DYNAMIC ECONOMETRIC MODELS
Vol. 9 – Nicolaus Copernicus University – Toruń – 2009

Roman Huptas
Cracow University of Economics

Intraday Seasonality in Analysis of UHF Financial Data: Models and Their Empirical Verification

A b s t r a c t. The aim of this paper is to outline the typical characteristics of the ultra-high-frequency financial data and to present estimation methods of intraday seasonality of trading activity. Ultra-high-frequency financial data (transactions data or tick-by-tick data) is defined to be a full record of transactions and their associated characteristics. We consider two nonparametric estimation methods: cubic splines and a Nadaraya-Watson kernel estimator of regression. Both approaches are compared empirically and applied to financial data of stocks traded at the Warsaw Stock Exchange.

K e y w o r d s: financial UHF data, intraday seasonality, diurnal pattern, cubic splines, kernel estimation.

1. Introduction

The last dozen or so years has witnessed mounting global interest in analyses of the microstructure of financial markets. Research into the microstructure of financial markets centres around explaining the process behind the shaping of the price of financial instruments and analyses of individual trade events. The impact of various transaction factors and mechanisms on the way in which instruments prices are shaped was captured by the so-called theoretical microstructure models. A review of those models and numerous issues collectively referred to as market microstructure effects was incorporated into O'Hara (1995) and Dacorogna et al. (2001) (cf. Doman, Doman, 2004; Bień, 2006).

Analysis of financial market processes and empirical verification of hypotheses arising from theoretical microstructure models were made possible by the newly-gained access to transactional databases. These databases became the source of specific financial time series referred to as ultra-high-frequency or tick-by-tick data.

New modelling tools for financial time series anticipate specific qualities of transaction data. They include, above all, asynchronous distributions of obser-

vations relative to time units and discrete price changes. Additionally, the individual events of the transaction process manifest themselves with varying frequency from one time period to another. Consequently, there may be – from one day to the next – certain repeat pattern of intensity with which transactions are concluded. In pertinent literature, this pattern is referred to as intraday seasonality of durations, with durations or waiting times standing for the time spans between trade events. However, before one can use econometric models in analyses of ultra-high-frequency time series, it is essential to eliminate intraday seasonality, which strongly manifests itself in the time series. In estimating intraday seasonality use is mostly made of selected nonparametric statistical methods.

The aim of this paper is to outline the typical characteristics of UHF financial data and further to present modelling and estimation methods of intraday seasonality of transaction activities. Within the framework of these methods, two nonparametric approaches will be presented: cubic splines interpolation and kernel estimations of regression functions. Both approaches will be verified and compared empirically on the basis of data extracted from the Polish share market.

2. Characteristics of UHF Financial Data

Several years ago a new term became operative – „ultra-high frequency data", also known as „tick-by-tick data" or „transaction data". These are time series composed of trade event features to which the exact time of their appearance was assigned. Thus observations are recorded asynchronously on time units. UHF financial data have a few characteristic qualities, which do not manifest themselves at lower frequencies.

The characteristic features of time series of transaction data include: nonsynchronous distribution of observations over time units, discrete transaction price changes, appearance of a number of transactions in the same single second, bid-ask bounce of transaction prices and intraday seasonality i.e. transactions reveal a daily periodic pattern.

The most important quality of UHF financial data is the nonsynchronous i.e. erratic distribution of observations over time units. Those data can, for instance, be aggregated so that they correspond to equal sequential time units (multiples of minutes, hours, days) and then – in analyses – enable use of a whole range of the GARCH models. On the other hand, such aggregation of transaction data and their analysis as observations made and selected at equal intervals leads to a loss of information furnished by the transaction process itself. Transactions or changes in the share price do not happen at equal intervals. Consequently, the durations between transactions involving shares of a given company may provide relevant information as to the intensity of their trading. Thus the assumption that changes in prices or transactions are equidistant in terms of time may

cause us to draw false conclusions. The problem of nonsynchronous trading and relevant examples are dealt with more extensively in Tsay (2002, p. 207) (cf. Doman, Doman, 2004; Osińska, 2006).

