Univerzita Karlova v Praze
Fakulta sociálních věd

Institut ekonomických studií

DIPLOMOVÁ PRÁCE

Modeling the Czech Stock Market with High-Frequency Time-Series Methods

Modelování českého kapitálového trhu pomocí metod vysokofrekvenčních časových řad

Vypracoval: **Bc. Vít Bubák**
Vedoucí: **PhDr. Filip Žikeš**
Akademický rok: **2004 - 2005**

Prohlášení

Prohlašuji, že jsem diplomovou práci vypracoval samostatně a použil pouze uvedené prameny a literaturu.

Except where reference is made to the work of others, the work described in this thesis is my own or was done in collaboration with my thesis supervisor.

V Praze dne 30. 6. 2005

Abstract

The thesis is concerned with the empirical modeling of high-frequency data (HF data) under the predictive framework of information based models of market microstructure. The first part of the thesis (Chapter 1) provides a detailed description of microstructure of the Prague Stock Exchange (PSE). In the same chapter we also analyze basic properties of HF data using a sample of securities traded on the Exchange. Chapter 2 focuses on price duration (PD) process. Using a set of three of the most liquid securities traded on PSE's main market, we first examine whether the intensity of bid-ask quote arrivals carries any information about the state of the market. We provide evidence of clustering effect in PDs where short (long) durations tend to be followed by short (long) durations, resp. In addition, we show that larger autocorrelations in PDs tend to persist even after the time-of-day effects have been removed. We then test the predictions of market microstructure models using several proxies: intensity of trading, avg. volume per trade, and avg. spread. Although any straightforward judgements remain at best ambiguous, our results tend to favor the conclusions of information based models. In Chapter 3, we turn to the information content of a trade and try to measure the ultimate impact on a stock price of the part of the trade that is unexpected. We find that (a) full impact of a trade on the security price is not felt instantaneously, (b) as a function of trade innovation size, the impact is nonlinear, positive, increasing, and convex, and (c) the order flow does not seem to be affected by prior quote revisions.

Keywords: market microstructure, high-frequency data, duration analysis, price impact
JEL Classification: C22, C32, C41, E44, G14, G18

Abstrakt

Hlavním cílem práce je empirické modelování vysokofrekvenčních (VF) dat na pozadí informačních modelů tržní mikrostruktury. V první části práce je nejprve podrobně popsána tržní mikrostruktura pražské Burzy cenných papírů (BCCP). V této části jsou také analyzovány základní vlastnosti VF dat na vzorku několika společností z BCCP. Ve 2. kapitole obracíme naši pozornost na proces cenových durací (CD). Na vzorku tří akcií obchodovaných na hlavním trhu BCCP (SPAD) nejprve ověřujeme, zda intenzita kótovaných cen nese nějakou informaci o mikrostruktuře trhu. Předběžná analýza dokazuje, že CD mají tendenci shlukovat se, kdy krátké/dlouhé durace následují krátké/dlouhé durace. Statisticky významné autokorelace v CD navíc přetrvávají i poté, co jsou durace ošetřeny o intradenní efekty. Ve druhé části kapitoly testujeme předpovědi tržní mikrostruktury za použití několika proxy proměnných. Závěry této analýzy jsou v souladu s teoriemi informačních modelů tržní mikrostruktury, jejich interpretace však zůstává přinejmenším nedokonalá. Ve 3. kapitole zkoumáme informační hodnotu obchodu a pokoušíme se změřit konečný dopad jeho neočekávané části na cenu akcie. Naše analýza ukazuje, že (a) konečný dopad obchodu na cenu akcie není jednorázový, (b) jako funkce velikosti neočekávané části obchodu se konečný dopad na kótovanou cenu ukazuje být nelineární, kladný, rostoucí a konvexní, a (c) tok obchodů se nazdá být ovlivněn předcházejícími revizemi kotací.

Klíčová slova: tržní mikrostruktura, vysokofrekvenční data, informační modely
JEL klasifikace: C22, C32, C41, E44, G14, G18

Contents

Acknowledgements

I would like to thank my supervisor, Filip Žikeš, for an immense help and encouragement in writing this thesis. There is no doubt that I would have never successfully accomplished the task had it not been for his eager and enthusiastic approach to the subject. I also owe a great deal to everyone from the Sales and Trading Department at Wood and Company (Prague) for giving me the chance to fully understand the market maker's behavior as well as the general framework of the Prague Stock Exchange's microstructure. My thanks also go to Petr Krejčí of the Prague Stock Exchange for valuable discussion on the details of PSE's main market.

At last but not least, I would not want to forget to thank my family - it is to them that I owe the style of work I always favored. The financial support of the Grant Agency of Charles University under grant no. 297/2004/A-EK/FSV is also greatly appreciated.

Introduction

Most studies published in the financial literature deal with low-frequency, regularly spaced data. While these data have proved to be useful in their own right, there are questions in finance that they could never be expected to answer. The fact that they are always aggregated up to some fixed interval makes them simply too coarse and hence inadequate for the study of questions that arise at much higher frequencies.

Recent advances in computer technology and data collection have made it possible to observe and record the data at the finest of scales. On many of the financial markets today every single transaction can be recorded so that an ultimate frequency or, *ultra-high frequency* (ENGLE, 2000), is a commonplace. Hand in hand with the increasing availability and accessibility of large data sets on ultra-high frequency data (i.e. real time recordings of trades, order arrivals and quote updates) goes their popularity. The reason is obvious: since the ultra-high frequency or transaction-by-transaction data represent the original form of market prices, they can be directly used to study a variety of issues related to the details of price generation process.[1]

The part of finance that studies the transaction processes is called market microstructure analysis. Over the past twenty years, a significant amount of literature has appeared that tries to model the ways how market mechanisms and market designs affect the price formation process. Much of this literature has focused on the price-setting problem confronting market intermediaries.

The Walrasian auctioneer provides the simplest (as well as the oldest) characterization of the price-setting process. The auctioneer announces a potential trading range, and traders determine their optimal order at that price. If there are imbalances in traders' demands and supplies, a new potential price is suggested, and traders then

[1] By appropriately editing the data, we have since realized that it is possible to define almost any event of interest: for example, we can nowadays test the theories of market microstructure (described further in the text) related to the intra-day phenomena such as the real-time dynamics of price adjustment processes, the efficiency of different trading systems in price discovery, or the consequences of strategic behavior of different groups of market participants. Moreover, the ultra-high frequency data can be useful for studying the statistical properties, volatility in particular, of asset returns at lower frequencies. In words of ENGLE AND RUSSELL (2004), with high frequency financial data we do at last "stand to empirically address such questions".

revise their orders. No trading takes place until a market-clearing price is found. Still, most of the markets today differ dramatically from the Walrasian framework. In particular, specific market participants play roles far removed from the passive one of the auctioneer.

Demsetz (1968) was one of the first economists to analyze how the behavior of traders affects the formation of prices. Demsetz argued that while a trader willing to wait might trade at the single price envisioned in the Walrasian framework, a trader not wanting to wait could pay a price for immediacy, i.e. liquidity. This results in two equilibrium prices. Moreover, since the size of the price concession needed to trade immediately depends on the number of traders, the structure of the market could affect the cost of immediacy and thus the market-clearing price.

The price-setting problem examined by Demsetz has been investigated more formally using *inventory based models* of market microstructure. These models view the trading process as a matching problem in which the market maker - or price-setting agent - must use prices to balance supply and demand across time. There are several distinct approaches to modeling how prices are set by market makers: Garman (1976) focused on the nature of order flow; Stoll (1978) and Ho and Stoll (1981) examined the optimization problem facing dealers; and Cohen, Maier, Schwartz and Whitcomb (1981) analyzed the effects of multiple providers of immediacy. Common to each of these approaches are uncertainties in order flow, which can result in inventory problems for the market maker and execution problems for traders.

Still, there is as yet one other major class of models that tries to explain the mechanism of price adjustment process. Based on asymmetric information, these so called *information based* models provide a significantly different explanation of the price generation process compared to the one offered by inventory based models. In these models, three types of transactors are distinguished: the information-motivated transactors with superior information, the liquidity-motivated transactors without the superior information, and the transactors who believe they possess the information but have none. The market maker, who is in the middle of all trades, does not know with which type of transactor he trades and will consistently lose money to the information-motivated group of transactors. To compensate for this loss, the market maker will need to constantly fix different buy and sell prices so that his buy price (the bid) is then smaller than his sell price (the ask). This way, the market-maker fixes his prices conditionally on the type of trade and, effectively, makes money from the liquidity-motivated transactors who are willing to pay the spread for immediacy. The information based models of market microstructure thus explain why even in competitive markets a bid-ask spread will exist as long as there are information-motivated traders.

Starting with Kyle (1985), Glosten and Milgrom (1985) and Easley and OHara (1987), market microstructure research has paid a significant attention to the

effect of asymmetric information on market prices. These studies generally propose that if some traders have superior information about the underlying value of an asset, their trades could reveal what this underlying value is and so affect the behavior of prices. Put differently, each trade could have a *signal value*. Glosten and Milgrom (1985) were the first to incorporate the information that the trade itself can reveal in their own microstructure model. They showed that since the trade has a signal value, following the trade the market maker will revise her beliefs based on what she has just learned from the trade outcome and set new trading prices.[2] The study of Glosten and Milgrom, while important in its own right, further served as an important foundation for the models that focused on the role of time in the price generating process.

Diamond and Verrecchia (1987), Easley and OHara (1992), and Easley et al. (1996), among others, were among the first to suggest models in which the times between transactions enters as endogenous. These models not only extend the information-based models such as that of Glosten and Milgrom (1985), but also the inventory-based models of market microstructure due to Garman (1976) and Stoll (1978).[3]

Overview

Underlying many of the studies on market microstructure is the desire to know how prices are formed in the economy. As already noted, the behavior of prices can be best described with the use of high-frequency data. In other words, only through empirical investigations of the data obtained at the finest of scales can we effectively abord the issues put forward by the theories of market microstructure. We also mentioned that over the years an extensive empirical literature has emerged that exploits the market behavior over fine intervals. Still, a vast majority of these studies is based on the data taken from the Western stock markets such as NYSE or Paris Bourse. As a result, for many *peripheral markets* the relevant empirical research is still missing. The Prague Stock Exchange (PSE), perhaps the most important of all Central European stock markets, is one of these markets.[4] It is for this reason in particular that we choose to

[2] Glosten and Milgrom (1985) use standard inferences of Bayesian learning to determine the bid and ask prices. They show that the bid-ask spread is determined by the simple fact that someone wishes to buy, as well as the nature of the underlying information, the number of informed traders, the trader's elasticity, and the trade size.

[3] We leave further discussion of these models for Chapter (2).

[4] We need to mention that a handful of studies has already analyzed the high-frequency data from the PSE. For example, Hanousek and Nemecek (2002) investigated the relationship between liquidity and information based trading and the possible impact of market microstructure changes on this relationship. Similarly, Hanousek and Podpiera (2003) explored the impact of informed trading on the composition of the bid-ask spread from an information based point of view. Still, in

describe, analyze, and ultimately try to model the high-frequency data coming from this exchange.

Building upon a set of data from the PSE, we thus defy the existing empirical literature on the high-frequency data that concentrates only on the Western stock markets. Our thesis contributes to the empirical literature on market microstructure in the following three original ways. First, it is the first study to give a stand-alone description of the basic characteristics of high-frequency transaction data (incl. price duration data) obtained from the Prague Stock Exchange. Second, it is the first study to model the intraday behavior of price durations and the impact of the corresponding trade-related variables on price generation process using the data from the local stock exchange. This way, the thesis also examines the relevance of information based models of market microstructure for the local stock exchange. And finally, it is also the first study to empirically assess the price impact of stock trades using a robust empirical model based on microeconomic theory. This way, the thesis hopes to examine and consequently better understand the dynamics of trade and quote process on a representative central European stock market.

Outline of the Thesis

The setup of this thesis is as follows. In Chapter (1) we briefly review the most important features of the Prague Stock Exchange. The knowledge of particular characteristics of the PSE such as its organization and trading mechanism is a necessary prerequisite for the analysis that comes later in the thesis. Using a sample of securities from the PSE's main market, in this chapter we also examine the basic features of high-frequency data. A proper understanding of unique as well as more general features of tick-by-tick data such as the discreteness or heterogeneity comes in again as inevitable when dealing with the empirical models in Chapters (2) and (3).

In Chapter (2) we focus on transaction durations and examine the intensity of trading on the PSE. The primary objective here is to provide an empirical evidence that the intensity of transaction arrivals as measured and forecasted by the time between quotes carries information about the state of the market. Merging the trade and quote datasets provided by the PSE and thus linking the price duration process to the trade characteristics obtained over the price durations (i.e., the intensity of trading, the average volume per trade, and the average spread), we also provide evidence of the relevance of information based models for the local stock market. For a sample of frequently traded stocks listed on the PSE's main market, we use a logarithmic extension of the autoregressive conditional duration (log-ACD) model due to Bauwens

none of these studies has the HF data been analyzed as extensively as in this thesis.

AND GIOT (2000). With the autoregressive equation being specified on the logarithm of the conditional expectation of the durations, the log-ACD model allows to avoid the non-negativity constraints on the coefficients implied by the original specification and thus greatly facilitates the testing of market microstructure hypotheses put forward in this chapter.

In Chapter (3) we examine the information content of stock trade as revealed in its effect on stock price. We borrow the model proposed by Hasbrouck (1998) which is based on the assumption that the information content of a trade may be meaningfully measured as the persistent impact of the unexpected component of the trade, i.e. trade's innovation. By focusing on the trade innovation rather than the trade itself, we may avoid misleading inferences due to inventory control or other transient liquidity effects. By considering the persistent impact of the innovation, we concentrate on the information ultimately impounded in the price after transient liquidity effects have died out.

Finally, in Chapter (4) we conclude the thesis by summarizing on the main results found in the previous chapters.

There are five Appendices following the final Chapter. In Appendix (A), we provide a general description of the Trade and Quote data used in the analysis. Appendix (B) contains a detailed descriptive summary of the price durations examined in Chapter (2). In Appendix (C), we provide the code for merging the trade and quote transaction data into a single dataset. Appendix (D) contains the estimation code for the vector autoregressive model used in Chapter (3) and Appendix (E) the analytical derivation of the general calculation of impulse response function.

Chapter 1

Czech Stock Market and HF Data

Most of the empirical research[1] in high-frequency finance is based on datasets from the NYSE, a specialist market, and reflects its particular institutional features. A relatively small number of studies has directed its attention to the European stock exchanges, let alone to the stock exchanges in Central and Eastern Europe.

In this chapter, we will first examine the major characteristics of the Prague Stock Exchange. Representative of the most important stock markets in the region, the knowledge of particular features of the PSE such as its organization and trading mechanism is necessary for a proper understanding of the analysis presented later in the thesis. Given this understanding, we next describe some of the basic properties of high-frequency transaction data using a sample of data from the PSE's main market.

Namely, we will use data for two of the most liquid securities traded on the PSE in 2004 in order to make ourselves familiar with what we could call the inherent properties of tick-by-tick data. These characterisitcs include, among others, heterogeneity, discreteness, periodic/diurnal patterns, and/or the existence of price reversals. We will examine whether any or each of these characteristics is in fact as pronounced in case of the most-active securities listed on the PSE as it is in high-frequency data in general.

1.1 Prague Stock Exchange

Founded in 1992, the Prague Stock Exchange (PSE) is now the leading securities market organizer[2] in Central and Eastern Europe (CEE), covering more than 99% of the total

[1] Studies by Coughenour and Shastri (1999) and Madhavan (2000) both belong to often-cited summaries of empirical papers in the area.

[2] According to European Federation of Stock Exchanges, during the first six months of 2005 the PSE was the most active exchange in Central and Eastern Europe. With nearly EUR 17.5bil of total equity volume traded, the PSE overpassed Vienna (EUR 16,9bil), as well as Warsaw and Budapest.

trade value in the Czech Republic and at times up to 50% of the total trade value in the CEE countries. Securities registered on the PSE are traded on three markets: **main**, **secondary**, and **free markets**, where the main market is the most prestigious market on the Exchange.[3]

Only members of the PSE are allowed to trade directly on the stock exchange, either on their own account or on the account of their clients. Other persons can only trade indirectly through a member of the stock exchange.

Table 1: Characteristics of PSE

	2000	2001	2002	2003	2004
Total Trade Value (bil CZK)	264,145	128,799	197,398	257,442	479,662
Main + Secondary Markets	259,564	124,053	181,281	238,114	450,332
Daily Average (bil CZK)	1,060.8	515.2	789.6	1,025.7	1,903.4
Main + Secondary Markets	1,042.4	496.2	725.1	948.7	1787.0
Trade Volume (mil. pieces)	822,911	546,544	804,105	830,771	1,179,107

Total trade value, average daily trade value, and total number of shares traded on Prague Stock Exchange during the years 2000 to 2004. The years correspond to the period assessed later in the study. Source: PSE.

1.2 Market's Microstructure

The Prague Stock Exchange is a fully electronic exchange with the trading based on automated processing of its member's orders and instructions for the purchase and sale of securities. Only members are allowed to trade on PSE. Trading on the PSE is segmented into **two distinct sub-systems with distinct prices**[4]:

(a) quote driven system (referred to simply as SPAD), and

(b) order driven system (described by automatic trades)

In general, PSE members send electronic buy or sell instructions to the PSE and if conditions for matching opposite instructions within the above subsystems are met, a trade gets immediately recorded. In addition, PSE members must also report their *over-the-counter* (OTC) trades concluded without direct usage of either of the above price-determining mechanisms (usually over the phone). OTC trades must be registered with the PSE in order to preserve transparency of prices on the market.[5] Dissemination

[3] The main market has the most stringent conditions regarding the admission of securities to trading.

[4] This does not mean that the *same* instruments cannot be traded in either of the two sub-systems.

[5] All trades (both those concluded on the PSE within the two subsystems and OTC trades registered with the PSE) enter a central PSE database of trades that were concluded on a given day and are

of *trading information* is provided in real time to all participants of the market; namely, the information is immediately sent to the members of the PSE, the Czech Securities Commission (CSC), and the data vendors mentioned. It is also available for free on the PSE's internet pages (www.pse.cz). *Settlement* of all trades in the PSE database of trades occurs through UNIVYC, a 100% subsidiary of the PSE used for clearing. Trading at the PSE generally conforms to the $T+3$ standard settlement cycle. Trades in the SPAD system and automatic trades (see below) are settled in $T+3$ by UNIVYC. Block trades can be settled in a period from $T+0$ to $T+15$, and guarantees by the PSE's Guarantee Fund do not apply.

Each of the above-mentioned trading subsystems operates on its own *timetable*. Trading through the PSE occurs only during open phase hours. However, even outside open trading session hours, PSE members are still required to report their OTC trades. The time period after 17:00h technically belongs to the following business day; therefore, all OTC trades concluded from 17:00h to 20:00h and registered within the PSE database of trades are designated as belonging to the following trading day rather than to the day just ended at 16:00h.

1.2.1 Stock and Bond Market Support System

The Stock and Bond Market Support System (SPAD) is a price-driven trading system based on the activity of market makers. As already mentioned, the whole system is screen-based. This way, all the market makers as well as other members of PSE can see all the quotes and trades. Today, the system accommodates eight of the most liquid Czech securities (called blue-chips) supported by ten market makers[6].

The SPAD system operates in two phases: an **open phase** and a **closed phase**. During an open phase (from 9:30h to 16:00h), all market makers are obliged to publish their quotations (buying and selling prices) for issues for which they act as market makers. As the actual trading occurs during this period only, we assume just this phase as directly relevant to our study. We will describe the closed phase later.

A member of PSE member who supports trading in assigned securities traded in SPAD and thus increases liquidity of the securities within SPAD is called a **market maker**. For each security in SPAD, there has to be a minimum of three market

immediately published on anonymous basis (that is, no name of the involved members is visible) to PSE members and to external data vendors such as Bloomberg or Reuters. Daily summaries of trades are published, too.

[6] Hanousek and Podpiera (2003) studied the functioning of SPAD since its launch in 1998. They provide evidence that the new system has succeeded in increasing the transparency of the market and that it has improved the price discovery function of the exchange by attracting a large portion of order-flow to the main market.

makers[7] quoting prices. Each market maker is required to quote a buy and a sell price at all times during the open phase for a standardized number of shares or its multiple (maximum quadruple) to be delivered on a $T + 3$ basis. The Trading Committee also sets a maximum bid-offer spread for quotes of the same market maker to bring prices of purchases and sales closer.

Table 2: SPAD Issues

Issue	ISIN	Std Qt	Max Spread	AL Qt
Cesky Telecom	CZ0009093209	5,000	6	95,000
CEZ	CZ0005112300	10,000	2	100,000
Erste Bank	AT0000652011	2,000	8	34,000
Komercni Banka	CZ0008019106	1,000	20	11,000
Orco	LU0122624777	500	10	31,000
Philip Morris CR	CS00008418869	100	200	2,200
Unipetrol	CZ0009091500	10,000	3	300,000
Zentiva	NL0000405173	3,000	8	51,000

Parameters of issues in SPAD as of April 5, 2005. Values for Std Qt (Standard Quantity) and AL Qt (Above-Limited Quantity) are in pieces. Source: PSE.

All quotes of all market makers take the form of a buy or sell instruction sent to the PSE and are immediately displayed via electronic means to all PSE members. Any member can immediately take the best quote displayed within SPAD, resulting in instructions being matched and a trade being recorded and published. Quotes worse than the current best quotes within SPAD are informative only. Once a given market maker's quote becomes the best bid or offer available within SPAD (so called **best quote**), such quote becomes binding for that market maker. This means that the market maker is obligated to conclude a trade at such bid or offer price quoted should any other PSE member choose to accept it. On the other hand, the market maker is free to adjust his quotes up or down based on its assessment of demand and supply in the stock at any time. Once the best quote is accepted and a trade is thus recorded, the respective market maker is allowed up to a 3-minute recovery period to recalculate its positions and reset quotes. The second best indicative quote of another market maker immediately becomes the best quote and is binding from that moment on, until that quote is accepted too, or until a better quote appears within SPAD.

During the open phase, PSE members are allowed to conclude OTC trades with respect to SPAD securities with each other (OTC SPAD trade) only within a narrow

[7] Non-market-making members are subject to limitations as to their daily maximum trading volume with the market makers, as a protection measure for market makers against excessive settlement risks.

price range often referred to as a **permitted range**. The permitted range is constantly changing as market makers publish their quotes and it is defined as an up-to-the-minute price range limited by the price 0.5% below the best SPAD bid at the bottom and by the price 0.5% above the best SPAD offer at the top. However, OTC SPAD trades with large blocks of shares with total value larger than CZK 40 million can be registered with any price even outside the current permitted range. In addition, OTC SPAD trades must be reported to the PSE within a 5-minute deadline after such trade is concluded in order to preserve market transparency. The reason for mandatory reporting of OTC SPAD trades is to direct PSE members to SPAD and curb OTC trading among them, as OTC is considered less transparent for the market.

The **opening price** of an instrument traded in SPAD is equal to the midpoint of the permitted range as of the start of the open phase; in other words, it is calculated based on the quotes of market makers at the start of the session. The closing price of an instrument traded in SPAD is equal to the midpoint of the permitted range at the close of the open phase (16:00h). If during the open phase the arithmetical midpoint of the up-to-the-minute permitted range deviates by more than 20% from the arithmetical midpoint of the permitted range as of the start of the open phase, all quotes by the market makers become informative only, including the best quotes. Naturally, trades already concluded are not affected by this rule. The reason for this rule is to protect market makers against excessive losses once dramatic price moves occur.

Closed phase stands for the period between two open phases (or, more exactly, from 17:00h to 16:00h and from 7:30h to 9:30h Central European Time). During the closed phase, market makers are not obliged to quote any prices. Technically, the closed phase allows the market makers to clear the trades that they did not manage to conduct during the preceding (open) phase. However, over-the-counter SPAD trades must still be reported (and get published), although the reporting limitations are softer than during an open phase.

1.2.2 Automatic Trades

Automatic trades occur in an order-driven system. Here, the trades are concluded on the basis of matching orders for the purchase and sale of securities in a PSE order book as entered via electronic means into relevant PSE subsystem by member firms. Matching of orders takes place under two regimes: the auction regime (with price priority queueing), followed immediately by the continuous regime (queue priority: price first, then order entry time).

Due to different price defining strategies, securities within the SPAD subsystem can have different prices from those within the automatic trades subsystem, i.e., there is no direct technical interaction between the two subsystems as to prices. Trading hours for

automatic trades are from 7:30h through 15:45h. As the transactions resulting from the automatic trades are not directly relevant to our analysis, we will not describe this part of the market any further.

1.2.3 Block Trades

All OTC trades of members must be registered within the PSE database of trades. Members are obliged to report not only OTC trades between themselves but also any trades that they make with non-PSE members. These OTC trades are generally referred to as *block trades* and must be registered within one hour after such trades are concluded (the PSE system is open for registration from 7:30h to 20:00h). No price limitations apply in this case (except for the OTC SPAD trade limitations described above). For all OTC trades, members are free to agree on a settlement different from the $T+3$ standard (including OTC SPAD trades) within the range of $T+0$ to $T+15$ ($T+1$ to $T+15$ in case of OTC SPAD trades).

OTC trades registered with the PSE are regarded as a tool to protect market transparency rather than as a special trading subsystem of its own. Thus, through the OTC trades, PSE members effectively engage in "off-exchange" transactions with securities listed on the PSE, meaning that these transactions are concluded outside the PSE price-setting systems. For transparency reasons, these transactions are reported to the PSE and are registered in the relevant PSE databases.

1.3 High Frequency Data

1.3.1 Basic Characteristics

As already noted, the high frequency data[8] possesses unique features absent in data measured at lower frequencies. Consequently, analysis of this kind of data poses interesting and unique challenges to econometric modelling and statistical analysis.

First, the number of observations in high-frequency data sets can be overwhelming. In a single day, the most liquid markets may generate the number of transactions equivalent to the number of daily data within thirty years. The corresponding figures are, of course, much smaller for the markets such as the PSE[9]; nevertheless, even

[8] For the sake of brevity, we will from now on refer to ultra-high frequency data as simply high-frequency data although this term has been widely used to refer also to data observed on daily basis.

[9] Considering the number of quotes for the most traded stock on the Prague Stock Exchange in 2004, *Cesky Telecom*, and dividing this number by the number of trading days within the given year, the resulting average daily number of observations of an actively traded PSE stock can go to hundreds. Still, the number of transactions on an actively traded stock on the New York Stock Exchange (NYSE) can reach tens of thousands.

then some important implications for the statistical modeling arise that do not appear when the models are based on aggregated observations. For example, a relatively higher number of independently measured observations implies more degrees of freedom and hence more precise estimators. As DACOROGNA ET AL. (2001, p. 6) point out, "[] large amount of data allows us to distinguish between different models (model validation) with a higher statistical precision".

Second, virtually all high-frequency transactions are inherently irregularly spaced in time. Since most econometric models are specified for homogenous (i.e., equally spaced) time series, this poses an apparent complication as to what time intervals to use to analyze the data. The problem of irregular temporal spacing is also connected with that of non-synchronous trading. (TSAY, 2002, p. 176) As different stocks have different trading frequencies, using interpolation during the fixed interval analysis may introduce spurious correlations when dealing with multiple series each with its own transaction rate (ENGLE AND RUSSEL, 2004).

Third, the high-frequency data are discrete. There is, however, one important difference between the discreteness exhibited by normal (low-frequency) data and that of the data measured at the highest of frequencies. When viewed over long time horizons the variance of the process is usually quite large relative to the magnitude of the minimum movement. For high-frequency data, however, this is not the case as most of the times the transaction price changes may take only a handful of valued called ticks.[10]

Fourth, high-frequency data typically exhibit periodic (intra-day and/or intra-week) patterns in market activity. For example, it is well known a fact that trading activities on many of the developed exchanges around the world tend to be more dense in the beginning and closing of the trading day than in the lunch hours. Volatility, as well as the frequency of trades, volume, and spreads (we define spreads shortly) all exhibit a similar U-shaped pattern over the course of the day.

Fifth, other particular features of high-frequency data, such as temporal dependence, may distort inferences based on standard statistical models. The dependence is largely the result of price discreteness and the fact that there is often a spread between the prices for which the seller is willing to buy (bid) and/or sell (ask) in the given trade. Inventory effects such as splitting of large orders to smaller sizes may also play their role in temporal dependence.

