December future contracts expire. Eight days before the contract ends - the second Thursday of March, June, September and December - the liquidity (in volume) of the contract that is going to expire is switched to the contract that will expire only in the subsequent quarter. This follows another dynamic which would most likely spoil the calibration of the Hawkes Poisson model.

Calibrating the Hawkes Poisson model requires that the stationarity holds. It has been shown that stationarity cannot hold for time windows during which major macro-economic news announcements occur [2]. Just before these announcements, nearly all trading comes to halt in anticipation of the news. After the announcement trading rates shoot up very quickly. To be able to calibrate the Hawkes Poisson models effectively, we exclude those days.

Lastly, known days, such as where new federal rates were published, were not used for the experiment. In 2011 these were 26-Jan-2011, 15-Mar-2011, 27-Apr-2011, 22-Jun-2011, 09-Aug-2011, 21-Sep-2011, 02-Nov-2011 and 13-Dec-2011.

The E-mini market of 2011 had the following behavior:

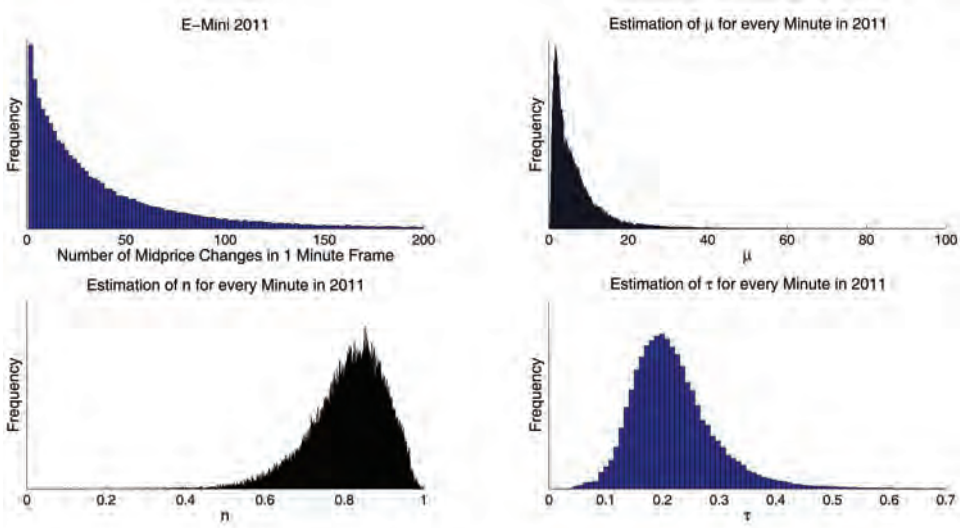

Figure 8: Analysis of every minute during trading hours on the E-Mini S&P500 market in 2011. The top left figure shows the distribution of actual number of mid-price changes for every minute during trading hours in 2011. The other three figures show for the same time windows the distribution of the estimated parameters for the exponential kernel.
Most sample minutes in 2011 had less than 50 trades. We would assume, that these sample minutes would dominate the behavior of μ, n and τ. Appendix 2 was attached to this thesis to show, how the distribution of mu, n and tau differ for a different subset of sample minutes with different number of midprice changes per minute.

	Number of Mid-price Changes	μ	n	τ
First Quartile	10	2.2	0.7	0.17
Median	26	4.6	0.82	0.208
Third Quartile	54	8.6	0.88	0.255
Mean	45	6.9	0.81	0.219
Variance	6105	61.9	0.01	0.007
Variance (w/o 5% outliers)	1015			

Table 1: Detailed statistics of distribution of "Number of Mid-price Changes", "Estimated μ", "Estimated n" and "Estimated τ" for figure 8

The estimation of Hawkes Poisson parameters was done with the experimental set-up 4, which will be described further in table 4.3.2. The estimation algorithm is being described in chapter 3.3. As they are estimations, they are afflicted with a (negligible) bias and are included here only for qualitative purposes.

We see that most minutes during trading hours in 2011 have had less than 150 mid-price changes. However during some extreme times, many more mid-price changes occurred. The 90% Quantile of number of mid price changes per minute was 95. The 95% Quantile 136 mid-price changes. These values are of importance if we want to define "extreme" events in the future. Furthermore, μ looks like bell curve but skewed towards larger values. This stems from the fact that we cannot have negative μ. From the comparison of low μ and higher number of mid-price changes we can already see that the endogenous intensity is higher than the exogenous intensity.
Similar to μ, τ looks like a normal distribution around 0.2 with a light skew towards larger values. Which is not surprising because tau, per definition, could not be larger than zero.
N is consistently under 1, but closer to criticality, with a mean n of 0.8. This coincides with findings of [3]

4.3.2 Performance Indicators

The different set-ups were ranked according to the following performance indicators:

- Mean Relative Deviation ($\mathrm{E}[\frac{N_e - N_r}{N_e}]$)
- Median Relative Deviation

The *median of the relative deviation* is the median for all relative deviations of all predictions made for the year 2011 on the S&P500 E-Mini market. We must choose how we value a low mean relative deviation compared to a low median relative deviation depending on the trading situation. Does one prefer to have most predictions correct, accepting that some predictions are way off (preference for low median relative deviation), or does one prefer a prediction model which tries to minimize the number of extremely wrong predictions (preference for low mean relative deviation)? Obviously an optimal algorithm has both low median relative deviation as well as low mean

relative deviation.

One possible application of the mid-price change prediction algorithm is its ability to predict extreme events, e.g. minutes with very high trading activity. We defined extreme events as any minute with more than 95 mid-price changes. (This value is the 90% Quantile of the year 2011; tests were also conducted for extreme events defined by having more than 136 mid-price changes per minute, which corresponds to the 95% Quantile in 2011. Those additional tests can be seen in the appendix.) The performance benchmarks for this extreme-event-prediction are as follows:

False positive rate is defined as

$$\frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} = \frac{FP}{FP + TN} = \frac{FP}{N} \tag{18}$$

The true positive rate (also called sensitivity) is defined as

$$\frac{\text{True Positive}}{\text{False Negative} + \text{True Positive}} = \frac{TP}{FN + TP} = \frac{TP}{P} \tag{19}$$

Here false positive, false negative, true positive and true negative are the number of minutes for the year 2011 for which the prediction algorithm incorrectly, respectively correctly predicted the following minute to have more/less than 95 mid-price changes.

We tested the following 12 configurations. For easier referencing they are labelled from 1 to 12.

In the following section, we will analyze the most interesting results for these configurations. A table containing the complete set of results can be seen in the appendix.

4.3.3 Test Result

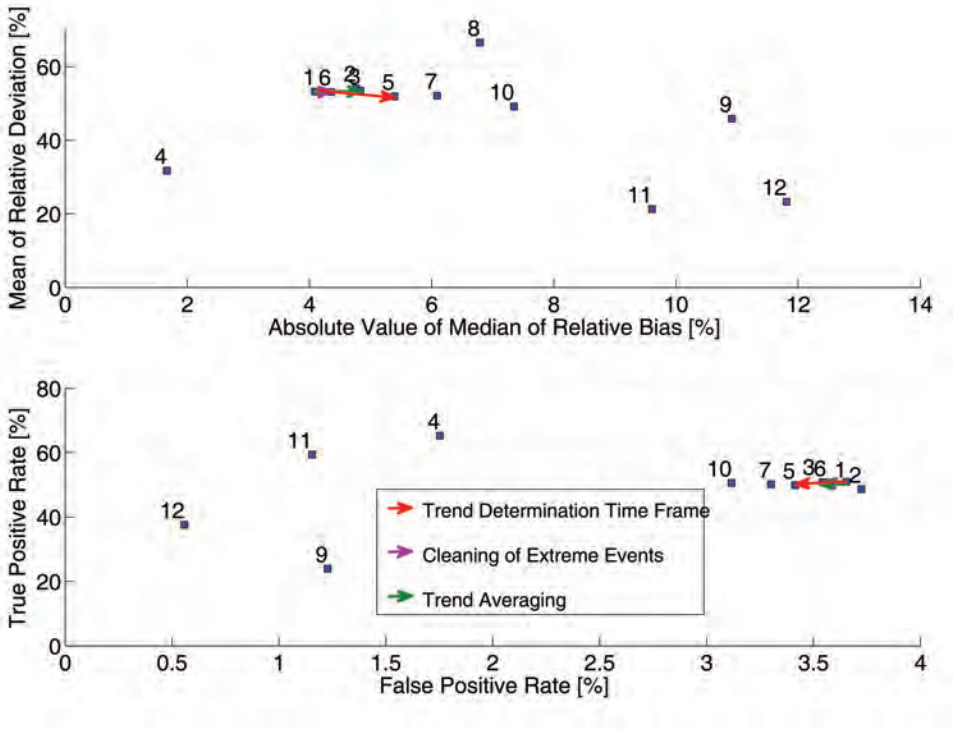

Figure 9: Top Figure: Median vs. Mean Relative Deviation for all different experimental configurations. Bottom Figure: False positive rate versus true positive rate for all different experimental configurations. The three arrows always point from a setup to a similar setup with just one changed parameter. Red arrow: 3 month trend determination time frame → 1 month trend determination time frame. Purple arrow: Exclusion of 30% of days in the trend determination time frame due to extreme trading behavior → Exclusion of 50% of days in the trend determination time frame due to extreme trading behavior. Green arrow: Equally weighted averaging method for trend determination → Exponential weighted moving average for trend determination. On the following pages the difference of prediction performance for the three setup differences are discussed. Setup 8 is not shown in the bottom figure, as it lies outside the visible axis.