An alternative way of analysis of financial data distributed asynchronously over time units involves using the so-called transaction-time models (ACD, UHF-GARCH models etc). With those models, raw data are used for analysis. Owing to that, information inherent in the duration of the time between selected trade events can also come into focus. Duration analyses may furnish information on the microstructure of the financial market, affording a more accurate insight into various market interdependencies. In pertinent literature, the most frequently modelled durations between trade events are trade durations, price durations, volume durations (cf. Engle, Russell, 1997, p. 1149).

3. Intraday Seasonality of Durations

Under normal economic conditions, transactions reveal a daily seasonality factor. It appears that the number of transactions is higher immediately after the opening of business than prior to the close of the session (when the time gaps between transactions are the shortest), and markedly smaller during midday hours, i.e. in the middle of the session (so- called "lunchtime effect" when durations between transactions are also the longest). Thus, there exists a certain repetitious pattern of transaction intensity for each day. This is termed „intraday seasonality of durations". Consequently, Engle and Russell (1997) recommend decomposition of duration into a deterministic component $\phi(t_i)$ depending on moment t_i of the commencement of a given duration, and a stochastic component $\hat{x}_i$, which is free from the seasonality effect and which models process dynamics. Pertinent literature (cf. Engle, Russell, 1997) recommends that data be transformed as follows:

$$\hat{x}_i = \frac{x_i}{\phi(t_i)}, \tag{1}$$

where: $x_i = t_i - t_{i-1}$ - duration between transactions at time t_i and t_{i-1}, $\hat{x}_i$ - duration purged of the seasonality effect, $\phi(t_i)$ - multiplied factor of intraday seasonality at time t_i.

The seasonal factor $\phi(t_i)$ is construed as the average duration of each time unit during which we made data observations (most often denoting the average duration for each second). The diagram which illustrates intraday seasonality pattern, also known as diurnal pattern or time-of-day function mostly has the shape of the letter U turned upside down.

In numerous situations researchers lack adequate information to fully specify a parametric function of intraday seasonality. Despite the fact that intraday

cyclicality is not the key issue of investigation, it still cannot be ignored, but much rather needs to be included in analyses. Thus, in order to estimate the time-of-day function use can be made of selected nonparametric statistical methods such as splines, Fourier series, neural networks, wavelet analyses or kernel methods. In most works on duration modelling use is made of cubic splines or kernel estimations.

In pertinent literature the time-of the day function is determined most commonly by means of splines. This is a method which allows smoothing of average durations between events in subsequent time periods. Firstly, all durations during all the subsequent hours of the sessions on each day are averaged. Then a cubic spline with knots on every full hour of the session is determined. Knots correspond to previously determined average durations. This version of approximation of the daily period factor was presented in the paper (Engle, Russell, 1997). With a view to ensuring enhanced elasticity, the authors added a knot on the half-hour of the last hour of the session to capture the fast growing trade activity prior to the close of the stock market. A slightly different approach to cubic spline approximation can be observed in paper (Bauwens, Giot, 2000, p. 135). The authors note that the intraday seasonality factor may vary from one day of the week to the other, i.e. the shape of the periodic factor for Monday can be distinct from that for Tuesday etc. Consequently, the estimation of the intraday seasonality function was conducted separately for each day of the week to allow for possible seasonality within the week. In the first step, durations for the subsequent half-hours of the session were averaged separately for each day of the week and then the parameters of cubic splines were estimated for knots at full hours and half hours.

An alternative and second most commonly practised method of estimation of intraday seasonality function is the kernel estimation method. The intraday seasonality pattern is estimated as the Nadaraya-Watson kernel estimator of regression of raw durations on the time of the day (cf. Bauwens, Veredas, 2004, p. 398):

$$\phi(t) = \frac{\sum_{i=1}^{n} x_i K\left(\frac{t - t_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{t - t_i}{h_n}\right)},$$

where: t - number of seconds since the midnight of each day (or since the start of a session), x_i - durations corresponding to moments t_i (x_i is a dependent variable), t_i - number of seconds since the midnight of each day (or since the start of a session) until the moment of a given transaction, K - kernel function, h_n - bandwidth, s - standard deviation of sample t_i, n - number of observations.