At last, we should mention the problem of erroneous datasets of high-frequency data. For various reasons, just like any other economic data even the high-frequency data tend to be recorded with errors (e.g., missing observations, data gaps, or disor-

[10] A *tick* corresponds to the smallest allowable price change by which the price is allowed to change during trading on an exchange. The minimum value of price change is set institutionally. The ticks will be described in more detail in Section (1.3.3).

dered sequences) that need to be cleaned and corrected prior to direct analysis. The dimensions of high-frequency datasets render the problem more demanding than if we used the data of lower frequences, however, as they require the knowledge of data manipulation techniques unique to dealing with large quantities of data. The same holds true of data handling during estimations. It is often the case that we need to program an applications that can handle the manipulation for the given dataset.

In the following sections, we will take a more detailed look at each of the characteristics of high-frequency data just mentioned. To facilitate the discussion and enhance our understanding of the subject, it is only natural that we use the data from the Prague Stock Exchange, as the empirical exploration of this data is the main subject of the thesis. We will postpone the description of the source database until later in the study as it is not relevant for the current analysis.

1.3.2 Heterogeneity in Temporal Spacing

Irregular temporal spacing is perhaps the most important feature of high-frequency data. To show this feature graphically, we first consider the transactions data for Cesky Telecom and CEZ, currently the most liquid companies listed on PSE. These data come from the Trade database of PSE and will be the subject of several examples throughout this and the following three sections.

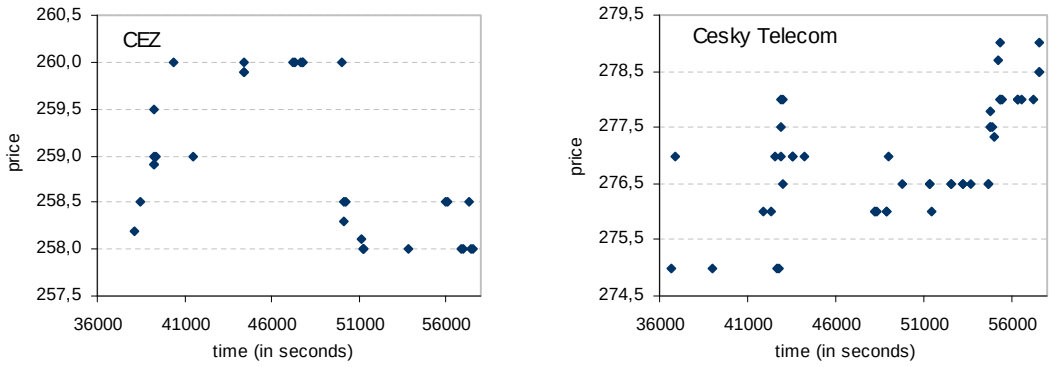

Figure 1: Plot of a sample of transaction prices for CEZ and Cesky Telecom stocks. The data are for the interval from 10:00h to16:00h, October 1, 2004. Source: TQ Database (PSE), own calculations.

Figure (1) plots six hours of transaction prices as recorded on October 1, 2004. The horizontal axis measures in seconds the time of the day (e.g. 10:00h is equivalent to 36,000 sec) and vertical axis is the price. Each miniature square denotes a single transaction. The irregular spacing of data is immediately evident as some transactions appear to occur only seconds apart while others may be tens of minutes apart. This is

most evident for Cesky Telecom, where for example between 12:30h and 13:30h (45,000 sec to 48,600 sec) only a minimal number of trades took place.

We now turn to other properties of high-frequency data. In order to make our demonstration more reliable, to say the least, we use a much larger sample than the one used to show the irregular temporal spacing. We first consider the transactions data for CEZ and Cesky Telecom from January 5, 2004 to November 12, 2004. There are 215 trading days within the period corresponding to 14,084 transactions for CEZ and 15,380 for Cesky Telecom. To make the analysis simple, we ignore any trades that took place between trading days[11] and focus only on the transactions that occurred during normal trading hours from 9:30 to 16:00 CET (open phase).

1.3.3 Discreteness in Price Changes

In general, every stock exchange has a rule that restricts the prices to fall on a pre-specified set of values. In other words, the price changes which generally occur on transaction by transaction basis must fall on multiples of the smallest allowable price change called a tick. The ticks may vary from asset to asset and may even change over time for the same asset. In case of the PSE, the minimum increment by which the stock price can move is derived from the price of the stock. That is, one tick equals CZK 0.01 for the stock whose price p_i falls in the interval $p_i \in (0\ to\ 1,000]$, CZK 0.10 for the stock with $p_i \in (1,000\ to\ 10,000]$, and so on.

Table (3) gives the frequencies in percentages of price change measured in the multiples of tick size of CZK 0.01. The tick size of CZK 0.01 is appropriate as for both CEZ and Cesky Telecom, $p_i \in (0\ to\ 1,000]$. In order to make the changes more visible, we use the multiples corresponding to up to 75 times the tick size for each of the two securities. This is a relatively high multiple, considering what Engle and Russell point out in one of their recent studies. In particular, according to Engle and Russell (2004, p. 3) in a market for an actively traded stock it is generally not common for the price "[] to move a large number of ticks from one transaction to another." We conclude the large multiples are a particular feature of even the most active stocks traded on the PSE.

From the table, we observe that: a) in case of CEZ , slightly more than 30% of the intra-day transactions were without price change, while the number was much larger for Cesky Telecom (Telecom), b) the price underwent a change larger than zero but smaller than 25 times the tick size in approximately 34% of intra-day transactions in case of CEZ and only about 13% in case of Telecom, and c) only about 9,4% of price changes resulted in price changes of 75 ticks or more for CEZ and 17% for Telecom. Finally,

[11] In any case, it is well known a fact that the overnight stock returns differ substantially from intraday returns. Refer to Stoll and Whaley (1990) and the references therein.

the distribution of positive and negative price changes was approximately symmetric.

Table 3: Frequencies of Price Change

	CEZ		Telecom	
Interval	Number	in (%)	Number	in (%)
($-\infty$ to -0,75]	610	4,38	1,312	8,63
(-0,75 to -0,50]	803	5,77	1,043	6,86
(-0,50 to -0,25]	985	7,08	551	3,62
(-0,25 to 0)	2,218	15,94	979	6,44
no change	4,262	30,62	7,342	48,27
(0 to 0,25)	2,488	17,88	985	6,48
[0,25 to 0,50)	1,043	7,49	513	3,37
[0,50 to 0,75)	805	5,78	1173	7,71
[0,75 to $+\infty$)	703	5,05	1311	8,62

Frequencies of price change in multiples of tick size for CEZ and Cesky Telecom (Telecom) stock from January 5 to November 12, 2004. The price changes between trading days are ignored. Source: TQ Database (PSE), own calculations.

Clearly, the frequency distribution shows a much larger propensity to no-change in case of Telecom, effectively implying fatter tails. The observation is even more apparent if we use a graphical representation for the distribution of price changes as in the histogram in Figure (2) below.

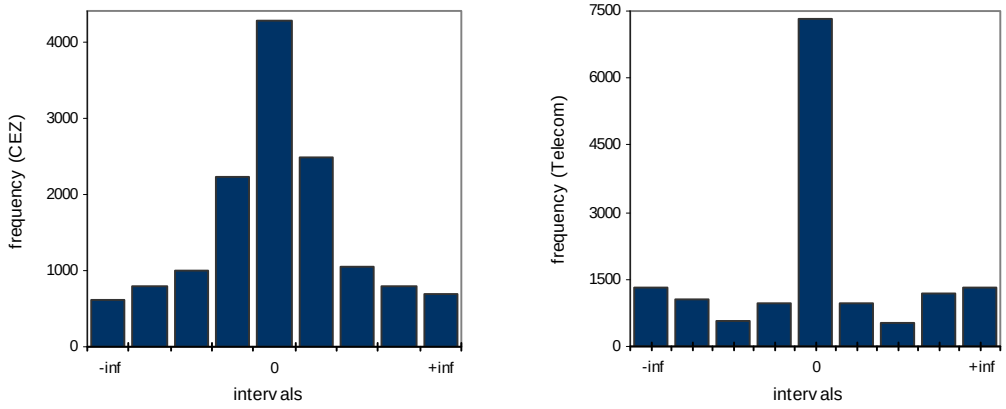

Figure 2: Histogram of price changes measured in multiples of tick size for CEZ and Cesky Telecom (Telecom) stock from January 5 to November 12, 2004. The price changes between trading days are ignored. Note that '0' on x-axis corresponds to no price change. Source: TQ Database (PSE), own calculations.

The discreteness just discussed has an impact on measuring volatility, dependence, or any other characteristic of prices that is small relative to the size. The same discreteness in transaction data also induces a high degree of kurtosis which is in fact typical

of high frequency data. Using the sample of data at hand, CEZ has the kurtosis of 62 and Cesky Telecom of 44.

1.3.4 Diurnal Patterns in Transaction Data

Another feature of high-frequency data that is well documented in the empirical literature is that they tend to exhibit a strong diurnal pattern.[12] In other words, volatility, volume, as well as bid-ask spreads[13] are generally higher near the open and the close of the market when the traders open and close their positions, respectively, and shorter around lunchtime.

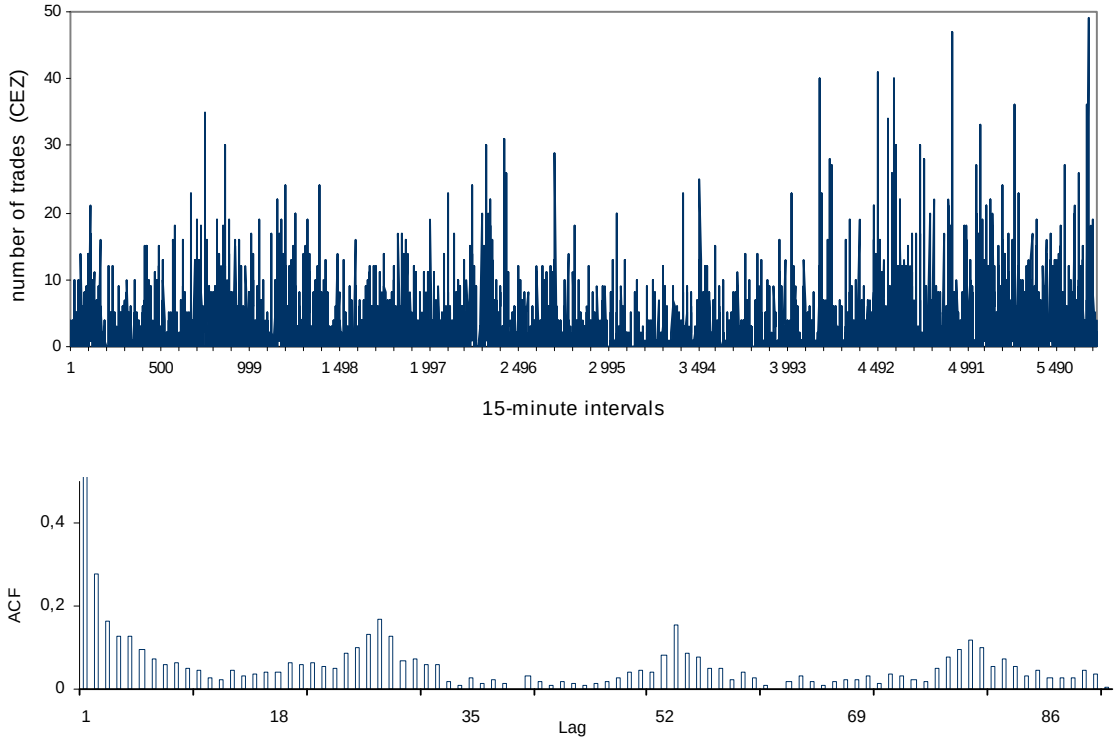

Figure 3: CEZ trade transaction data for the period from January 5, 2004 to November 12, 2004: (top) the number of trades in 15-minute time intervals, and (bottom) the sample autocorrelation function for the same series. The y-axis for the ACF function was made shorter in order to render the autocorrelation pattern for the series more apparent. Source: TQ Database (PSE), own calculations.

Consider, for example, the number of (trade) transactions in a 15-minute time interval for CEZ stock. Denote the series by x_t, where x_1 is the number CEZ transactions

[12] For an early reference see McInish and Wood (1992).

[13] The bid-ask spread is computed by subtracting bid quotes (i.e., prices at which the market maker or other traders in the market are willing to buy assets), from ask quotes (i.e., prices at which the market maker or other traders in the market are willing to sell assets).

from 9:30h to 9:45h on January 5, 2004, x_2 is the number of transactions from 9:45h to 10:00h, and so on. Again, we ignore the time gaps between trading days. Figure (3) shows the time plot of x_t (top), as well as the sample ACF of x_t for lags 1 to 90 (bottom). The upper figure clearly illustrates that the number of (trade) transactions exhibits a daily pattern. The cyclical pattern of the ACF only confirms the finding. The periodicity of 26 lags corresponds with the number of 15-minute intervals in a trading day.

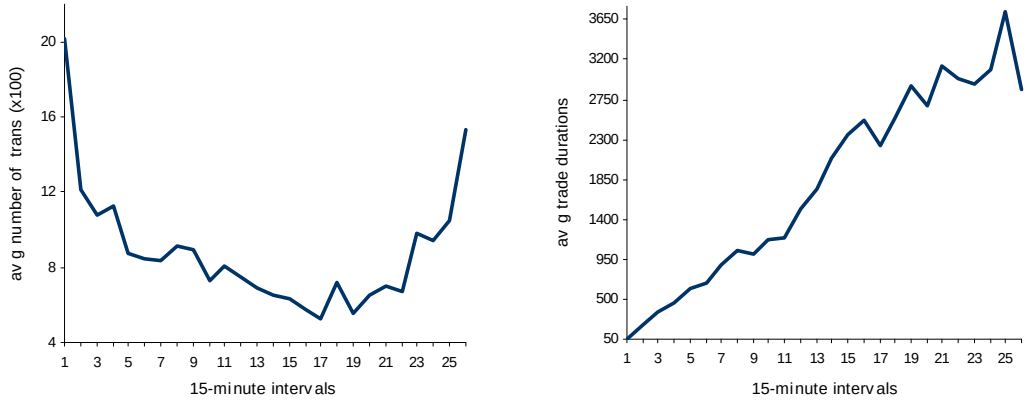

Figure 4: CEZ trade transaction data for the period from January 5, 2004 to November 11, 2004: (left) time plot of the average number of transactions in 15-minute time intervals, and (right) the average trade durations. There are 26 observations, averaging over the 220 trading days for the period in question. Source: TQ Database (PSE), own calculations.

To further illustrate the diurnal pattern present in high-frequency data, in Figure (4, left) we plot the average number of transactions within 15-minute time intervals for CEZ stock. Calculated over the 215 trading days, there are 26 such averages, the number again reflecting on the fact there as many 15-minute intervals in one trading day. For one more time, the plot clearly shows a "smiling" U-shape, indicating heavier trading in CEZ stock at the opening and closing of the market and thinner trading during the lunch hours.

Figure (4, right) also shows a different graphics: the average trade durations. Defined over the 15-minute intervals in the same way as the average number of transactions, the average trade durations tend to follow the same U-shape pattern as many other kinds of high-frequency data. Although it is not really clear from the graphics, the transaction durations are also the shortest near the market's open and just prior to its close.[14]

The time durations deserve one more note with respect to the existence of multiple transactions within a single second. Indeed, it is possible that multiple transactions,

[14]This finding was first documented by Engle and Russel (1998).

even with different prices, occur at the same time. This is partly due to the fact that time is measured in seconds that may be just too long a time scale in periods of heavier trading. Of the total of 14,084 observations on CEZ stock, there was a total of 1,538 zero-time intervals. In other words, during the normal trading hours of the 250 trading days under analysis, multiple transactions in a second occurred 1,538 times, which is about 10.92% of all trades. Owing to this in particular, the especially large number of zero-time durations become an important issue in statistical modeling of the time durations between trades.[15] The modeling of time durations represents the main subject of the following chapter.

1.3.5 Price Reversals and Other Patterns

In Section (1.2.1), we described the basic mechanism by which the market makers on the PSE post different prices for purchases and sales of security. The market makers buy at the bid price P_b and sell at a higher ask price P_a.[16] The difference between the two prices, $P_b - P_a$, called the bid-ask spread,[17] is the primary source of compensation for market makers who stand read to buy or sell the security whenever the public whishes to sell or buy. Obviously, since the prices of securities move in multiples of ticks, the bid-ask spread must also move in multiples of ticks. Typically the spread is small - namely one or two ticks. Still, repeating what we said in Section (1.3.3), due to a low tick base (e.g., CZK 0.01 per share if the price of the share $p_i \in (0\ to\ 1,000]$) the spread is often quite large for most of the stocks traded on the PSE.

The existence of bid-ask spread has several important consequences in time series properties of asset returns. One important implication is that the bid-ask spread generally introduces a negative lag-1 serial correlation in the series of observed price changes. This is referred[18] to as the *bid-ask bounce* in the finance literature.

There are several ways to examine whether the bid-ask bounce exists in the sample of CEZ and Cesky Telecom data analyzed so far. One possibility is to classify the intraday (trade) transactions in terms of price movements between two consecutive trades (i.e., from the $[t-1]$th transaction to the ith transaction) as in Tables (4a) and (4b). The price movements are classified into *up*, no *change*, and *down*, denoted $(+)$, (0),and $(-)$, respectively. Given it is a common practice to exclude the first

[15] As Tsay (2002) notes, the discretness and concentration on no change "[] make it difficult to model intraday price changes". Campbell, Lo, and MacKinlay (1997) discuss several econometric models that have been proposed in the literature to deal with these features.

[16] We may note that for the public, P_b is the sale price and P_a is the purchase price.

[17] Leaving aside the simple analysis in this section, the bid-ask spread is not the main concern of this study. Still, we may note that its determination is one of the most successful areas of the market microstructure research.

[18] See Roll (1984) for one of the first discussions of the subject.

transaction of each day, we do the same here. Consequently, this leaves us with 13,915 transactions for CEZ and 15,207 for Cesky Telecom.

The trade-by-trade data show that: a) consecutive price increases or decreases are not rare, accounting for $1,674 / 13,915 = 12\%$ in case of up-up movements and $1,408 / 13,915 = 10.1\%$ in case of down-down movements, b) there is an important tendency to move from up to down (and, similarly from down to up) than to no change, and c) there is a relatively high tendency to remain unchanged: compared to about 14.3% of consecutive price changes for CEZ, this tendency is even more pronounced in case of Cesky Telecom (38.3%) where it confirms the observations from Section (1.3.3).

Table 4a: Two-Way Price Movements (CEZ)

	ith trade			
$(i-1)$th trade	(+)	(0)	(−)	Margin
(+)	1,674	1,225	2,139	5,038
(0)	1,205	1,987	1,069	4,261
(−)	2,160	1,048	1,408	4,616
Margin	5,039	4,260	4,616	13,915

Table 4b: Two-Way Price Movements (Telecom)

	ith trade			
$(i-1)$th trade	(+)	(0)	(−)	Margin
(+)	966	1,277	1,739	3,982
(0)	1,306	4,838	1,196	7,340
(−)	1,710	1,226	949	3,885
Margin	3,982	7,341	3,884	15,207

Two-way classification of price movements in consecutive intraday trades for CEZ and Cesky Telecom (Telecom) stock from January 5 to November 12, 2004. The price movements are classified into *Up*, *Unchanged*, and *Down*. The price movements resulting from interday transactions are ignored. Source: TQ Database (PSE), own calculations.

The first observation on both CEZ and Cesky Telecom stocks does not confirm the existence of price reversals - and hence the bid-ask bounce - in the sample of intraday transaction data under analysis. To confirm that this is indeed the case, we consider a directional series D_i for price movements, where D_i assumes the value $+1$, 0, and -1 for *up*, *unchanged*, and *down* price movements, respectively, for the ith transaction. Constructing the ACF of $\{D_i\}$, we observe several significant negative spikes distributed over many of the lags. As a result, we see that the existence of price reversals in intraday transaction data for CEZ and Cesky Telecom stocks is at best ambiguous.

Similarly to lower frequency returns, high frequency data tends to exhibit volatility clustering. Large price changes tend to follow large price changes and vice-versa. The

sample ACF for the absolute value of the transaction price change (lags 1 to 200) is shown in Figure (5) below.

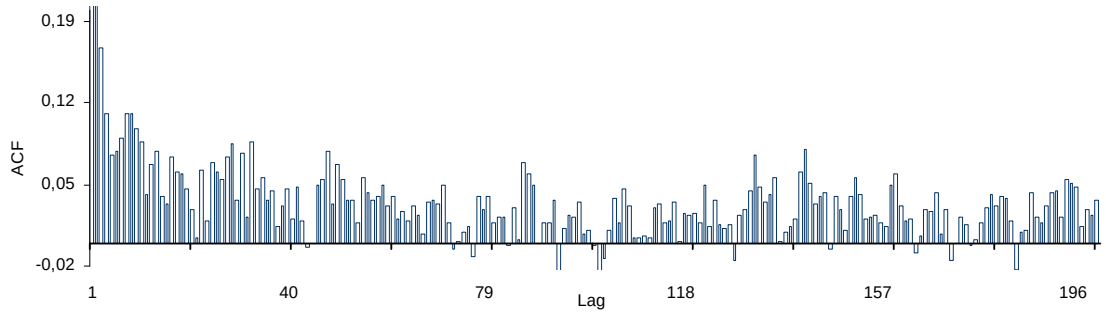

Figure 5: Autocorrelation function for the squared midquote price changes for CEZ (trade) transaction data for the period from January 5, 2004 to November 11, 2004. Source: TQ Database (PSE), own calculations.

Although the usual set of positive autocorrelations can be observed, the autocorrelation function is likely to be influenced by the diurnal pattern present in the data which we do not remove here. Instead, we would refer to sections (2.4.1) and (2.5.2) where the temporal dependence in both raw and diurnally adjusted transactions durations is examined. In fact, the transaction rates is yet another kind of transaction data that exhibits long sets of positive autocorrelation even after the deterministic component has been removed.

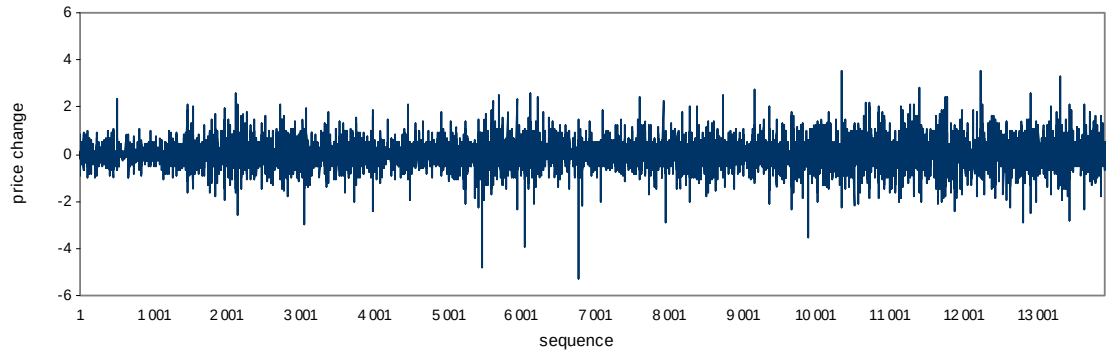

Figure 6: The plot of price changes for CEZ stock for the period from January 5, 2004 to November 12, 2004. The figure shows the time plot of price changes in consecutive trades measured in multiples of tick size of (CZK 0,01 x 100). Only regular (*open phase*) trading hours are considered. Source: TQ Database (PSE), own calculations.

One last note concerns unusual price movements. Figure (6) shows the time plot of price changes in consecutive intraday dates, measured in multiples of tick size of CZK 0.01 times 100. Again, the figures shows the data for CEZ for the period from January

5 to November 12, 2004. We can observe several unusual price movements for CEZ stock, particularly around the 5,500th, 6,000th, and 6,800th observation. They were a drop of approximately 450, 400, and 520 ticks, respectively. Unusual price movements like these occur infrequently in intraday transactions.

1.3.6 Low Frequency vs High Frequency

There exists at once important and interesting relationship between the data obtained at various frequencies and the models that are typically developed and tested with them. Graphically, this relationship can be depicted as in Figure (7). As already noted, the high-frequency data have opened great possibilities to test the market microstructure models,[19] while traditionally low-frequency data are used for testing macroeconomic models. In between the two extremes lies the whole area of financial and time series modeling, which is typically studied with daily or monthly data as, for example, option pricing or GARCH models.

The graphics below clearly shows that we have a continuum of both types of samples and models. In other words, there is no place for antagonism between the time series and microstructure approaches. In fact, as noted by (Dacorogna et al., 2001, p.5), the antagonism should "[] slowly vanish with more and more studies combining both with high-frequency data."

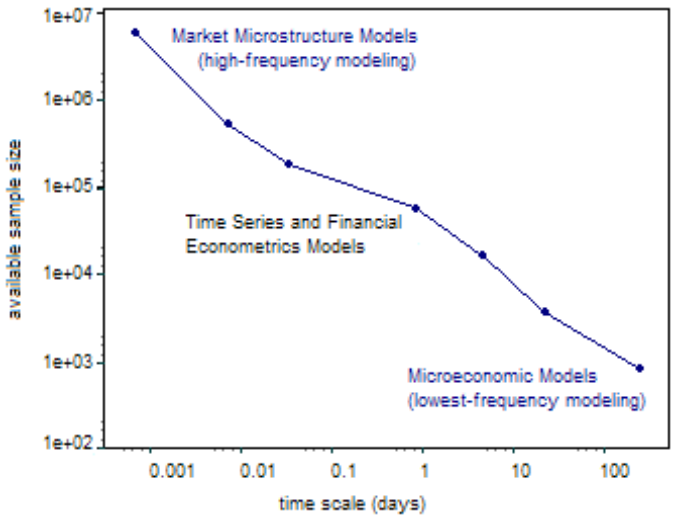

Figure 7: A graphical representation of sample size vs time-scale relationship. Reproduced from Dacorogna et al. (2001).

[19] Excellent surveys on the use of high-frequency financial data sets in financial econometrics are provided by Andersen (2000), Campbell, Lo and MacKinlay (1997), Dacorogna et al. (2001), Ghysels (2000), Goodhart and OHara (1997), Gourieroux and Jasiak (2001), and Tsay (2001).

Chapter 2

Intensity of Trading and Market Information

As emphasized in the studies by Diamond and Verrecchia (1987), Easley and OHara (1992), and Easley et al. (1996), among others, the waiting times (or *durations*) between events such as trades, quote updates, and price changes play a key role in understanding the processing of private and public information in financial markets. These models further extend the previous inventory-based models of market microstructure due to Garman (1976) and Stoll (1978), as well as other information-based models such as that of Glosten and Milgrom (1985).

In Diamond and Verrecchia (1987), the notion of time as a signal is considered from the short-sale constraints point of view. They show that if short sale restrictions do indeed affect information-motivated transactor's behavior, two results are possible: First, observing an absence of trade serves as a signal of bad news. Relatively long durations are then likely to be associated with bad news, inducing a negative re-evaluation of the asset value. Consequently, long durations are more likely to appear when informed traders would sell the asset but short-sell constraints prevent them from doing so. Second, since prices adjust more slowly to information in their model, it takes them longer to reflect full information values.[1]

Easley and OHara (1992) focused on the role of time in price adjustment process. Unlike their predecessors who assumed the time between trades to be exogenous to the price formation process, Easley and OHara explicitly argue that in the markets characterized by the presence of traders with different levels of information, the duration between two trades conveys information and hence plays a plausible role

[1] As mentioned by OHara (1995), this hypothesis has been investigated before in several empirical studies by examining how markets with and without options adjust to information (e.g., Jennings and Starks (1985). The notion here is that options replace the inability to short sell and thereby should increase efficiency.

in leading markets to price discovery. Specifically, their model implies that long durations are likely to be associated with no news, whereas short durations, and hence high trading activity, will most probably reveal the presence of asymmetric information in the market. In case of long durations the probability of dealing with an information-motivated transactor is then small and hence the market-maker is to decrease the bid-ask spread. On the contrary, the release of news should lead to an increase in the intensity of trading and hence more frequent revisions of the bid-ask prices posted by the market-makers.