Trend Determination Timeframe We tested the prediction algorithm for the trend determination timeframe of one month as well as three months. All other parameters were kept the same.

Label	1	5
Kernel	Exp	Exp
Trend	**3 month** (15%-85%)	**1 month** (15%-85%)
Sample Frame (min)	30	30
Absolute Value of Median of Relative Deviation	4.1%	5.4%
Mean of Relative Deviation	53.2%	51.9%
False Positive Rate Q90	3.7%	3.4%
True Positive Rate Q90	51.0%	49.9%

Table 2: Forecasting performance for two configurations with different trend determination timeframes. See figure 9, red arrow

In figure 9 we see that the mean relative bias is already quite high at 50%. This means that the average prediction states the number of events as 50% too high. While the mean deviation drops ever so slightly for shorter trend determination periods, the median prediction accuracy improves by 25% from 4% overestimation to a little over 5% overestimation.

If we look at the distribution of the performance indicators for all configurations, we see that the difference in the true positive rate and false positive rate for the different trend determination frame lengths is benign. Thus we cannot definitively state that one trend determination frame length performs better than the other trend determination frame length.

Cleaning of Extreme Days for Finding a Trend As discussed in chapter 4.1.3, we do not include days with an extremely high or extremely low total number of mid-price changes when finding the trend. We tested the prediction algorithm once for a trend where the bottom and top 15% quantiles were excluded and once where the top and bottom 25% quantiles were excluded.

Label	1	6
Kernel	Exp	Exp
Detrending (Quantile)	3 month (**15%-85%**)	3 month(**25%-75%**)
Sample Frame (min)	30	30
Absolute Value of Median of Relative Deviation	4.1%	4.4%
Mean of Relative Deviation	53.2%	53.2%
False Positive Rate Q90	3.7%	3.6%
True Positive Rate Q90	51.0%	50.7%

Table 3: Forecasting performance for two configurations with different cleaning of extreme days methods. See figure 9, purple arrow

Again, we see in figure 9 that the difference in the mean of relative bias as well the change of the absolute value of the median of relative bias is negligible between the two averaging methods.

Changing the Quantiles of days included in the trend determination frame did not improve or worsen the prediction performance of the algorithm as seen in the fpr and tpr performance indicator. As such, we cannot conclusively state which method works better.

Trend-Detection: Non-Weighted Averaging versus EWMA We tested two averaging methods for finding the trend.

- Averaging equally over all days in the trend determination frame, i.e. not including weights.
- Exponentially-weighted moving average (EWMA), where days closer to the present are weighted more for the trend to be detected. Every day was weighted $p = \exp(-0.1\Delta t)$ where Δt is the time difference (in days) counted from the prediction day.

The results are as follows:

Label	1	3
Kernel	Exp	Exp
Detrending (Quantile)	**3 month** (15%-85%)	**EWMA** 3 month (15%-85%)
Sample Frame (min)	30	30
Absolute Value of Median of Relative Deviation	4.10%	4.8%
Mean of Relative Deviation	53.20%	53.5%
False Positive Rate Q90	3.7%	3.5%
True Positive Rate Q90	51.0%	50.8%

Table 4: Forecasting performance for two setups with different trend averaging methods. See figure 9, green arrow

We see here a very similar picture to the previous trend determination settings. The prediction performance measures in mean relative deviation, median relative deviation, true positive rate or false positive rate do not change substantially with the choice of averaging method. Based on these data we cannot conclusively recommend EMWA over normal averaging.

Detrending versus No Detrending Lastly, we tested how the prediction algorithm performed if we did not at all detrend the data of the day for which we would like to predict the number of mid-price changes. Though the necessity of detrending was stated in the theory section, it is sensible to assume that for sufficiently short sample timeframes, the unconditional intensity can be considered constant.

Label	1	3	5	6	4
Kernel	Exp	Exp	Exp	Exp	Exp
Trend	**3 month (15%-85%)**	**EWMA**	**1 month (15%-85%)**	**3 month (25%-75%)**	**No Detrending**
Sample Frame (min)	30	30	30	30	30
Absolute Value of Median of Relative Deviation	4.1%	4.8%	5.4%	4.4%	1.7%
Mean of Relative Deviation	53.2%	53.5%	51.9%	53.2%	31.7%
False Positive Rate Q90	3.7%	3.5%	3.4%	3.6%	1.8%
True Positive Rate Q90	51.0%	50.8%	49.9%	50.7%	65.3%

Kernel	Exp	Exp
Trend	**3 month (15%-85%)**	**No Detrending**
Sample Frame (min)	90	90
Absolute Value of Median of Relative Deviation	7.3%	9.6%
Mean of Relative Deviation	49.2%	21.3%
False Positive Rate Q90	3.1%	1.2%
True Positive Rate Q90	50.6%	59.4%

Label	9	12
Kernel	Pow	Pow
Trend	**3 month (15%-85%)**	**No Detrending**
Sample Frame (min)	90	90
Absolute Value of Median of Relative Deviation	10.9%	11.8%
Mean of Relative Deviation	45.9%	23.3%
False Positive Rate Q90	1.2%	0.6%
True Positive Rate Q90	23.9%	37.6%

Table 5: Forecasting performance of multiple configurations to compare forecasting algorithms where detrending was used with forecasting algorithms where no detrending was used. See figure 10

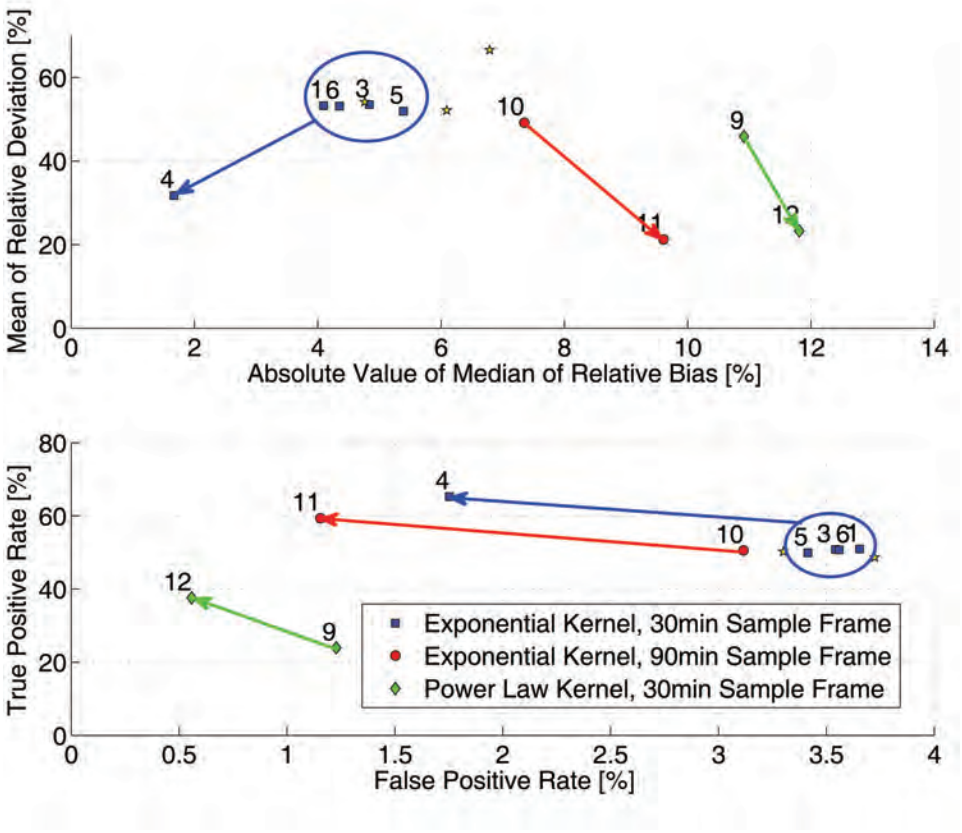

Figure 10: Top figure: Median versus Mean Relative Deviation for all different experimental configurations. Bottom figure: False positive rate versus true positive rate for all different experimental configurations. The arrows point from setups with detrending to a similar setup without detrending.