As far as the kernel function is concerned, the paper (Bauwens, Veredas, 2004, p. 398) makes use of the quartic kernel (with optimal bandwidth of $2{,}78sn^{-1/5}$) which has the following shape:

$$K(x) = \begin{cases} \frac{15}{16}(1-x^2)^2, & dla \quad |x| \le 1 \\ 0, & dla \quad |x| > 1 \end{cases},$$

and in paper (Bauwens, Giot, 2002, p. 13) use is made of the gamma kernel function.

In the case of paper (Bauwens, Veredas, 2004, p. 398) the estimation of the intraday seasonality function is made separately for each day of the week to incorporate possible seasonality arising from the transaction repetition patterns also over a week-long period.

4. Empirical Example

The empirical verification of the methods presented above was carried out on the basis of time series involving trades in the shares of three companies listed in the WIG20 index: Telekomunikacja Polska S.A. (TPSA)/Polish Telecom/, Agora S.A. (Agora) and CEZ S.A. (CEZ) over a period between 22 March 2009 and 25 June 2009. The analysis covers transactions closed during the continuous quotation phase. On the basis of such time series, durations between each transaction were determined. Additionally, the time lags between the close of the session and the opening of next day's trading were removed.

Tabel 1. Descriptive statistics of transaction durations

	CEZ	Agora	TPSA
Number of observations	13919	19183	65166
Mean	98.930	84.840	25.110
Standard deviation (SD)	224.720	176.560	43.640
Dispersion index (=Mean/SD)	2.270	2.080	1.740
Minimum	1	1	1
Maximum	4196	4003	833
ACF(1)	0.220	0.225	0.212
ACF(2)	0.157	0.176	0.168
Q(5)	1805.430	2799.010	8948.820
Q(10)	2569.120	4111.970	13294.120
Q(15)	3056.570	5179.310	16546.610
Q(20)	3298.350	5866.790	19483.450

Note: ACF(k) – the value of the *k*-th order autocorrelation coefficient, Q(k) – the value of the Ljung-Box Q-statistic of *k*-th order, descriptive statistics in seconds.

The basic descriptive statistics of transaction durations for the shares in question are illustrated in Table 1. In our example, we witness three companies experiencing different trading activity patterns. The majority of the transactions involved TPSA, for which the average duration between transactions is 25 seconds. CEZ reported the fewest transactions and the average transaction time is 99 seconds. Agora, in turn, is a company of average liquidity and the average duration is around 85 seconds. In an analysis of the features of the distribution of durations our attention is momentarily attracted to marked overdispersion, i.e. the standard deviation exceeds the mean. The dispersion indexes (the ratio standard deviation to mean) are generally very high, which may imply great dynamics of the series in question. It is worth noting that the greater the frequency of trades, the lower the value of the dispersion index.

The values of the Q Ljung-Box statistics in Table 1 formally test the null hypothesis whereby there is no autocorrelation of durations respectively for the fifth, the tenth, the fifteenth and the twentieth order. Clearly, on the basis of the determined test statistics, the null hypothesis whereby there is no autocorrelation is definitely rejected for all three companies. Duration autocorrelation is thus extremely strong.

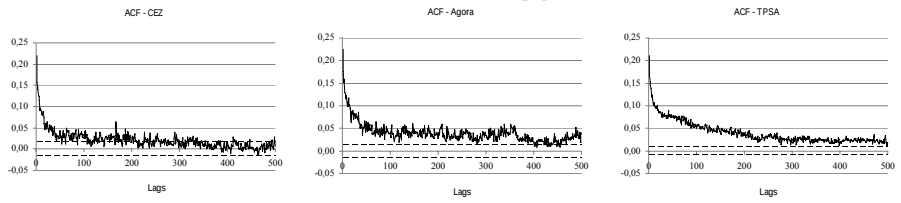

Figure 1. Autocorrelation functions of transaction durations for CEZ, Agora and TPSA

Graphs representing the autocorrelation function of the durations for the companies in question are to be found in Figure 1. Regardless of the company, the first values of the autocorrelation function are surprisingly low and stand at around 0.22. For CEZ shares the ACF function fairly soon shrinks to zero for the first several dozen delays, only to level off. As far as TPSA shares are concerned, the autocorrelation function takes much more time to decrease, approximately at hyperbolic speed (rate), which is typical of long memory processes. This evidences high „stability" (persistence) of the process. Moreover, the high values of lower order autocorrelations indicate stronger clustering of transaction activities. The dynamics of duration processes and the effect of clustering of transaction activities may be noted in figure 2 containing graphs of the series of the first 5000 observations.