The primary objective of this chapter is to provide an empirical evidence that the intensity of transaction arrivals as measured and forecasted by the time between quotes carries information about the state of the market. Merging the trade and quote datasets and thus linking the price duration process to the characteristics of the trade process obtained over the price durations (i.e., the intensity of trading, the average volume per trade, and the average spread), we also provide evidence of the relevance of information-based models for the local stock market, the Prague Stock Exchange (PSE).

In the analysis, we make use of the logarithmic autoregressive conditional duration (log-ACD) model as first introduced by Bauwens and Giot (2000) to model the durations between successive trades (quotes). The log-ACD model is more flexible than the original version of the ACD model developed by Engle and Russell (1998). With the autoregressive equation being specified on the logarithm of the conditional expectation of the durations, the log-ACD model allows the next conditional mean duration to be asymmetrically affected by durations respectively shorter and longer than the current conditional mean. In other words, the logarithmic version of the ACD model avoids the non-negativity constraints on the coefficients implied by the original specification and thus greatly facilitates the testing of market microstructure hypotheses put forward in our study.

The organization of this chapter is as follows. In Section (2.1) we describe two major classes of self-exciting processes (i.e. the processes where the past information impacts the probability structure of future events). The theory of self-exciting processes provides a basis for the autoregressive conditional duration (ACD) model whose general framework is described in the Section (2.2). In Section (2.3), we explain the basic methodology behind the price duration processes. Section (2.4) provides basic description of the dataset used in the analysis. The empirical results are presente in the Section (2.5) that follows. Finally, Section (2.6) summarizes the most important empirical findings.

2.1 Transaction Process

Broadly speaking, two main classes of high-frequency models exist. They include the extensions of ARCH type of models and ACD type of models. The question of which model to use depends on whether we want to use the transaction data as they normally appear (i.e., irregularly spaced in time), or as *artificially* transformed so that they effectively appear as if they were regularly spaced in time.

Regularly Spaced Data

As stressed in the last section, the transaction data inherently arrive in irregular time intervals, while standard econometrics techniques are base on fixed time interval analysis. A natural inclination on part of the econometrician then arises to transform the data in such a way as to make possible the use of standard time-series techniques. Again, there are two possibilities how we can do this:

First, we can treat the high-frequency time-series process as a collection of *numbered observations*, thus "forgetting" the information given by the time between the observations and overlooking the fact that they are inherently not regularly spaced in time. More precisely, we could define the intraday returns[2] directly in *transaction time* simply as a difference of the natural logarithms of two succeeding prices, or $r_i = \ln(p_i) - \ln(p_{i-1})$. However, as Giot (1999) points out, if these returns are directly used in a standard time-series model, "[it] can lead to meaningless estimation results as it treats all observations as being equidistantly spaced in time, which they are not." Furthermore, transaction time assumes that all observations convey the same information, whatever their spacing in time and associated characteristics such as volume. With these drawbacks, it is perhaps not that surprising that transaction time returns are not much used in the empirical literature.

Second, we can aggregate the transaction data to some fixed time interval using the *equidistant sampling*. By sampling the data at a given frequency we can once again use one of the now well documented time series techniques. In the same time, however, we effectively face the dilemma of what interval to use. Intuitively, for purchases of consumer durables by an individual, a natural interval might be months or even years. Frequently traded stock, on the other hand, have transactions every few seconds; hence a much shorter interval is appropriate. Still yet, as Engle and Russell (1998, p. 1128) note, if a short interval is chosen, "[] there will be many intervals with no new information and heteroskedasticity of a particular form will be introduced into the data." On the other hand, if a long interval is chosen, the micro structure features

[2] When working with bid and ask quotes, the returns are usually defined on the mid-point, defined as $p_i = (b_i + a_i)/2$.

of the data will be lost. In particular, multiple transactions will be averaged and the characteristics and timing relations of individual transactions will be lost, mitigating the advantages of moving to transaction data in the first place.
The problem becomes more complicated when one realizes that the rate of arrival of transaction type of data may vary over the course of the day, week, or year making the choice of an "optimal" interval more difficult.[3] We could see in Section (1.3.4) that for stocks, activity is higher near the open and the close then in the middle of the day. Even more intriguing is the case of transactions that are generally infrequent but that may suddenly exhibit very high activity. This may be due to some observable event such as a news release (e.g., financial results for the previous year) or to an unobservable event which may best be thought of as a stochastic process. In these cases the choice of a fixed interval for data analysis is very perilous as it may leave the investigator with many uninformative points, or disguise the periods of most interest. In other words, as data aggregation necessarily removes the time between events from the dataset, we have to face losing very importing information.
Regardless of what we have just said, the financial literature dealing with intraday characteristics of an asset (price process, liquidity, behavior of market agents trading the asset), has for a long time considered the time as being exogenous, with the implication that time between market events does not matter. See, for example, Kyle (1985) or Glosten and Milgrom (1985). For a broad range of empirical studies, such as Glosten and Harris (1988) or Hasbrouck (1991), the time between market events does not enter the analysis either.

Irregularly Spaced Data

Following our discussion of the adjusted high-frequency data where the time between events is modeled as fixed and assumed to be at most *insignificant*, the treatment of raw transaction data is much different. Clearly, as the time between events becomes at least as important as their price, using ARCH type of models does not make any more.

An alternative to fixed interval analysis of high-frequency data was offered by Engle and Russell (1998). The authors proposed a model that treats the time between events as a stochastic process with the arrival times of the events treated as random variables which follow a point process. In fact, Engle and Russell formulated a completely new model for dependent point processes, where the conditional intensity function is parametrized as a function of the time between past events. In addition,

[3] The presence of day of week seasonalities on the Prague Stock Exchange (as well as on Polish and Hungarian stock markets) have been documented before. See, for example, Bubak and Zikes (2004).

numerous natural extensions to the model include other effects such as characteristics associated with past transactions. The dependence of the conditional intensity on past durations then naturally suggested that the authors called their model the *autoregressive conditional duration* (ACD) model. In the next section we will provide a short overview of the point process framework. We will develop the particular parametrization of the ACD model in the section that follows. In both cases we will closely follow the framework offered by Engle and Russell.

2.2 Point Processes

Consider a stochastic process that is simply a sequence of times $\{t_0, t_1, ..., t_n, ...\}$ with $t_0 < t_1 < \cdots < t_n$ .... Associated with the arrival times is the counting function $N(t)$ which is the number of events that have occurred by time t. Clearly, it is a step function that is continuous from the left with limits from the right. If there are characteristics associated with the arrival times, such as a price or volume, the process is called a *marked point process*.

Two general characterizations of a point process can be introduced here following Snyder and Miller (1991). A point process on $[t, \infty)$ is said to evolve *without after effects* if, for any $t > t_0$, the realization of points during $[t, \infty)$ does not depend in any way on the sequence of points during the interval $[t_0, t)$. A counting process is said to be *conditionally orderly* at time $t \geq t_0$ if for a sufficiently short interval of time and conditional on any event P defined by the realization of the process on $[t_0, t)$, the probability of two or more events occurring is infinitessimal relative to the probability of one event.

We focus on point processes which evolve with after-effects and which, in the same time, are conditionally orderly. A complete description of such processes is naturally formulated in terms of the intensity function conditional on all available past information which must minimally include the arrival times and the count. This conditional intensity process is therefore defined[4] by

$$\lambda\left(t \mid N(t), t_1, ..., t_{N(t)}\right) = \lim_{\triangle t \to 0} \frac{P\left(N(t+\triangle t) > N(t) \mid N(t), t_1 ..., t_{N(t)}\right)}{\triangle t} \tag{2.1}$$

In many applications this will equivalently be called the hazard function, particularly in a context where there may be many individuals rather than a point process under study. As discussed in Lancaster (1990) and Snyder and Miller (1991), for example, the conditional intensity, the conditional intensity of the durations, and the

[4]Similarly to Engle and Russell (1998, p. 1130) we will further assume that when $N(t) = 0$, there are no further arguments to the function.

conditional survivor function each are complete descriptions of a conditionally orderly stochastic process. Letting p_i be a family of conditional probability density functions for arrival time t_i, the log likelihood can be expressed in terms of the conditional densities (or *intensities*) as

$$L = \sum_{i=1}^{N(T)} \log p_i \left(t_i \mid t_0, ..., t_{i-1} \right), \tag{2.2}$$

$$L = \sum_{i=1}^{N(T)} \log \lambda \left(t_i \mid i-1, t_0, ..., t_{i-1} \right) - \int_0^T \lambda \left(u \mid N\left(u\right), t_0, ..., t_{N(u)} \right) \, du. \tag{2.3}$$

Equation (2.2) is a general statement of the intensity function of a *self-exciting* point process, which is a process where the past information impacts the probability structure of future events. It was originally proposed by HAWKES (1971) and by RUBIN (1972). These are sometimes called Hawkes self-exciting processes. The success in using such processes depends upon the parametrization of the conditional intensity.

The simplest point process in this class is the Poisson process for which λ is a constant. A more flexible process is then the inhomogenous Poisson process for which the intensity varies only with t itself so that the arrival rate of an event is assumed to be a deterministic function of time. In neither case, however, do past events influence the future arrival rates; they evolve without after-effects.

When the intensity depends on the number of events but not the timing of these events, then the process is a *pure birth process*. RUBIN (1972) introduced limited memory self-exciting processes. A process is called an *m-memory self-exciting counting process* if only the m most recent arrival times are present in the conditional intensity. In this notation, a zero memory self-exciting process is a *Markov birth process*, and a homogeneous 1-memory process is a *renewal process*.

With longer memory there are many suggestions on how to parametrize the conditional intensity. We will briefly describe two existing classes of point process models that are characterized by equation (2.1) and leave the definition of the self-exciting process underlying the ACD model for the next section.

The first class of models is formulated in calendar time. Linear representations of this class of models can be expressed as

$$\lambda \left(t \mid N(t), t_1, ..., t_{N(t)} \right) = \omega + \sum_{i=1}^{N(T)} \pi \left(t - t_i \right), \tag{2.4}$$

where each past arrival time t_i contributes $\pi \left(t - t_i \right)$ to the intensity at time t. π is called an infectivity measure as motivated by epidemiology, as well as population dynamics

and earthquake prediction.[5] These types of specifications were initially proposed by HAWKES (1971a, b) and are used in OGATA AND AKAIKE (1982). Still, these calendar time models are inappropriate in the study of transaction types of data since they imply that the marginal effect of an event that occurred, say 20 minutes in the past is independent of the intervening history; there may have been 0 events or 100 events in the interval.

The second class of conditional intensity parametrizations focuses on the intervals between events and are formulated in event time. In these models, the conditional intensity could be parametrized in terms of

$$\lambda\left(t \mid N(t), t_1, ..., t_{N(t)}\right) = \omega + \sum_{i=1}^{N(T)} \pi\left(t - t_i\right), \tag{2.5}$$

so that the impact of a duration between successive events depends upon the number of intervening events. Such models were first studied by WOLD (1948) and later by COX (1955). Wold proposed a model for correlated intervals using an autoregressive structure similar to standard ARMA techniques. The model was subsequently reformulated[6] as an *exponential autoregressive moving average EARMA(p,q) model.* These models assume that the durations are conditionally exponentially distributed with a mean that follows an ARMA process. Nevertheless, as ENGLE AND RUSSELL (1998, p. 1131) point out, the formulation of an additive error process that has this property "[] is complex and the resulting maximum likelihood procedures are virtually unworkable, at least in general settings."

In some cases, the conditional intensity can be derived from the more fundamental assumptions. For example, the *Cox model* (1955), also called *doubly stochastic,* typically assumes that there is a latent independent process which governs the arrival rate. Suppose this is a counting process $M(t)$ which might be called "information" for financial applications. Thus the intensity is conditional on $M(t)$ as well as t itself. SNYDER AND MILLER (1991, p. 134, theorem 7.2.2) prove that such a process is itself a self-exciting process with a form given by (2.1), which is simply the expectation of the intensity over M conditional on the past history of N. Hence, the class of self-exciting processes also includes the *Cox process* although it is not generally easy to determine the form of the intensity process from the conditional expectation.

COX (1955, 1972a) formulated a model for duration analysis with covariates and later generalized the model to the popular proportional hazards framework. The con-

[5] See OGATA AND KATSURA (1986) for an extensive review of the subject.

[6] Refer to GAVER AND LEWIS (1980), LAWRENCE AND LEWIS (1980) for further details concerning the model.

ditional intensity can be written as

$$\lambda\left(t \mid z_{N(t)}, ..., z_1\right) = \lambda(t) \exp\left\{\beta' z_{N(t)}\right\} \tag{2.6}$$

where z_i is a vector of explanatory variables associated with the arrival time i. One suggestion mentioned in Cox (1972b) was to include lagged durations as an explanatory variable. Duration models were introduced into econometrics by Lancaster (1981), and given a dynamic focus in Heckman and Borjas (1980) among others, to examine the impact of past unemployment on current spells. In these models, the data are typically short time series on many individuals, so that the question of whether this truly reflects state dependence or merely unmeasured heterogeneity becomes very important.

2.3 Autoregressive Conditional Duration

Having just described the two major classes of self-exciting processes (i.e. the processes where the past information impacts the probability structure of future events), we may now introduce the general framework for the ACD model, itself based on a new family of self-exciting processes formulated by Engle and Russell (1998). We will see that these processes have conditional densities different from both of the representations mentioned in the last section. For this reason, we will also follow the same notation as previously.

The ACD model is most conveniently specified in terms of the conditional density of the durations. Letting x_i be the duration between two market events that happened at times t_{i-1} and t_i, i.e. $x_i = t_{i-1} - t_i$, the density of x_i conditional on past x's is specified directly as:

$$E(x_i \mid I_{i-1}) = \psi_i(I_{i-1}; \theta) \equiv \psi_i, \tag{2.7}$$

where ψ_i is the expectation of the ith duration. I_{i-1} denotes the information set available at time t_{i-1}, supposed to contain at least $\tilde{x}_{i-1}$ and $\tilde{\psi}_{i-1}$, where $\tilde{x}_{i-1}$ denotes x_{i-1} and its past values, and likewise for $\tilde{\psi}_{i-1}$. The ACD class of models then consists of parametrizations of (2.7) and the assumption that

$$x_i = \psi_i \varepsilon_i, \tag{2.8}$$

where $\{\varepsilon_i\} \sim i.i.d.$with density $p(\varepsilon; \zeta)$ with non-negative support, and θ and γ are variation free.[7]

From expression (2.7) and (2.8) it is apparent that we now have a host of poten-

[7] In other words, if $\theta \in \Theta$ and $\gamma \in \Gamma$, then $(\theta, \gamma) \in \Theta \otimes \Gamma$.

tial specifications for the ACD model, each defined by different specifications for the expected durations and for the distribution of ε. To derive a general expression for the conditional intensity, let p be the density function of ε and let S be the associated survival function.[8] Then define

$$\lambda_0 = \frac{p(\varepsilon;\zeta)}{S(\varepsilon;\zeta)} \tag{2.9}$$

as the baseline intensity, or baseline hazard, using the name popularized by the proportional hazard literature. The conditional intensity function of an ACD model is then given by

$$\lambda(t \mid N(t), t_{i-1}, t_{i-2}, ..., t_0) = \lambda_0\left(\frac{t - t_{N(t)-1}}{\psi_{N(t)}}\right)\frac{1}{\psi_{N(t)}} \tag{2.10}$$

Since the past information influences the rate at which time passes (i.e. ψ_i enters the baseline hazard), this type of model is referred to as an *accelerated failure* time model in the duration literature.[9] Indeed, the past history influences the conditional intensity by both a multiplicative effect and a shift in the baseline hazard.

In addition, as the rate at which time progresses through the hazard function is dependent upon ψ_i, it can be also viewed in the context of time deformation models. As Engle and Russell (2004) mention, during some periods the pace of the market is more rapid rather than other periods.

Associated with the intensity in (2.10) is also the conditional expectation of the waiting time until the next event. The ACD model therefore has an interesting interpretation in the context of time deformation models because the model is formulated in transaction time but models the frequency and distribution of calendar time between events.[10] The ACD formulation can use but does not require auxiliary data or assumptions on the causes of time flow; it is simply a time-series model of time.

The simplest version of the ACD model assumes that the durations are conditionally exponential so that the baseline hazard is simply one and the conditional intensity is

$$\lambda\left(t \mid x_{N(t)}, ..., x_1\right) = \psi^{-1}_{N(t)+1}. \tag{2.11}$$

An m-memory conditional intensity would imply that only the most recent m durations

[8] The survival function is defined as $S_0(\varepsilon,\zeta) = \int_{\varepsilon}^{\infty} p(u;\zeta)\,du$.

[9] Such models are natural in some medical examples where patients with particular characteristics move through a disease more rapidly than others.

[10] See, for example, Tauchen and Pitts (1983), or more recently Muller et al. (1990), and Ghysels and Jasiak (1994).

influenced the conditional duration, suggesting a possible specification:

$$\psi_i = \omega + \sum_{j=1}^{m} \alpha_j \, x_{i-j}. \tag{2.12}$$

A more general model without the limited memory characteristic is

$$\psi_i = \omega + \sum_{j=1}^{p} \alpha_j \, x_{i-j} + \sum_{j=1}^{q} \beta_j \, \psi_{i-j}. \tag{2.13}$$

This in fact is an ACD(p, q) model where the p and q refers to the orders of the lags. Since durations are by definition non-negative we require that $\omega > 0$, $\alpha_j \geq 0$, and $\beta_j \geq 0$ for all j.

This model is convenient because it allows for various moments to be calculated by expectation regardless of the form of the baseline hazard. For example, the conditional mean of x_i is ψ_i, the conditional duration, but the unconditional mean is

$$E\left(x_i\right) = \mu = \frac{\omega + \boldsymbol{\gamma}^T E\left[\mathbf{z}_i\right]}{\left[1 - \sum(a_j + \beta_j)\right]}. \tag{2.14}$$

This is most easily seen by taking expectations of both sides of equation (2.14), although Engle and Russell (1998) establish that this exists only when all the roots of the associated difference equations lie outside the unit circle.

The simplest and often successful member of this family is the EACD$(1, 1)$ where E signifies the exponential distribution for the error terms:

$$\psi_i = \omega + \alpha x_{i-1} + \beta \psi_{i-1} \tag{2.15}$$

with the following constraints on the coefficients: α, $\beta \geq 0$, $\omega > 0, \forall i$ and $(i = 1...N)$. In this model, the conditional variance of x is ψ_i^2 but the unconditional variance is given by

$$\sigma^2 = \mu^2 \left(\frac{1 - \beta^2 - 2\alpha\beta}{1 - \beta^2 - 2\alpha\beta - 2\alpha^2} \right). \tag{2.16}$$

Thus whenever $\alpha > 0$ the unconditional standard deviation will exceed the mean, exhibiting *excess dispersion* (or overdispersion) as is so often noticed in duration data sets.

The general specification of the ACD model as introduced in (2.13) is limited by the positivity constraints on the coefficients of the model. Such constraints may be quite restrictive. In particular, if we want to include additional explanatory variables in the autoregressive equation (2.13) , we must ensure that the righ-hand side of (2.13)

remains strictly positive. As it will become apparent in Section (2.5.3), this can be a problem when additional variables suggested by market microstructure theories are included in the equation. With these concerns in mind, BAUWENS and GIOT (1997) introduce a logarithmic ACD model where

$$\log(\psi_i) = \omega + \sum_{j=1}^{m} \alpha_j \log(x_{i-j}) + \sum_{j=1}^{q} \beta_j \log(\psi_{i-j}) + \boldsymbol{\gamma}^T \mathbf{z}_i. \tag{2.17}$$

In this equation for ψ_i the non-negativity of duration(s) is ensured regardless of parameter values. The equation also shows the vector of exogenous variables $\mathbf{z}_i$.

In the empirical part of the paper, we focus on the log-ACD model as this specification is more suitable for testing market microstructure hypotheses. To see why, consider a hypothesis that says that a bid-ask spread is negatively related to expected duration due to the presence of informed traders. Since the bid-ask spread is non-negative, we expect the partial effect γ to be negative. Clearly, had we used the equation (2.13), the large bid-ask spreads might, ceteris paribus, predict negative durations which would be inconsistent with the theoretical predictions.

2.3.1 Relationship to ARMA (p,q) Models

The ACD(p, q) specification in (2.13) appears very similar to an ARCH(p, q) models of ENGLE (1982) and BOLLERSLEV (1986) and indeed the two models share many of the same properties. From (2.8) and (2.13) it follows that the durations x_i follow an ARMA $(\max(p, q), q)$. Letting $\eta_i \equiv x_i - \psi_i$ which is a Martingale difference sequence by construction, the duration process can be expressed as

$$x_i = \omega + \sum_{j=1}^{\max(p,\, q)} \left(\alpha_j + \beta_j\right) x_{i-j} - \sum_{j=1}^{q} \beta_j \eta_{i-j} + \eta_i, \tag{2.18}$$

which is an ARMA(m,q) process with highly non-Gaussian innovations. Forecasts of waiting times can be computed directly from this representation using the conventional ARMA analytics. If $\alpha(L)$ and $\beta(L)$ denote polynomials in the lag operator of orders p and q respectively then the persistence of the model can be measured by $\alpha(1) + \beta(1)$. For most duration data this sum is very close to (but less than) one[11] indicating strong persistence but stationarity (ENGLE AND RUSSELL, 2004). It also becomes clear from this representation that restrictions must be placed on parameter

[11] If all roots of the associated polynomial are less then unity, then the duration process will be mean reverting and the impact of a given duration on future expected durations will die out exponentially. Since the transactions to be analyzed occur within seconds of each other, the persistence of shocks will be very limited in calendar time unless roots are very close to unity.

values to ensure non-negative durations. These restrictions impose that the infinite AR representation implied by inverting the MA component must contain non-negative coefficients for all lags. These conditions are identical to the conditions derived in NELSON AND CAO (1992) to ensure non-negativity of GARCH models. For example, for the ACD(1,1) model this reduces to $\omega \geq 0$ and α, $\beta \geq 0$.

2.3.2 Extensions of the Model

The specifications in (2.11) and (2.13) can be generalized in many ways. The baseline hazard can be given many parametric shapes. The most popular is to assume that the conditional distribution is Weibull which is equivalent to assuming that x^{ϵ} is exponential. Other popular alternatives are the generalized gamma, log logistic, and log normal, as discussed in LANCASTER (1990).

There is a number of possible specifications for the density of ε, the most common are exponential, Weibull, generalized gamma and Burr[12]. The choice of $p_0(t)$, of course, affects the shape of the conditional intensity. For exponential distribution, the hazard function is constant, and thus for given expected duration the probability that an event occurs in an interval $t + \Delta t$ given that it has not occurred by t is constant in t. The Weibull distribution exhibits monotonic hazard function and hence $\lambda(t \mid N(t), t_1, ..., t_{N(t)})$ is either increasing or decreasing in t depending upon the value of the parameter of the Weibull distribution. The most flexible choice is the generalized gamma of Burr distribution which both have hazard functions that can be either constant, monotonic or U-shaped, depending upon the parameters[13]. Using these distributions we can therefore model situations where the conditional probability of an event occurring first increases in t, but eventually reaches a maximum and starts to decline. In this paper, we focus on the exponential model.

A further step is to estimate the hazard semiparametrically using the methods of piecewise constant, spline, kernel, or other smoothers. In the ARCH context ENGLE AND GONZALEZ-RIVERA (1991) estimated the density using a spline, while in the duration context ENGLE (1996) used a k-nearest neighbor estimator which we also apply in this thesis.

[12] See e.g. ENGEL and RUSSEL (1998) for the exponential and Weibull models, LUNDE (1999) for the generalized gamma model and GRAMMIG and MAURER (2000) for the Burr model.

[13] Both generalized gamma and Burr distribution nest the exponential and Weibull distributions as special cases.

2.4 Thinning the Point Process

The most fundamental application of the ACD model suggested by its authors is to measure and forecast the intensity of transaction arrivals which is essentially the instantaneous quantity of transactions. In this case, the variable of concern is the transaction duration defined as the time between the i^{th} and $(i-1)^{th}$ transactions. Figure (8) shows the time plot of time intervals between trades for the sample of CEZ transaction data between January 5 and November 12, 2004.

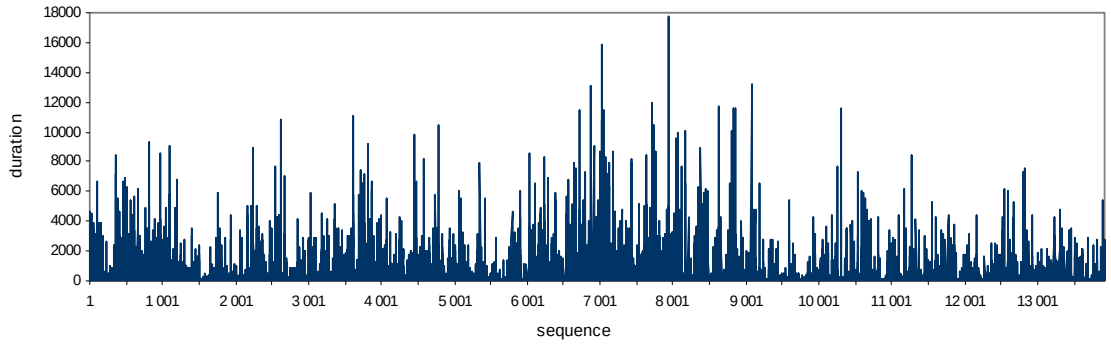

Figure 8: Time plot of (trade) transaction durations for CEZ stock for the period from January 5, 2004 to November 12, 2004. Only regular (*open phase*) trading hours are considered. Source: TQ Database (PSE), own calculations.

Sometimes, however, we may not be interested in modeling the time between transactions but rather in studying the marks associated with the arrival times such as volume, bid-ask spread, or price. Therefore, Engle and Russell (1998) proposed a method for modeling the rate of change of other variables by selectively thinning the point process. The resulting ACD model can accommodate exogenous explanatory variables, effectively providing a framework for testing various market microstructure hypotheses.

2.4.1 Adjusting the Raw Durations

As shown in many empirical studies[14], transactions durations exhibit significant diurnal patterns. The deterministic component of expected durations can be either estimated jointly with the ACD model or it can be first removed from raw durations and then the ACD model can be fitted to the seasonally adjusted durations. In this paper, we take the second approach. We write the raw duration as

$$x_i = \phi(t_i)\psi_i\varepsilon_i,$$

[14]See, for example, Engle and Russel (1998), Bauwens and Giot (1997), and the references therein.

and assume that the seasonal component $\phi_i(t)$ can be approximated by cubic splines

$$\phi(t_i) = \sum_{j=1}^{K} I_j \left[c_j + d_{1j}(t_i - k_{j-1}) + d_{2j}(t_i - k_{j-1})^2 + d_{3j}(t_i - k_{j-1})^3 \right], \tag{2.19}$$

where I_j is an indicator for the jth segment of the spline, i.e. $I_j = 1$ if $t_i \in \langle k_{j-1}, k_j)$ and zero otherwise, and the parameters $c_j, d_{1j}, d_{2j}, d_{3j}, j = 1...K$ are restricted by the usual differentiability conditions. We set the nodes $(k_0, ..., k_K)$ at 9:30, 10:30, 11:30, 12:30, 13:30, 14:30, 15:30 and 16:00. The reason for having two nodes over the last trading hour is to allow for more flexibility during the period of overlapping trading with the U.S. market The cubic splines are estimated by ordinary least squares and the raw durations are standardized according to

$$\tilde{x}_i = \frac{x_i}{\hat{\phi}(t_i)}. \tag{2.20}$$

Using the Berndt-Hall-Hall-Hausman algorithm, we will estimate the ACD(m, q) model for diurnally adjusted durations $\{\tilde{x}_i\}$ by the method of maximum likelihood. As shown in Gourieroux, Monfort and Trognon (1984), if the conditional model is properly specified the ML parameter estimators are consistent if and only if the distribution used to obtain them belongs to the linear exponential family, regardless of the true density. Since the exponential distribution belongs to this family, we can interpret the ML estimates of the exponential model as quasi maximum likelihood estimates (QMLE). This is, however, not the case with generalized gamma distribution. We thus face the common trade-off between consistency and efficiency when estimating the generalized gamma model[15].