Not performing any detrending improves the mean performance for prediction remarkably. Judging only on the median of relative bias, we cannot clearly state if detrending or detrending yields better results. However, if we look at the false positive rate and the true positive rate, we see that NOT detrending actually massively lowers the false positive rate and increases the true positive rate, in turn increasing the quality of our predictions. This was especially surprising as the U-shape trend can be clearly seen in all the data and all past research on this topic assumed detrending would increase the accuracy of the Hawkes Poisson algorithm.

This result could imply that the behaviour of the time-dependent exogenous intensity $\mu(t)$ may be regular over long periods of time but have very individual specific behaviour for the individual day. Detrending works with the assumption that the trend equals the time dependent behaviour of the time dependent exogenous intensity. Detrending with the wrong trend would mean to detrend endogenous behavior at some points whilst not detrending actual exogenous time dependent behaviour at other points in time. Further trend determination methods are proposed, especially for much shorter trend determination time frames as we had to assume stationarity of the exogenous intensity when using long trend determination time frames.

We choose not to use detrending for the number of mid-price changes prediction algorithm.

4.4 Calibration: Sample Frame

Theoretically the more input data we feed into the algorithm, the better the prediction for the Hawkes Poisson parameters and the number of mid-price changes should become. However, the longer the sample frame, the more computational power is needed. We tested the prediction algorithm with different length of input frame length (10min, 30 min, 60min and 90min). Recall that we expect the assumption of a constant unconditional intensity to be truer for shorter sample timeframes.

4.4.1 Test Setup

The test setup and performance indicators were the same as for the trend deciding test. (See page 26)

4.4.2 Test Result

Label	8	1	7	10
Kernel	Exp	Exp	Exp	Exp
Trend	3 month (15%-85%)	3 month (15%-85%)	3 month (15%-85%)	3 month (15%-85%)
Sample Frame (min)	**10**	**30**	**60**	**90**
Absolute Value of Median of Relative Deviation	6.8%	4.1%	6.1%	7.3%
Mean of Relative Deviation	66.6%	53.2%	52.1%	49.2%
False Positive Rate Q90	9.2%	3.7%	3.3%	3.1%
True Positive Rate Q90	57.4%	51.0%	50.2%	50.6%

Label	4	11
Kernel	Exp	Exp
Trend	No Detrending	No Detrending
Sample Frame (min)	**30**	**90**
Absolute Value of Median of Relative Deviation	1.7%	9.6%
Mean of Relative Deviation	31.7%	21.3%
False Positive Rate Q90	1.8%	1.2%
True Positive Rate Q90	65.3%	59.4%

Label	2	9
Kernel	Pow	Pow
Trend	3 month (15%-85%)	3 month (15%-85%)
Sample Frame (min)	**30**	**90**
Absolute Value of Median of Relative Deviation	4.8%	10.9%
Mean of Relative Deviation	54.1%	45.9%
False Positive Rate Q90	3.7%	1.2%
True Positive Rate Q90	48.6%	23.9%

Table 6: Forecasting performance of multiple setups to compare how the length of the sample frame impacts forecasting performance. See figure 11

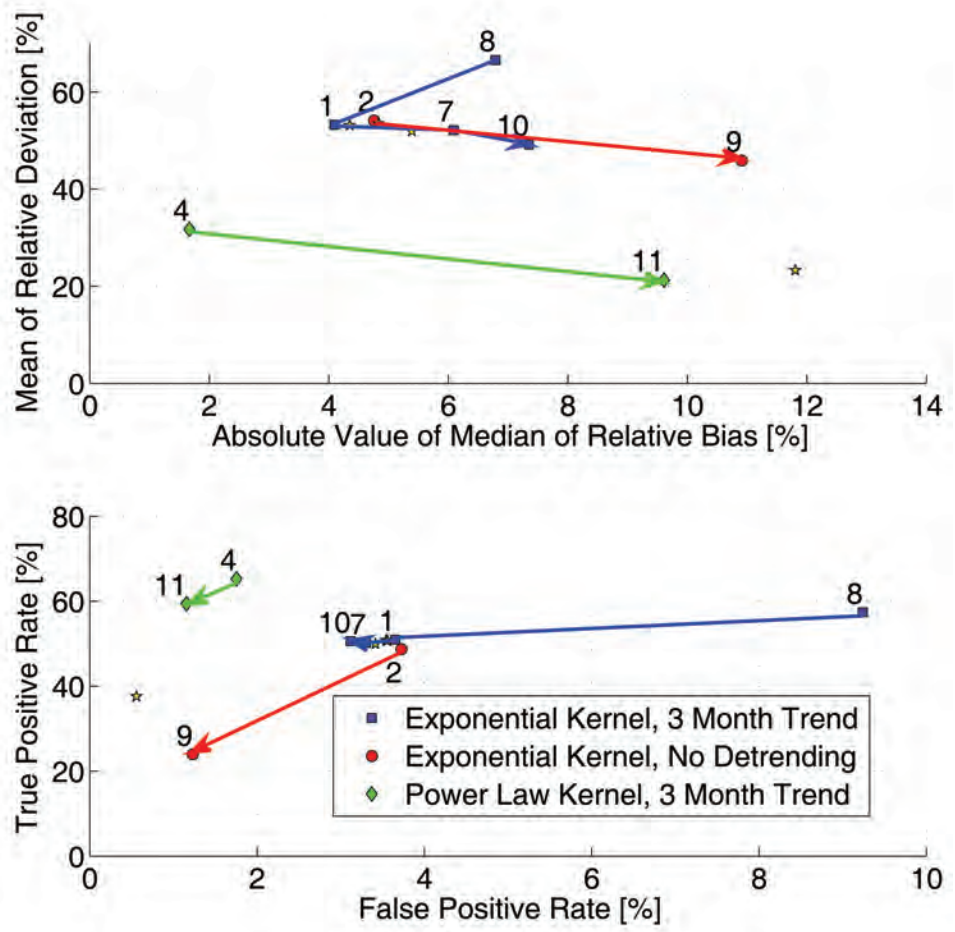

Figure 11: Top figure: Median versus Mean Relative Deviation for all different experimental configurations. Bottom figure: False positive rate versus true positive rate for all different experimental configurations.The arrow points from short sample frames to larger sample frames.

As expected, the longer the sample frame, the lower the mean of relative deviation simply because we have a larger number of data points to perform the calibrations on. We see a big increase in mean deviation error as well as median of relative deviation from 10 minutes to 30 minutes. While the mean deviation further decreases from 30 minutes to 60 minutes and 90 minutes respectively, the median of relative bias worsens.

We see a similar picture for the false positive rate and true positive rate analysis. The false positive rate more than halves from 9.2% to 3.7% when increasing the sample frame from 10 minutes to 30 minutes for the exponential kernel configuration. The further improvements from 30 minutes to 60 minutes and to 90 minutes are much smaller.

Increasing the input frame generally lowers the numbers of predicted mid-price changes. While there is big reduction in the false positive rate for the power law kernel configuration when increasing the input frame, the true positive rate falls to the extremely low value of under 25%. This is not acceptable for a prediction algorithm.

As the pre-trading hours dynamic is different from the trading hours dynamic, the sample frame must fully lie during trading hours. Hence we cannot predict the number of mid-price changes in the 'heat up phase' in the morning. One option would be to gradually increase the sample frame length during the early morning with the knowledge that the first predictions are less accurate. Further tests are suggested in the future with such a system.

We chose a sample frame length of 30 minutes.