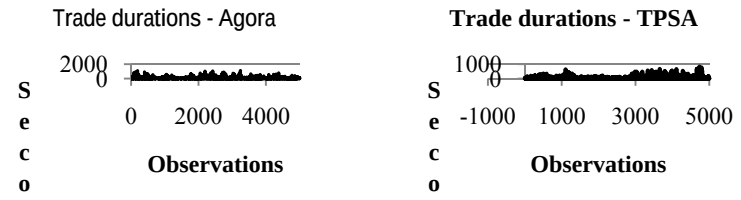

Figure 2. Plots of trade durations for Agora and TPSA – first 5000 observations

The existence of very powerful autocorrelation of durations may result from the midday seasonality of transactional activity. With this in mind, patterns of intraday seasonality were estimated by means of four methods, and further verified empirically to determine whether the method selected translates into effective elimination of high autocorrelation of the series under research. The following approaches, described extensively in point 3, were applied:

1. Nadaraya-Watson estimator of regression of the duration on the time of the day, determined separately for each day of the week (*NW_days*);
2. Nadaraya-Watson estimator of regression of the duration on the time of the day, ignoring possible seasonality over a week-long period (*NW*);
3. cubic splines with knots on every full hour and every half-hour of the session, determined separately for each day of the week (*CS_days*);
4. cubic splines with knots on every full hour and every half-hour of the session, ignoring possible seasonality over a week-long period (*CS*).

In the event of kernel estimators, use was made of the quartic kernel with the optimal bandwidth of $2{,}78sn^{-1/5}$. Next, after the intraday seasonality factor was estimated, durations purged of the seasonality effect were determined pursuant to formula (1). All algorithms and procedures were implemented using the GAUSS software.

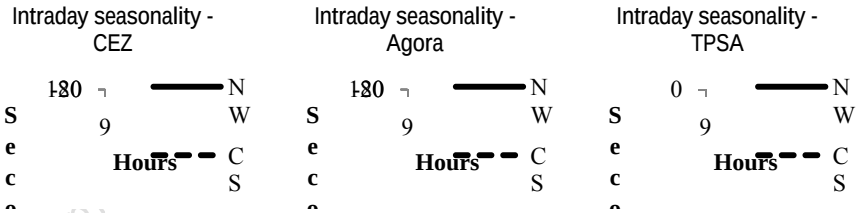

Figure 3. Intraday seasonality patterns for CEZ, Agora and TPSA

Figure 3 depicts the shape of intraday seasonality patterns (*NW* and *CS* methods) for the three companies under analysis, without regard for the effect that the different days of the week have. Figure 4 shows plots of the estimated time-of-day functions for subsequent days of the week for CEZ, Agora and

TPSA respectively, obtained by means of the kernel estimator (*NW*) and cubic splines (*CS*).