2.4.2 Price Durations and Volatility

Other then transactions durations, however, our interest lies rather in investigating the behavior of price durations which are defined as duration between events for which the price has changed. With each transactions time is associated a price and such process is called market point process with price being the mark. By selecting only those points for which price has changed we perform thinning of the point process for transaction arrival times and obtain a point process for price arrival times. To avoid small price changes caused by recording or quoting errors and the impact of large trades that temporarily move the price by a small amount, we consider as a price change a movement in the mid-price by at least c, a constant. Then the probability that the

[15] Gramming and Maurer (2000) show in a Monte Carlo analysis that a misspecification in the error distribution can severely deteriorate the accuracy of duration forecast.

price changes by at least c in an interval Δt is by equation (2.1) $\lambda(t|t_{i.1}, ..., t_0)\Delta t + o(\Delta t)$ and otherwise there is no change.

As ENGLE and RUSSELL (1998) demonstrate, by applying the ACD model to the price durations, we obtain a model for the inverse of volatility. To see this, define the instantaneous volatility as

$$\sigma^2(t) = \lim_{\Delta t \to 0} E \left\{ \frac{1}{\Delta t} \left[\frac{P(t + \Delta t) - P(t)}{P(t)} \right]^2 \right\}. \tag{2.21}$$

Substituting $c(\Delta t)$ for $P(t + \Delta t) - P(t)$ in (2.21) and taking limits we obtain a model for the expected instantaneous conditional volatility per second given by

$$\sigma^2(t|t_{N(t)}, ..., t_1) = \left(\frac{c}{P(t)} \right)^2 \lambda(t|t_{N(t)}, ..., t_1). \tag{2.22}$$

The relation between instantaneous volatility and price duration is simple: in a given time period, high price intensity $\lambda(t|t_{N(t)}, ..., t_1)$ implies high number of transactions with price change of at least c, and by equation (2.22) high volatility. High number of transactions is associated with short price durations and thus there is an inverse relationship between price durations and volatility. Since the ACD model can capture the clustering of durations, it can therefore also capture volatility clustering, which has been observed ever since the advent of the ARCH model.

2.5 Data Description

We base the analysis on the data from the Prague Stock Exchange (PSE). Similarly to other data of this kind, the data come in two datasets: one that carries the intra-day *trade data* and one that consists of the intra-day *quotes data*. The former set contains a detailed record of every single trade in a specific stock, including the time and the price at which the trade took place, the amount of the stock traded, as well as other information relevant to the trade such as type of trade. The latter set includes the time of the quotes posted, the corresponding bid and ask prices, and the depth - that is, how many shares the market maker is willing to buy/sell at the given bid/ask quote. Both datasets contain the information for all stocks traded on PSE's main market (SPAD) from January 2000 to November 2004.

Of the four years of data available, we select the most recent year for the analysis. The period assessed then begins with the first trading day of 2004 (January 5) and ends with the last day the data were available for that year (November 12). The 218 day trading sample is long enough to allow reasonably precise estimations (see EASLEY,

KIEFER, AND OHARA, 1993). We work with three securities: two non-financial (CEZ, Cesky Telecom) and one financial (Komercni Banka). These stocks were the most actively traded of the total of eight titles present on SPAD in 2004.

Two days were deleted from the 218 day trading sample. A halt occurred on Tuesday, May 16, when the trading in SPAD was interrupted for nearly three hours due to technical error. Another interruption took place in case of Komercni Banka on Friday, June 10, when the trading was postponed for four hours owing to several wrong order submissions. After deleting these two days from each of the two datasets, we adjust the data as follows. First, by focusing solely on open trading hours (from 9:30h to 16:00h), we remove the overnight transactions. This way, we effectively ignore the overnight duration and treat the data consecutively from day to day.[16] Furthermore, in case of quote data, we only consider unique quotation times and hence regard the simultaneously recorded quotes as a single quotation. Despite these adjustments the sample remains very large. For CEZ, we are left with 14,092 open-phase observations on trades and 39,609 unique open-phase observations on quotes. Komercni Banka (KB) and Cesky Telecom (Telecom) show much higher number of trades versus quotes, as it appears from Table (5).

Table 5: General Description

	CEZ	KB	Telecom
- average price	323.1	2,884.7	206.8
# of trade observations	14,297	20,133	15,615
- open phase	14,092	19,896	15,390
# of quote observations	42,004	49,534	26,215
- open phase	41,763	49,273	25,980
- w/out multiple records	39,609	46,424	24,846

The data summary data contains the information on the number of trade and quote observations for CEZ, Komercni Banka (KB), and Cesky Telecom (Telecom) from January 5 to November 12, 2004. 'Open phase' shows number of transactions during open phase (from 9:30 to 16:00). In case of quote observations, the line *w/out multiple records* relates to number of transactions ignoring overnight duration and without simultaneous trades.

2.5.1 Summary Statistics

In order to estimate the model for the price process and ultimately test the market microstructure hypotheses discussed in Introduction, we also compute the price dura-

[16] To be specific, if the last transaction on day j takes place at 15:55:40, and the first transaction of the next day at 9:30:35, the duration between these two events is not used, so that the first duration for day $j+1$ will effectively be the one between 9:30:35 and the next transaction.

tions for each of the three stocks under analysis.[17] As described earlier in the chapter, we define price durations by filtering the bid-ask quote durations and retaining only those leading to a significant cumulated change in the midpoint (or midprice)[18] of the bid-ask quotes. Specifically, in our study we calculate a significant cumulated change in the midprice as a change leading to at least a CZK 0.25 cumulated change in the midprice for CEZ, CZK 5.00 for KB, and CZK 0.50 for Cesky Telecom. Illustrating the calculation on example of CEZ, a cumulated change of CZK 0.25 may result from two successive positive changes of CZK 0.125, just as it can follow from four successive changes of CZK 0.125 with the first two going in opposite directions and the next two in the same direction.

Defining the price durations on the midpoint of the bid-ask quotes and thinning the process with respect to a given cumulative change allows to eliminate the problem of "bid-ask" bounce, an inherent feature of quote transactions data that is nevertheless annoying to work with as it gives relatively few information. Thinning the process by filtering the numerous small changes as in the example with CEZ (CZK 0.125 changes) can be justified on the assumption of the transitory nature of such changes. In fact, Biais, Hillion and Spatt (1995) show that the bid-ask quote process is often characterized by information events that lead to similar (successive) changes in the quotes and thus move significantly the mid-point. In addition to what has been just said, thinning the bid-ask quote process also allows to avoid small price changes caused by recording or quoting errors and the impact of large trades that temporarily move the price by a small amount, effectively extracting a process where only meaningful price changes are retained in the end.

In Table (6), we provide characteristics of price durations computed for CEZ, KB and Cesky Telecom using different thresholds. In case of CEZ, the sample size is reduced to 7,728 after thinning with threshold of CZK 0.25. This corresponds to about 19.5% of the size of the original sample. The mean duration is 500 seconds (or 5.3 minutes) with a standard deviation of 1,240.5. The minimum price duration is 1 second while the maximum is 18,778 (5h 13 min). A relatively high value for overdispersion (2.48) as well as strong autocorrelation associated with CEZ's price durations are both characteristic of duration data filtered at small thresholds. Increasing the relative threshold,[19] the values for overdispersion as well as the Ljung-Box statistics for the first ten autocorrelations on price durations become smaller.

[17] Prior to constructing the price duration process, we need to merge the trade and quote transaction data into a single dataset. We merge the files using a data merging program provided in the Appendix (C).

[18] The midprice is defined as $p_i = (bid_i + ask_i)\ /\ 2$.

[19] We have examined five different thresholds for CEZ and Cesky Telecom (with c = CZK 0.25, 0.50, 1.00, 1.50, and 2.00) as well as for Komercni Banka (with c = CZK 2.5, 5.0, 10.0, 15.0, and 20.0). A detailed summary is available in Appendix (B).

The summary statistics are very similar for other two titles, Komercni Banka (KB) and Cesky Telecom (Telecom). As in the case of CEZ, for each of these two stocks we choose the price thinning thresholds based on the properties of the underlying stock (i.e., price, median spread, mean, tick size) as well as with regard to consistency of results in the sample (i.e., similar percentage of thinning and overdispersion).

Table 6: Summary Statistics for Thinned Price Duration Data

	CEZ	KB	Telecom
# of Adjusted Quotes	39,609	46,424	24,846
# of Price Durations	7,728	4,901	7,001
Thinning (in %)	19.51	10.56	28.17
c (CZK)	0.25	5.00	0.50
Mean	500.2 (0.99)	671.9 (0.99)	520.6 (0.99)
Std. Deviation	1,240.5 (1.98)	1,429.2 (1.64)	1,252.8 (2.07)
Overdispersion	2.48 (2.01)	2.14 (1.66)	2.41 (2.09)
Minimum	1 (0.001)	1 (0.001)	1 (0.001)
Maximum	18,778 (32.22)	16,648 (21.67)	21,439 (37.99)
$Q(10)$	594.3 (1,047.6)	171.4 (299.1)	428.1 (586.9)

Summary statistics for the quote data for CEZ, Komercni Banka (KB), and Cesky Telecom (Telecom) from January 5 to November 12, 2004. The number (#) of price durations shows the number of durations after thinning the original set of quote observations.

2.5.2 Intraday Behavior of Price Durations

As explained in the previous section, transactions durations exhibit significant diurnal patterns. Price durations - which effectively define an intraday volatility process - are not different in this regard. Indeed, as discussed by McInish and Wood (1992), volatility tends to be systematically higher near the market's open and generally just prior to the market's close. With this in mind, we transform the raw price durations prior to estimation by removing the (deterministic) intraday seasonality.

Figure (9) plots the estimated diurnal components for the price durations as functions of time of the day, where the vertical axis is measured in seconds. The well known inverted U-shape is clearly apparent. The opening of the market (9:30h) is very active with price quotations occurring, on average, every 10 to 20 seconds.[20] The middle of the day tends to be less volatile with peaks in the expected durations being most pronounced just before 14:00h for CEZ and KB and 12:00h and 13:00h for Telecom. At these periods, the quotations tend to occur least frequently, with the durations

[20] The plot of expected price durations in Figure (9) on the next page seems to start at zero expected duration. Given the size of the picture, however, this is only an optical illusion.

between quotations of just over 15 minutes (900s) for CEZ and 23 minutes (1,400s) for KB. We may notice an important jump in the activity just before the close (16:00h) corresponding to opening of U.S. markets at 9:00h in the morning. In short, we observe the pattern is very similar for all three stocks. It should be noted, however, that the plots depicted are merely for illustrating a general pattern, and that the diurnal patterns may differ quite strongly from day to day as well as from week to week.

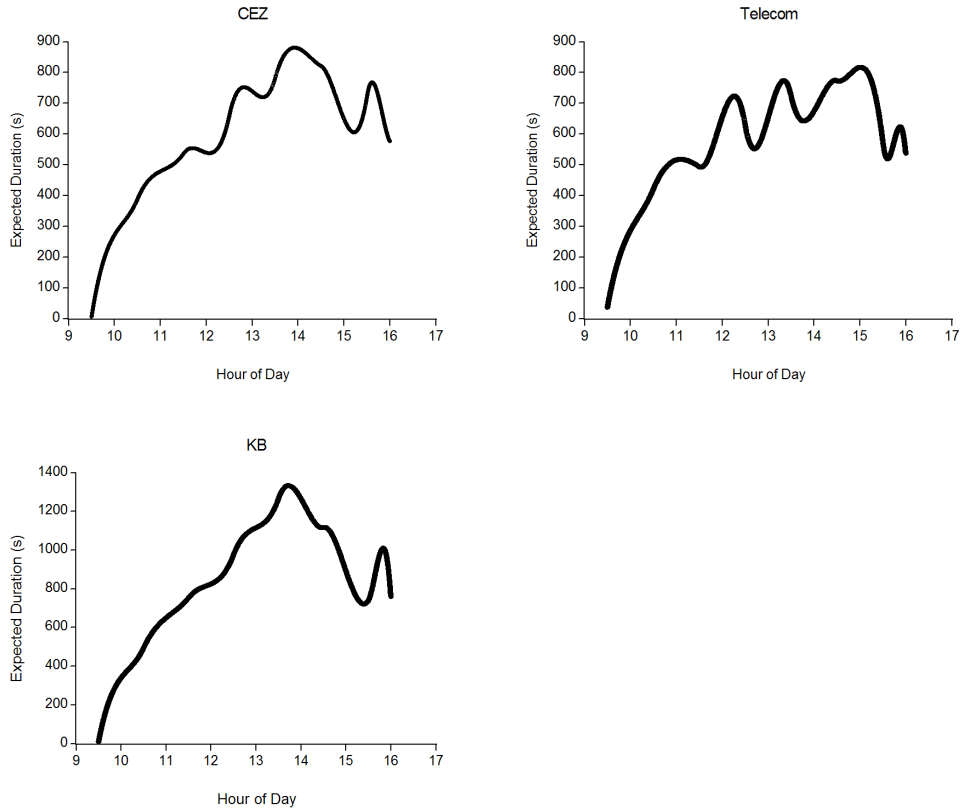

Figure 9: Diurnal components for the price durations of CEZ, price thinning threshold of c = CZK 0.25, Cesky Telecom (Telecom), threshold of c = CZK 0.25, and Komercni Banka (KB), threshold of c = CZK 5.00. The durations are based on the sample period January 5, 2004 to November 12, 2004.

Figure (10) plots the autocorrelation functions for each of the three stocks. The ACFs are presented for both raw and diurnally adjusted durations. The series exhibit long sets of positive autocorrelation spanning many quotes even after the deterministic component has been removed. These autocorrelations indicate clustering of durations.

Figure (11) on page (37) plots the density functions for diurnally adjusted price durations. The estimated densities are similar to the exponential distribution with most of the mass close to zero.

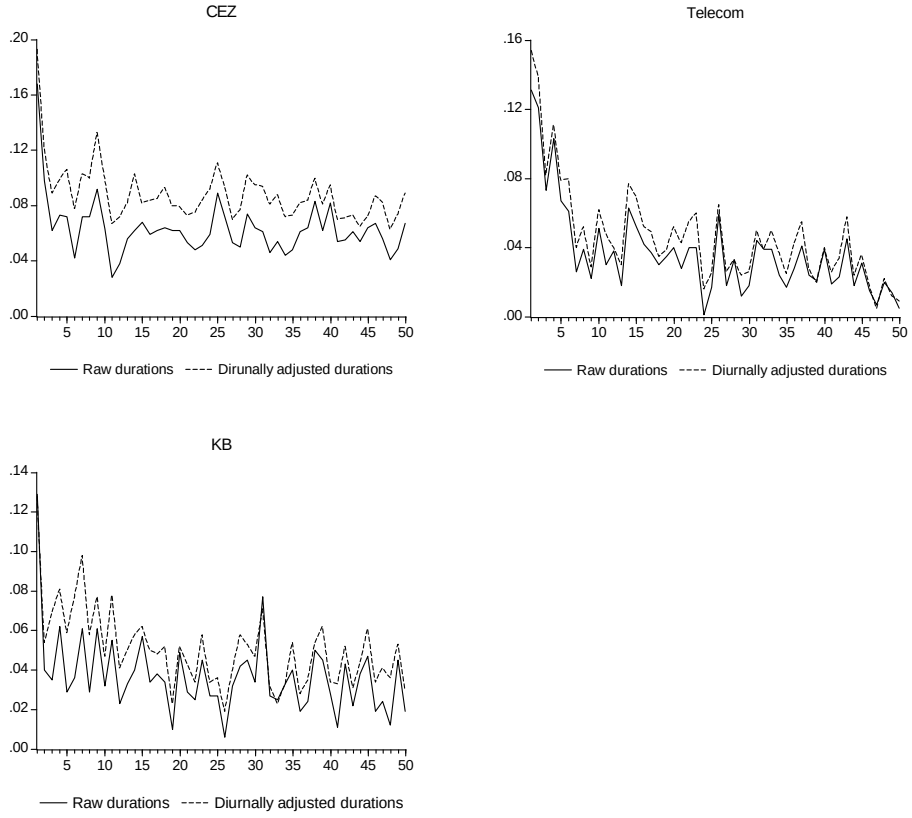

Figure 10: Autocorrelations functions for the raw and diurnally adjusted price durations of CEZ, price thinning threshold of c = CZK 0.25, Cesky Telecom (Telecom), threshold of c = CZK 0.25, and Komercni Banka (KB), threshold of c = CZK 5.00. The price durations are based on the sample period January 5 to November 12, 2004.

2.5.3 Additional Explanatory Variables

In order to test the hypotheses of market microstructure, we need to link the price durations just discussed to characteristics of the trade process. As explained by Bauwens and Giot (2000), with price durations this becomes possible as they are relatively "long" with respect to trade durations and thus allow for the definition of market characteristics for the trading process over the price durations. We focus on three variables related to characteristics of the trade process as suggested by information-based models discussed in the introduction: a) measure of the intensity of trading, b) average volume per trade, and c) existing spread when the past trades were made. All of these variables can be computed relatively easily based on the information contained in the merged trade and quote datasets. Moreover, the variables are meaningful as they are constructed based on a relatively large number of trades. We shortly describe each of them separately in the following paragraphs.

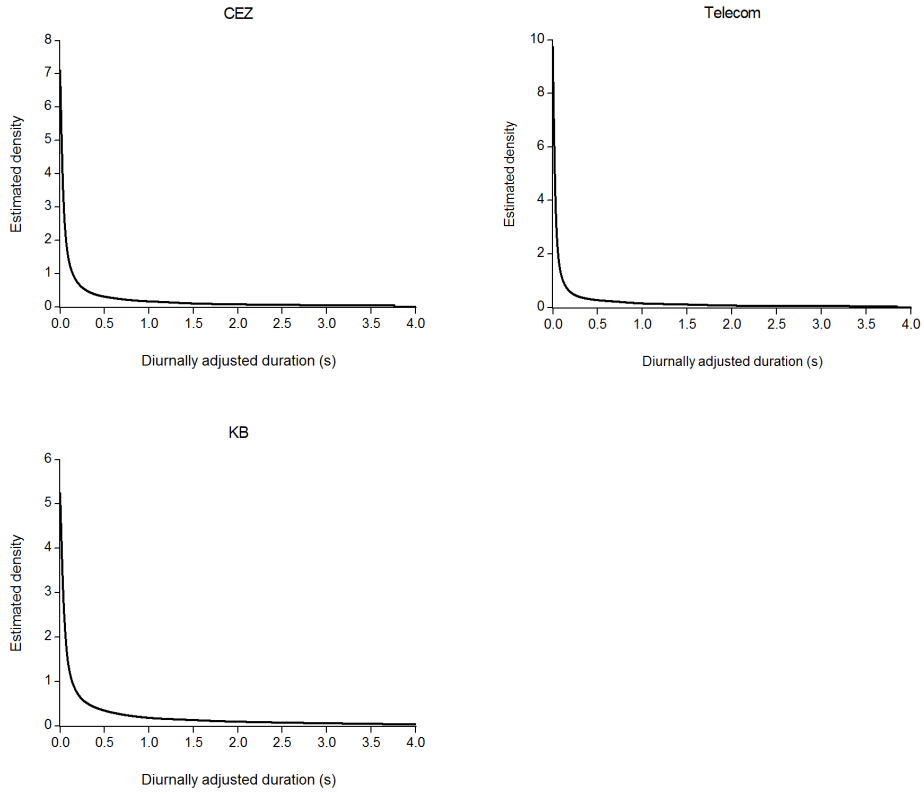

Figure 11: Density functions of diurnally adjusted price durations for CEZ, price thinning threshold of c = CZK 0.25, Cesky Telecom (Telecom), threshold of c = CZK 0.25, and Komercni Banka (KB), threshold of c = CZK 5.00. The price durations are based on the sample period January 5 to November 12, 2004.

Over each price duration, a measure of *trading intensity* is constructed using the number of trade transactions (or simply trades) per second recorded over each price duration divided by the length of the duration. This way, a small number of trades over a long duration leads to a low trading intensity and vice versa. According to the information-based model due to Easley and OHara (1992), discussed in more detail in the Introduction, an increase in the intensity of trading should lead to more frequent quote revisions. Put differently, according to the model the number of transactions would influence the price process through information based clustering of transactions. Nevertheless, a different model by Admati and Pfleiderer (1988) predicts that the number of transactions would have no impact on the intensity of trading. While casual evidence may be ambiguous, using the variable just discussed we are able to test the hypothesis that the number of transactions would influence the price process through information based clustering of transactions statistically.

Further evidence about the validity of predictions of numerous information-based microstructure models can be found using *average volume per trade* variable. Defined

as the average, over previous price duration, of the volume of the trades made during that duration, the average volume per trade is indicative of possible informed trading as shown in the model of EASLEY AND OHARA (1992). The important role of volume has been highlighted in several other studies as well (e.g., EASLEY, KIEFER AND OHARA, 1997). Still, in all cases the volume is found to exhibit an informational content that is not contained in the price process.

The last variable that we include in the regressions is the *spread*. Defined as the average spread, over previous price duration, corresponding to the trades made during that duration, a high spread is once again indicative of possible informed trading and should be linked to short durations. In addition, the asymmetric information models (e.g., GLOSTEN AND MILGROM, 1985) generally predict that spreads tend to become wider as the probability increases of informed agent trading. As a particular kind of such situations, we can imagine a specialist reacting to transaction clustering that results from information-based trading. In such cases, the specialist can set narrow spreads when the fraction of informed traders is small and vice versa when the fraction of traders is large.

2.6 Empirical Results

The log-ACD model is used as the base model in our analysis. We use log-ACD (2,2) model for CEZ and Cesky Telecom (Telecom) and log-ACD (3,2) model for Komercni Banka (KB). In each case, we use the same model to estimate the price process without explanatory variables (Model 1) as well as with additional explanatory variables related to microstructure hypotheses (Models 2 to 4). In the next paragraph, we first describe the results for the model without explanatory variables. Subsequently, we turn our attention to the price models with additional variables.

For CEZ, the estimated coefficients and t-statistics for the transaction duration model (Model 1) are presented in Table (7). The parameters for the stochastic factor are all highly significant. The sum of coefficients on lagged durations (α_i) and lagged conditional durations (β_i) is 0.998, indicating the duration process has strong persistence as measured in transaction time. Similar results can be found for KB (Table 8) and Telecom (Table 9). The Ljung-Box statistics associated with the standardized residuals lie between 12.61 for CEZ and 15.55 for KB. These statistics suggest that the log-ACD model performs relatively well when accounting for the intertemporal dependence in transaction arrival rates.

Turning our attention to price duration models, we can immediately observe that a great majority coefficients is again statistically significant. The coefficients on lagged durations (α_i) tend to exhibit the same pattern with lag-1 duration having positive and lag-2 duration negative impact on the expected price duration. β_1's are generally

positive, ranging from close to 1.19 for CEZ to 1.74 in case of KB.

Table 7: Estimated Price Models for CEZ

	Model 1	Model 2	Model 3	Model 4	Model 5
ω	0.024 (22.03)	0.024 (20.03)	0.024 (21.88)	0.031 (18.17)	0.034 (17.74)
α_1	0.202 (65.21)	0.202 (65.17)	0.203 (64.95)	0.203 (65.05)	0.203 (64.66)
α_2	−0.185 (−61.57)	−0.185 (−61.47)	−0.185 (−61.39)	−0.184 (−60.81)	−0.183 (−59.66)
β_1	1.187 (84.02)	1.187 (83.55)	1.192 (84.53)	1.172 (80.74)	1.166 (78.21)
β_2	−0.206 (−15.15)	−0.206 (−15.08)	−0.211 (−15.56)	−0.194 (−13.90)	−0.190 (−13.43)
γ_1		0.001 (1.11)			0.003 (3.82)
γ_2			0.001 (3.71)		0.002 (5.86)
γ_3				−0.005 (−5.11)	−0.009 (−7.59)
$Q(10)$	12.61	12.55	12.03	12.39	10.85

Coefficient estimates and t-statistics (in parentheses) for estimated transaction duration model (model 1) and price duration models (model 2 to 5) for CEZ. Coefficient $\gamma_1$corresponds to first explicative variable (intensity of trading), γ_2 to average volume per trade, and γ_3 to spread. Robust t-statistics are shown in parentheses. Q(10) is the Ljung-Box Q statistics for serial correlation in stadardized residuals for up to ten lags. The sample period runs from January 5 to November 12, 20004.

Model 2 includes the lagged intensity of trading as an additional explanatory variable to test for the hypothesis that the number of transactions would influence the price process through information based clustering of transactions. This coefficient is statistically insignificant in case of CEZ. For the other two stocks, the coefficients are positive for KB and negative for Telecom. The negative sign means that the expected price durations are shorter, and equivalently the volatility is higher, following the periods of higher transaction rates. Thus, only Telecom is in accordance with the hypotheses put forward by Easley and OHara (1992).

The same results holds true when the second explanatory variable - lagged average volume per trade - is added to the log-ACD model (Model 3). Although this time all coefficients are statistically significant, they are positive in two out of three cases (CEZ and KB). Consequently, only Telecom's result is in agreement with the hypothesis that higher average volume shortens the next expected duration.

Finally, the coefficients on third explanatory variable - lagged average spread - are statistically significantly negative only for CEZ and Telecom (Model 4). Thus, the hypothesis that higher average spread shortens the next expected duration is validated with these two stocks. For KB, the coefficient is insignificant and the result remains inconclusive.

When all three additional variables are included into the regression model at the same time (Model 5), the results are again mixed. As the coefficients keep their signs, it is only in case of Cesky Telecom that they are all negative. This time, however, all of them are highly significant. Obviously, in CEZ and KB only the coefficients on the average spread variable are negative as in the previous Model 4.

Table 8: Estimated Price Models for KB

	Model 1	Model 2	Model 3	Model 4	Model 5
ω	0.003 (5.77)	0.032 (7.89)	0.002 (5.60)	0.003 (3.82)	0.005 (4.24)
α_1	0.138 (29.84)	0.138 (29.96)	0.135 (29.18)	0.137 (29.65)	0.133 (28.58)
α_2	−0.197 (−16.97)	−0.004 (−0.39)	−0.193 (−16.76)	−0.196 (−16.91)	−0.187 (−16.07)
α_3	0.062 (7.44)	−0.102 (8.71)	0.060 (7.39)	0.061 (7.41)	0.057 (6.90)
β_1	1.737 (51.79)	0.090 (1.21)	1.740 (53.81)	1.737 (51.84)	1.729 (49.42)
β_2	−0.739 (−22.30)	0.872 (11.89)	−0.741 (−23.24)	−0.740 (−22.34)	−0.732 (−21.26)
γ_1		0.008 (3.01)			0.000 (−0.25)
γ_2			0.001 (4.88)		0.001 (4.68)
γ_3				−0.000 (−0.29)	−0.002 (−2.67)
$Q(10)$	15.55	28.36	11.87	15.35	11.78

Coefficient estimates and t-statistics (in parentheses) for estimated transaction duration model (model 1) and price duration models (model 2 to 5) for Komercni Banka. Coefficient $\gamma_1$corresponds to first explicative variable (intensity of trading), γ_2 to average volume per trade, and γ_3 to spread. Robust t-statistics are shown in parentheses. Q(10) is the Ljung-Box Q statistics for serial correlation in stadardized residuals for up to ten lags. The sample period runs from January 5 to November 12, 20004.

From these results it would seem that only spread seems to have a consistent negative impact on the next expected duration. In absolute value, the influence of this variable is also the most pronounced of all three additional explanatory variables included in the model. We also note that the likelihood ratio test statistics for joint significance of explanatory variables are 17.44 (p-value of 0.0006) for CEZ and 2.20 (p-value of 0.5319) for KB. Interestingly, it is the highest for Telecom (67.59 with p-value of 0.0000).

The log-ACD model shows an evident success in removing the autocorrelation in the price duration data. For example, in case CEZ, the Ljung-Box Q-statistic of order 10 has been reduced from 594.3 (see Table 7) to around 12 although in this case, the residuals are somewhat influenced by the additional explanatory variables. The results are similar to the other two stocks in this regard. The duration process also continues

to exhibit a strong persistence throughout the stocks although such conclusion should be considered carefully as price durations are much longer on average than transaction durations.