4.5 Calibration: Choice of Kernel

4.5.1 Test Setup

The test setup and performance indicators were the same as for the trend deciding test. (See page 26)

4.5.2 Test Result

Label	1	2	Label	10	9
Kernel	**Exp**	**Pow**	**Kernel**	**Exp**	**Pow**
Trend	3 month (15%-85%)	3 month (15%-85%)	Trend	3 month (15%-85%)	3 month (15%-85%)
Sample Frame (min)	30	30	Sample Frame (min)	90	90
Absolute Value of Median of Relative Deviation	4.1%	4.8%	Absolute Value of Median of Relative Deviation	7.3%	10.9%
Mean of Relative Deviation	53.2%	54.1%	Mean of Relative Deviation	49.2%	45.9%
False Positive Rate Q90	3.7%	3.7%	False Positive Rate Q90	3.1%	1.2%
True Positive Rate Q90	51.0%	48.6%	True Positive Rate Q90	50.6%	23.9 %

Label	11	12
Kernel	**Exp**	**Pow**
Trend	No Detrending	No Detrending
Sample Frame (min)	90	90
Absolute Value of Median of Relative Deviation	9.6%	11.8%
Mean of Relative Deviation	21.3%	23.3%
False Positive Rate Q90	1.2%	0.6%
True Positive Rate Q90	59.4%	37.6%

Table 7: Forecasting performance of multiple setups to compare how the choice of kernel impacts forecasting performance. See figure 12

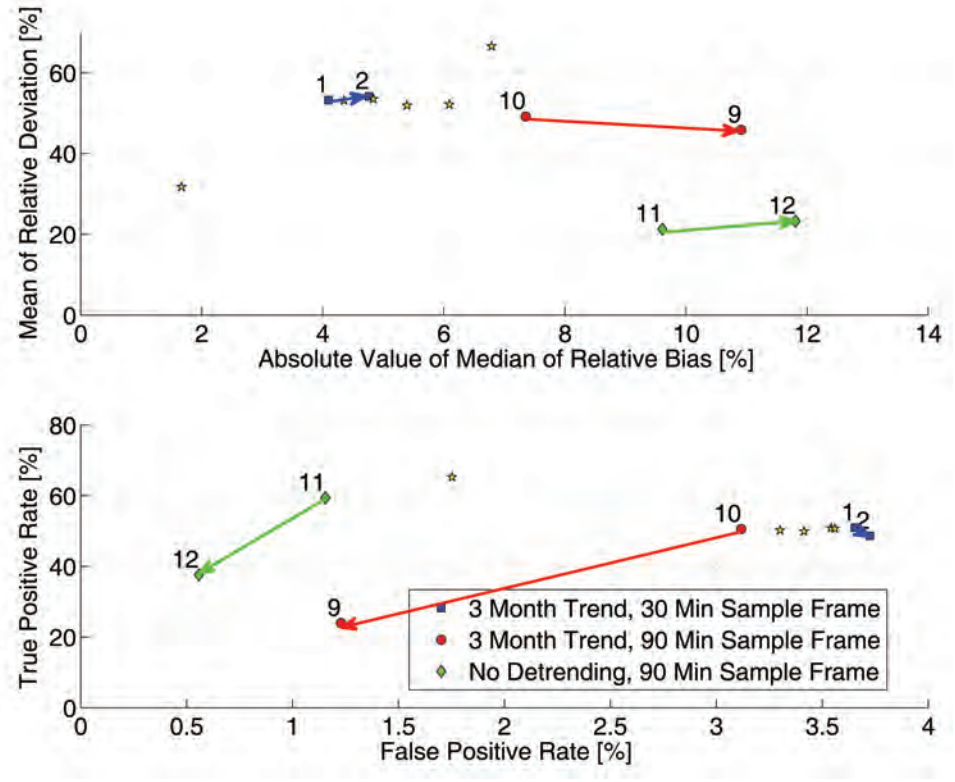

Figure 12: Top figure: Median versus Mean Relative Deviation for all different experimental configurations. Bottom figure: False positive rate versus true positive rate for all different experimental configurations. The arrow points from exponential kernel to power law kernel.

The median of relative bias is higher for the power law kernel than for the exponential kernel.

The mean relative bias is slightly higher or slightly lower for power law kernel compared to exponential kernel depending on which other parameters were chosen. While the absolute median of relative bias is worse for long memory power law kernel, we cannot yet justify the choice of one kernel over the other just from the top figure of figure 12.

Logically the difference between the exponential kernel and the power law

kernel for shorter sample frames is less pronounced than for longer sample frames. Power law kernels are long memory kernels and would need long sample frames to correctly estimate parameters and the number of mid-price changes. On the other hand, for longer sample frames the power law kernel estimates a much lower number of events, and as such, lowers the false positive rate and lowers the true positive rate. We expect our prediction to find at least 50% of the extreme trading minutes and thus chose the exponential kernel over the power law kernel.

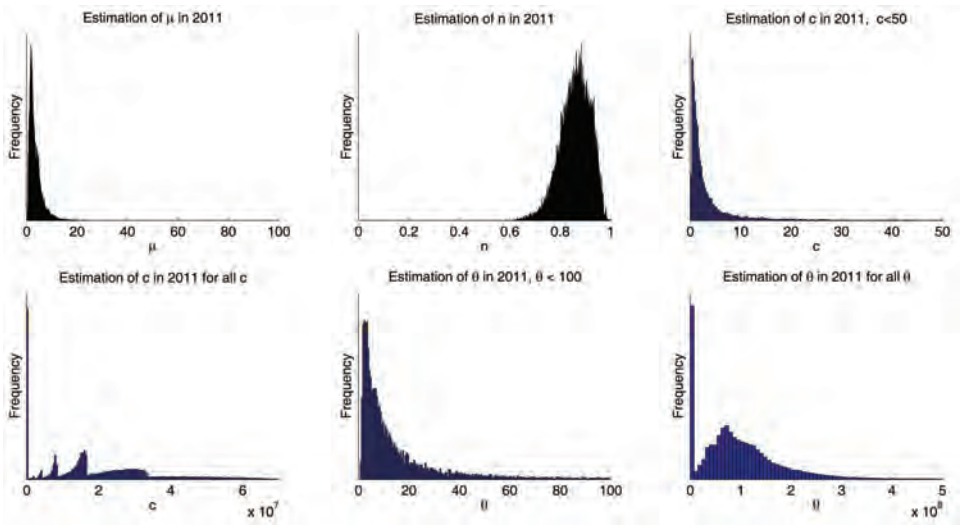

Figure 13: Distribution of power law kernel parameters in 2011 estimated through configuration 10 (90 minutes sample frame, no detrending used, power law kernel). Estimation of parameters was done using the estimation algorithm described in chapter 3.3 using the maximum likelihood estimator.

In figure 13 we examine all minutes sampled in 2011 for a sample frame of 90 minutes with no detrending to find the distribution of estimated power law Hawkes Poisson parameters (μ, n, c and θ). For μ we see a peak at five. As can be seen in figure 8 the number of mid-price changes goes much higher than five, implying that the majority of events were essentially endogenous. This is supported by the near-critical value for n, as can be seen in the respective subplot of Figure 13.

If we look at the distribution of power law kernel parameters in 2011 estimated through configuration 10, we can see extremely high values for c and θ. This implies a short memory, similar to an exponential short term memory kernel. This is a further indication that the market follows a exponential kernel dynamics. The peaks on the bottom left graph for c are due to numerical issues in the estimation algorithm. The bottom subplots are specific to the parameters of the power law kernel. The central plot simply magnifies the last subplot.

This could be a further indicator that the real life data may not follow the Hawkes Poisson dynamics with power law long memory kernel. For that reason, and the performance indicators comparing power law kernel setups with exponential law setups, **we chose to use the exponential kernel for our prediction algorithm**.

4.6 Test Results Combined

To recapitulate the last chapters, we decided to use **300 simulations**, **not use detrending**, **chose an exponential kernel** and use a **sample frame length of 30 minutes**.

Not only would we like to calibrate the model to its optimum, we also need to compare our algorithm to different prediction methods. We chose to compare the prediction performance of our algorithm with the prediction performance of the naive approach where we use the average arrival rate $\langle\lambda\rangle$ to predict the average number of events in the prediction time frame: E[Midprice Changes] = $\langle\lambda\rangle\Delta t$. The derivation for the average intensity is the following

$$\langle \lambda(t) \rangle = \langle \mu + \int_{-\infty}^{t} \varphi(t-s) dN(s) \tag{20}$$

$$= \langle \mu \rangle + \int_{-\infty}^{t} \varphi(t-s) \langle \lambda(s) \rangle ds \tag{21}$$

$$= \mu + \langle \lambda(t) \rangle \int_{0}^{\infty} \varphi(t) dt \tag{22}$$

$$= \mu + \langle \lambda(t) \rangle n \tag{23}$$

which leads to (24)

$$\langle \lambda \rangle = \frac{\mu}{1-n} \tag{25}$$

where φ expresses the positive influence of the past events t_i on the current value of the intensity.