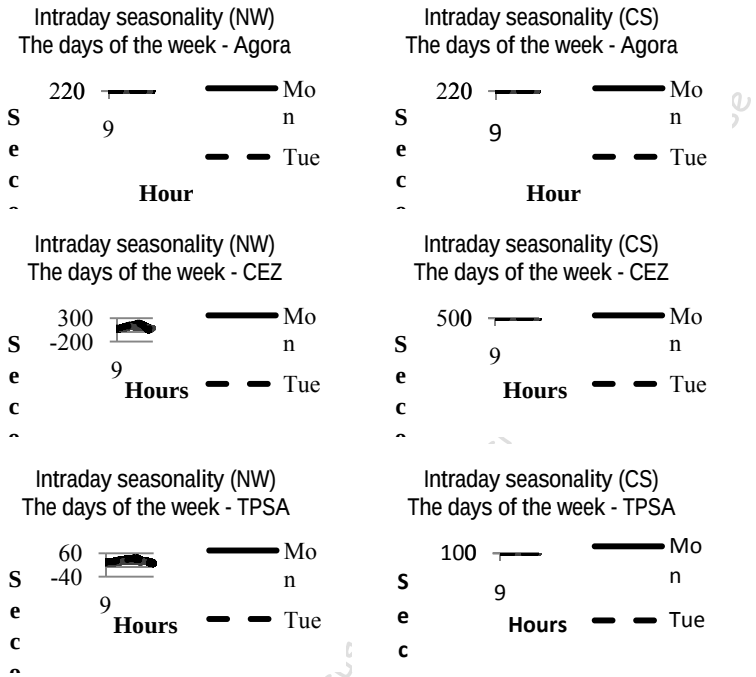

Figure 4. Intraday seasonality patterns for the days of the week for CEZ, Agora, TPSA

The graphs presenting the estimated time-of-day functions have the shape of the letter U turned upside down and reveal unequivocally that durations are subject to daily seasonality. The durations between transactions are markedly shorter after the opening and before the close of the session than at midday. The extent of trading activity between 12:00 a.m. and 2:00 p.m. is noticeably less, due, amongst others, to the lunchtime effect. It should be noted that in the case of companies listed on the American and West European Stock markets, the effect is less manifest than in the case of their Australian counterpart, where the effect is more pronounced, i.e. the hump in the graph is manifestly spikier (cf. (Bauwens, Giot, 2001; Hautsch 2004)). Similarly, trading activity at the opening of the session is more intense as traders begin to accommodate the information of the night before (macroeconomic data, etc). Trading activity at the close of trade can be explained in terms of some investors' attempt to close their open positions. It is worth noting that intraday seasonality will vary from one day of the week to the next (Figure 4). It seems that regardless of the type of company, trading activity is most intense on Tuesdays and Wednesdays than on any of the other days.

Duration statistics purged of intraday seasonality by means of the four above-named methods were included in Table 2. Deseasonalisation of the data

partly reduced the autocorrelation of transaction durations. From the point of view of effective elimination of seasonality impact on autocorrelation "measured" in terms of the values of the Ljung-Box test statistics, the *NW_days* (Table 2) appears to be the most effective method. In the case of TPSA and Agora it was definitely the most successful, i.e. the values of test statistics are the lowest of all four methods used. On the other hand, in the case of CEZ company, it ranked number two, with the *CS_days* approach ranked the highest.

Tabel 2. Descriptive statistics of adjusted transaction durations after deseasonalisation by means of the four methods

Stock	Method	Mean	SD	Disp. Index	ACF(1)	Q(5)	Q(10)	Q(15)	Q(20)
CEZ	*NW_days*	0.990	2.320	2.340	0.205	1360.150	1897.080	2101.860	2240.300
	NW	0.990	2.200	2.220	0.215	1537.830	2111.380	2475.990	2676.680
	CS_days	0.980	2.280	2.320	0.190	1290.340	1782.540	1972.950	2121.000
	CS	0.970	2.130	2.190	0.209	1535.590	2137.890	2530.380	2759.580
Agora	*NW_days*	0.990	1.980	2.000	0.215	2518.120	3521.420	4333.470	4886.400
	NW	0.990	1.980	2.000	0.211	2675.970	3719.670	4600.870	5169.360
	CS_days	0.990	1.950	1.970	0.209	2599.490	3632.710	4520.890	5097.310
	CS	0.990	1.950	1.970	0.215	2553.890	3604.400	4480.420	5078.150
TPSA	*NW_days*	0.990	1.660	1.670	0.195	7357.560	10568.660	12843.060	14864.480
	NW	0.990	1.680	1.700	0.197	7614.770	10964.070	13401.270	15598.530
	CS_days	0.990	1.670	1.690	0.197	7527.590	10874.810	13294.540	15400.190
	CS	0.990	1.680	1.700	0.200	7793.400	11248.550	13802.000	16085.570

Note: SD – standard deviation; ACF(k) – the value of the *k*-th order autocorrelation coefficient, Q(k) – the value of the Ljung-Box Q-statistic of *k*-th order, descriptive statistics in seconds.