Table 9: Estimated Price Models for Cesky Telecom

	Model 1	Model 2	Model 3	Model 4	Model 5
ω	0.024 (18.45)	0.027 (17.45)	0.033 (20.71)	0.047 (18.91)	0.051 (19.32)
α_1	0.173 (68.11)	0.174 (67.50)	0.174 (68.97)	0.173 (66.93)	0.175 (66.70)
α_2	−0.158 (−62.42)	−0.160 (−63.48)	−0.158 (−61.60)	−0.156 (−60.15)	−0.158 (−60.03)
β_1	1.314 (105.5)	1.330 (112.7)	1.267 (100.91)	1.268 (95.66)	1.260 (99.59)
β_2	−0.340 (−29.91)	−0.354 (−32.53)	−0.300 (−26.34)	−0.302 (−25.12)	−0.294 (−25.50)
γ_1		−0.004 (8.73)			−0.003 (−6.13)
γ_2			−0.007 (−17.91)		−0.005 (−13.20)
γ_3				−0.018 (−15.77)	−0.015 (−13.17)
$Q(10)$	14.25	12.55	16.87	12.81	13.78

Coefficient estimates and t-statistics (in parentheses) for estimated transaction duration model (model 1) and price duration models (model 2 to 5) for Cesky Telecom. Coefficient $\gamma_1$corresponds to first explicative variable (intensity of trading), γ_2 to average volume per trade, and γ_3 to spread. Robust t-statistics are shown in parentheses. Q(10) is the Ljung-Box Q statistics for serial correlation in stadardized residuals for up to ten lags. The sample period runs from January 5 to November 12, 20004.

2.7 Concluding Remarks

In this chapter, we have applied the logarithmic version of the original ACD model developed by Engle and Russell (1998) to price duration process of three of the most liquid securities traded on the Prague Stock Exchange in order to examine whether the intensity of bid-ask quote arrivals carries any information about the state of the market. The preliminary empirical analysis provides evidence of clustering effect in price durations: that is, short (long) durations tend to be followed by short (long) durations, respectively. In fact, we show that large autocorrelations in price durations tend to persist even after the time-of-day effects have been removed from the process.

We take the duration analysis a step further when we empirically test the predictions of market microstructure as demonstrated by Bauwens and Giot (2000) and Engle and Russell (1998). Given the price durations are "longer" than transaction durations, we can define three market characteristics as suggested by the information-based models of market microstructure in order to test for the hypotheses. The vari-

ables concerned are the intensity of trading, the average volume per trade, and the average spread. We obtain the following empirical results: of the three variables, only the average spread seems to have a consistent negative impact on the next expected duration among all stocks. With the other two variables, the results are ambiguous. Both the coefficients on the intensity of trading and the average volume per trade remain positive when tested individually/jointly in two of the three stocks analyzed. In abstract, our results tend to favor the conclusions of information models, however any straightforward judgements remain at best ambiguous.

Several extensions of the model that we used in the empirical analysis are possible. First, a more general distribution for the error term of the model could be used; for example, the Burr distribution which has a non-monotonic hazard function (see GRAMMIG AND MAURER,1999). Second, the specification of the model could include non-linear transformations of the explanatory variables (see BAUWENS AND GIOT, 2000). Or third, while in the model we focused only on the variables suggested in the EASLEY AND OHARA (1992) model, the literature on market microstructure suggests other possibilities as well, such as the depth of the bid and ask, or the changes in price or in spread (see ENGLE AND LUNDE, 1998). This and similar topics are left for further empirical research.

Chapter 3

Price Impact of Stock Trades

Starting with KYLE (1985), GLOSTEN AND MILGROM (1985) and EASLEY AND OHARA (1987), market microstructure research has paid a significant attention to the effect of asymmetric information on market prices. These studies generally propose that if some traders have superior information about the underlying value of an asset, their trades could reveal what this underlying value is and so affect the behavior of prices.

The magnitude of the price effect for a given trade size is generally held to be a positive function of the proportion of potentially informed traders in the population, the probability that any of these traders has in fact observed the private information signal, and the precision of the corresponding private information (see OHARA, 1995). Given the close dependence of the price impact on these factors, we are provided with a strong motivation for the determination of such impact using a real-world transaction data. In this section we strive to assess the size of the price impact using the high-frequency trade and quote transaction data from the Prague Stock Exchange (PSE). Entertaining a framework that is robust to deviations from the assumptions of the formal models of market microstructure, it is thence one of the more important purposes of our analysis to examine and consequently better understand the dynamics of trade and quote process on a representative central European stock market.

The analysis in this chapter is based on an empirical investigation of six of the most active securities traded on the PSE's main market. The market considered here (as in the previous sections) is thus an order driven specialist market in which the market-maker exposes bid and ask quotes to the trading public.[1] Over the past twenty years, a large body of theory has evolved that analyzes the market maker's exposure to trader's with superior information. Other than the studies mentioned earlier, the relevant literature has been described in more detail earlier in the Introduction. With regard to the extent of the information assymetry, this body of theory yields two

[1] We desribed the market-making mechanism in detail in Section (1.1).

important empirical predictions.

The first prediction assumes that the asymmetry is positively related to the spread. Empirical researchers have since sought to find measurable proxies to examine posted bid-ask spreads.[2] Although such procedure has its advantages (e.g., the bid-ask spreads are relatively easy to observe) there are difficulties connected with the price discretness or the existence of the clearing fees that effectively render the examination of the asymmetric information using the bid-ask spread nontrivial. In other words, it is always difficult to fully resolve which components of the bid-ask spread reflext the asymmetric information and which mirror other information such as the transaction costs. On the Czech market, Hanousek and Podpiera (2002) explored the impact of informed trading on the composition of the bid-ask spread. One of the major conclusions of their study is that the Czech market-maker-based trading system is rather efficient in dealing with informed trading. In fact, according to their study less than 20% of the bid–ask spread is explained by informed trading, which corresponds roughly to the share of the adverse-selection component in developed markets.

Still a different group of studies has concentrated on the price impact of the trade. The works of Glosten and Harris (1988) and Foster and Viswanathan (1988) were the original contributions to have started the line of analyses on the potential impact of the trade on quoted price. Nevertheless, even these analyses were originally based on a number of tenuous assumptions such as serial independence of transaction, no delay in the effect of a trade on the price, and a linear trade-price relationship with the intercept corresponding to fixed transaction costs. There are good reasons for questioning these assumptions. In their studies, Garman (1976) and Stoll (1976) were among the first to show that inventory control considerations induce serial dependencies in trades, as do price pressure effects and order fragmentation. In other models, lagged adjustment to new information and exchange-mandated price smoothing[3] may lead to distribution over time of the information impact as well, although we must say the latter is not relevant to the analysis of data from the PSE as (except for very extreme cases) the quotes may be adjusted freely. Finally, the form of the functional relationship between trade size and information is a consequence of a fundamentally unobservable cost and information structure, and many better performing structures thence necessarily entail nonlinear effects.

Hasbrouck (1988) has attempted to partially resolve information and inventory control effects according to the persistence of their impact on the security price. Hasbrouck have suggested that the inventory control effects be considered as inherently

[2] For one of the first studies on the subject, refer to McInish and Wood (1988).

[3] Price smoothing occurs when the market-maker is compelled to set the quotes in such a way as to ensure a smooth price adjustment path. If this happens, the market-maker may not be able to revise the quotes fully and as immediately as free use of the news would allow.

transient, while the information inferred from a trade due to asymmetric information is assumed to be permanently impounded in the stock price. In the study at hand, we follow the work of HASBROUCK (1991) by assuming that such transience characterizes not only inventory control effects but most other non-information imperfections (e.g. price discreteness, price pressure, or order fragmentation) as well. Our approach is thus both attractive and practical as it implies that the information effect of a trade be measured as that which persists over a substantial period of time.

One important feature of the approach adopted by HASBROUCK (1991) as well as in this study is its generality. In his earlier paper, HASBROUCK (1988) noted that if there were to be any private information inferred from a trade, it should be inferred "[] not from the total trade but from that component which was unanticipated - the trade innovation". Still, to investigate this proposition, the paper assumed that the component of the trade which was unexpected depended solely on knowledge of the past trade history. In other words, the paper effectively employed a univariate trade innovation. The present study generalizes the investigation of the implications of trade innovation to incorporate broader information sets (such as histories of the quote revisions and nonlinear functions of the trade variables) and thence achieves a broader picture of the asymmetry information effect. Only to summarize, in the present section we model the trades and quote revisions as a system characterized by auto- and cross-correlations of a very general nature. The information impact of a trade may be formally defined as the ultimate persistent impact on the price quoted resulting from the unexpected component of the trade. This persistent impact is preferred to the immediate impact because the latter may be contaminated by transient liquidity effects. Use of the trade innovation (rather than the total trade) as the driving force has the effect of excluding the portion of the trade which is predictable as (by definition) it conveys no information. Relating the ultimate price impact to just his trade innovation then effectively makes any concerns about market imperfections unnecessary.

This chapter is organized as follows. In Section (3.1) we present a heuristic development of the simple bivariate linear time-series model of trades and quotes revisions. We also motivate and interpret the vector autoregression model (VAR model) and show that the VAR modeling strategy applied to trades and quotes allows, at least in principle, a resolution between private information (trade innovation) and public information (quote revision innovation). As already mentioned, we will adopt the approach similar to the one used by HASBROUCK (1991) and use a similar notation as well for reason of easier reference. In the sections that follow we turn to empirical results. We describe the data in Section (3.2), and present estimations of a simple bivariate and a more advanced quadratic VAR models in Sections (3.3.1) and (3.3.2), respectively. The quadratic model is more elaborate in that it incorporates nonlinear trading effects. Section (3.4) concludes the chapter with a brief summary of most relevant results.

3.1 Modeling the Trade - Quote Revision

In this section, we develop the basic econometric specification underlying the regression model used further in this chapter. Given that no preestablished formula exists as to how to infer such specification, we entertain a heuristic approach according to HASBROUCK (1991). We will start with basic empirical predictions of the structural theoretical models of asymmetric information as described in the previous section. Relaxing the restrictive economic assumptions and adding statistical assumptions of a fairly general nature, we will arrive at a robust empirical specification in which the impact of trade on price due to asymmetric information is both meaningful and observable.

The notation and sequencing conventions in the model discussed are similar to that of HASBROUCK (1988). We denote the prices at which the market-maker is willing to buy as q_t^b and q_t^a, where b (a) stands for the bid (ask) quote, respectively. The timing convention is that these quotes are set after the trade has occurred at time t. Consequently, the quotes prevailing before trading has taken place are q_{t-1}^b and q_{t-1}^a.

A transaction is characterized by its signed volume x_t. Such volume is positive if the trade has been initiated by a buyer (purchase) or negative in case the trade has been initiated by a seller (sale). Based on the observation of x_t, the market-maker posts new quotes q_t^b and q_t^a. The formal models typically assume that these quotes satisfy a zero-expected-profit condition for the market-maker. It follows that if there are no transaction costs, and the only update to the public information set at time t is the trade announced, then the revision in the quotes at time t fully reflects and summarizes the information inferred from the observation of x_t.

The primary price variable used in this section is the midpoint[4] of the quotes. To justify the midpoint of the quotes as a meaningful variable, HASBROUCK (1991) makes an initial assumption that the quotes are symmetrical about the expected value of the security conditional on all public information[5]. Defining the value of the security at some terminal time τ in the future as Υ_t, and letting Φ_t be the public information set at time t, the symmetry assumption can be formally expressed as:

$$E\left[\left(q_t^b + q_t^a\right)/2 - \Upsilon_t \mid \Phi_t\right] = \left(q_t^b + q_t^a\right)/2 - E\left(\Upsilon_t \mid \Phi_t\right) = 0. \tag{3.1}$$

Under the symmetry assumption, the information inferred from x_t may be conveniently summarized as the subsequent revision in the quote midpoint; That is, as the difference

[4] The variable is defined the same way as in Section (2.5.1), note (18).

[5] We will relax the assumption shortly as it is incompatible with the presence of serial correlation in the quote revisions

between current and last (or prevailing) quote midpoint:

$$r_t = \left(q_t^b + q_t^a\right)/2 - \left(q_{t-1}^b + q_{t-1}^a\right)/2. \tag{3.2}$$

Obviously, such formalization makes it possible to identify r_t with the information impact of x_t without being affected by a cost-based component of the spread. Accordingly, the transaction costs, fixed and symmetric for purchases and sales, can then be easily accommodated. Still, the inference can be affected by the arrival of non-trade public information such as news announcements. In this case, the quote revision reflects public as well as private information, making it impossible to infer the price impact of a particular trade. To avoid this problem, we may assume that there is constancy over time in the function relating the trade and the quote revision. In other words, the quote revision is assumed to be a stable function of the trade. In the trivial case, the functional dependence is linear,

$$r_t = bx_t + \varsigma_{1,t} \ .$$

In the equation, the coefficient b measures the price impact of the trade and the disturbance $\eta_{1,t}$ reflects the public information. Obviously, this specification is characterized by a trade impact that is fully contemporaneous. Many microstructure imperfections cause lagged effects in the impact of trade, however, suggesting a more flexible structure of the equation of the following form[6]:

$$r_t = \sum_{j=1}^{\infty} a_j r_{t-j} + \sum_{j=0}^{\infty} b_j x_{t-j} + \eta_{1,t} \ . \tag{3.3}$$

Among the imperfections that cause the lagged effects, price discreteness[7], for example, may induce threshold effects, since a quote revision many not be optimal until a series of trades of the same direction has occurred (Harris, 1990). Inventory control effects, lagged adjustment to information, and exchange-mandated price smoothing also involve serial dependencies.

As previously noted, serial correlation in the quote process is incompatible with the assumption of quote symmetry in equation 3.1 To take care of this problem, the property of conditional symmetry may instead be replaced by a weaker version of symmetry that is valid for the expectation of quotes at some future time s conditional

[6] Although in principle this representation may be of infinite order, it is of practical matter to truncate it at some appropriate lag in the empirical studies.

[7] Price discreteness, or restriction of quotes to a fixed grid, was discussed in greater detail in Section (1.2.3).

on the information set at time t. Hence, as $s \to \tau$,

$$\left(E\left[\left(q_t^b + q_t^a\right)/2 - \Upsilon_t \mid \Phi_t\right]\right) \to 0. \tag{3.4}$$

The reasoning behind this requirement is intuitive: as time passes, all rational agents expect quotes to revert (on average) to the fair value of the security. Put differently, although a quote revision model as specified by equation (3.3) allows for the deviations between the quote mid-point and efficient prices, the deviations are effectively transient given the assumption holds. As will become clear, the analysis presently discussed relies heavily on the expectation of the future quotes implied by the dynamic model. The assumption just described ensures that by carrying the projection out sufficiently far into the future, we arrive at the current efficient price.

The same microstructure effects that we discussed in case of quote revisions will also lead to serial dependencies in trades. The trades, up to now assumed exogenous and wholly unanticipated, can be then modeled in the same fashion as the quotes process:

$$x_t = \sum_{j=1}^{\infty} c_j r_{t-j} + \sum_{j=1}^{\infty} d_j x_{t-j} + \eta_{2,t} \ . \tag{3.5}$$

In this case, the disturbance $\eta_{2,t}$ captures the unanticipated component of the trade: its innovative value-added on top of the expectation formed from linear projection on the trade and quote revision history. Hasbrouck (1988) makes it clear that if there is any private information to be inferred from a trade, it must in effect reside in this innovation.

Together, equations (3.3) and (3.5) comprise bivariate vector autoregressive model. It is assumed that the disturbance terms $\eta_{1,t}$ and $\eta_{2,t}$ have zero means and are jointly and serially uncorrelated:

$$E\eta_{1,t} = E\eta_{2,t} = 0,$$
$$E\eta_{1,t}\eta_{1,s} = E\eta_{2,t}\eta_{2,s} = E\eta_{1,t}\eta_{2,s} = 0, \qquad \text{for } s \neq t. \tag{3.6}$$

The model given by equations (3.3), (3.5), and (3.6) differs from a usual VAR specification in that the r_t specification includes contemporaneous value of x_t. As the coefficients in (3.3) and (3.5) are linear projection coefficients, this implies that $E\eta_{1,t}\eta_{2,t} = 0$. The zero cross-correlation reflects the fact that the quote revision follows the trade, and r_t cannot contemporaneously influence x_t. Related to this is one of the subsidiary purposes of this chapter - the analysis of Granger-Sims causality[8] patterns in the trade and quote data. As Hasbrouck (1988) points out, there is a strong presumption of

[8] When dealing with two jointly covariance stationary time-series x and y, x is said by Granger-Sims to cause y if knowledge of past x and y leads to better predictions of y than would result from knowledge of past y alone. See Geweke, Meese, and Dent (1983).

causality running from trades to quote revisions, and the structure permits this both contemporaneously and with lags. However, due to the above timing considerations, a similar structure for the causality running from lagged quote revisions to trades does not permit contemporaneous causality running from quote revisions to trades.

In order to more clearly assess the role of public and private information in the model, we first consider the trade representation as described by equation (3.5). The equation relates the knowledge of trade and price history to an agent's expectation of the *t-th* trade. If there is any new information contained in x_t, it must reside in its unanticipated component, the innovation $\eta_{2,t}$, since the remaining (historical) component is entirely known[9].

The information content of trade innovation can be assessed as follows. Suppose that at time $t = 0$ the system is in a stable state: that is, all lagged values of r_t and x_t are zero, and the prevailing quotes $\left(q^b_{-1}, q^a_{-1}\right)$ are set to the unconditional expectation of the security's value, $E\left[\left(q^b_{-1} + q^a_{-1}\right)/2\right] = E\left[\Upsilon_t\right]$. At $t = 0$, a signed order $\eta_{2,0}$ arrives and the current trade is set to $x_0 = \eta_{2,t}$. Letting $\eta_{1,t} = 0$ for $\forall t$ and $\eta_{2,t} = 0$ for $t > 1$, and iterating on equations (3.3) and (3.5), we can compute $E\left[r_t|\ \eta_{2,0}\right]$ and $E\left[x_t|\ \eta_{2,0}\right]$ for $t \geq 0$. The sum of the predicted quote revisions through step m,

$$\alpha_m\left(\eta_{2,0}\right) = \sum_{t=0}^{m} E\left[r_t|\ \eta_{2,0}\right] = 0, \qquad \text{for } s \neq t. \tag{3.7}$$

gives us the expected cumulative quote revision (CQR) through the *m-th* step. Given the symmetry condition (3.4) holds, then as m increases,

$$\alpha_m\left(\eta_{2,0}\right) = \sum_{t=0}^{m} E\left[\left(q^b_t + q^a_t\right)/2 - \left(q^b_{t-1} + q^a_{t-1}\right)/2|\ \eta_{2,0}\right] \rightarrow E\left[\Upsilon_t|\ \eta_{2,0}\right] - E\left[\Upsilon_t\right].$$

In other words, the expected CQR converges to the revision in the efficient price. For this reason, $\alpha_m\left(\eta_{2,0}\right)$ can also be interpreted as the information revealed by the trade innovation. Thence, CQR constitutes the basic construct underlying our analysis.

In Appendix (E), we describe the calculation of the impulse response function and thus the corresponding CQR in more detail.

[9]This does not mean that the innovation is a deterministic function of the new information. The presence of uninformed traders, for example, will introduce a noise component $\eta_{2,t}$ that is uncorrelated with private information. In fact, unless some noise is present, the market-maker will only lose to arriving traders. This is effectively one of the predictions of information asymmetry models.

3.2 Data Description

We use high-frequency data on six of the most actively traded stocks listed on PSE's main market (SPAD) between January 5, 2004, and November 12, 2004. Although there were two more securities listed on SPAD in the period in question, we do not consider these in our analysis[10] as their trading generated a relatively small number of transactions.

The data are extracted from the same Trade and Quote datasets that we analyzed in the previous sections. Prior to constructing the time series of trade and quote revisions in the manner consistent with the model's notation and methodology, we adjust the data as follows. First, we remove transactions that took place outside of normal trading hours. In addition, we delete the days on which we know that a significant interruption in trading had occurred. Second, we remove the first transaction of each day as it is likely to be influenced by the trading outside of normal trading hours. Finally, in case of quote data, we consider only unique quotation times and hence regard the simultaneously recorded quotes as a single quotation. Following these adjustments, we merge the files using a data merging program provided in the Appendix (C).

The model discussed in the previous section is specified in transaction time. In other words, the model's *internal clock* t is defined directly by the transaction sequence. In practice this means that the first observation in the sample (i.e., the first trade or quote on January 5, 2004) occurs at time t_0. Consequently, the t is incremented each time a new trade or quote is posted. There is one important exception to this rule: following Hasbrouck (1991), we assign the same t subscript to a quote revision that occurs within 15 seconds following a trade.

Since the transaction data provided by the PSE are not classified according to the nature of a trade (buy or sell), we use the Lee and Ready (1991) *midquote rule* to classify a trade. With this rule, the prevailing quote mid-point corresponding to a trade is used to decide whether a trade is a buy, a sell, or undecided. If the transaction price is higher (lower) than the quote mid-point, it is viewed as a buy (sell). If the price is exactly at the mid-point, the nature of the trade (buy or sell) remains undecided, and x_t is set to zero.

It sometimes occurs that multiple trades take place at the same second. We follow Engle and Russell (1998) and treat multiple transactions at the same time as one single transaction and aggregate their trade volume and average prices.

In Table 10, we present basic summary statistics for the trade data. This table shows that Komercni Banka (KB) is the most frequently traded stock in the sample, with the average duration equal to 51 seconds. Unipetrol is the least frequently traded

[10] The two other securities are Ceske Radiokomunikace and Zentiva. We include a detailed description of the dataset including these two as well as the other six securities in the Appendix (A).

stock of the sample with an average duration of 1,039 seconds, or 17.5 minutes. Average trading volume varies from 142 shares (PMCR) to 26,301 shares (Cesky Telecom).

Table 10: Summary Statistics for Trade Data

	CEZ	Erste	KB	PMCR	Telecom	Unipetrol
# of observations	14,297	8,415	20,133	7,843	15,615	3,949
open phase*	14,083	8,161	19,883	7,551	15,380	3,683
avg price	323	2,637	2,885	16,390	207	74
avg volume (shares)	10,283	1,402	1,327	142	22,510	26,301
avg volume (CZK mil)	4.510	2.618	3.816	2.343	3.311	1.963
mean duration	72.2	559.3	51.1	566.4	55.7	1,038.8
- std. deviation	2,798.7	1,556.3	2,419.2	1,668.6	2,659.0	2,260.5
- maximum	17,727	22,915	16,896	21,771	19,069	19,061

Summary statistics for the trade data for CEZ, Erste Bank (Erste), Komercni Banka (KB), Philip Morris CR (PMCR), Cesky Telecom (Telecom), and Unipetrol from January 5 to November 12, 2005. 'Open phase' shows the number of transactions during open phase (from 9:30 to 16:00) when cleaned of the first observation.

3.3 Empirical Results

We begin the empirical analysis with a simple bivariate model similar to the one outlined in section (3.1). The model is based on the assumption that quote revisions are linear in the trade variable (we use signed trade x_t variable) and hence, as already noted, it is only a tenuous approximation of the "real" quote revision process. As HASBROUCK (1991, p. 196) emphasizes, not only are the underlying cost and information functions unlikely to be linear but "[] the order entry and trade negotiation processes are almost certainly dependent on trade size". Consequently, a model which is more realistic but still amenable to linear estimation will be considered later in the section.

Prior to analyzing the results of the bivariate VAR model, we should emphasize one important difference between the model discussed so far and the model actually estimated: while the former is linear in the signed trade, in the latter we replace x_t with an indicator variable x_t^0 which simply measures the direction (purchase or sale) of the trade. This variable is effectively defined as discrete taking the values of $+1$ if $x_t > 0$, 0 if $x_t = 0$, and -1 if $x_t < 0$. Estimating the model with x_t^0, we effectively obtain a better-behaved model as with x_t we would risk that estimated coefficients on x_t would vary highly in magnitude among different firms. Our model therefore involves equations in which r_t and x_t^0 are regressed against lagged values.

This model as well as the ones that follow were estimated using ordinary least squares methodology of the multiple-equation system.[11]

[11] The estimation was performed by Ox, version 3.40 for Windows. The estimation code can be

3.3.1 Simple Bivariate VAR Model

Estimation results for the simple bivariate VAR model are presented separately for trade and quote equation in Tables (11) and (12), respectively. The model is estimated using twelve lags.[12] In the tables we also present the results of an ARCH LM-test (Engle, 1982). The test provides significant evidence of autoregressive conditional heteroscedasticity in the VAR disturbances $(\eta_{1,t}, \eta_{2,t})$, since the null hypothesis of no ARCH-effects is rejected at each reasonable significance level.[13]

We first examine the price equation. Of particular importance are the coefficient estimates of $x_t^0, \ldots, x_{t-12}^0$. With the exception of Erste Bank, the coefficients are all significant, although the significance decreases as we move to lower lags.[14] The coefficient estimates are predominantly positive with a hint of reversal around the tenth lag. Let us now observe the behavior of x_t^0 which implies how much, on average, the quote midpoint is raised immediately following the purchase order. The impact of x_t^0 is most pronounced in case of Philip Morris CR (PMCR). In fact, the number is almost eight times as large as the second highest estimate in the sample. For other securities in the sample, the impact of x_t^0 is much smaller.

The lagged values of r_t show a statistically significant negative impact of the first lag on the quote midpoint. The impact is positive starting with the second lag. In case of Unipetrol, the negative impact is slightly more persistent. From a purely descriptive point of view, this pattern implies negative lag-1 serial correlation in the quote revisions.

Turning to the volume equation (Table 12, p. 54), we notice a relatively strong positive autocorrelation in trades reflected by the lagged values of x_t^0 coefficients in the x_t^0 estimation. This is consistent with the observations of Hasbrouck and Ho (1987) and Hasbrouck (1988, 1991), and suggests simply that purchases tend to follow purchases, and similarly for sales. As Hasbrouck (1991, p. 194) points out, "[this] pattern of positive autocorrelation at low lags is highly typical". In particular, we observe a striking absence at low and moderate lags of the trade reversal consistent with inventory control mechanisms. In this regard, the short-run predominance of positive autocorrelation is more consistent with lagged adjustment to new information.

---

found in Appendix (D).

[12] Although we tested for the presence of serial correlation in the residuals using the standard LM-test, the results were inconclusive for smaller lags.

[13] An obvious extension of our analysis would be to subsequently define a bivariate GARCH-model that would allow to specifically model the conditional hetereskedastic effects found in both returns and trading volume. Still, we leave this possibility for further empirical research.