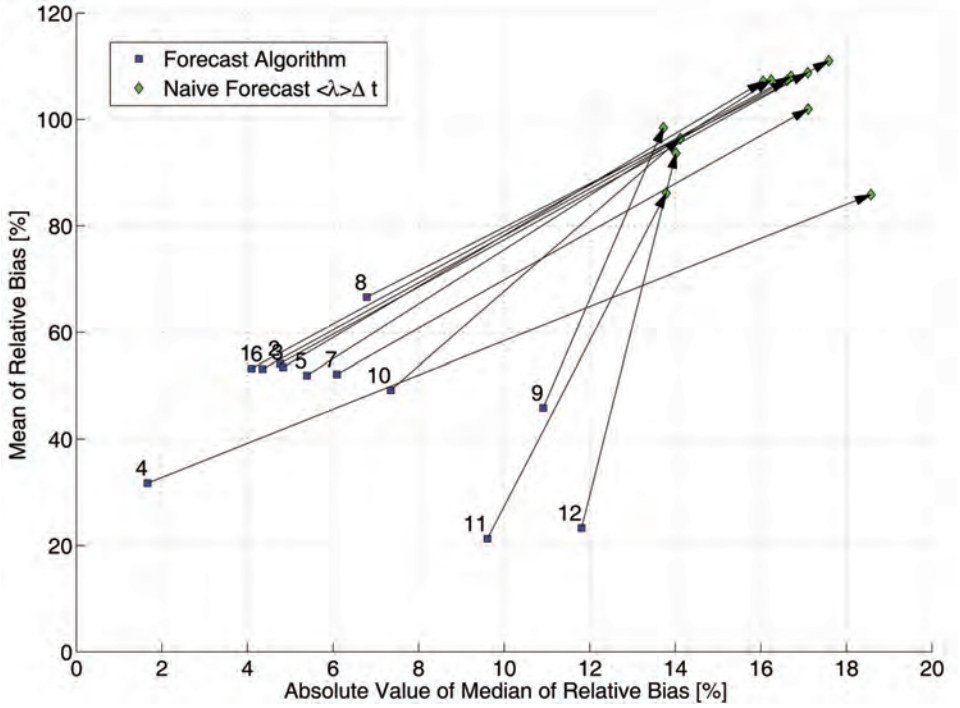

Figure 14: Median versus Mean Relative Deviation for all different experimental configurations. The arrow points from our prediction algorithm to the prediction using the averaged intensity $\langle\lambda\rangle$. Clearly, the naive approach consistently yields worse results.

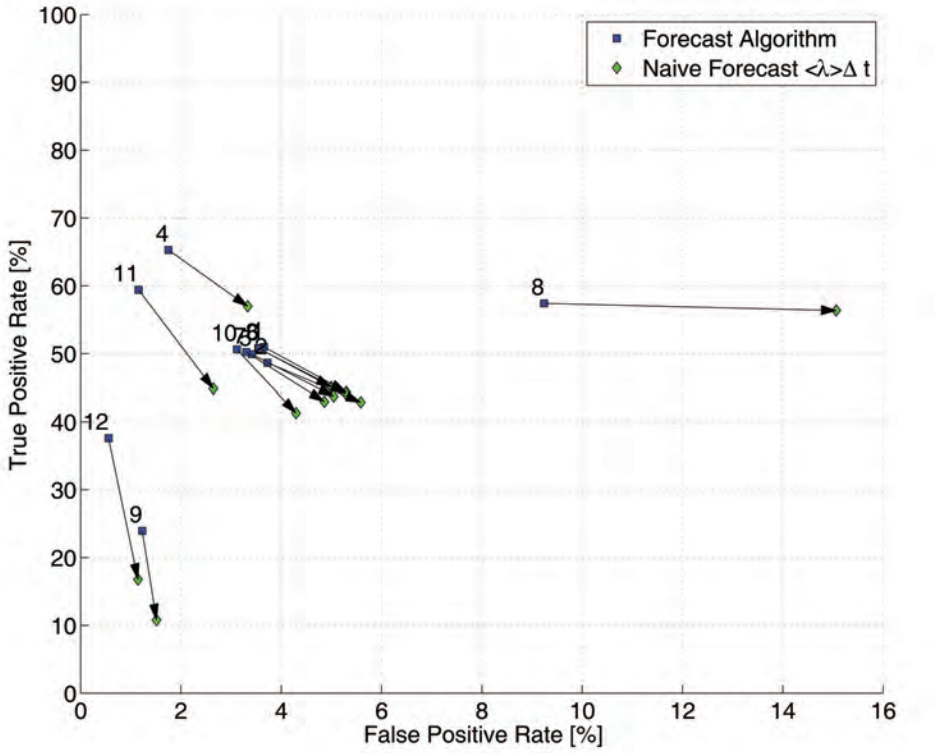

Figure 15: False positive rate versus true positive rate for all different experimental configurations. The arrow points from our prediction algorithm to the prediction using the naive approach via the averaged intensity $\langle\lambda\rangle$. Clearly, the naive approach consistently yields worse results.

We can clearly see that our algorithm outperforms the simple algorithm of $\langle\lambda\rangle\Delta t$

Furthermore figure 15 shows that setup **Setup 4 and 11 yield the best prediction**. These are the setups with exponential kernel and no detrending used:

Label	Kernel	Trend	Input Frame (min)
4	Exp	No Detrending	30
11	Exp	No Detrending	90

If a trader uses this algorithm as a tool, he needs to know the trustworthiness of a positive (extreme event) prediction or negative (non-extreme event) prediction. The positive predictive value

$$\text{PPV} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{26}$$

and negative predictive value

$$\text{NPV} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Negative}} \tag{27}$$

for those configurations in 2011 are:

Setup	4	11
Positive Predictive Value Q90	80.7%	85.9%
Negative Predictive Value Q90	96.2%	95.3%

5 Stability of Prediction Algorithm

After calibrating the prediction algorithm, we want to know how stable the Hawkes Poisson algorithm is in predicting the number of mid-price changes. How does the prediction performance change for different realities (different Hawkes Poisson parameters)? As we concluded in chapter 4.5.2 it is reasonable to assume the dynamics of the E-Mini market in 2011 are well described by an exponential short memory kernel Hawkes Poisson process. Thus we would like to test how our prediction would perform for different values of n as well as for τ.

5.1 Prediction Quality of Calibrated Algorithm Based on Changing Hawkes Poisson Parameters

Testing prediction performance for different parameter realities requires that we know the parameters of said reality, as well as be able to change just one parameter at a time. This is possible for a synthetically created time series as explained in chapter 4.2.1.

The configuration for a given parameter set for the exponential Hawkes Poisson process was as follows:

1. The length of the synthetic time series is defined from $[0\ t_{end}]$.
2. We define a time t_0 with $0 < t_0 < t_{end}$ from which on we would like to predict and the length of the time frame which ends at t_0 and will be the input to the algorithm described in chapter 3.3 to predict the parameters (μ, n, τ for short memory exponential Kernel and μ, n, c, θ for the long memory power law kernel.) (Sample frame)
3. A synthetic time series for the time window $[0\ t_{end}]$ length is simulated by applying the intensity λ of the Hawkes Poisson model (equation 3) for every event in the sample frame.
4. The parameters μ, n and τ are estimated on the sample frame by using the maximum likelihood estimator method described in chapter 3.3
5. Simulation for a prediction timeframe starting at time t_0 estimated parameters. We measured the discrepancy between the number of events our algorithm predicted in the timeframe $[t_0\ t_{end}]$ and the actual number of events in the (synthetic) reality. This step is repeated N_{sim} times and the results are averaged.
6. As the process of creating a reality is stochastic in itself, the entire experiment is repeated and results are averaged.