An analysis of the values of Ljung-Box statistics implies that the *NW_days* and *CS_days* approaches should be used. Consequently, in eliminating intraday seasonality, the "day of the week" effect i.e. possible seasonality arising from variations in trading activity over the entire week should be taken into account. Regardless of the deseasonalisation method used, the values of Ljung-Box test statistics for the companies in question dropped by approximately 15%-25%, but still continued to be very high. Thus, the null hypothesis implying lack of autocorrelation continues to be rejected on each reasonable level of significance. This bears witness to the fact that the dynamics of transaction durations are influenced by factors other than the purely deterministic seasonality effect, which in turn is due to the structure of the share market.

5. Summary

Based on the results of empirical data, the application of the kernel estimator of regression separately for each day of the week appeared to be the most effective method of elimination of intraday seasonality impact on the autocorrelation of transaction durations. It is noteworthy, though, that the results for all analytical methods used are highly similar. In the case of splines, their prelimi-

nary averaging of durations between events for the subsequent full or half-hours may become something of a drawback. The extent of data aggregation and the elasticity of the estimated function can be reduced by increasing the number of knots used in the spline. So it appears that the inclusion of intraday seasonality models in the base models is a natural step, and wins over earlier data filtrations and testing if the use of a two-step or one-step approaches will have identical impact on the quality of the estimators determined.

References

Bauwens, L., Giot, P. (2000), The Logarithmic ACD Model: An Application to the Bid-ask Quote Process of Three NYSE Socks, *Annales d'Économie et de Statistique*, 60, 117–149.

Bauwens, L., Giot, P. (2001), *Econometric Modelling of Stock Market Intraday Activity*, Kluwer Academic Publishers, Boston.

Bauwens, L., Giot, P. (2002), Asymmetric ACD Models: Introducing Price Information in ACD Models, CORE Discussion Paper 9844.

Bauwens, L., Veredas, D. (2004), The Stochastic Conditional Duration Model: A Latent Variable Model for the Analysis of Financial Durations, *Journal of Econometrics*, 119, 381–412.

Bień, K. (2006), Model ACD – podstawowa specyfikacja i przykład zastosowania (ACD Model – Basic Specification and Example of Application), *Przegląd Statystyczny (Statistical Survey),* t.53, z. 3, 83-97.

Dacorogna, M. M., Gençay, R., Müller, U., Olsen, R. B., Pictet, O. V. (2001), *An Introduction to High-Frequency Finance*, Academic Press, San Diego.

Doman, M., Doman, R. (2004), *Ekonometryczne modelowanie dynamiki polskiego rynku finansowego (Econometric Modelling of Dynamics of Polish Financial Market)*, Wydawnictwo AE w Poznaniu, Poznań.

Engle, R. F., Russell, J. R. (1997), Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data, *Econometrica*, 66, 1127–1162.

Hautsch N. (2004), *Modelling Irregularly Spaced Financial Data*, Springer-Verlag, Berlin, Heidelberg.

O'Hara, M. (1995), *Market Microstructure Theory*, Blackwell Inc., Oxford.

Osińska, M. (2006), *Ekonometria finansowa (Financial Econometrics)*, PWE, Warszawa.

Tsay, R.S. (2002), *Analysis of Financial Time Series*, Wiley Series in Probability and Statistics, John Wiley& Sons, New York.

Wewnątrzdzienna sezonowość w analizie danych finansowych UHF: modele i ich empiryczna weryfikacja

Z a r y s t r e ś c i. Celem artykułu jest krótkie przedstawienie cech charakterystycznych dla danych finansowych UHF oraz prezentacja metod modelowania i szacowania wewnątrzdziennej sezonowości aktywności transakcyjnej. W ramach tych metod są przedstawione dwa podejścia nieparametryczne: interpolacja za pomocą kubicznych funkcji sklejanych oraz estymacja jądrowa funkcji regresji. Oba prezentowane podejścia są zweryfikowane i porównane empirycznie na podstawie danych z polskiego rynku akcji.

S ł o w a k l u c z o w e: dane finansowe UHF, wewnątrzdzienna sezonowość, funkcje sklejane, estymator Nadaraya-Watsona.