[14] In these as well as the following estimations, we set the level of significance to 0.1%. Using such a small level is commonplace in high-frequency data where the number of observations is generally much larger than in low-frequency data (see Section 1.3.1)

Table 11: Estimates of Bivariate VAR Model - Price Equation

	CEZ	Erste	KB	PMCR	Telecom	Unipetrol
const	0.0022	−0.071	0.0285	0.161	0.0048	0.0023
	(3.05)	(−0.76)	(2.82)	(1.25)	(2.03)	(1.74)
a_1	−0.058	0.0008	−0.100	−0.053	−0.202	−0.057
	(−11.90)	(0.15)	(−22.34)	(−8.43)	(−34.25)	(−6.27)
a_2	0.0304	0.0015	0.0315	0.0049	0.0268	−0.012
	(6.27)	(0.28)	(7.01)	(0.77)	(4.44)	(−1.27)
a_3	0.0425	0.0014	0.0226	0.0612	0.0042	−0.017
	(8.77)	(0.27)	(5.03)	(9.64)	(0.69)	(9.74)
a_4	0.0316	0.0008	0.0295	0.0263	0.0244	0.0101
	(6.51)	(0.15)	(6.56)	(4.14)	(4.05)	(9.52)
a_5	0.0405	0.0010	0.0158	0.0229	0.0120	−0.004
	(8.34)	(0.18)	(3.50)	(3.61)	(1.99)	(9.74)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
a_{12}	0.0135	0.0002	0.0208	0.0350	0.0177	0.0151
	(2.81)	(0.04)	(4.71)	(5.58)	(3.00)	(1.65)
b_0	0.0358	0.6714	0.5646	5.6419	0.0638	0.0182
	(22.69)	(2.74)	(27.14)	(18.72)	(15.01)	(5.73)
b_1	0.0321	−0.261	0.5422	5.5324	0.0458	0.0111
	(22.12)	(−1.06)	(25.68)	(18.17)	(10.73)	(3.50)
b_2	0.0216	0.1921	0.3401	2.2699	0.0256	0.0101
	(13.44)	(0.78)	(15.89)	(7.37)	(5.95)	(3.18)
b_3	0.0114	0.0960	0.1387	1.7345	0.0109	0.0038
	(7.06)	(0.39)	(6.46)	(5.62)	(2.53)	(1.18)
b_4	0.0052	0.7267	0.0596	0.8473	0.0013	0.0028
	(3.21)	(2.95)	(2.77)	(2.74)	(0.31)	(0.88)
b_5	0.0038	0.6884	0.0388	0.7311	0.0069	0.0037
	(2.36)	(2.79)	(1.80)	(2.36)	(1.58)	(1.18)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
b_{12}	−0.005	0.7009	−0.065	−0.436	−0.003	−0.002
	(−2.90)	(2.86)	(−3.04)	(−1.42)	(−0.76)	(−0.77)
R^2	0.055	0.001	0.061	0.054	0.062	0.012
ARCH(4)	47.9	56.8	377.0	10.2	79.1	9.25
(p−value)	(0.00)	(0.02)	(0.00)	(0.04)	(0.00)	(0.00)

Coefficient estimates and t-statistics (in parentheses) for the price equation of the simple bivariate VAR model. The set of variables in the model include the price (quote midpoint) change (a) and trade indicator variable (b). A method proposed by White (1980) was used to obtain heteroskedasticity consistent standard errors. ARCH (4) refers to Engle's (1982) LM test for ARCH effects of order 4. The test is chi-2 (4) distributed. The period assessed runs from January 5 to November 12, 2004.

The nature and duration of the autocorrelation in trades is worth further discussion as it plays an important role in the specification and estimation of VAR models. For example, the trade reversal patterns tend to arise only in large pooled samples of relatively low market-value stocks.[15] But even in these samples the reversal is often found to be very weak and tends to be distributed over very long lags.

[15] This is certainly not the case of the sample analyzed in our study.

Table 12: Estimates of Bivariate VAR Model - Volume Equation

	CEZ	Erste	KB	PMCR	Telecom	Unipetrol
const	0.0027	0.0028	−0.006	−0.006	−0.012	−0.004
	(1.22)	(1.37)	(−2.84)	(−2.23)	(−3.57)	(−1.07)
c_1	0.1072	0.0002	0.0040	0.0004	−0.089	0.0035
	(7.25)	(1.39)	(4.17)	(2.68)	(−10.93)	(0.13)
c_2	0.1096	0.0001	0.0055	0.0007	0.0054	0.0293
	(7.40)	(0.88)	(5.71)	(4.91)	(0.65)	(1.10)
c_3	0.0399	*insig.*	0.0036	0.0003	−0.010	0.0241
	(2.69)	(0.34)	(3.69)	(1.92)	(−1.20)	(0.91)
c_4	0.0279	*insig.*	0.0029	*insig.*	−0.002	0.0651
	(1.88)	(0.36)	(2.98)	(−0.31)	(−0.22)	(2.46)
c_5	0.0405	*insig.*	0.0018	−0.001	−0.012	0.5610
	(2.73)	(0.21)	(1.81)	(9.74)	(−1.42)	(−2.12)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
c_{12}	0.0091	*insig.*	−0.002	*insig.*	0.0170	0.0213
	(0.62)	(2.66)	(−2.17)	(0.70)	(2.08)	(0.81)
d_1	0.0993	0.0796	0.1256	0.0837	0.0479	0.0548
	(20.44)	(14.65)	(27.86)	(13.13)	(8.09)	(5.97)
d_2	0.0851	0.0706	0.1222	0.0966	0.1348	0.1021
	(17.36)	(12.94)	(26.73)	(15.00)	(22.72)	(11.11)
d_3	0.0528	0.0600	0.0487	0.0539	0.0720	0.0449
	(10.71)	(10.97)	(10.56)	(8.32)	(12.01)	(4.86)
d_4	0.0324	0.0333	0.0325	0.0386	0.0670	0.0330
	(6.56)	(6.09)	(7.03)	(5.95)	(11.14)	(3.57)
d_5	0.0184	0.0314	0.0229	0.0398	0.0432	0.0390
	(3.71)	(5.73)	(4.95)	(6.13)	(7.18)	(4.21)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
d_{12}	0.0172	0.0156	0.0149	0.0051	0.0063	0.0078
	(3.50)	(2.87)	(3.25)	(0.79)	(1.06)	(0.85)
R^2	0.044	0.028	0.062	0.040	0.066	0.031
ARCH(4)	4,005	1,925	4,725	1,983	3,835	1,067
(p−value)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

Coefficient estimates and t-statistics (in parentheses) for the volume equation of the simple bivariate VAR model. The set of variables in the model include the price (quote midpoint) change (c) and trade indicator variable (d). A method proposed by White (1980) was used to obtain heteroskedasticity consistent standard errors. ARCH (4) refers to Engle's (1982) LM test for ARCH effects of order 4. The test is chi-2 (4) distributed. The period assessed runs from January 5 to November 12, 2004.

For individual stocks, the reversal pattern is simply too weak to obtain reliable estimates in samples of the present size (see HASBROUCK, 1991). In fact, in the empirical literature this consideration generally underlies the decision to truncate the model specifications at lags below those at which the reversal might be presumed to operate.

Still, finding a solution to one problem generates another. The question is how the trade effects at longer lags might affect the analysis. Given the weakness of the trade reversal, it is certainly unlikely that addition of trades at long lags would add

more significance to the trade or quote revision estimations. It is thus not likely that the estimated coefficients would be significantly altered. However, the omission of longer lags may have more serious implications for the implied cumulative quote revisions.[16] The cumulative nature of this function means that over a long period even the (significantly) very smallest of the effects may add up to significance. The problem may be illustrated (although not proved!) by analyzing the dynamic properties of the system. We may do this by examining the response of the quote midpoint to a purchase at time $t = 0$, where the purchase means that $x^0_{t=0} = 1$.

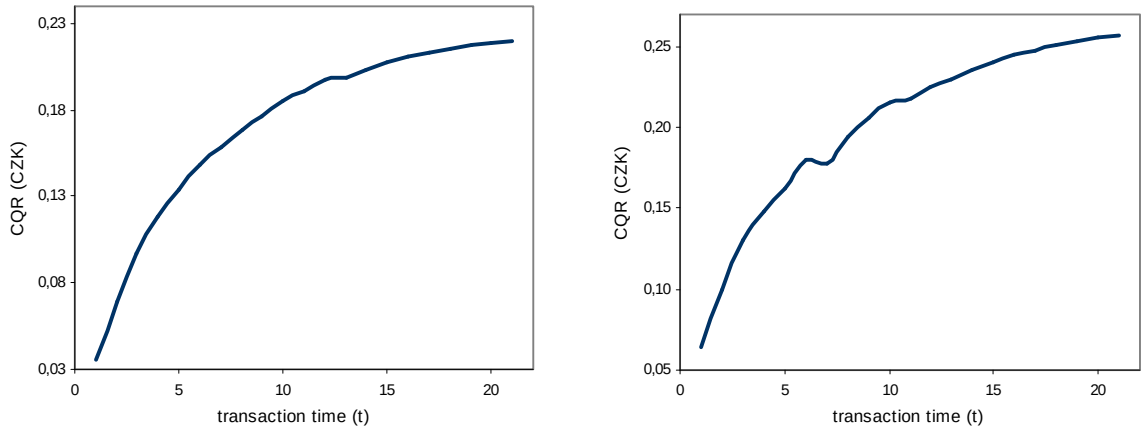

Figure 12: The graph of cumulative quote revision process (CQR) for CEZ (left) and Cesky Telecom (right) stocks. The figure depicts the CQR as implied by a bivariate VAR model presented in Tables (11) and (12), subsequent to an initial buy order. The index t refers to transactions (lags 1 through 20).

The cumulative quote revision is graphed in Figure (12) for two of the stocks assessed in the analysis, CEZ (left) and Cesky Telecom (right). The convergence seems to be quite slow. To be exact, it takes about 20 lags for either of the two securities to reach its convergence level: CZK 0.22 for CEZ and CZK 0.25 for Cesky Telecom. Quite interestingly, for both CEZ and Cesky Telecom, nearly 80% of this level is reached already by the $8th$ lag. This can be examined by plotting the individual impulse responses in transaction time as in Figure (13) on the next page.

The fact that the convergence just discussed is not instantaneous suggests that transient dynamic considerations - the same features to which VAR models in particular are well suited - are important. Still, as we had a chance to observe, the convergence levels are not straight lines and owing to estimation limitations we can only speculate about what might happen at much higher lags if a long-run trade reversal were present. In such a case, the quote impact calculations implied by VAR estimates of low order

[16] Refer to Equation (3.7) for a more detailed discussion.

would likely overstate the long-run price impact. In other words, the model would catch up the initial positive impact of a trade on the quote, but would miss the subsequent long-run reversion. HASBROUCK (1991) states that if this problem arises, it is most likely to be present in the estimations for stocks of low market values and may cause the estimates of the price impact to be biased upwards. This might be the case of Unipetrol and Philip Morris CR (PMCR), which had relatively lowest market capitalization of all stocks in the period considered.

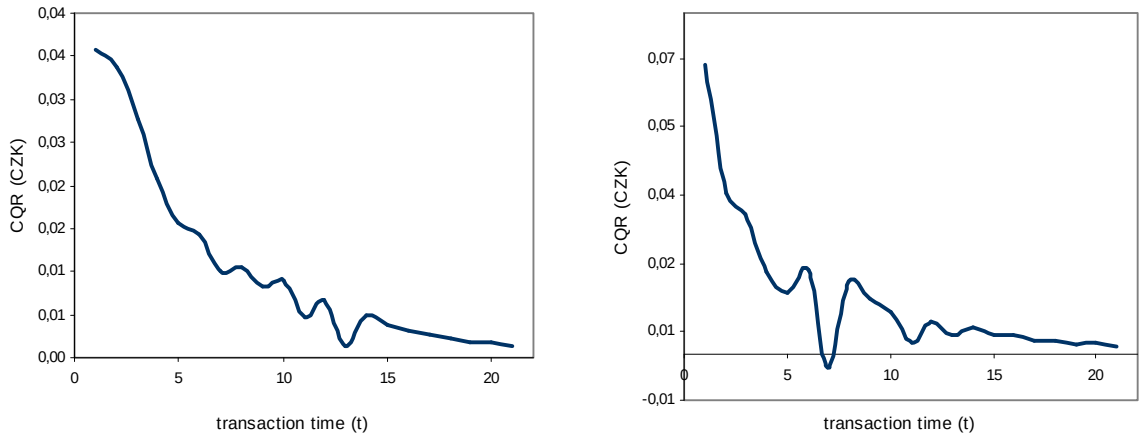

Figure 13: The graph of impulse response function (IR) for CEZ (left) and Cesky Telecom (right) stocks. The figure depicts the IR as implied by a bivariate VAR model presented in Tables (11) and (12), subsequent to an initial buy order. The index t refers to transactions (lags 1 through 20).

Other than the positive autocorrelation in trades, one of the primary factors in determining the model's dynamic adjustment path, the pattern of positive lagged r_t coefficients in the x_t^0 specification also stands out as noteworthy. In Section (3.1) we mentioned the possibility of Granger-Sims causality running from quote revisions to trades. It is obvious that the positive r_t coefficients do not imply the causality.[17]

The interpretation of these findings deserves a short note. HASBROUCK (1991) suggests that the behavior of lagged r_t coefficients could be explained by the measurement error. According to Hasbrouck, the most likely source of this error is the stale (transaction) process where the transactions are reported faster than the quotes are. Given that in our study the bid-ask prices are quoted at a much slower rate than it is the case of the securities analyzed by Hasbrouck, it should be safe to assume that such a stale process does not occur in the data from the PSE and thus should not influence the results in our study. Still, another source of the measurement error, the quote

[17] In the r_t estimation, a formal Wald test could not reject the null hypothesis that the quote revision coefficients are zero for any of the stocks; the significance level was set at 1%.

reporting errors, may be capable of generating the observed behavior and we take this into consideration in the concluding remarks.

3.3.2 Nonlinearities in Trades $\rightsquigarrow$ Quotes Relation

The VAR systems analyzed so far have been simple bivariate models based on the assumption that the dependence between quote revisions and the trades (as proxied by trade sign variable in the estimation) is linear. The quote revision process in particular was modelled as a linear function both of its own lagged values and contemporaneous and lagged values of trades.

Still, this assumed linearity is at best a questionable approximation of reality. As we already mentioned, not only are the underlying cost and information functions unlikely to be linear but the order entry and trade negotiation processes are almost certainly dependent on trade size (HASBROUCK, 1991). A more realistic group of models may be developed by moving to multivariate VAR specifications. In the following section we first develop what could be considered a representative example of such model and include the results of its estimation in the latter paragraphs as well.

One obvious specification for the (multivariate) model involves regressing the quote revision against current and lagged signed powers of the trade variable. The possibility stems from the fact that a function in general may be approximated by a polynomial expansion. The model analyzed here is quadratic. Other than the signed trade variable x_t and the indicator variable x_t^0 introduced earlier, we also add a large trade variable defined as $[x_t]^2$. The full quadratic VAR model thus includes four linear equations in which each of the four variables is regressed against lagged values of the entire set. The general specification for each equation is:

$$(*)_t = \sum_i a_i r_{t-i} + \sum_i b_i x^0_{t-i} + \sum_i c_i x_{t-i} + \sum_i d_i x^2_{t-i} + \eta_{*,t} \ ,$$

where $(*)$ indexes the set of variables in the model.

A final note on the model methodology relates to its estimation. Given that none of the variables included in the set were found to be statistically significant linear predictors of $[x_t]^2$, we follow HASBROUCK (1991) and do not estimate the equation for $[x_t]^2$; in effect, we treat the large-trade variable as exogenous.

As in the previous section, we estimate the quadratic VAR model for each of the six SPAD securities.[18] The results are provided in Table (13) on page 59 for CEZ and Cesky Telecom stocks and in Table (14) on page 60 for the other four stocks.

[18] As in the simple bivariate case, the estimation of the quadratic VAR model was performed by Ox version 3.40 for Windows. The estimation code is available in Appendix (D).

In each of the two tables, the estimated coefficients are summarized by equation and groups of like variables. In case of CEZ and Cesky Telecom several lags of estimated coefficients are included for reasons of clarity. For each equation and each type of variable the sum of the coefficients across all 12 lags and the t-statistics associated with this sum are also reported in the tables. The p-value associated with an $F-$test of the null hypothesis that all coefficients in the group are zero as well as other relevant parts of estimation are not included for brevity. As in the previous estimations, we include the LM(4) and ARCH-LM test statistics as well.

Again, we first discuss the estimates for r_t equation. For all of the six stocks under analysis, the coefficient sums for x_t^0 and x_t are all predominantly positive with the only exception coming from Erste Bank and Unipetrol stocks where the coefficients are significant but negative at the x_t^0 variable. The coefficient sum for $[x_t]^2$ is mostly statistically insignificant. In particular, for four of the six stocks under analysis the coefficient sums are negative but insignificant. In case of CEZ and KB, the two stocks with a relatively high t-statistics, the the coefficient sums for $[x_t]^2$ are positive. In other words, as a function of trade size, the price impact is generally positive, increasing, and convex.

In the trade variable regressions x_t^0, the pattern of positive autocorrelation is most pronounced in the indicator variable and less striking in the x_t. CEZ and KB are once again exceptional in this regard. The coefficients of lagged quote revisions in the x_t equation are either positive (CEZ, Erste Bank, PMCR, Unipetrol), or negative (Telecom, KB), in the x_t^0 equation they are all positive but for Telecom, and finally in the r_t equation they are again either positive or negative. As in the previous section, we investigated the Granger-Sims causality running from quote revisions to each of the trade variables. Using an F-test, the null hypothesis that r_t did not Granger-cause any of the trade variables could not be rejected for any of the stocks in the sample at 0.1 percent level of significance.

3.3.3 A Few Notes on Cumulative Quote Revision

The interpretation of the impulse response function $\alpha_m\left(\eta_{2,0}\right)$, defined in (3.4), as a measure of private information rests on the assumption that $\eta_{2,0}$ reflects no public information. This assumption follows directly from the dichotomy that pervades most of the formal models of asymmetric information as well as the derived empirical models: in all of these, the trade is driven partially by private information and partially by liquidity needs, but in no part by public information which is relevant to forecasting the value of the security.

Table 13: Estimates of the Quadratic VAR Model for CEZ and Telecom

	CEZ			Telecom		
const.	0.0021 (2.73)	0.0030 (1.26)	0.0233 (1.21)	0.0041 (1.67)	−0.011 (−3.27)	−0.089 (−2.64)
	r_t	x_t^0	x_t	r_t	x_t^0	x_t
a_1	−0.057 (−11.73)	0.1065 (7.18)	0.7114 (5.26)	−0.208 (−35.16)	−0.090 (−10.98)	−0.476 (−5.87)
a_2	0.0327 (6.70)	0.1103 (7.41)	0.7562 (5.58)	0.0271 (4.48)	0.0056 (0.67)	−0.142 (−1.71)
...	...	...	...	...	...	...
$\sum_{i=1}^{12} a_i$	0.2078 (2.52)	0.3691 (2.41)	1.9883 (0.69)	−0.036 (−0.95)	0.0336 (2.83)	−0.788 (−4.65)
b_0	0.0224 (10.00)	0.0906 (13.27)	0.6404 (10.29)	0.0516 (10.38)	0.0550 (8.01)	0.2138 (3.15)
b_1	0.0213 (9.50)	0.0880 (12.86)	0.4260 (6.82)	0.0338 (6.78)	0.1416 (20.57)	0.5648 (8.29)
...	...	...	...	...	...	...
$\sum_{i=0}^{12} b_i$	0.0672 (29.88)	0.3464 (50.59)	1.7615 (28.22)	0.1145 (22.95)	0.5018 (72.52)	1.9644 (28.70)
c_1	0.0021 (8.45)	0.0013 (1.78)	0.0245 (3.54)	−0.001 (−1.04)	−0.001 (−1.85)	−0.004 (−0.62)
c_2	0.0005 (1.94)	−0.001 (−0.76)	0.0080 (1.15)	-7×10^{-5} (0.14)	−0.002 (−2.15)	0.0044 (0.63)
...	...	...	...	...	...	...
$\sum_{i=1}^{12} c_i$	0.2062 (1.91)	0.3832 (2.89)	2.0877 (8.77)	−0.042 (−1.99)	0.0406 (5.41)	−0.657 (3.22)
d_1	2×10^{-5} (4.18)	-4×10^{-6} (−0.22)	-3×10^{-6} (−0.21)	-5×10^{-6} (0.78)	-4×10^{-6} (0.45)	-2×10^{-5} (−0.20)
d_2	5×10^{-6} (0.95)	-6×10^{-5} (−3.55)	−0.001 (−4.68)	-2×10^{-6} (−0.26)	-1×10^{-5} (−1.23)	−0.001 (−5.50)
...	...	...	...	...	...	...
$\sum_{i=1}^{12} d_i$	2×10^{-5} (2.15)	-1×10^{-5} (−0.65)	−0.002 (−3.10)	-6×10^{-6} (−2.87)	2×10^{-5} (−2.04)	−0.001 (−9.61)
R^2	0.046	0.045	0.028	0.055	0.067	0.019
ARCH(4) (p−value)	51.2 (0.00)	291.7 (0.00)	4,002 (0.00)	83.8 (0.00)	85.4 (0.00)	3,842 (0.00)

Estimates of the quadratic vector autoregressive model for Erste Bank (Erste), Komercni Banka (KB), Philip Morris CR (PMCR), and Unipetrol stocks. The set of variables in the model include the price (quote midpoint) change, the trade indicator variable, the signed trade volume, and the *large* volume variable. The table contains summary statistics for each group of regression coefficients including the sum of the coefficients in the group and a t-statistic for this sum (in parentheses). The results of an F-test of the null hypothesis that all coefficients in the group are zero are available on request. A method proposed by White (1980) was used to obtain heteroskedasticity consistent standard errors. ARCH (4) refers to Engle's (1982) LM test for ARCH effects of order 4 (the test is chi-2 (4) distributed). The period assessed runs from January 5 to November 12, 2004.

Table 14: Estimates of the Quadratic VAR Model (Other Stocks)

	Erste			KB		
	r_t	x_t^0	x_t	r_t	x_t^0	x_t
$\sum_{i=1}^{12} a_i$	0.0009 (1.86)	0.0011 (0.36)	0.0048 (7.12)	0.1366 (30.01)	0.0182 (18.69)	−0.021 (−2.43)
$\sum_{i=0}^{12} b_i$	0.1231 (0.34)	0.3839 (5.18)	0.9369 (21.11)	1.0350 (37.39)	0.4262 (71.97)	2.1201 (40.08)
$\sum_{i=1}^{12} c_i$	−0.296 (1.52)	0.0109 (2.68)	0.1411 (20.31)	0.1190 (39.59)	0.0270 (19.83)	0.1078 (14.88)
$\sum_{i=1}^{12} d_i$	−0.001 (1.56)	0.001 (1.56)	0.001 (1.56)	−0.001 (1.97)	−2.869 (−2.71)	−0.001 (−9.53)
R^2	0.003	0.028	0.015	0.052	0.063	0.063
ARCH(4) ($p-$value)	0.0 (1.00)	259.0 (0.00)	1,925 (0.00)	368.6 (0.00)	94.4 (0.00)	4,720 (0.00)
	PMCR			Unipetrol		
	r_t	x_t^0	x_t	r_t	x_t^0	x_t
$\sum_{i=1}^{12} a_i$	0.1720 (2.80)	0.0007 (0.50)	0.0019 (2.47)	−0.014 (−1.94)	0.0257 (1.20)	1.6269 (1.34)
$\sum_{i=0}^{12} b_i$	8.5921 (1.69)	0.4020 (4.42)	0.9711 (0.74)	0.0604 (4.51)	0.3759 (2.27)	0.7296 (1.19)
$\sum_{i=1}^{12} c_i$	0.7539 (0.96)	−0.013 (−1.57)	−0.002 (−1.41)	−0.014 (−4.11)	0.0371 (2.57)	1.703 (2.88)
$\sum_{i=1}^{12} d_i$	−0.001 (−1.55)	−0.001 (−1.72)	−0.001 (−1.13)	−0.001 (−1.02)	−0.001 (−2.11)	−0.003 (−5.86)
R^2	0.048	0.041	0.017	0.011	0.034	0.038
ARCH(4) ($p-$value)	11.3 (0.02)	23.2 (0.00)	1,975 (0.00)	0.26 (1.00)	178.4 (0.00)	1,058 (0.00)

Estimates of the quadratic vector autoregressive model for Erste, Komercni Banka (KB), Philip Morris CR (PMCR), and Unipetrol stocks. The set of variables in the model include the price (quote midpoint) change, the trade indicator variable, the signed trade volume, and the *large* volume variable. The table contains summary statistics for each group of regression coefficients including the sum of the coefficients in the group and a t-statistic for this sum (in parentheses). The results of an F-test of the null hypothesis that all coefficients in the group are zero are available on request. A method proposed by White (1980) was used to obtain heteroskedasticity consistent standard errors. ARCH (4) refers to Engle's (1982) LM test for ARCH effects of order 4 (the test is chi-2 (4) distributed). The period assessed runs from January 5 to November 12, 2004.

In the model that we adopted from Hasbrouck (1991), the dichotomy is not less apparent. In fact, the equations (3.3) and (3.5) identify all public information with the quote revision innovation ($\eta_{1,t}$) and all private information with the trade innovation ($\eta_{2,t}$). This renders the model very approximative with respect to reality as in practice: a pure dichotomy simply does not hold. The estimated values of α_m may then either capture the response of the quote to private information inadequately or else may

reflect the response of the quote to public information as well. We will now consider a few of the imperfections that might upset the dichotomy.

One aspect of the dichotomy implies that the quote revision innovation reflects only public information. Since on the PSE the quotes are provided by dealers, this assertion may be violated if the dealers possess (and ultimately use) superior information. In that case, the prevailing quotes may reflect the private information (as they are set by the dealers), which will of course not be captured by the impulse response function.

The other aspect of the dichotomy implies that the trade innovation contains no public information. We may formalize this as the requirement that public information is not useful in predicting the trade innovation. In reality, this requirement is violated if any of the following takes place at the market: a) the specialists get involved in smoothing the price transition path, and/or b) the quotes and/or limit orders are reported with a lag (stale quotes/limit orders). We assume only these two "imperfections" as they are most often cited in the empirical literature.

If the market-maker is required to show a smooth price transition path, then quotes may not be immediately revised to reflect public information. The implications of this are clear: if the market maker is compelled to set the quotes in a such a way as to ensure a smooth price adjustment path (this is indeed the case of most of the exchanges including the PSE), then he may not be able to revise the quotes as fully and as immediately as unconstrained use of the news would imply. As noted in the Introduction, however, the stale quotes do not present a problem for the correct interpretation of the impulse response function for the data from the PSE.

3.4 Concluding Remarks

The aim of this chapter has been to analyze the information content of a trade as implied by its effect on the quote revision. In the analysis, we used the data for all stocks traded on the PSE's main market (SPAD) in the first eleven months of 2004.

Entertaining a heuristic approach, we first arrived at a robust empirical specification in which the impact of trade on price due to asymmetric information is both meaningful and observable. In such a model, the information content of a trade may be meaningfully measured as the persistent impact of the unexpected component of the trade; that is, as the ultimate impact of the trade innovation. By focusing on the trade innovation rather than the trade itself, we avoid misleading inferences due to inventory control or other transient liquidity effects. By considering the persistent impact of the innovation, we concentrate on the information ultimately impounded in the price after transient liquidity effects have died out.

Our findings for the six securities traded on the PSE are as follows. First, the full impact of a trade on the security price is not felt instantaneously but a with a protracted

lag. This conclusion effectively sets an important benchmark for further studies of the PSE. Any analyses of the stocks traded on the PSE which would assume that the full impact of a trade on price is instantaneous would be seriously incomplete. Second, as a function of trade innovation size, the ultimate impact of the innovation on the quote is nonlinear, positive, increasing, and convex. This is a tentative characterization of the trade size-price impact relation. The convexity seems to be a particular feature of the stocks traded on the PSE as, in general, the relationship tends to be concave. Still, results regarding the convex pattern in trade $\rightsquigarrow$ quote relation should be accepted only carefully, as the conclusion is based on the coefficients of only two of the six securities under analysis. Finally, we show that the order flow does not seem to be affected by prior quote revisions. In other words, there does not seem to be Grange-Sims causality running from quote revisions to trades. This finding is of marginal importance although had the causality been in fact proved, it would have presented an intriguing reason for practical experimentation.

A number of directions exist for further research along the lines presented in the chapter. The potential studies fall in two groups. In the first group lie comparative analyses of the trade impact across firms which have different size (market capitalization) and perhaps also trading patterns. Another important set of issues that belong to this group concerns interday behavior of the trade impact such as time-of-day and day-of-week seasonalities. The second class of studies concerns refinements. For example, what characterization appears to best describe the price impact as a function of trade size? We leave these and other extensions to further research in the area.

Conclusion

This thesis contains empirical research on market microstructure theory and covers several closely related topics: the investigation of trading intensity, the impact of trade-related characteristics on the speed of the quote revision at time $(t + 1)$, and the information content of stock trades as revealed by their effect on stock price. The analysis builds on the works of Engle and Russell (1998), Bauwens and Giot (1999), and Hasbrouck (1998), among others.

Being well aware of the fact that the actual mechanism used to set prices is not merely a channel to an inevitable income, but rather is an input into the equilibrium price itself (OHara, 1990), in the thesis we first described the basic features of the PSE including its trading structure and processing of instruction information. Given this understanding, we next described some of the basic characteristics of high-frequency transaction data for a sample of data from the PSE's main market. Namely, we made ourselves familiar with the heterogeneity, discreteness, diurnal patterns, or price reversals. We noticed that some of these characteristics were not as pronounced in case of the Czech (most-active) securities as they are in case of high-frequency data in general. We found that - at least for the two most liquid securities present on the Czech stock market during the period - the stocks tend to exhibit large tick multiples. Alsom, the bid-ask bounce is perhaps not a case of the Czech high-frequency data.