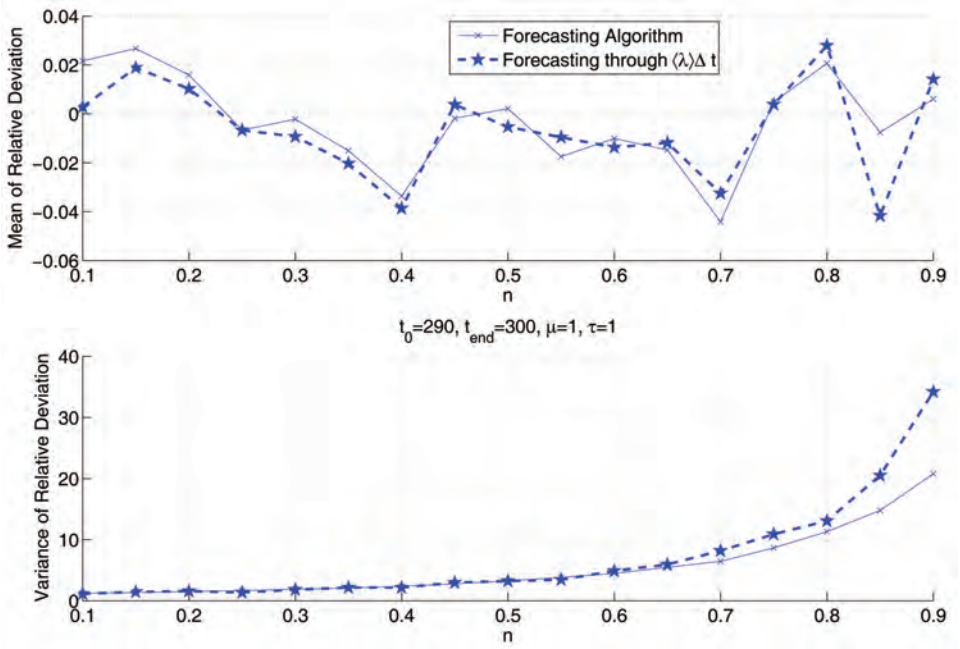

Figure 16: Relative Deviation and Variance for constant $\mu = 1$ and $\tau = 1$ but varying n both for the Hawkes Poisson prediction algorithm as well as the naive assumption of $\langle\lambda\rangle\Delta t$ for prediction.

As we work within a reality perfectly created with a Hawkes Poisson process, the deviation is much smaller than on the test performed on real market data. We are, however, only interested in the behavior of the prediction for changing parameters, not its value.

The relative mean deviation from events predicted and actual number of events is not dependent on the endogenous parameter, the branching ratio n. The change we see in the top of figure 16 can be attributed to noise. This is even more surprising, as the values of n approach criticality, the system becomes more and more complex.

The variance of the relative deviation, however, increases with n. This is particularly true for the prediction with the naive assumption and a further argument to use the Hawkes Poisson prediction algorithm to predict the number of events (i.e. mid-price changes)

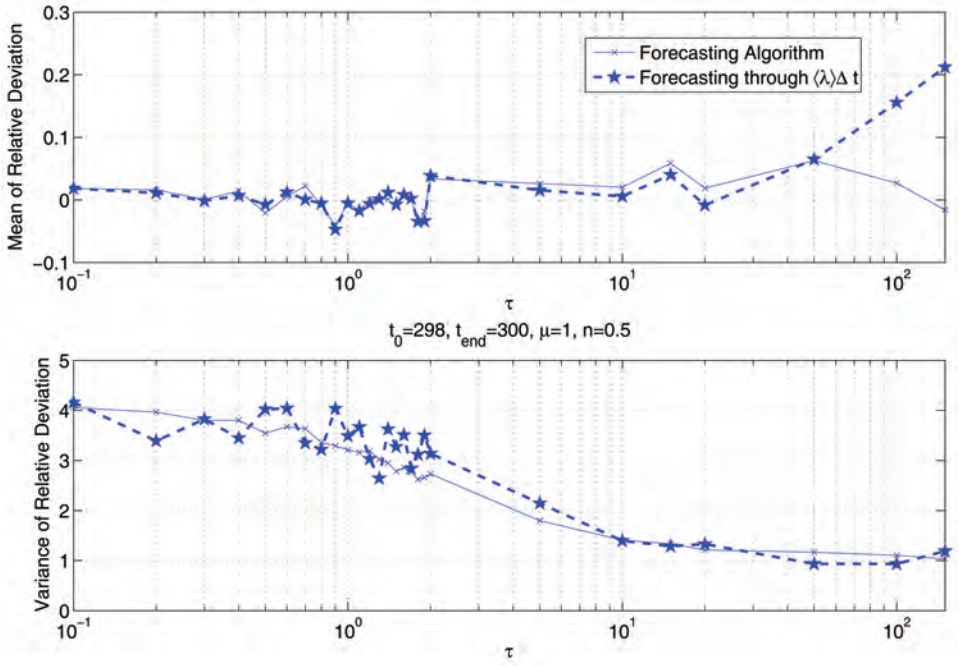

Figure 17: Relative Deviation and Variance for constant $\mu = 1$ and $\tau = 1$ but varying τ both for the Hawkes Poisson prediction algorithm as well as the naive assumption of $\langle\lambda\rangle\Delta t$ for prediction.

For very high values of τ, which implies that the probability of an event having a daughter event at a later time is much higher, the relative deviation for the naive prediction gets worse, while the detailed prediction algorithm based on the Hawkes Poisson model performs equally well. However, the variance of deviation improves, gets smaller, for longer lifetimes τ. We show in figure 8 that the E-mini market has rather short lifetimes between 0.1 and 0.4 minutes, so we can assume good mean prediction but a lot of variance from our algorithm.

5.2 Analysis of the Evolution of Selected Hawkes Poisson Parameters

We calibrated our algorithm for trading data of the E-Mini $S\&P500$ market in the year 2011. As a further test we analysed how the prediction algorithm performed for different years. We chose to include the years 2007 to 2012

since this includes the turbulent years of the stock market crash 2008 as well as 2010 and 2011 flash crashes.

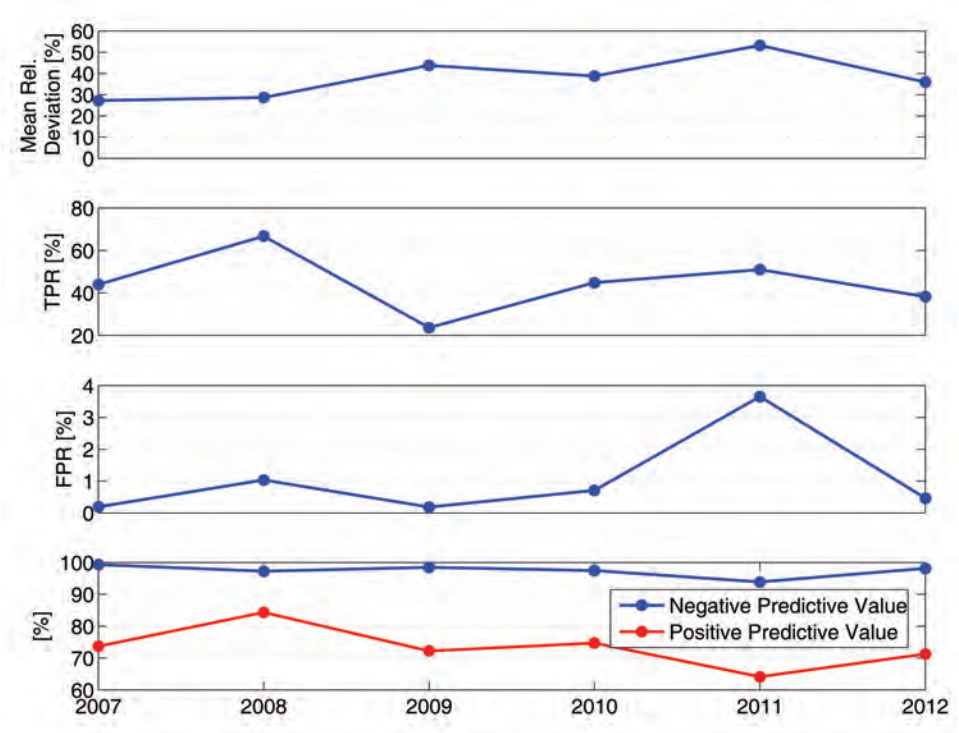

Figure 18: Performance indicators for Hawkes Poisson prediction algorithm on $S\&P500$ E-Mini market data from 2007 to 2012. From top to bottom. First graph: Mean of relative deviation of prediction. Second graph: True Positive Rate TPR is the percentage of all extreme events the algorithm could predict of each respective year. Third graph: False positive rate FPR is the percentage of all non-extreme events, which are falsely flagged as extreme events by the prediction algorithm. Fourth graph: True positive value TPV tells the ratio of correctly predicted extreme events to the total number of predicted extreme events. Negative Predictive Value NPV tells the ratio of correctly predicted non-extreme events to the total number of predicted non-extreme events.

Incidentally, for most performance indicators, the prediction algorithm performed better on other years than for 2011. We believe that this is caused by the chaotic behavior before and during the flash crash of 2011. In 2011,

we had quite high exogenous intensity as well as a high endogenous intensity (see figure 20). In 2011 the prediction algorithm overestimated the number of mid-price changes by 50%. In 2007 this nearly half, with just over 25%. Future studies may test if the prediction algorithm applies to other years in the past in order to see if this trend continues. However, using data from the past may lead to some issues. Firstly, the volume of trades was lower, so the prediction algorithm would have much less data to work with. Secondly, the endeogenity was much lower in the past [3], thus most events occurred due to exogenous factors, which cannot as easily be predicted with market data.

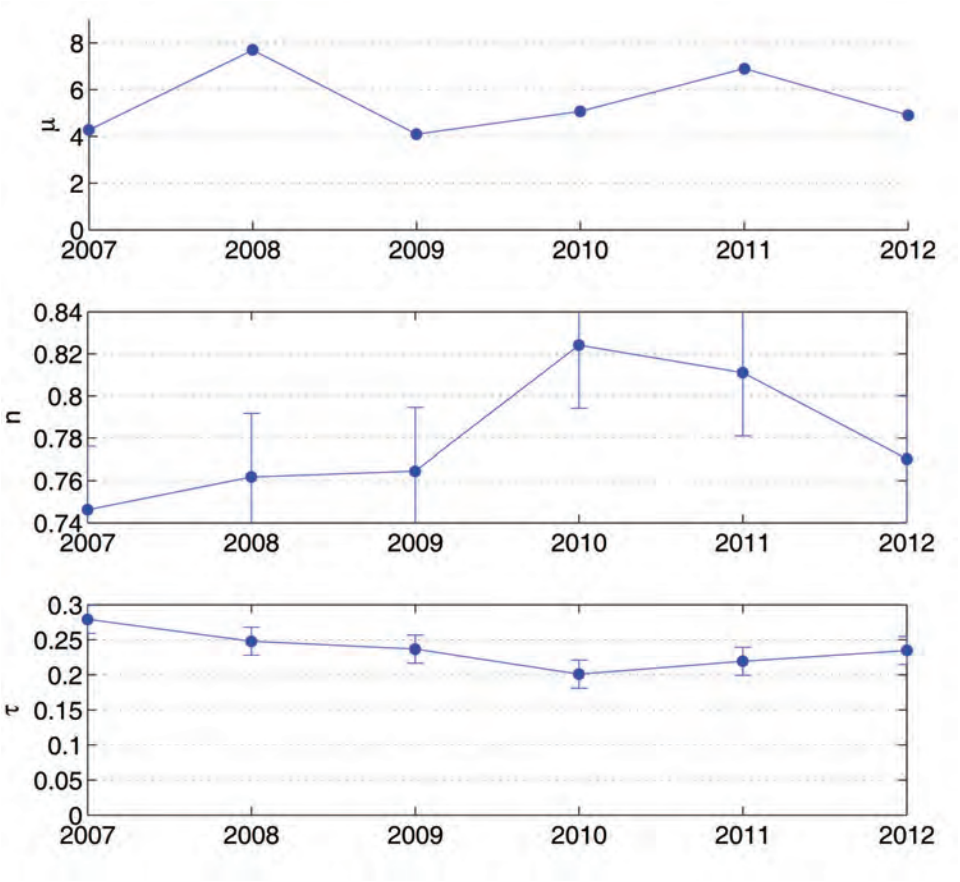

Figure 19: Estimated average yearly unconditional intensity(μ), endeogenity (n) and over time from 2007 to 2012.

There is no overall clear trend to be seen for the exogenous intensity μ over the years. However, the two peaks in 2008 and 2011, respectively, coincide with the stock market crash in 2008 and the the August 2011 stock market crash. This crude observation implies that crises and crashes do have not only an endogenous but also exogenous component.

There is a strong increase from 2007 to 2010 with a decrease in the following years for the endogenous factor n. The branching ratio ranges between 0.75 to 0.83, which is approaching the critical value of 1. What is striking is that the peaks do not resemble those from the unconditional intensity. The peak of the branching ratio as the direct measure of endogeneity implies that the crash or more generally the activity that occurred in 2011 are mostly of endogenous nature.

The slight downward trend of τ over the years is very crude and can, in fact, not be concluded as such. A downward trend would imply that, over the years, the memory kernel became less influential, meaning that the tendency of an event to trigger a single daughter decreased. A high branching ratio, however, could, in principle, more than compensate for this.

6 Conclusion

We were able to successfully implement and calibrate a forecasting algorithm to predict the number of mid-price changes. The two most successful setups were based on the Hawkes Poisson process with exponential kernel.

Kernel	Trend	Input Frame (min)
Exp	No Detrending	30
Exp	No Detrending	90

A strong increase in forecasting performance when choosing **not** to detrend the data of the in-frame was observed. This could yield the following two interesting conclusions. Either the individual underlying exogenous intensity is quite unique for each day, even though, over time, we can see a trend emerge. Or the way we measured the trend was not optimal. Trend for

shorter trend determination time frames are proposed for further analysis.

In this work, the events marked by time stamps were mid-price changes. However, in theory it is possible to apply the methodology developed in this work onto other events, such as number of trades or price changes. In a next step, the prediction algorithm, calibrated with the settings discussed above, can be applied to a bi-variate setup where we distinguish positive price changes from negative price changes. This leads to a forecasting algorithm which can, in theory, predict future price levels and can be used as a basis in high frequency trading for buy/sell choices.

7 Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Dr. Vladimir Filimonov for his continuous support of my Master Thesis and research. Dr. Filimonov's guidance, as well as his patience, motivation, enthusiasm, and immense knowledge, helped me throughout the research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Master Thesis.

My sincere thanks also goes to Stefan Rustler, who inspired my research with his great insights about the thesis topic as well as through our interesting discussions about the Hawkes Poisson process, even late into the night at the university.

Lastly, I would like to express my gratitude to Suzanne Greene for her review of my thesis.

References

[1] V. Filimonov, D. Sornette. *Quantifying reflexivity in financial markets: towards a prediction of flash crashes*, Phys. Rev. E 85 (5), 056108 (2012). 1, 3.2

[2] V. Filimonov, D. Bicchetti, N. Maystre, D. Sornette *Quantification of the high level of endogeneity and of structural regime shifts in commodity markets*, Journal of international Money and finance 42 (2014) 174-192 4.3.1

[3] V. Filimonov, D. Sornette. *Apparent criticality and calibration issues in the Hawkes self-excited point process model: application to high-frequency financial data.*, arXiv:1308.6756 [q-fin.ST](2013). 2, 3.1, 3.2, 3.2, 3.3, 3.3, 2, 4.1.3, 4.1.3, 4.1.3, 4.3.1, 5.2

[4] S. J. Hardiman, N. Bercot, J. P. Bouchaud. *Critical reflexivity in financial markets: a Hawkes process analysis*, The European Physical Journal B 86 (10), 1-9 (2013).

[5] I. M. Toke. *An Introduction to Hawkes Processes with Applications to Finance*, http://lamp.ecp.fr/MAS/fiQuant/ioane_files/HawkesCourseSlides.pdf. Accessed June 30, 2014. 4, 4, 5

[6] S Rustler, Dr. V. Filimonov, Prof. D. Sornette *Towards Quantifying Self-Excitation in Twitter Messaging.* (2014) 3.4

[7] P. Hewlett *Clustering of order arrivals, price impact and trade path optimisation , In Workshop on Financial Modeling with Jump processes*, Ecole Polytechnique (2006) 3.2

[8] C. G. Bowsher *Modelling security market events in continuous time: Intensity based, multivariate point process models*, Journal of Econometrics 141 (2) (2007) 876-912 3.2

[9] R. Cont *Statistical Modeling of High Frequency Financial Data: Facts, Models and Challenges*, IEEE Signal Processing 28 (5) (2011) 16-25 3.2

[10] D. Vere-Jones, T. Ozaki *Some examples of statistical estimation applied to earth-quake data I. Cyclic Poisson and self exciting models*, Annals of the Institute of Statistical Mathematics 34 (1) (1982) 189-207 3.2

[11] D. Vere-Jones *Stochastic Models for Earthquake Occurrence*, Journal of the Royal Statistical Society. Series B (Methodological) 32 (1) (1970) 1-62 3.2

[12] Y. Ogata *Statistical models for earthquake occurrences and residual analysis for point processes*, Journal of the American Statistical Association 83 (401) (1988) 9-27 3.2, 3.3, 3.3

[13] A. Helmstetter, D. Sornette *Subcritical and supercritical regimes in epidemic models of earthquake aftershocks*, Journal of Geophysical Research 107 (B10) (2002) 2237 3.2

[14] T. Utsu *A statistical study of the occurrence of aftershocks*, Geophyiscial Magazine 30 (1961) 521-605 3.2

[15] T. Utsu, Y. Ogata *The centenary of the Omori formula for a decay law of aftershock activity*, Journal of Physics of the Earth 41 (1) (1995) 1-33 3.2