We next assessed the behavior of high-frequency transaction durations. Such durations play a key role in understading the processing of information in the financial markets and are equally relevant in the theories of market microstructure. We first investigated the basic properties of both raw and adjusted price durations to see whether they tend to exhibit the same clustering effect that can be generally found in high-frequency data. We found long sets of positive autocorrelation spanning many quotes even after the deterministic component has been removed. These autocorrelations indicated clustering of durations. Still a more important part of the analysis was concerned the examination of the marks associated with the price durations where the price durations were defined as the time needed to witness a given cumulative price change in the price of an asset. Among these marks, the intensity of trading, the volume per trade, and the corresponding spread play the key roles in the information

based theories of market microstructure and we investigated their impact on the speed of the quote revision at time $(t+1)$ using their respective proxies: the intensity of trading, the average volume per trade, and the average spread.

Based on a sample of three most liquid stocks traded on PSE's main market in 2004, we found that of the three proxy variables considered, only the average spread seemed to have a consistent negative impact on the next expected duration among the stocks. With the other two variables, the results were ambiguous. Both the coefficients on the intensity of trading and the average volume per trade remained positive when tested individually/jointly in two of the three stocks analyzed. In abstract, our results tend to favor the conclusions of information models, however we should be careful in making any straightforward judgements in this regard.

In the next part of the thesis, we studied the price impact of stock trades. Here, our analysis followed the studies of Kyle (1985), Glosten and Milgrom (1985) and Easley and OHara (1987) who had examined the effect of asymmetric information on market prices. One of the major propositions of their studies was that if some traders have superior information about the underlying value of an asset, their trades could reveal what this underlying value is and so affect the behavior of prices. We used a vector autoregressive (VAR) model for trades and returns similar to Hasbrouck (1991) to investigate the price impact of (stock) trades for a sample of six of the most actively traded stocks listed on the PSE's main market in 2004. We concluded that (a) for a sample of stocks assessed, the full impact of a trade on the security price is not felt instantaneously but a with a protracted lag, and (b) as a function of trade innovation size, the ultimate impact of the innovation on the quote is nonlinear, positive, increasing, and convex. Finally, we showed that the order flow does not seem to be affected by prior quote revisions. In other words, there does not seem to be Grange-Sims causality running from quote revisions to trades. This finding is of marginal importance although had the causality been in fact proved, it would have presented an intriguing reason for practical experimentation.

Overall, our findings have several important implications for the future research of the Czech stock market. Representative of these is for instance the fact that the trade impact is not immediate. In fact, it takes some time for a trade to fully transmit its innovative part on the stock price. Obviously, any analyses of the stocks traded on the PSE that would from now on assume the full impact of a trade on price instantaneous would be seriously incomplete.

Further Research

Several opportunities exist for further empirical research along the lines presented in the thesis. For instance, similarly to Chapter (2) where we investigated the information

relevant to the price durations process, we might also focus on the information given by the traded volume. Thinning the transaction process with respect to a cumulative traded volume equal to some threshold c_v, we could obtain the volume durations and investigate the market's liquidity (process) in a much the same way as we examined the volatility (process) in case of price durations.

Another interesting possibility how to extend the current empirical analysis would be to combine the VAR model of HASBROUCK (1991) as described in Chapter (3) with the ACD model of ENGLE AND RUSSELL (1998) employed in Chapter (2). By allowing the price impact of trades to depend upon the trading intensity, we could this way examine the information content of stock trades in relation to market activity.

Still more opportunities exist if we assume the analysis of high-frequency data in general. Provided the data were easily accessible, we might find it interesting to compare the stock markets in the Central and Eastern Europe with regard to their trading protocols and market designs. We might ask, for example, what would be the impact of such characteristics on the price effects of trades, market liquidity, and/or market efficiency on each of these markets.

Appendix A - Trade and Quote Database

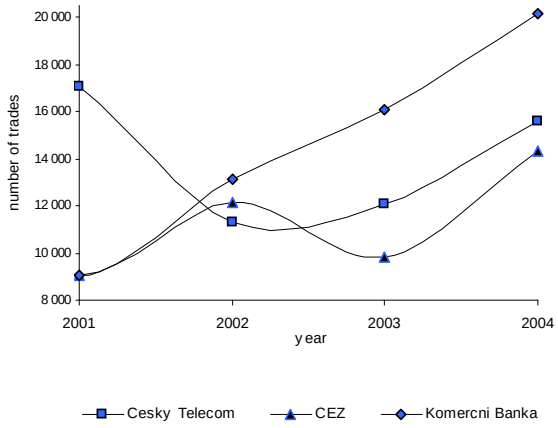

Figure A1: Time plot of the number of trades (left) and quotes (right) in the years 2001 to 2004 for the stocks analyzed in the study.

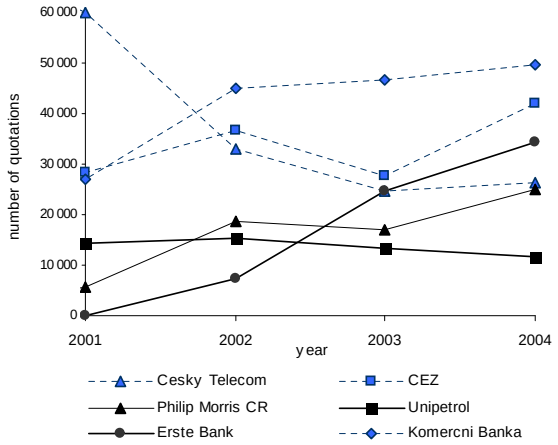

Figure A1: Time plot of the number of trades (left) and quotes (right) in the years 2001 to 2004 for the stocks analyzed in the study.

Table A1: Trade and Quote Database (Number of Trades)

Nr.	Company Name	2000	2001	2002	2003	2004	Total
	Non-Financial						
1	Ceske Radiokom.	14,170	8,750	3,925	2,688	2,623	32,156
2	**Cesky Telecom**	20,998	17,027	11,281	12,048	15,615	76,969
3	**CEZ**	12,088	9,021	12,130	9,801	14,297	57,337
4	IPS Skanska	2,314	*n.t.*	*n.t.*	*n.t.*	*n.t.*	2,314
5	**Philip Morris CR**	250	1,824	3,983	4,423	7,843	18,323
6	**Unipetrol**	4,840	3,517	3,161	3,033	3,949	19,401
7	Zentiva	*n.t.*	*n.t.*	*n.t.*	*n.t.*	2,915	2,915
	Financial						
8	Ceska Sporitelna	6,484	3,768	2,309	*n.t.*	*n.t.*	12,561
9	**Erste Bank**	*n.t.*	*n.t.*	1,655	5,419	8,415	15,489
10	**Komercni Banka**	12,555	9,020	12,132	16,099	20,133	70,939
11	IP Banka	1,014	*n.t.*	*n.t.*	*n.t.*	*n.t.*	1,014
12	RIF	1,242	*n.t.*	*n.t.*	*n.t.*	*n.t.*	1,242

The table shows the number of trade observations in a given year for all twelve companies traded on SPAD (PSE's order driven part of the main market) during the period from January 5 to November 12, 2004. The year when the company was not traded is denoted *n.t.*. Companies in bold have been analyzed in the study.

Table A2: Trade and Quote Database (Number of Quotes)

Nr.	Company Name	2000	2001	2002	2003	2004	Total
	Non-Financial						
1	Ceske Radiokom.	44,651	31,399	15,908	7,454	7,497	106,909
2	**Cesky Telecom**	74,500	60,135	33,055	24,558	26,216	218,464
3	**CEZ**	30,933	28,424	36,578	27,590	42,004	165,529
4	IPS Skanska	8,864	*n.t.*	*n.t.*	*n.t.*	*n.t.*	8,864
5	**Philip Morris CR**	779	5,632	18,633	17,044	25,146	67,234
6	**Unipetrol**	13,932	14,371	15,250	13,485	11,529	68,567
7	Zentiva	*n.t.*	*n.t.*	*n.t.*	*n.t.*	6,445	6,445
	Financial						
8	Ceska Sporitelna	18,079	12,066	5,914	*n.t.*	*n.t.*	36,059
9	**Erste Bank**	*n.t.*	*n.t.*	7,446	24,599	34,384	66,429
10	**Komercni Banka**	36,461	27,086	44,990	46,610	49,534	204,681
11	IP Banka	3,879	*n.t.*	*n.t.*	*n.t.*	*n.t.*	3,879
12	RIF	3,352	*n.t.*	*n.t.*	*n.t.*	*n.t.*	3,352

The table shows the number of trade observations in a given year for all twelve companies traded on SPAD (PSE's order driven part of the main market) during the period from January 5 to November 12, 2004. The year when the company was not traded is denoted *n.t.*. Companies in bold have been analyzed in the study..

Appendix B - Price Durations (Thresholds)

Price durations for CEZ, Cesky Telecom and Komercni Banka stocks for the period from January 5 to November 12, 2004. The number of bid-ask (B-A) quotes was obtained after filtering the adjusted dataset of the total of 39,608 quotes for CEZ, 24,845 quotes for Cesky Telecom, and 46,425 quotes for Komercni Banka at the respective thresholds given by c_p. $Q(10)$ denotes the Ljung-Box Q-statistic for the first ten autocorrelations of the price duration $x_{p,i}$. Source: Bid-Ask Quote Dataset (PSE); own calculations.

Table A1: Price Durations for CEZ

	$c_p = 0{,}25$	$c_p = 0{,}50$	$c_p = 1{,}00$	$c_p = 1{,}50$	$c_p = 2{,}00$
# of B-A Quotes	7,728	3,076	1,264	779	576
(% of Adj. Total)	19.51	7.77	3.19	1.97	1.45
Maximum	18,778	23,366	23,396	23,396	23,396
Mean	500.2	1,175.5	2,832.4	4,836.9	6,726.2
St. Deviation	1,240.5	2,392.2	4,457.0	6,469.2	7,537.3
Overdispersion	2.48	2.04	1.57	1.34	1.12
Ljung-Box Q (10)	594.4	359.1	223.6	147.8	85.3

Table A2: Price Durations for Cesky Telecom

	$c_p = 0{,}25$	$c_p = 0{,}50$	$c_p = 1{,}00$	$c_p = 1{,}50$	$c_p = 2{,}00$
# of B-A Quotes	15,072	7,001	2,427	1,261	856
(% of Adj. Total)	60.66	28.18	9.77	5.08	3.45
Maximum	16,934	21,439	23,043	23,250	23,388
Mean	295.4	520.6	1,377.0	2,768.5	4,018.8
St. Deviation	794.9	1,252.8	2,620.2	4,140.0	5,114.1
Overdispersion	2.69	2.41	1.90	1.50	1.27
Ljung-Box Q (10)	522.3	428.1	296.9	173.5	116.6

Table A3: Price Durations for Komercni Banka

	$c_p = 2{,}5$	$c_p = 5{,}0$	$c_p = 10{,}0$	$c_p = 15{,}0$	$c_p = 20{,}0$
# of B-A Quotes	12,308	4,901	2,106	1,285	879
(% of Adj. Total)	26.51	10.56	4.54	2.77	0.19
Maximum	15,490	16,648	21,855	22,238	23,398
Mean	303.8	671.9	1,519.5	2,513.5	5,265.7
St. Deviation	782.3	1,429.3	2,663.2	3,703.2	7,537.3
Overdispersion	2.57	2.13	1.75	1.47	1.38
Ljung-Box Q (10)	263.4	171.4	131.6	71.4	58.2

Appendix C - Trade and Quote Data Merger

The program allows to merge trade and quote transaction data files into a single entity in which the most recent bid and ask quotation (i. e., quotes as of $t_i < t$ for every i) are set next to the price of the trade that occurs at time t. The names of the two input input files as well as the name of the output file need to be provided on the command line as follows: (trade trans. file) (quote trans. file) (output file)

The program was created (and is compatible) with Java 2 SE, build 1.3.1-b24.

```
/**
 * FileMrg.java
 * @author Vit Bubak
 * @version 1.1, May 2005
 */

import java.io.*;
import java.lang.Math.*;
import java.text.*;
import java.util.*;

public class FileMrg2 {

     static StringTokenizer tokenT = null;
     static StringTokenizer tokenQ = null;
     static String lineTrade = null;
     static String lineQuote = null;

     // trade token variables
     static String dateT = null;
     static String timeT = null;
     static String isinT = null;
```

```
static String nameT = null;
static String numbT = null;
static String pricT = null;
static String voluT = null;

// quote token variables
static String dateQ = null;
static String timeQ = null;
static String isinQ = null;
static String nameQ = null;
static String bBidQ = null;
static String bAskQ = null;
static String lowBQ = null;
static String uppBQ = null;

// common variables:  (time of event)
static int tmExT = 0;
static int tmExQ = 0;

// common variables:  (time) and (quote)
static int xT = 0;
static int xQ = 0;

// file variables
static String flTrades = null;
static String flQuotes = null;
static String flOutput = null;

// array of "final" strings
static String[] finalStr = new String[30000];

// static variable for size
static int s = 0;

// the 'main' function
public static void main(String[] args) throws IOException {
// both in- and output files were specified on the command line
BufferedReader flTrades = new BufferedReader
                          (new FileReader(args[0]));
BufferedReader flQuotes = new BufferedReader
```

```
                          (new FileReader(args[1]));
PrintWriter flOutput    = new PrintWriter
                          (new FileWriter(args[2]));
readIt (flTrades, flQuotes);
writeIt(flOutput);
}
// reads the input from the file
public static void readIt(BufferedReader inTrades,
                          BufferedReader inQuotes)
throws IOException {
int i = 1;              // count variable
int tLast = 34200;      // starting value for 'last' time
int tNew = 34200;       // starting value for 'new' time

String helpStr = null;  // help string for storage
int thisTrade = 34200;  // help var for this trade's time
int prevTrade = 0;      // help var for previous trade's time
int thisQuote = 34200;  // help var for this quote's time
int prevQuote = 0;      // help var for previous quote's time

// previous values
String prevDtL = " "; String dtL = " ";
String prevTmL = " "; String tmL = " ";
String prevBBL = " "; String bBL = " ";
String prevBAL = " "; String bAL = " ";
String prevLBQ = " "; String lBQ = " ";
String prevUBQ = " "; String uBQ = " ";

lineTrade = inTrades.readLine();

while (lineTrade != null) {
tokenT = new StringTokenizer(lineTrade);
dateT = tokenT.nextToken();
timeT = tokenT.nextToken();
tmExT = Integer.parseInt(tokenT.nextToken());
isinT = tokenT.nextToken();
nameT = tokenT.nextToken();
numbT = tokenT.nextToken();
// reads the price of the trade
```

```
pricT = tokenT.nextToken();
voluT = tokenT.nextToken();
// saves the last time
prevTrade = thisTrade;
thisTrade = tmExT;

// takes care of changing date
if (thisTrade < prevTrade) {

while (thisQuote > thisTrade) {
lineQuote = inQuotes.readLine();
tokenQ = new StringTokenizer(lineQuote);
dateQ = tokenQ.nextToken();
timeQ = tokenQ.nextToken();
tmExQ = Integer.parseInt(tokenQ.nextToken());
isinQ = tokenQ.nextToken();
nameQ = tokenQ.nextToken();
bBidQ = tokenQ.nextToken();
bAskQ = tokenQ.nextToken();
lowBQ = tokenQ.nextToken();
uppBQ = tokenQ.nextToken();

// reassigns the time
prevQuote = thisQuote;
thisQuote = tmExQ;
// reassigns other vars
prevDtL = dtL; dtL = dateQ;
prevTmL = tmL; tmL = timeQ;
prevBBL = bBL; bBL = bBidQ;
}
}

while (thisQuote < thisTrade) {

lineQuote = inQuotes.readLine();
tokenQ = new StringTokenizer(lineQuote);
dateQ = tokenQ.nextToken();
timeQ = tokenQ.nextToken();
tmExQ = Integer.parseInt(tokenQ.nextToken());
isinQ = tokenQ.nextToken();
```

```
        nameQ = tokenQ.nextToken();
        bBidQ = tokenQ.nextToken();
        bAskQ = tokenQ.nextToken();
        lowBQ = tokenQ.nextToken();
        uppBQ = tokenQ.nextToken();

        // reassigns the time variable
        prevQuote = thisQuote;
        thisQuote = tmExQ;
        // reassigns other variables
        prevDtL = dtL; dtL = dateQ;
        prevTmL = tmL; tmL = timeQ;
        prevBBL = bBL; bBL = bBidQ;
        prevBAL = bAL; bAL = bAskQ;
        }

        helpStr = ( dateT   + " " + timeT   + " " + tmExT   + " " +
                    tmExT   + " " + prevDtL + " " + prevTmL + " " +
                    nameT   + " " + voluT   + " " + pricT   + " " +
                    prevBBL + " " + prevBAL );

        finalStr[i] = helpStr;
        i++;
        // reads in a new line of trade data
        lineTrade = inTrades.readLine();
        }

        // assigns the value of local var 'i' to static 's'
        s = i;

        // closes the trades and quotes inputs
        inTrades.close();
        inQuotes.close();
    }

    public static void writeIt (PrintWriter out) throws IOException {
    // inserts the finalStr's content into the file
    for (int k = 1; k < s; k++) {out.println(k + " " + finalStr[k]);}
    out.close();
    }
```

Appendix D - VAR Estimation Code

The model is used to investigate the impact of (unexpected) trade volume on subsequent quote revision on a market-maker driven financial market. Set in transaction time, the model is written as $Ax_t = c + B(L)\, x_t + \varepsilon_t$, where c is a vector of constants, $B(L)$ is a matrix polynomial in the lag operator and ε is an orthogonal vector of white noise terms.

The estimation code may be easily adjusted to estimate bivariate VAR models as well as the models of higher order. The program is written in Ox programming language developed by J. A. Doornik.

Additional features of the estimation are:
a) sum of the coefficients in VAR(L) subequations,
b) heteroskedasticity-consistent standard error,
c) heteroskedasticity-consistent LM test for autocorrelation in residuals,
d) ARCH LM test for conditional heteroskedasticity,
e) impulse-response function (IRF) with standard error
f) random-walk decomposition

```
/**
 * Quadratic Struct.  VAR(L) by OLS
 * @author Vit Bubak and Filip Zikes
 * @version 1.1, June 2005
 */

#include <oxstd.h>
#include <oxprob.h>
#include <oxdraw.h>
#import <database>

const decl L  = 12; // number of lags in VAR equation
const decl AT =  4; // number of lags in ARCH (x) test
```

```
main() {
decl time;  time  = timer();
decl dbase; dbase = new Database();

// Loading the set of data
// cez:42909, cte:28784, ers:33860, kob:49962, pmc:24956, uni:11937
dbase.Load("C:/...");
dbase.Info();

decl data, N, L;
data=dbase->GetAll();

N=42908; // number of observations
L=12;    // number of lags in VAR equation
decl v, q, v1, v2;
decl lv, lq, lv1, lv2, c;
decl s, ls;
v = data[][0];
q = data[][1];

// Rescaling linear and sqr.  volumes, resp.
v1 = 0.000001*data[][2];
v2 = 0.000001*data[][3];

// instrum.  variable (for additional purposes)
// s = data[][4];

// Definition of lagged variables
//(the number of lags must be a constant)
lv  = lag0(v,<1:12>);
lq  = lag0(q,<1:12>);
lv1 = lag0(v1,<1:12>);
lv2 = lag0(v2,<1:12>);

// Definition the vector of ones
 c = ones(N,1);

// Direct.  volume, quote, volume, and squared volume eq.'s/matrices
decl yvol, yq, yvol1, yvol2;
decl xmatvol, xmatq, xmatvol1, xmatvol2;
```

```
decl resv, resq, resv1, resv2;

// b(*)[n]:  matrices of OLS coefficients
decl bv, bq, bv1, bv2;
decl sv, sq, sv1, sv2;

// Auxiliary matrices
decl xtxv, xtxq, xtxv1, xtxv2;
decl sigma2v, sigma2q, sigma2v1, sigma2v2;

// Matrices of OLS and White standard errors
decl stdOLSv, stdOLSq, stdOLSv1, stdOLSv2;
decl stdWHCv, stdWHCq, stdWHCv1, stdWHCv2;

// t-statistics, p-values
decl tstatOLSv, tstatOLSq, tstatOLSv1, tstatOLSv2;
decl pvalOLSv, pvalOLSq, pvalOLSv1, pvalOLSv2;

// Robust t-stat and p-values
decl tstatWHCv, tstatWHCq, tstatWHCv1, tstatWHCv2;
decl pvalWHCv,  pvalWHCq,  pvalWHCv1,  pvalWHCv2;
decl R2vol, R2quote, R2vol1, R2vol2;

// Defining the dependent matrices for (direct.)  vol, q and vol
yvol  = v [L:][];
yq    = q [L:][];
yvol1 = v1[L:][];

// Defining the corresp.  (direct.)  vol, q, and (sqr.)  vol matrices
xmatvol  = c[L:][]~lv[L:][]~lq[L:][]~lv1[L:][]~lv2[L:][];
xmatq    = c[L:][]~lv[L:][]~lq[L:][]~lv1[L:][]~lv2[L:][];
xmatvol1 = c[L:][]~lv[L:][]~lq[L:][]~lv1[L:][]~lv2[L:][];

// OLS estimation of (sig.)  vol equation
olsc(yvol,xmatvol,&bv,&xtxv);
resv=yvol-xmatvol*bv;
sigma2v=(1/(N-2*L-1))*resv'resv;
stdOLSv=diagonal((xtxv.*sigma2v).^(1/2))';
tstatOLSv=bv./stdOLSv;
pvalOLSv=2*tailn(fabs(tstatOLSv));
R2vol=varc(xmatvol*bv)/varc(yvol);
```

```
//OLS estimation of q equation
olsc(yq,xmatq,&bq,&xtxq);
resq=yq-xmatq*bq;
sigma2q=(1/(N-2*L-2))*resq'resq;
stdOLSq=diagonal((xtxq.*sigma2q).^(1/2))';
tstatOLSq=bq./stdOLSq;
pvalOLSq=2*tailn(fabs(tstatOLSq));
R2quote=varc(xmatq*bq)/varc(yq);

//OLS estimation of vol equation
olsc(yvol1,xmatvol1,&bv1,&xtxv1);
resv1=yvol1-xmatvol1*bv1;
sigma2v1=(1/(N-2*L-2))*resv1'resv1;
stdOLSv1=diagonal((xtxv1.*sigma2v1).^(1/2))';
tstatOLSv1=bv1./stdOLSv1;
pvalOLSv1=2*tailn(fabs(tstatOLSv1));
R2vol1=varc(xmatvol1*bv1)/varc(yvol1);

/* Robust standard errors, t-stat, p-val
Note:  loop is needed to estimate the matrix x'diag(e^2)x,
since diag(e^2) has over billion of elements and requires
an enormous amount of memory, which is usually not available.
*/

decl xexv, xexq, xexv1;
decl i, sumv;
sumv = zeros(4*L+1,4*L+1);

for(i=0;i<N-L;++i)
{
  xexv = ((resv[i])^2).*xmatvol[i][]'xmatvol[i][]+sumv;
  sumv = xexv;
}

stdWHCv   = diagonal((xtxv*xexv*xtxv).^(1/2))';
tstatWHCv = bv./stdWHCv;
pvalWHCv  = 2*tailn(fabs(tstatWHCv));

decl j, sumq;
sumq      = zeros(4*L+1,4*L+1);
```

```
for(j=0;j<N-L;++j)
{
  xexq = ((resq[j])^2).*xmatq[j][]'xmatq[j][]+sumq;
  sumq = xexq;
}

stdWHCq   = diagonal((xtxq*xexq*xtxq).^(1/2))';
tstatWHCq = bq./stdWHCq;
pvalWHCq  = 2*tailn(fabs(tstatWHCq));

decl el, sumv1;
sumv1     = zeros(4*L+1,4*L+1);

for(el=0;el<N-L;++el)
{
  xexv1 = ((resv1[el])^2).*xmatvol1[el][]'xmatvol1[el][]+sumv1;
  sumv1 = xexv1;
}

stdWHCv1   = diagonal((xtxv1*xexv1*xtxv1).^(1/2))';
tstatWHCv1 = bv1./stdWHCv1;
pvalWHCv1  = 2*tailn(fabs(tstatWHCv1));

decl outv, outq, outv1;
outv  = bv~stdOLSv~stdWHCv~tstatOLSv~pvalOLSv~pvalWHCv;
outq  = bq~stdOLSq~stdWHCq~tstatOLSq~pvalOLSq~pvalWHCq;
outv1 = bv1~stdOLSv1~stdWHCv1~tstatOLSv1~pvalOLSv1~pvalWHCv1;

// The sum of the coefficients in VAR sub-equations
decl cDvV, cQtV, cLvV, cSqV;
decl cDvQ, cQtQ, cLvQ, cSqQ;
decl cDvL, cQtL, cLvL, cSqL;


// Initializing the sum variables
cDvV = 0; cQtV = 0; cLvV = 0; cSqV = 0;
cDvQ = 0; cQtQ = 0; cLvQ = 0; cSqQ = 0;
cDvL = 0; cQtL = 0; cLvL = 0; cSqL = 0;

// Coefficients at lagged (direct.)  vol(s)
decl sm1, xvV, xvQ, xvL;
```

```
for(sm1=1;sm1<=L;++sm1)
{
    xvV=bv[sm1]+cDvV;
    xvQ=bq[sm1]+cDvQ;
    xvL=bv1[sm1]+cDvL;
    cDvV=xvV;
    cDvQ=xvQ;
    cDvL=xvL;
}

// Coefficients at lagged quote(s)
decl sm2, xqV, xqQ, xqL;
for(sm2=(L+1);sm2<=(L*2);++sm2)
{
    xqV=bv[sm2]+cQtV;
    xqQ=bq[sm2]+cQtQ;
    xqL=bv1[sm2]+cQtL;
    cQtV=xqV;
    cQtQ=xqQ;
    cQtL=xqL;
}

// Coefficients at lagged (lin.)  vol(s)
decl sm3, xvLvV, xvLvQ, xvLvL;
for(sm3=((L*2)+1);sm3<=(L*3);++sm3)
{
    xvLvV=bv[sm3]+cLvV;
    xvLvQ=bq[sm3]+cLvQ;
    xvLvL=bv1[sm3]+cLvL;
    cLvV=xvLvV;
    cLvQ=xvLvQ;
    cLvL=xvLvL;
}

// Coefficients at lagged (sqr.)  vol(s)
decl sm4, xvSvV, xvSvQ, xvSvL;
for(sm4=((L*3)+1);sm4<=(L*4);++sm4)
{
    xvSvV=bv[sm4]+cSqV;
    xvSvQ=bq[sm4]+cSqQ;
```

```
    xvSvL=bv1[sm4]+cSqL;
    cSqV=xvSvV;
    cSqQ=xvSvQ;
    cSqL=xvSvL;
}

// ARCH(x) LM test
decl resvy,resvx,resqy,resqx,resv1y,resv1x;
resvy=resv.^2;
resqy=resq.^2;
resv1y=resv1.^2;
resvx =lag0(resvy,<1:AT>);
resqx =lag0(resqy,<1:AT>);
resv1x=lag0(resv1y,<1:AT>);

decl r2vy,r2vx,r2qy,r2qx,r2v1y,r2v1x;
r2vy=resvy[AT:][];
r2vx=resvx[AT:][];
r2qy=resqy[AT:][];
r2qx=resqx[AT:][];
r2v1y=resv1y[AT:][];
r2v1x=resv1x[AT:][];

decl dv, dq, dv1;
olsc(r2vy,r2vx,&dv);
olsc(r2qy,r2qx,&dq);
olsc(r2v1y,r2v1x,&dv1);

decl R2v, R2q, R2v1;
decl testv, testq, testv1;
decl pv, pq, pv1;
R2v =varc(r2vx*dv)/varc(r2vy);
R2q =varc(r2qx*dq)/varc(r2qy);
R2v1=varc(r2v1x*dv1)/varc(r2v1y);
testv =(N-L-AT)*R2v;
testq =(N-L-AT)*R2q;
testv1=(N-L-AT)*R2v1;
pv =tailchi(testv, AT);
pq =tailchi(testq, AT);
pv1=tailchi(testv1,AT);
```

```
// AR(4) robust LM test
decl arL=4;
decl ar1v,ar2v,ar3v,ar4v;
decl ar1q,ar2q,ar3q,ar4q;
decl ar1v1,ar2v1,ar3v1,ar4v1;
decl lresv,lresq,lresv1;
lresv =lag0(resv,<1:4>);
lresq =lag0(resq,<1:4>);
lresv1=lag0(resv1,<1:4>);

// First-pass regression
olsc(lresv[arL:][0],xmatvol[arL:][],&ar1v);
olsc(lresv[arL:][1],xmatvol[arL:][],&ar2v);
olsc(lresv[arL:][2],xmatvol[arL:][],&ar3v);
olsc(lresv[arL:][3],xmatvol[arL:][],&ar4v);
olsc(lresq[arL:][0],xmatq[arL:][],&ar1q);
olsc(lresq[arL:][1],xmatq[arL:][],&ar2q);
olsc(lresq[arL:][2],xmatq[arL:][],&ar3q);
olsc(lresq[arL:][3],xmatq[arL:][],&ar4q);
olsc(lresv1[arL:][0],xmatvol1[arL:][],&ar1v1);
olsc(lresv1[arL:][1],xmatvol1[arL:][],&ar2v1);
olsc(lresv1[arL:][2],xmatvol1[arL:][],&ar3v1);
olsc(lresv1[arL:][3],xmatvol1[arL:][],&ar4v1);

decl matresv, matresq, matresv1;
decl erv, erq, erv1;

matresv  = (lresv[arL:][0]-xmatvol[arL:][]*ar1v)~
                 (lresv[arL:][1]-xmatvol[arL:][]*ar2v)~
                 (lresv[arL:][2]-xmatvol[arL:][]*ar3v)~
                 (lresv[arL:][3]-xmatvol[arL:][]*ar4v);

matresq  = (lresq[arL:][0]-xmatq[arL:][]*ar1q)~
                 (lresq[arL:][1]-xmatq[arL:][]*ar3q)~
                 (lresq[arL:][2]-xmatq[arL:][]*ar3q)~
                 (lresq[arL:][3]-xmatq[arL:][]*ar4q);

matresv1 = (lresv1[arL:][0]-xmatvol1[arL:][]*ar1v1)~
                 (lresv1[arL:][1]-xmatvol1[arL:][]*ar3v1)~
                 (lresv1[arL:][2]-xmatvol1[arL:][]*ar3v1)~
```

```
                (lresv1[arL:][3]-xmatvol1[arL:][]*ar4v1);

erv =resv[arL:][].*matresv;
erq =resq[arL:][].*matresq;
erv1=resv1[arL:][].*matresv1;

// Second-pass regression
decl one, gv, gq, gv1;
decl artestv, artestq, artestv1;
decl arpvv, arpvq, arpvv1;
one=ones(N-L,1);

olsc(one[arL:][],erv,&gv);
olsc(one[arL:][],erq,&gq);
olsc(one[arL:][],erv1,&gv1);
artestv =(N-L-arL)-sumsqrc(one[arL:][]-erv*gv);
artestq =(N-L-arL)-sumsqrc(one[arL:][]-erq*gq);
artestv1=(N-L-arL)-sumsqrc(one[arL:][]-erv1*gv1);
arpvv =tailchi(artestv,arL);
arpvq =tailchi(artestq,arL);
arpvv1=tailchi(artestv1,arL);

println(" ");
println("Estimation Results");
println("----------------------");

// (Signed) volume equation estimate
println("Signed volume equation");
print(" coeff"," OLSstdErr"," WHCstdErr",
         " OLSt-stat"," OLSp-val"," WHCp-val");
print(outv); println(" ");

println("R2",R2vol); println(" ");
println("direct vol cf's:  ",cDvV);
println("quote cf's:  ",cQtV);
println("linear vol cf's:  ",cLvV);
println("(sqrt) vol cf's:  ",cSqV);
println(" ");

// Quote (price) equation estimate
println("Quote equation");
```

```
print(" coeff"," OLSstdErr"," WHCstdErr",
         " OLSt-stat"," OLSp-val"," WHCp-val");
print(outq); println(" ");

println("R2", R2quote); println(" ");
println("direct vol cf's:  ",cDvQ);
println("quote cf's:  ",cQtQ);
println("linear vol cf's:  ",cLvQ);
println("(sqrt) vol cf's:  ",cSqQ);
println(" ");

// (Linear) volume equation estimates
println("(Linear) volume equation");
print(" coeff"," OLSstdErr"," WHCstdErr",
         " OLSt-stat"," OLSp-val"," WHCp-val");
print(outv1); println(" ");

println("R2", R2vol1); println(" ");
println("direct vol cf's:  ",cDvL);
println("quote cf's:  ",cQtL);
println("linear vol cf's:  ",cLvL);
println("(sqrt) vol cf's:  ",cSqL);
println(" ");

// Results of ARCH(4) LM test
println(" ");
println("ARCH(4) LM test");
println(" test stat.  "," p-val.");
print((testv~pv)|(testq~pq)|(testv1~pv1));

// Results of AR(4) HET-Robust LM test
println(" ");
println("AR(4) heteroskedasticity-robust LM test");
println(" test stat.  "," p-val.");
print((artestv~arpvv)|(artestq~arpvq)|(artestv1~arpvv1));

// Matrices of coefficients
// (the vector of endogenous variables is given by y=(q, vol))
decl A,B1,B2,B3,B4,B5,B6,B7,B8,B9,B10,B11,B12,C;
decl phi0, phi1, phi2, phi3, phi4, phi5, phi6;
decl phi7, phi8, phi9, phi10, phi11, phi12;
```

```
// Structural parameters
A=(1~-bq[1])|(0~1);

B1=(bq[L+2]~bq[2])|(bv[L+1]~bv[1]);
B2=(bq[L+3]~bq[3])|(bv[L+2]~bv[2]);
B3=(bq[L+4]~bq[4])|(bv[L+3]~bv[3]);
B4=(bq[L+5]~bq[5])|(bv[L+4]~bv[4]);
B5=(bq[L+6]~bq[6])|(bv[L+5]~bv[5]);
B6=(bq[L+7]~bq[7])|(bv[L+6]~bv[6]);
B7=(bq[L+8]~bq[8])|(bv[L+7]~bv[7]);
B8=(bq[L+9]~bq[9])|(bv[L+8]~bv[8]);
B9=(bq[L+10]~bq[10])|(bv[L+9]~bv[9]);
B10=(bq[L+11]~bq[11])|(bv[L+10]~bv[10]);
B11=(bq[L+12]~bq[12])|(bv[L+11]~bv[11]);
B12=(bq[L+13]~bq[13])|(bv[L+12]~bv[12]);

C=bq[0]|bv[0];

// Reduced-form parameters
phi0 =invert(A)*C;
phi1 =invert(A)*B1;
phi2 =invert(A)*B2;
phi3 =invert(A)*B3;
phi4 =invert(A)*B4;
phi5 =invert(A)*B5;
phi6 =invert(A)*B6;
phi7 =invert(A)*B7;
phi8 =invert(A)*B8;
phi9 =invert(A)*B9;
phi10=invert(A)*B10;
phi11=invert(A)*B11;
phi12=invert(A)*B12;

// Matrices of MA parameters
decl mu;
decl psi1, psi2, psi3, psi4, psi5;
decl psi6, psi7, psi8, psi9, psi10;

// (20 steps in impulse response function)
```

```
decl psi11, psi12, psi13, psi14, psi15,
       psi16, psi17, psi18, psi19, psi20;
mu = invert(unit(2,2)-phi1-phi2-phi3-phi4-phi5-
       phi6-phi7-phi8-phi9-phi10-phi11-phi12)*phi0;

// Recursive calculation of cum.  response
psi1 =phi1;
psi2 =phi1*psi1+phi2;
psi3 =phi1*psi2+phi2*psi1+phi3;
psi4 =phi1*psi3+phi2*psi2+phi3*psi1+phi4;
psi5 =phi1*psi4+phi2*psi3+phi3*psi2+phi4*psi1+phi5;
psi6 =phi1*psi5+phi2*psi4+phi3*psi3+phi4*psi2+phi5*
        psi1+phi6;
psi7 =phi1*psi6+phi2*psi5+phi3*psi4+phi4*psi3+phi5*
        psi2+phi6*psi1+phi7;
psi8 =phi1*psi7+phi2*psi6+phi3*psi5+phi4*psi4+phi5*
        psi3+phi6*psi2+phi7*psi1+phi8;
psi9 =phi1*psi8+phi2*psi7+phi3*psi6+phi4*psi5+phi5*
        psi4+phi6*psi3+phi7*psi2+phi8*psi1+phi9;
psi10=phi1*psi9+ phi2*psi8+ phi3*psi7+ phi4*psi6+
        phi5*psi5+ phi6*psi4+ phi7*psi3+ phi8*psi2+
        phi9*psi1+ phi10;
psi11=phi1*psi10+phi2*psi9+ phi3*psi8+ phi4*psi7+
        phi5*psi6+ phi6*psi5+ phi7*psi4+ phi8*psi3+
        phi9*psi2+ phi10*psi1+ phi11;
psi12=phi1*psi11+phi2*psi10+phi3*psi9+ phi4*psi8+
        phi5*psi7+phi6*psi6+ phi7*psi5+ phi8*psi4+
        phi9*psi3+ phi10*psi2+ phi11*psi1+ phi12;
psi13=phi1*psi12+phi2*psi11+phi3*psi10+phi4*psi9+
        phi5*psi8+phi6*psi7+ phi7*psi6+ phi8*psi5+
        phi9*psi4+ phi10*psi3+ phi11*psi2+ phi12*psi1;
psi14=phi1*psi13+phi2*psi12+phi3*psi11+phi4*psi10+
        phi5*psi9+phi6*psi8+ phi7*psi7+ phi8*psi6+
        phi9*psi5+ phi10*psi4+ phi11*psi3+ phi12*psi2;
psi15=phi1*psi14+phi2*psi13+phi3*psi12+phi4*psi11+
        phi5*psi10+phi6*psi9+ phi7*psi8+ phi8*psi7+
        phi9*psi6+ phi10*psi5+ phi11*psi4+ phi12*psi3;
psi16=phi1*psi15+phi2*psi14+phi3*psi13+phi4*psi12+
        phi5*psi11+phi6*psi10+phi7*psi9+ phi8*psi8+
```

```
         phi9*psi7+ phi10*psi6+ phi11*psi5+ phi12*psi4;
psi17=phi1*psi16+phi2*psi15+phi3*psi14+phi4*psi13+
         phi5*psi12+phi6*psi11+phi7*psi10+phi8*psi9+
         phi9*psi8+ phi10*psi7+ phi11*psi6+ phi12*psi5;
psi18=phi1*psi17+phi2*psi16+phi3*psi15+phi4*psi14+
         phi5*psi13+phi6*psi12+phi7*psi11+phi8*psi10+
         phi9*psi9+ phi10*psi8+ phi11*psi7+ phi12*psi6;
psi19=phi1*psi18+phi2*psi17+phi3*psi16+phi4*psi15+
         phi5*psi14+phi6*psi13+phi7*psi12+phi8*psi11+
         phi9*psi10+phi10*psi9+ phi11*psi8+ phi12*psi7;
psi20=phi1*psi19+phi2*psi18+phi3*psi17+phi4*psi16+
         phi5*psi15+phi6*psi14+phi7*psi13+phi8*psi12+
         phi9*psi11+phi10*psi10+phi11*psi9+ phi12*psi8;

// Vector of structural impulses
decl eps, impulse;
impulse=1;
eps=0|impulse;

// Vector of reduced form impulses
decl e;
e=invert(A)*eps;

decl response, psi0, k;
response=zeros(21,2);
psi0=unit(2,2);

response[0][]=(psi0*e)';
response[1][]=(psi1*e)';
response[2][]=(psi2*e)';
response[3][]=(psi3*e)';
response[4][]=(psi4*e)';
response[5][]=(psi5*e)';
response[6][]=(psi6*e)';
response[7][]=(psi7*e)';
response[8][]=(psi8*e)';
response[9][]=(psi9*e)';
response[10][]=(psi10*e)';
response[11][]=(psi11*e)';
response[12][]=(psi12*e)';
```

```
response[13][]=(psi13*e)';
response[14][]=(psi14*e)';
response[15][]=(psi15*e)';
response[16][]=(psi16*e)';
response[17][]=(psi17*e)';
response[18][]=(psi18*e)';
response[19][]=(psi19*e)';
response[20][]=(psi20*e)';

// Impulse-response (IR) function var's
decl qs, vols, grid;
// Cumulative response (CR) function var's
decl qqs;

qs =response[][0];
qqs=cumulate(qs);

println(" ");
print("Impulse-response function");
print(qs);
println(" ");
println("Cumulative response function");
println(qqs);
println(" ");

//Graph of IR and CR functions (only seen in GiveWin)
grid=<0:20>;
DrawXMatrix(0, qs', {"Response of quotes
                  to a shock in volume"}, grid, "period");
DrawXMatrix(1, qqs', {"Cumulative response of quotes
                  to a shock in volume"},grid,"period");
ShowDrawWindow();


print("Time needed for the calculations:  ",timespan(time)," s");
}
```

Appendix E - Impulse Response Function

The calculation of impulse response function as well as the corresponding cumulative response function (CRF) follows from the definition of the VAR model as defined by equations (3.3) and (3.5). Given the symmetry condition (3.4) holds, the resulting cumulative CQR through the m-th step can be obtained as follows,

$$\alpha_m\left(\eta_{2,0}\right) = \sum_{t=0}^{m} E\left[r_t|\ \eta_{2,0}\right] = 0, \qquad \text{for } s \neq t.$$

On the following lines we present a standalone derivation of CQR for a practical calculation.

Assuming only the simplest bivaritate case, we first expand the equations (3.3) and (3.5) to obtain a full form through m-th steps:

$$r_t = const_r + a_1 r_{t-1} + a_2 r_{t-2} + \ldots + a_m r_{t-m} + (b_0 x_t) + b_1 x_{t-1} + \ldots + b_m x_{t-m} + \eta_{1,t}, \quad (8)$$

$$x_t = const_x + c_1 r_{t-1} + c_2 r_{t-2} + \ldots + a_m r_{t-m} + \_\_\_\_\_ + b_1 x_{t-1} + \ldots + b_m x_{t-m} + \eta_{1,t}. \quad (9)$$

In equation (9), we put the term $b_0 x_t$ in parentheses as in the real estimation the quote does not depend only on the lagged values of itself r_t and x_t both also on the contemporaneous value of trade x_t.

Rewriting the equations (8) and (9) using lag operators and rearranging the sides, we obtain

$$r_t\left[1 - a_1 L - a_2 L^2 - \ldots - a_m L^m\right] + x_t\left[-b_0 - b_1 L - \ldots - b_m L^m\right] = const_r + \eta_{1,t} \quad (10)$$

$$r_t\left[- c_1 L - c_2 L^2 - \ldots - c_m L^m\right] + x_t\left[1 - d_1 L - \ldots - b_m L^m\right] = const_x + \eta_{2,t} \quad (11)$$

In matrix format, the equations (10) and (11) can be written as

$$\underset{(A)}{\begin{bmatrix} 1 & -b_0 \\ 0 & 1 \end{bmatrix}} z_t - \underset{(B_1)}{\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix}} L z_t - \ldots - \underset{(B_m)}{\begin{bmatrix} a_m & b_m \\ c_m & d_m \end{bmatrix}} L^m z_t =$$

$$\underset{(C)}{[const_r, const_x]'} + \underset{(D)}{\left[\eta_{1,t}, \eta_{2,t}\right]'},$$

where L is a lag operator and $z_t = [r_t, x_t]'$.

Dividing both sides by (A) matrix, we can then re-write the preceding equation as

$$z_t = A^{-1}B_1 + A^{-1}B_2 + \ldots + A^{-1}B_m + A^{-1}C + A^{-1}D \tag{12}$$

Letting $\varphi_1 = A^{-1}B_1, \varphi_2 = A^{-1}B_2$, et cetera, we can immediately derive the price impact at time t_1 through t_m subsequent to a trade at time t_0. At time $t = 0$, a trade order $\eta_{2,0}$ arrives, making the current trade $x_0 = \eta_{2,0}$.

By letting $\eta_{1,t} = 0$ for all t, and $x_0 = 1$, we effectively assume that $[A^{-1}D] = A^{-1}\left[\eta_{1,t}, \eta_{2,t}\right]' = A^{-1}[0,1]'$. Thus, substituting $\varsigma = A^{-1}[0,1]'$ and following (12), the first three impulses (times t_1, t_2, and t_3) are

$$\begin{aligned}
[t_0] &= \varsigma, \\
[t_1] &= \varphi_1\varsigma, \\
[t_2] &= \varphi_1[\varphi_1\varsigma] + \varphi_2\varsigma = \varphi_1^2\varsigma + \varphi_2\varsigma = (\varphi_1^2 + \varphi_2)\varsigma, \\
[t_3] &= \varphi_1[\varphi_1[\varphi_1\varsigma] + \varphi_2\varsigma] + \varphi_2\varphi_1\varsigma + \varphi_3\varsigma = (\varphi_1^3 + \varphi_1\varphi_2 + \varphi_2\varphi_1 + \varphi_3)\varsigma.
\end{aligned}$$

The corresponding cumulative response function (CQR) can be obtained by cumulating the impulse responses over successive (transaction) time periods. For example, the CQR for lags 1 to 3 would be

$$CRQ\ [t \in (0;3)] = [\varphi_1 + (\varphi_1^2 + \varphi_2) + (\varphi_1^3 + \varphi_1\varphi_2 + \varphi_2\varphi_1 + \varphi_3)]\varsigma.$$

References

[1] Admati, A. R. and P. Pfleiderer (1988): "A Theory of Intraday Patterns: Volume and Price Variability", *The Review of Financial Studies*, **1**, 3-40.

[2] Bauwens, L. and P. Giot (1998): "Asymmetric ACD model: Introducing Price Information in ACD Models with a Two-State Transition Model", CORE Discussion Paper, **9844**.

[3] Bauwens, L. and P. Giot (2000): "The Logarithmic ACD Model: An Application to the Bid-Ask Quote Process of Three NYSE Stocks", *Annales D'Économie et de Statistique*, no. **60,** 117-149.

[4] _____ , Giot, P., Grammig, J. and D. Veredas (2000): "A Comparison of Financial Duration Models via Density Forecasts", *Working Paper*, Université Catholique de Louvain and University of Frankfurt.

[5] Biais, B., Hillion, P. and C. Spatt (1995): "An Empirical Analysis of the Limit order Book and the Order Flow in the Paris Bourse", *Journal of Finance*, 50, 1655-1689.

[6] Bollerslev, T. (1986): "Generalized Autoregressive Conditional Heteroskedasticity", *Journal of Econometrics*, **31**, 307-327.

[7] Bubak, V. and F. Zikes (2004): "Seasonality and the Non-trading Effect on Central European Stock Markets", NUS Working Paper Series.

[8] Campbell, J. Y., Lo, A. W. and A. C. MacKinlay (1997): *The Econometrics of Financial Markets*, Princeton University Press: New Jersey.

[9] Cohen, K.J., Maier, S. F., Schwartz, R. A. and D. K. Whitcomb (1981): "Transaction Costs, Order Placement Strategy, and the Existence of the Bid-Ask Spread", *Journal of Political Economy*, **89**, 287-305.

[10] Cox, D. R. (1972a): "Regression Models and Life Tables", *Journal of the Royal Statistical Society* (Series B), **34**, 187-220.

[11] _____ , (1972b): "The Statistical Analysis of Dependencies in Point Processes", in *Symposium on Point Processes*, ed. by P. A. W. Lewis, New York: John Wiley, 55-66.

[12] Coughenour, J. and K. Shastri (1999): "Symposium on Market Microstructure: A Review of Empirical Research", *Financial Review*, **34**, 1-20.

[13] Dacorogna, M. M. et Al. (2001): *An Introduction to High-Frequency Finance*, London: Academic Press.

[14] Demsetz, H. (1968): "The Cost of Transacting", *Quarterly Journal of Economics*, **82**, 33-53.

[15] Diamond, D. W. and R. E. Verecchia (1987): "Constraints on Short-selling and Asset Price Adjustments to Private Information", *Journal of Financial Economics*, **18**, 227-311.

[16] Easley, D., Kiefer, N. M. and M. OHara (1993): "One Day in the Life of a Very Common Stock", *Review of Financial Studies*, **10**, 805-835.

[17] _____ , Kiefer, N. M. and M. OHara (1997): "The Information Content of the Trading Process", *The Journal of Empirical Finance*, **12**, 159-186.

[18] _____ , Kiefer, N. M., OHara, M. and J. P. Paperman (1996): "Liquidity, Information and Infrequently Traded Stocks", *J. of Finance*, **51**, 1405-36.

[19] _____ , and M. O'Hara (1992): "Time and the Process of Security Price Adjustment", *Journal of Finance*, **19**, 69-90.

[20] Engle, R. F. (1982): "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U. K. Inflation", *Econometrica*, **50**, 987-1008.

[21] Engle, R. F. and G. Gonzales-Rivera (1991): "Semiparametric ARCH Models", *Journal of Business and Economic Statistics*, 9, 345-359.

[22] Engle, R. F. and . Lunde (1998): "Trades and Quotes: A Bivariate Point Process", Discussion paper, **7**, University of California, San Diego.

[23] Engle, R. F. and J. Russell (1998): "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data", *Econometrica*, **66**, 1127-62.

[24] Engle, R. F. and J. Russell (2004): "Analysis of High Frequency Financial Data", Working Paper, Stern School, New York University.

[25] Fernandes, M. and J. Grammig (2000): "Non-parametric Specification Tests for Conditional Duration Models", European University Institute and University of Frankfurt.

[26] Foster, D. F. and S. Viswanathan (1987): "Variations in Volumes, Spreads and Variances", Working Paper, Futures and Options Research Center, Duke University.

[27] Garman, M. (1976): "Market Microstructure", *Journal of Financial Economics*, 3, 257-275.

[28] Gaver, D. P. and P. A. W. Lewis (1980): "First Order Autoregressive Gamma Sequences and Point Processes", *Advances in Applied Probability*, **12**, 727-745.

[29] Ghysels, E. and J. Jasiak (1994): "Stochastic Volatility and Time Deformation: An Application to Trading Volume and Leverage Effects", C. R. D. E., Université de Montreal, unpublished manuscript.

[30] Giot, P. (1999): "Time Transformations, Intraday Data and Volatility Models", CORE Discussion Paper, **9944**.

[31] Glosten, L. and L. Harris (1988): "Estimating the Components of the Bid-Ask Spread, Journal of Financial Economics", **21**, 123-142.

[32] Glosten, L. R. and P. Milgrom (1985): "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents", *Journal of Financial Economics*, 14, 71-100.

[33] Gourieroux, C., Monfort, A. and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Theory", *Econometrica*, **52**, 681-700.

[34] Grammig, J. and K.-O. Maurer (2000): "Non-monotonic Hazard Functions and the Autoregressive Conditional Duration Model", *Econometrics J.*, **3**, 16-38.

[35] Hanousek, J. and J. Nemecek (2002): "Market Structure, Liquidity, and Information Based Trading at the PSE", *Emerging Markets Review*, **3**, 293-305.

[36] _____ , and R. Podpiera (2003): "Informed Trading and the Bid-Ask Spread: Evidence from an Emerging Market, *Journal of Comparative Economics*, **2**, 275-296.

[37] _____ , (2003): "Czech Experience with Market-Maker Trading System", *Economic Systems*, **2**, 177-191.

[38] HARRIS, L. (1990): "Estimation of Stock Price Variances and Serial Covariances from Discrete Observations", *J. of Fin. and Quant. Analysis,* **25**, 291-306.

[39] HASBROUCK, J. (1988): "Trades, Quotes, Inventories and Information", *Journal of Financial Economics*, **22**, 229-252.

[40] _____ , (1990): "The Summary Informativeness of Stock Trades: An Econometric Analysis", *Review of Financial Studies* (vol. 4), **3**, 571-595.

[41] _____ , (1991): "Measuring the Information Content of Stock Trades", *Journal of Finance*, **1**, 179-207.

[42] _____ , T. S. Y. HO (1991): "Order Arrival, Quote Behavior and the Return Generating Process, *Journal of Finance*, **42**, 1035-1048.

[43] HAWKES, A. G. (1971): "Spectra of Some Self-Exciting and Mutually Exciting Point Processes", *Biometrika*, **58**, 83-90.

[44] HECKMAN, J. J. AND B. J. BORJAS (1980): "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers from a Continuous Time Model of Heterogeneity and State Dependence", *Economica*, **47**, 247-283.

[45] HO, T. AND H. STOLL (1981): "Optimal Dealer Pricing Under Transactions and Return Uncertainty", *Journal of Financial Economics*, **9**, 47-73.

[46] KYLE, A. S. (1985): "Continuous Auctions and Insider Trading", *Econometrica*, **53**, 1315-36.

[47] LANCASTER, T. (1990): *The Econometric Analysis of Transition Data*, Econometric Society Monographs, Cambridge: Cambridge University Press.

[48] LAWRENCE, A. J. AND P. A. W. LEWIS (1980): "The Exponential Autoregressive-Moving Average ERMA(p, q) Process", *Journal of the Royal Statistical Society* (Series B), **42**, 150-161.

[49] LEE, CH. AND M. READY (1991): "Inferring Trade Direction from Intraday Data", *Journal of Finance* (vol. 26), **2**, 733-746.

[50] LUNDE, A. (1999): "A Generalized Gamma Autoregressive Conditional Duration Model", Working Paper, Aalborg University.

[51] MADHAVAN, A. (2000): "Market Microstructure: A Survey", Journal of Financial Markets (vol. 3), **3**, 205-258.

[52] McInish, T. and R. Wood (1992): "An Analysis of Intradaily Patterns in Bid/Ask Spreads for NYSE Stocks", *Journal of Finance*, 47, 753-764.

[53] Muller, U. A. et Al. (1990): "Statistical Study of Foreign Exchange Rates, Empirical Evidence of a Price Change Scaling Law, and Intraday Analysis", *Journal of Banking and Finance*, 14, 1189-1208.

[54] Nelson, D. and C. Cao (1992): "Inequality Constraints in the GARCH(1, 1) Model, Journal of Business and Economic Statistics, **10**, 229-235.

[55] Ogata, Y. and H. Akaike (1982): "On Linear Intensity Models for Mixed Doubly Stochastic Poisson and Self-Exciting Point Processes", *Journal of the Royal Statistical Society* (Series B), **44**, 102-107.

[56] Ogata, Y. and K. Katsura (1986): "Point-Process Models with Linearly Parametrized Intensity for the Application of Earthquake Catalogue", *Journal of Applied Probability*, **23A**, 231-240.

[57] OHara, M. (1995): *Market Microstructure Theory*, Oxford: Basil Blackwell.

[58]

[59] Rubin, I. (1972): "Regular Point Processes and Their Detection", *IEEE Transactions on Information Theory*, **ITT-18**, 547-557.

[60] Snyder, D. L. and M. I. Miller (1991): *Random Point Processes in Time and Space*, Second Edition, New York: Springer-Verlag.

[61] Stoll, H. R. (1976): "Dealer Inventory Behavior: An Empirical Investigation of Nasdaq/NMS Stocks", *Journal of Financial and Quantitative Analysis*, 359-380.

[62] _____ , (1978): "The Supply of Dealer Services in Securities Markets", *Journal of Finance*, **33**, 1133-1151.

[63] _____ , and R. Whaley (1990), "Stock Market Structure and Volatility", *Review of Financial Studies*, **3**, 37-71.

[64] Tauchen, G. and M. Pitts (1983): "The Price Variability-Volume Relationship on Speculative Markets", *Econometrica*, 51, 485-505.

[65] Teraswirta, T. and M. Meitz (2004): "Evaluating Models of Autoregressive Conditional Duration", SSE/EFI Working Paper Series in Economics and Finance, **557**, Stockholm School of Economics.

[66] TSAY, R. S. (2002): *Analysis of Financial Time Series*, New York: John Wiley.

[67] WOLD, H. (1948): "On Stationary Point Processes and Markow Chains", *Skandinavisk Aktuarietidskrist*, 31, 229-240.