[16] L. Knopoff, Y. Y. Kagan *Stochastic synthesis of earthquake catalogs*, Journal of Geophysical Research 86 (B4) (1981) 2853-2862

[17] Y. Y. Kagan, L. Knopoff *Statistical short-term earthquake prediction*, Science 236 (4808) (1987) 1563

[18] F. Papangelou *Integrability of Expected Increments of Pint Processes and a Related Random Change of Scale*, Transactions of the American Mathematical Society 165 (1972) 483-506 3.3

[19] E.J. Fama *Efficient Capital Markets: A Review of Theory and Empirical Work*, The Journal of Finance 25, 383 (1970) 2

[20] E.J. Fama *Efficient Capital Markets: II*, The Journal of Finance 46, 1575 (1991) 2

[21] P.A. Samuelson *Proof that Properly Anticipated Prices Fluctuate Randomly*, Industrial Management Review 6 41 (1965) 2

[22] P.A. Samuelson *Proof that Properly Discounted Present Values of Assets Vibrate Randomly*, Bell Journal of Economics, The Rand Corporation, vol. 4(2), 369 (1973) 2

[23] T. Ozaki *Maximum likelihood estimation of Hawkes' self-exciting point processes*, Annals of the Institute of Statistical Mathematics 31 1 (1979) 145-155 3.3, 3.3

[24] S. J. Hardiman, N. Bercot, J.-P. Bouchaud *Critical reflexivity in financial markets: a Hawkes process analysis* arXiv:1302.1405 [q-fin.ST] 2, 4.1.3

[25] Francine Gresnigt, Erik Kole, Philip Hans Franses*Interpreting Financial Market Crashes as Earthquakes: A New Early Warning System for Medium Term Crashes*, Tinbergen Institute Discussion Paper 14-067/III 3.2

[26] D. Sornette, S. Utkin *Limits of Declustering Methods for Disentangling Exogenous from Endogenous Events in Time Series with Foreshocks, Main shocks and Aftershocks*, Physical Review E79, 061110 (2009) arXiv:0903.3217 [q-fin.ST] 3.3

8 Appendix

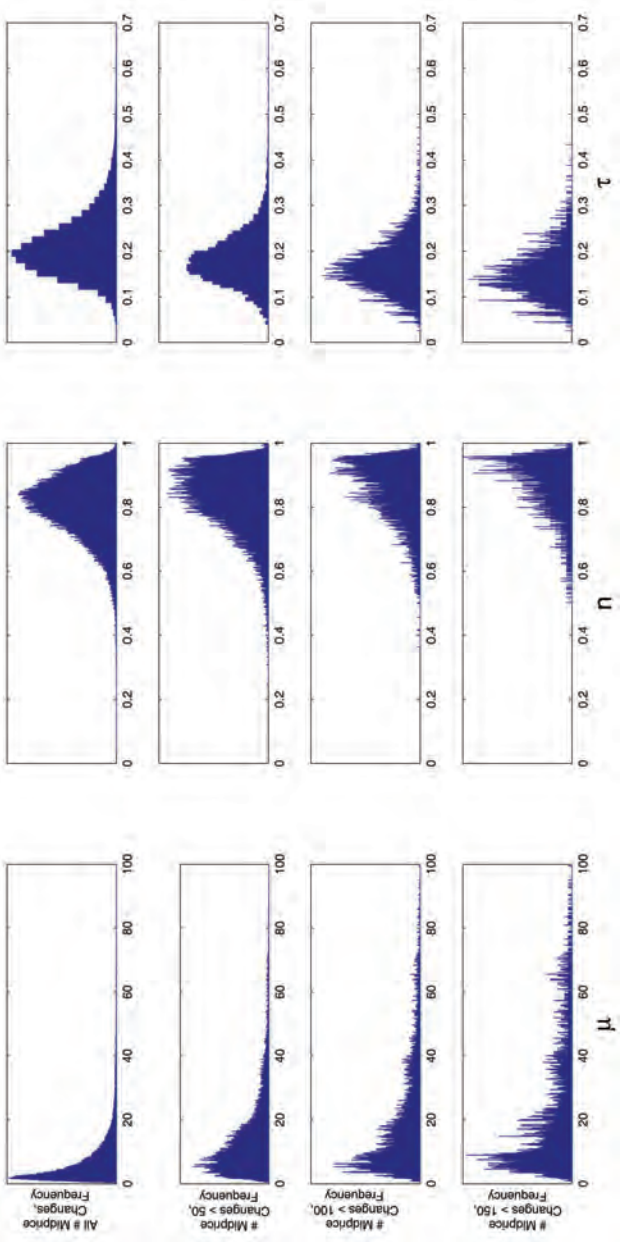

Figure 20: Distribution of parameters μ, n and τ in the year 2011 for different subset of sample minutes with different amount of midprice changes.

Label	1	2	3	4	5	6	7	8	9	10	11	12
Kernel	Exp	Pow	Exp	Exp	Exp	Exp	Exp	Exp	Pow	Exp	Exp	Pow
Trend	3 month (15-85%)	3 month (15-85%)	EWMA	No Detrending	1 month (15-85%)	1 month (25%-75%)	3 month (15-85%)	3 month (15-85%)	3 month (15-85%)	3 month (15-85%)	No Detrending	No Detrending
Sample Frame (min)	30	30	30	30	30	30	60	10	90	90	90	90
First Quartile	-47.3%	-47.6%	-45.6%	-33.8%	-46.2%	-47.1%	-45.2%	-37.7%	-46.4%	-43.7%	-41.2%	-44.5%
Median	-4.1%	-4.8%	-4.8%	-1.7%	-5.4%	-4.4%	-6.1%	6.8%	-10.9%	-7.3%	-9.6%	-11.8%
Third Quartile	71.5%	72.6%	67.5%	45.1%	67.7%	70.9%	65.8%	79.5%	59.5%	60.5%	34.8%	37.8%
Q75-Q25	118.8%	120.2%	113.1%	78.9%	113.9%	118.0%	111.0%	117.2%	105.9%	104.2%	76.0%	82.3%
Mean	53.2%	54.1%	53.5%	31.7%	51.9%	53.2%	52.1%	66.6%	45.9%	49.2%	21.3%	23.3%
Variance	641.2%	636.5%	693.0%	203.4%	579.8%	646.3%	606.0%	729.8%	428.5%	508.4%	175.5%	202.6%
Variance (w/o 5% outlier)	102.9%	106.3%	101.0%	39.2%	103.3%	103.0%	81.5%	99.8%	75.9%	72.4%	36.4%	41.9%
Sensitivity Q90	51.0%	48.6%	50.8%	65.3%	49.9%	50.7%	50.2%	57.4%	23.9%	50.6%	59.4%	37.6%
False Negative Rate Q90	49.0%	51.4%	49.2%	34.7%	50.1%	49.3%	49.8%	42.6%	76.1%	49.4%	40.6%	62.4%
Specificity Q90	96.3%	96.3%	96.5%	98.2%	96.6%	96.4%	96.7%	90.8%	98.8%	96.9%	98.8%	99.4%
False Positive Rate Q90	3.7%	3.7%	3.5%	1.8%	3.4%	3.6%	3.3%	9.2%	1.2%	3.1%	1.2%	0.6%
Positive Predictive Value Q90	64.1%	61.7%	64.9%	80.7%	64.3%	64.7%	66.0%	64.4%	52.3%	68.4%	85.9%	78.7%
Negative Predictive Value Q90	93.9%	93.8%	93.8%	96.2%	94.0%	93.8%	93.8%	88.0%	95.8%	93.6%	95.3%	96.7%
Sensitivity Q95	46.8%	43.3%	47.0%	62.8%	45.8%	46.6%	45.8%	52.3%	14.9%	45.0%	57.7%	30.3%
False Negative Rate Q95	53.2%	56.7%	53.0%	37.2%	54.2%	53.4%	54.2%	47.7%	85.1%	55.0%	42.3%	69.7%
Specificity Q95	98.0%	98.0%	98.0%	99.2%	98.1%	98.0%	98.2%	95.0%	99.5%	98.3%	99.5%	99.8%
False Positive Rate Q95	2.0%	2.0%	2.0%	0.8%	1.9%	2.0%	1.8%	5.0%	0.5%	1.7%	0.5%	0.2%
Positive Predictive Value Q95	59.3%	55.0%	59.6%	80.6%	58.6%	59.2%	61.7%	58.7%	39.2%	64.3%	86.8%	80.1%
Negative Predictive Value Q95	96.7%	96.8%	96.7%	98.0%	96.9%	96.7%	96.6%	93.6%	98.2%	96.4%	97.6%	98.6%

Table 8: