EKONOMETRIA

26

Zastosowanie matematyki w ekonomii

Redaktor naukowy **Janusz Łyko**

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2009

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 76
Ekonometria 26 2009

Roman Huptas
Uniwersytet Ekonomiczny w Krakowie

METODY SZACOWANIA WEWNĄTRZDZIENNEJ SEZONOWOŚCI W ANALIZIE DANYCH FINANSOWYCH POCHODZĄCYCH Z POJEDYNCZYCH TRANSAKCJI

Streszczenie: Celem artykułu jest krótkie omówienie cech charakterystycznych dla danych finansowych o ultrawysokiej częstotliwości oraz prezentacja metod modelowania i szacowania wewnątrzdziennej sezonowości aktywności transakcyjnej. W ramach tych metod przedstawione zostaną dwa podejścia: interpolacja za pomocą kubicznych funkcji sklejanych oraz estymacja jądrowa.

W danych finansowych UHF wyznaczana jest kubiczna funkcja sklejana z węzłami umieszczonymi w każdej pełnej godzinie trwania sesji. Węzłom odpowiadają średnie czasy trwania między zdarzeniami procesu transakcyjnego dla kolejnych godzin sesji. W przypadku estymacji jądrowej odwzorowanie wewnątrzdziennej sezonowości jest szacowane jako estymator jądrowy Nadaraya-Watsona funkcji regresji czasów trwania pomiędzy zdarzeniami procesu transakcyjnego względem odpowiadających im sekund w ciągu dnia.

Słowa kluczowe: dane finansowe UHF, wewnątrzdzienna sezonowość, kubiczne funkcje sklejane, estymator jądrowy Nadaraya-Watsona.

1. Wstęp

W ciągu ostatnich lat w ekonometrii finansowej na świecie obserwuje się wzrost zainteresowania analizą mikrostruktury rynków finansowych. Badania mikrostruktury rynków finansowych dotyczą wyjaśniania procesu kształtowania się cen instrumentów finansowych oraz analizy pojedynczych zdarzeń procesu transakcyjnego. Wpływ różnych czynników i mechanizmów transakcyjnych na sposób kształtowania cen instrumentów ujęto w tzw. teoretycznych modelach mikrostruktury. Przegląd teoretycznych modeli mikrostruktury oraz wielu problemów określanych wspólną nazwą efektów mikrostruktury rynku prezentowany jest szczegółowo w pracach [O'Hara 1995; Dacorogna i in. 2001] (por. [Bień 2006; Doman, Doman 2004]). Do szczegółowych aspektów mikrostruktury należą: płynność, koszty transakcyjne, zmienność cen, stopień odzwierciedlenia informacji przez ceny, asymetria informacji, intensywność procesu napływu informacji, wpływ działań animatorów rynku, niesynchroniczny handel oraz zyski transakcyjne (zob. [Doman, Doman 2004]).

Analiza procesów zachodzących na rynkach finansowych oraz empiryczna weryfikacja hipotez wynikających z teoretycznych modeli mikrostruktury możliwa jest w ostatnich dekadach dzięki dostępowi do baz danych transakcyjnych. Bazy transakcyjne to zbiory danych zawierające szczegółowe informacje dotyczące pojedynczych transakcji i zleceń na rynku finansowym (zob. [Hautsch 2004]). Bazy danych transakcyjnych są źródłem specyficznych finansowych szeregów czasowych ze względu na to, że obserwacje rejestrowane są w miarę ich pojawiania się, czyli niesynchronicznie względem jednostek czasu. Takie szeregi czasowe określa się jako „dane o ultrawysokiej częstotliwości" (*ultra-high-frequency data*) lub „dane tikowe" (*tick-by-tick data*) (zob. [Bień 2006; Engle 2000]).

Dostęp do szeregów danych finansowych o ultrawysokiej częstotliwości przyczynił się do rozwoju w ekonometrii finansowej nowego obszaru badań naukowych. Nowe narzędzia modelowania finansowych szeregów czasowych wychodzą naprzeciw specyficznym własnościom danych transakcyjnych. Należą do nich przede wszystkim niesynchroniczne rozmieszczenie obserwacji względem jednostek czasu oraz dyskretne zmiany cen. Ponadto poszczególne zdarzenia procesu transakcyjnego występują z różną częstotliwością w różnych porach dnia. Dlatego też efektem tego może być występowanie pewnego powtarzalnego dla każdego dnia schematu intensywności zawierania transakcji. Jest on w literaturze przedmiotu określany nazwą „wewnątrzdziennej sezonowości" (*intraday seasonality*) czasów trwania, przy czym „czasy trwania" (*durations*) lub „czasy oczekiwania" (*waiting times*) to odstępy czasu między zdarzeniami procesu transakcyjnego. Powyższe własności danych tikowych wymagają zastosowania do badań specjalistycznych narzędzi ekonometrycznych (m.in. modeli ACD – modeli autoregresyjnego warunkowego czasu trwania; modeli ACD-GARCH, SCD itp.). Zanim jednak w analizie szeregów czasowych o ultrawysokiej częstotliwości wykorzysta się ekonometryczne modele w czasie transakcyjnym, należy wyeliminować z tych szeregów zaobserwowaną silną wewnątrzdzienną sezonowość. Do szacowania wewnątrzdziennej sezonowości stosuje się najczęściej wybrane nieparametryczne metody statystyczne.

Celem tego opracowania jest krótkie przedstawienie cech charakterystycznych dla danych finansowych o ultrawysokiej częstotliwości oraz prezentacja metod modelowania i szacowania czynnika wewnątrzdziennej sezonowości aktywności transakcyjnej. W ramach tych metod przedstawione zostaną: interpolacja za pomocą kubicznych funkcji sklejanych oraz estymacja jądrowa funkcji regresji.

2. Własności danych finansowych o ultrawysokiej częstotliwości

Do pewnego czasu w literaturze związanej z finansami i ekonometrią finansową pojawiało się jedynie pojęcie szeregów czasowych o wysokiej częstotliwości. Szeregi danych o wysokiej częstotliwości to szeregi czasowe obserwacji synchronicznie rejestrowanych względem małych przedziałów czasu (dni, minut, wielokrotności minut). Od kilku lat zaczęło także funkcjonować nowe określenie – „dane o ultrawysokiej częstotliwości”, inaczej „dane tikowe” lub „dane transakcyjne”. Są to szeregi czasowe zbudowane z charakterystyk zdarzeń procesu transakcyjnego z przyporządkowanym dokładnym czasem ich pojawienia się. Zatem obserwacje są tu rejestrowane niesynchronicznie względem jednostek czasu. Dane finansowe o ultrawysokiej częstotliwości mają kilka charakterystycznych własności, które nie pojawiają się w przypadku niższych częstotliwości.

Do charakterystycznych cech szeregów czasowych danych transakcyjnych należą (por. [Tsay 2002, s. 212]):

1) niesynchroniczne rozmieszczenie obserwacji względem jednostek czasu,
2) dyskretne wartości cen akcji i dyskretne zmiany cen,
3) występowanie wielu transakcji w ciągu jednej sekundy,
4) odbicie *bid-ask* w przypadku cen transakcyjnych,
5) występowanie wewnątrzdziennej sezonowości (transakcje wykazują dzienny składnik periodyczny).

Jedną z najważniejszych cech danych finansowych o ultrawysokiej częstotliwości jest niesynchroniczne (nieregularne) rozmieszczenie obserwacji względem jednostek czasu. Tego typu dane można np. zagregować, tak aby odpowiadały kolejnym równym jednostkom czasu (wielokrotnościom minut, godzinom, dniom) i skorzystać w analizie z szerokiej gamy modeli GARCH. Jednak tego typu agregacja danych transakcyjnych i ich analiza jako obserwacji pobieranych w równych odstępach czasu wiąże się z utratą informacji, jaką niesie proces transakcyjny (por. [Hautsch 2004]). Transakcje czy też zmiany ceny akcji nie zachodzą w równych odstępach czasu. Zatem długości czasów trwania między transakcjami dotyczącymi akcji spółki mogą być ważną informacją na temat intensywności ich obrotów.

Przyjmowanie więc założenia, że zmiany cen albo transakcje są równoodległe w czasie może prowadzić do błędnych wniosków. Gdy handel odbywa się w nierównych odstępach czasu, a obserwacje są czynione w odstępach równych, rozkłady zwrotów obserwowanych mogą mieć zupełnie inne własności niż rozkłady „prawdziwych” zwrotów. Efektem nieregularności handlu mogą być pozorne korelacje rzędu pierwszego między stopami zwrotu z różnych akcji, autokorelacje rzędu pierwszego zwrotów portfela i, w niektórych przypadkach, ujemne autokorelacje rzędu pierwszego w szeregach zwrotów poszczególnych akcji (zob. [Tsay 2002]; por. [Doman, Doman 2004; Osińska 2006]). Problem niesynchroniczności transakcyjnej (*nonsynchronous trading*) wraz z odpowiednimi przykładami jest szerzej poruszany w [Tsay 2002, s. 207] (por. [Doman, Doman 2004; Osińska 2006]).

Alternatywnym sposobem analizy danych finansowych niesynchronicznie rozmieszczonych względem jednostek czasu jest wykorzystanie tzw. modeli w czasie transakcyjnym (rodzina modeli ACD – autoregresyjnego warunkowego czasu trwania, modele ACD-GARCH, SCD itp.). W tym przypadku analizie poddawane są surowe dane, które nie uległy żadnej agregacji ani transformacji. Badane są długości czasów trwania między wybranymi zdarzeniami procesu transakcyjnego. Analiza czasów trwania dostarczyć nam może informacji na temat mikrostruktury rynku finansowego, pozwalając na dokładniejszy wgląd w różnego rodzaju zależności występujące na rynku. W literaturze przedmiotu do najczęściej modelowanych czasów trwania między zdarzeniami procesu transakcyjnego należą (zob. [Engle, Russell 1997, s. 114; Hautsch 2004]):

– transakcyjne czasy trwania (*trade durations*),
– cenowe czasy trwania (*price durations*),
– wolumenowe czasy trwania (*volume durations*).

Transakcyjnymi czasami trwania nazywamy długości czasu upływającego pomiędzy kolejnymi transakcjami. Dzięki ich analizie można badać intensywność procesu transakcyjnego (por. [Hautsch 2004, s. 32]). Cenowe czasy trwania to odstępy czasu potrzebne do zarejestrowania wzrostu lub spadku ceny instrumentu finansowego o wartość równą co najmniej c_p ($c_p > 0$). Wartość c_p jest ustalana arbitralnie i jest ona wielokrotnością tiku cenowego (wynika to z faktu, że ceny akcji mają charakter dyskretny). W praktyce zarówno transakcyjne czasy trwania, jak i cenowe czasy trwania wykorzystuje się do tworzenia tzw. miar wewnątrzdziennych chwilowych zmienności cen transakcyjnych (zob. [Engle 2000; Giot 2005]). Wolumenowe czasy trwania to przedziały czasu pomiędzy zdarzeniami procesu transakcyjnego, w których dokonano obrotu co najmniej c_w akcjami. Wartość c_w ($c_w > 0$) jest ustalana arbitralnie przez badacza. W praktyce wolumenowe czasy trwania służą do badania płynności rynku (por. [Hautsch 2004, s. 33-34]).

Kolejną istotną cechą danych transakcyjnych są dyskretne zmiany cen transakcyjnych. Ceny akcji mają charakter wartości dyskretnych i zmieniają się w sposób dyskretny. Minimalną wartością dozwolonej zmiany ceny transakcyjnej jest tzw. tik cenowy (*tick*). Ceny transakcji przyjmują zatem wartości będące wielokrotnościami pojedynczego tiku cenowego. Na giełdach amerykańskich od 29 stycznia 2001 r. ceny akcji można wyznaczać z dokładnością do jednego centa (zob. [Tsay 2002]). Jeśli chodzi o giełdę europejską Euronext i Giełdę Papierów Wartościowych w Warszawie, to wielkość tiku cenowego uzależniona jest od wartości akcji. Okazuje się, że wielkość tiku cenowego ma duży wpływ na funkcjonowanie rynku. Zmniejszenie wielkości tiku redukuje bowiem rozpiętość spreadu *bid-ask*. Niewielki spread wpływa na spłycenie rynku, czyli na zmniejszenie liczby akcji oferowanych przez dealera.

Kolejna istotna cecha danych transakcyjnych związana jest ze spreadem *bid--ask* i obecnością na rynku finansowym animatorów rynku (dealerów). Na niektórych giełdach animatorzy rynku odgrywają ważną rolę w „ułatwianiu i usprawnianiu” handlu instrumentami finansowymi. Ich obowiązkiem jest podtrzymywanie płynności rynku, tzn. stoją w gotowości do kupna bądź sprzedaży akcji w zależności od zleceń, jakie napływają na rynek. Poprzez płynność rynku rozumie się zdolność do kupna lub sprzedaży odpowiedniej ilości instrumentu: szybko, anonimowo i bez ingerencji w cenę. W zamian za to animatorzy rynku mają prawo do ustalania różnych cen kupna i sprzedaży danego instrumentu, tzn. skupują po tzw. cenie zakupu P_b (*bid*) i sprzedają po cenie wyższej, tzw. cenie sprzedaży P_a (*ask*). Różnica między tymi cenami $P_a - P_b$ jest nazywana spreadem *bid-ask*. W praktyce spread *bid--ask* jest niewielki, tzn. rzędu kilku tików cenowych. Efektem istnienia spreadu może być pozorna ujemna autokorelacja rzędu pierwszego zmian cen transakcyjnych, nawet jeżeli podstawowa, fundamentalna wartość instrumentu finansowego nie ulega zmianie. Zjawisko to jest w literaturze przedmiotu nazywane **odbiciem *bid-ask*** (zob. [Roll 1984; Fantazzini 2004, s. 146]; por. [Tsay 2002, s. 212]). Ogólnie ceny transakcyjne nieustannie wahają się pomiędzy cenami *bid* i *ask* w zależności od napływających zleceń. Transakcje, które są inicjowane przez napływające zlecenia sprzedaży (przez stronę podaży), zawierane są po cenie zakupu (cenie *bid*), a te, które inicjowane są przez napływające zlecenia kupna (przez stronę popytu), zawierane są po cenie sprzedaży (cenie *ask*).

W przypadku danych obserwowanych w czasie transakcyjnym nie można zapomnieć o jeszcze jednej bardzo ważnej własności – obecności składnika periodycznego. W normalnych warunkach gospodarczych transakcje wykazują dzienny składnik sezonowy. Okazuje się, że transakcji jest więcej tuż po otwarciu oraz tuż przed zamknięciem sesji (wtedy odstępy czasu między transakcjami są najkrótsze), natomiast zdecydowanie mniej w godzinach południowych, czyli w środku sesji (wówczas czasy trwania między transakcjami są najdłuższe). Efektem tego jest zatem występowanie pewnego powtarzalnego dla każdego dnia schematu intensywności transakcyjnej. Zjawisko to określane jest mianem „wewnątrzdziennej sezonowości" (*intraday seasonality*) czasów trwania. W związku z tym sugeruje się wyeliminowanie z danych efektu deterministycznego, wynikającego z dziennej sezonowości procesu transakcyjnego. W literaturze przedmiotu (zob. [Engle, Russell 1997]) proponuje się więc transformować dane w następujący sposób:

$$\hat{x}_i = \frac{x_i}{\varphi(t_i)},$$

gdzie: $x_i = t_i - t_{i-1}$ – czas trwania między transakcjami z chwil t_i i t_{i-1},

$\hat{x}_i$ – czas trwania pozbawiony efektu sezonowości,

$\varphi(t_i)$ – multiplikatywny czynnik wewnątrzdziennej sezonowości w chwili t_i.

Czynnik sezonowy $\varphi(t_i)$ jest rozumiany jako średni czas trwania dla każdej jednostki czasu, w którym zaobserwowaliśmy dane, najczęściej średni czas trwania dla każdej sekundy (por. [Osińska 2006]). Wykres odwzorowujący wewnątrzdzienną sezonowość (*intraday seasonality pattern, diurnal pattern, daily periodic pattern, time-of-day function*) ma najczęściej kształt odwróconej litery U. Do estymacji odwzorowania wewnątrzdziennej sezonowości stosuje się w literaturze przedmiotu wybrane metody statystyczne. W większości prac dotyczących modelowania czasów trwania wykorzystuje się do tego kubiczne funkcje sklejane albo estymację jądrową.

3. Interpolacja za pomocą funkcji sklejanych w modelowaniu wewnątrzdziennej sezonowości

Do estymacji odwzorowania wewnątrzdziennej sezonowości można wykorzystać kubiczne funkcje sklejane (splajny kubiczne). Podejście to jest jednym z najczęściej stosowanych w literaturze przedmiotu (zob. [Engle, Russell 1997; Giot 2005]). Przeanalizujmy krótko, na czym polega idea interpolacji za pomocą funkcji sklejanych i w jaki sposób można oszacować parametry takich splajnów (zob. [Stoer 1979]).

Interpolacji za pomocą funkcji sklejanych używa się głównie w tym celu, aby połączyć dane punkty krzywą „możliwie gładką" (zob. [Stoer 1979]). Niech przez $\Delta := \{a = x_0 < x_1 < \ldots < x_n = b\}$ będzie dany podział przedziału $[a, b]$.

Definicja. ***Przez kubiczną funkcję sklejaną odpowiadającą podziałowi*** Δ***, gdzie*** x_i ***(***$i = 0, 1, \ldots, n$***) nazywają się węzłami, rozumiemy funkcję rzeczywistą*** $S_\Delta : [a,b] \to \Re$ ***mającą własności:***

1) $S_\Delta \in C^2[a,b]$: S_Δ ma na $[a,b]$ ciągłą pochodną rzędu drugiego,

2) w każdym podprzedziale $[x_i, x_{i+1}]$ ($i = 0, 1, \ldots, n-1$) S_Δ jest wielomianem stopnia trzeciego.

Funkcja sklejana jest zatem kawałkami złożona z n wielomianów stopnia trzeciego tak, że sama funkcja S_Δ i jej obydwie (tzn. pierwsza i druga) pochodne nie mają w punktach x_i ($i = 1, \ldots, n-1$) punktów nieciągłości.

Niech $Y := \{y_0, y_1, \ldots, y_n\}$ jest zbiorem $n+1$ liczb rzeczywistych, odpowiadających wartościom x_i ($i = 0, 1, \ldots, n$), czyli węzłom. Oznaczmy funkcję sklejaną S_Δ, dla której zachodzi dodatkowo $S_\Delta(Y; x_i) = y_i$ ($i = 0, 1, \ldots, n$), przez $S_\Delta(Y; \cdot)$. Dalej mówić będziemy o tzw. naturalnych splajnach kubicznych, tzn. funkcjach sklejanych, które mają następującą własność: $S''_\Delta(Y; x_0) = S''_\Delta(Y; x_n) = 0$.

Niech $h_{j+1} := x_{j+1} - x_j$ ($j = 0, 1, \ldots, n-1$). Jako momenty M_j, $M_j := S''_\Delta(Y; x_j)$ ($j = 0, 1, \ldots, n$) oznaczmy drugą pochodną szukanej funkcji sklejanej $S_\Delta(Y; \cdot)$ w punktach $x_j \in \Delta$. Momenty M_j można obliczyć jako rozwiązanie układu równań liniowych i za pomocą M_j można łatwo podać samą interesującą nas funkcję sklejaną $S_\Delta(Y; \cdot)$.

Zauważmy, że $S_{\Delta}(Y;\cdot)$ w każdym przedziale $[x_j, x_{j+1}]$ $(j = 0, 1, \ldots, n-1)$ jest funkcją stopnia trzeciego, więc $S''_{\Delta}(Y;\cdot)$ jest funkcją liniową, którą można opisać za pomocą M_j:

$$S''_{\Delta}(Y; x) = M_j \frac{x_{j+1} - x}{h_{j+1}} + M_{j+1} \frac{x - x_j}{h_{j+1}} \text{ dla } x \in [x_j, x_{j+1}].$$

Dla $S_{\Delta}(Y;\cdot)$ mamy zatem po odpowiednich obliczeniach następujące przedstawienie:

$$S_{\Delta}(Y; x) = \alpha_j + \beta_j(x - x_j) + \gamma_j(x - x_j)^2 + \delta_j(x - x_j)^3 \text{ dla } x \in [x_j, x_{j+1}],$$

gdzie:

$$\alpha_j := y_j,\ \ \beta_j = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{2M_j + M_{j+1}}{6} h_{j+1},\ \ \gamma_j := \frac{M_j}{2},\ \ \delta_j := \frac{M_{j+1} - M_j}{6h_{j+1}}.$$

Zatem $S_{\Delta}(Y;\cdot)$ jest wyrażone za pomocą M_j i pozostaje jeszcze tylko kwestia obliczenia M_j $(j = 0, 1, \ldots, n)$. Wykorzystując ciągłość $S'_{\Delta}(Y;\cdot)$ w punktach $x = x_j$ $(j = 1, \ldots, n-1)$, otrzymujemy po odpowiednich przekształceniach układ równań liniowych dla M_j:

$$M_0 = 0,$$

$$a_j M_{j-1} + 2M_j + b_j M_{j+1} = d_j \quad j = 1, \ldots, n-1,$$

$$M_n = 0,$$

gdzie:

$$a_j = \frac{h_j}{h_j + h_{j+1}},\quad b_j = \frac{h_{j+1}}{h_j + h_{j+1}},\quad d_j = \frac{6}{h_j + h_{j+1}}\left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j}\right).$$

Do estymacji wielomianowej funkcji sklejanej można także podejść w trochę inny sposób. Krótko opisany zostanie sposób estymacji wielomianowej funkcji sklejanej złożonej z wielomianów stopnia q. Przyjmuje się, że estymator będący wielomianową funkcją sklejaną można zapisać jako kombinację liniową jego funkcji bazowych postaci (zob. [Ćwik, Koronacki 2005; Gatnar, Walesiak 2009; Hastie, Tibshirani, Friedman 2001]):

$$1,\ \{x^j\}_{j=1}^{q-1},\ \{(x - x_i)_+^q\}_{i=0}^{n-1},$$

gdzie: $x_i, \quad i = 0, 1, \ldots, n$ – węzły funkcji sklejanej,

$$a_+ = \begin{cases} a, & a \geq 0 \\ 0, & a < 0. \end{cases}$$

Splajn ma zatem następującą postać:

$$\tilde{f}(x) = \sum_{j=1}^{q} \alpha_j x^{j-1} + \sum_{i=0}^{n-1} \beta_i (x - x_i)_+^q.$$

Należy zwrócić uwagę na fakt, że lokalnie dopasowywany jest tylko wyraz w potędze q. Składniki wielomianu o niższych potęgach są dopasowywane „globalnie”.

Wybór wartości n, czyli liczby przedziałów oraz węzłów, może odbywać się adaptacyjnie, np. na drodze kroswalidacyjnej albo jest ustalany przez badacza. Dla zadanych węzłów $x_i, i = 0, 1, \ldots, n,$ pozostałe parametry funkcji sklejanej szacuje się metodą najmniejszych kwadratów, minimalizując sumę:

$$\sum_{i=0}^{n} (y_i - \tilde{f}(x_i))^2, \tag{1}$$

gdzie $y_i, i = 0, 1, \ldots, n$ są wartościami odpowiadającymi poszczególnym wartościom $x_i, i = 0, 1, \ldots, n,$ czyli węzłom. Ostatecznie otrzymujemy regresyjną funkcję sklejaną.

W literaturze znaleźć można także pojęcie wygładzonej funkcji sklejanej (zob. [Ćwik, Koronacki 2005; Gatnar, Walesiak 2009]). Zamiast rozwiązywać zadanie (1) najczęściej minimalizuje się kryterium kwadratowe z uwzględnieniem kary za oscylacyjność estymatora,

$$\sum_{i=0}^{n} (y_i - \tilde{f}(x_i))^2 + \lambda \int_R [\tilde{f}''(x)]^2 dx,$$

gdzie λ jest nieujemnym **współczynnikiem kary**, nazywanym też **współczynnikiem wygładzającym**, a $\tilde{f}''(\cdot)$ jest drugą pochodną funkcji $\tilde{f}(\cdot)$. Oczywiście całka z kwadratu drugiej pochodnej funkcji $\tilde{f}(\cdot)$ jest tym większa, im bardziej oscylacyjny jest przebieg tej funkcji. Ogólnie wyrażenie $\lambda \int_R [\tilde{f}''(x)]^2 dx$ jest karą za niegładkość estymatora, czyli za gwałtowne, oscylacyjne zmiany jego wartości. Powyższa kara za niegładkość określana jest mianem regularyzacji estymatora. Taka regularyzacja ma za zadanie wygładzenie estymatora, a otrzymany w ten sposób estymator nazywany jest **wygładzoną funkcją sklejaną**. Najczęściej przyjmuje się, że stopień wielomianów funkcji sklejanej wynosi $q = 3$.

Wyznaczanie odwzorowania wewnątrzdziennej sezonowości za pomocą funkcji sklejanych jest w literaturze podejściem najbardziej powszechnym. Oczywiście dzięki tej metodzie następuje wygładzenie średnich czasów trwania pomiędzy zdarzeniami dla kolejnych okresów. Na początek następuje uśrednienie wszystkich czasów trwania ze wszystkich dni z kolejnych godzin sesji. Następnie wyznaczana jest kubiczna funkcja sklejana z węzłami umieszczonymi w każdej pełnej godzinie trwania sesji. Węzłom odpowiadają oczywiście wyznaczone wcześniej uśrednione czasy trwania. Taka wersja aproksymacji dziennego czynnika periodycznego została zaprezentowana w pracy [Engle, Russell 1997]. W celu uzyskania lepszej elastyczności autorzy dodatkowo umieścili węzeł w połowie ostatniej godziny trwania sesji ze względu na szybko rosnącą aktywność transakcyjną przed zamknięciem giełdy. Trochę inne podejście w ramach aproksymacji splajnami kubicznymi można zobaczyć w pracy [Bauwens, Giot 2000, s. 135]. Autorzy zwracają uwagę, że czynnik wewnątrzdziennej sezonowości może być różny dla różnych dni tygodnia, tzn. inaczej może kształtować się odwzorowanie składnika periodycznego dla poniedziałku, a inaczej dla wtorku itd. Zatem oszacowanie funkcji wewnątrzdziennej sezonowości przeprowadzono osobno dla każdego dnia tygodnia, aby uwzględnić ewentualną okresowość wynikającą z aktywności transakcyjnej również w okresie tygodniowym. W pierwszym kroku uśredniono czasy trwania z kolejnych połówek godzin sesji dla każdego dnia tygodnia osobno, a potem oszacowano parametry splajnów kubicznych z węzłami umieszczonymi w każdej godzinie i połówce godzin. Z kolei w pracy [Strickland, Forbes, Martin 2003] wykorzystano do estymacji odwzorowania wewnątrzdziennej sezonowości wygładzone kubiczne funkcje sklejane (*cubic smoothing splines*), gdzie współczynnik kary za niegładkość estymatora wyznaczono metodą uogólnionej kroswalidacji.

4. Estymacja jądrowa w modelowaniu wewnątrzdziennej sezonowości

Alternatywną i drugą najczęściej stosowaną w praktyce metodą szacowania funkcji wewnątrzdziennej sezonowości jest metoda estymacji jądrowej. Estymacja jądrowa należy do nieparametrycznych metod statystycznych.

W pkt 2 wspomnieliśmy, że multiplikatywny czynnik wewnątrzdziennej sezonowości $\varphi(t_i)$ jest rozumiany jako średni czas trwania dla każdej jednostki czasu, w którym zaobserwowaliśmy dane (najczęściej średni czas trwania dla każdej sekundy). Interesuje nas zatem wyznaczenie funkcji regresji. Jeżeli przynajmniej przybliżona (ogólna) postać modelu regresyjnego opisującego interesujące badacza zjawisko nie jest znana *a priori* (oczywiście tak jest w naszym przypadku), można odwołać się do pomocy jednego z wielu podejść nieparametrycznych, czyli takich, gdzie o estymowanej funkcji nie zakładamy, że jest znana z dokładnością do skończenie wielu (estymowanych) parametrów. Chcemy jedynie, by estymowana funkcja była ciągła. Formalnie estymatorem funkcji regresji (nieparametrycznym) jest wielomianowa funkcja sklejana, zwana splajnem. Teraz zajmiemy się jednak estymatorem jądrowym funkcji regresji.

W XX wieku niezależnie od siebie Nadaraya i Watson zaproponowali taki sam nieparametryczny estymator funkcji regresji przy założeniu, że oprócz zmiennej objaśnianej również i zmienne objaśniające są losowe. Estymatorem Nadaraya-Watsona wartości funkcji regresji $r(x)$ (funkcji regresji pierwszego rodzaju) zmiennej losowej Y względem zmiennej losowej X jest zmienna losowa postaci (zob. [Domański, Pruska 2000], por. [Ćwik, Koronacki 2005]):

$$r_n(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i K\left(\dfrac{x - X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\dfrac{x - X_i}{h_n}\right)}, & \text{gdy } \sum_{i=1}^{n} K\left(\dfrac{x - X_i}{h_n}\right) \neq 0, \\ 0, & \text{gdy } \sum_{i=1}^{n} K\left(\dfrac{x - X_i}{h_n}\right) = 0, \end{cases}$$

gdzie h_n jest pewnym ciągiem liczb dodatnich nazywanym **współczynnikiem wygładzającym**, a funkcja $K(t)$, zwana **jądrem**, jest funkcją spełniającą warunki:

$$\sup_{-\infty < t < \infty} K(t) < \infty,$$

$$\lim_{|t| \to \infty} |t| K(t) = 0,$$

$$K(t) = K(-t),$$

$$\int_{-\infty}^{\infty} t^2 K(t) dt < \infty.$$

Estymator Nadaraya-Watsona łączy w sobie idee lokalnego wygładzania i estymacji jądrowej. Zauważmy, że estymacja opiera się w tym przypadku tylko i wyłącznie na wygładzaniu obserwowanych wartości y_i za pomocą funkcji jądra, bez lokalnej aproksymacji estymowanej funkcji regresji wielomianem lub inną prostą funkcją (tak jak to ma miejsce w przypadku kubicznej funkcji sklejanej).

Przyjrzyjmy się teraz, jak w literaturze przedmiotu wykorzystywany jest estymator jądrowy do wyznaczenia dziennego składnika periodycznego dla czasów trwania. Odwzorowanie wewnątrzdziennej sezonowości jest szacowane jako estymator jądrowy funkcji regresji czasów trwania względem odpowiadających im sekund w ciągu dnia (zob. [Bauwens, Veredas 2004, s. 398]):

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

gdzie: t – liczba sekund od północy każdego dnia (ewentualnie od początku sesji),
x_i – czasy trwania odpowiadające momentom t_i (x_i jest zmienną zależną),
t_i – liczba sekund od północy każdego dnia (ewentualnie od początku sesji) do momentu wystąpienia danej transakcji,
K – funkcja jądra,
h_n – współczynnik wygładzania (*bandwidth*),
s – odchylenie standardowe wartości t_i,
n – liczba obserwacji.

Jeśli chodzi o funkcję jądra, to w pracy [Bauwens, Veredas 2004, s. 398] użyto jądra *quartic* (z optymalnym współczynnikiem wygładzania równym $2{,}78sn^{-1/5}$) postaci:

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

a w pracy [Bauwens, Giot 2002, s. 13] posłużono się funkcją jądra gamma.

W przypadku pracy [Bauwens, Veredas 2004, s. 398] oszacowanie funkcji wewnątrzdziennej sezonowości przeprowadzono osobno dla każdego dnia tygodnia, aby uwzględnić ewentualną sezonowość wynikającą z okresowości transakcyjnej również w okresie tygodniowym. Zastosowano więc analogiczne podejście jak w pracy [Bauwens, Giot 2000].

5. Uwagi końcowe

Celem pracy było omówienie najważniejszych własności finansowych danych transakcyjnych oraz prezentacja metod modelowania i oszacowywania czynnika wewnątrzdziennej sezonowości aktywności transakcyjnej. W praktyce istnieje wiele możliwości estymacji odwzorowania wewnątrzdziennej sezonowości, ale żadna z nich nie dominuje nad innymi w sensie własności statystycznych. Dwa najczęściej stosowane podejścia to estymacja kubicznymi funkcjami sklejanymi (ewentualnie kubicznymi wygładzonymi funkcjami sklejanymi) oraz estymacja jądrowa. Obie te metody szczegółowo zostały omówione w opracowaniu. Oprócz tych podejść w literaturze przedmiotu do szacowania dziennego składnika periodycznego proponuje się zastosowanie aproksymacji szeregami Fouriera (interpolacji trygonometrycznej). Deterministyczny czynnik wewnątrzdziennej sezonowości przyjmuje wówczas postać (zob. [Fantazzini 2004, s. 168]):

$$\varphi(t_i^*) = \delta \cdot t_i^* + \sum_{q=1}^{Q}[\delta_{c,q} \cdot \cos(t_i^* \cdot 2\pi q) + \delta_{s,q} \cdot \sin(t_i^* \cdot 2\pi q)],$$

gdzie: $t_i^* \in [0,1]$ – liczba sekund od otwarcia sesji aż do momentu pojawienia się i-tej transakcji podzielona przez czas trwania sesji w sekundach,

$\delta, \delta_{c,q}, \delta_{s,q}$ – parametry do oszacowania,

Q – stopień wielomianu.

Jeszcze inne podejście można znaleźć u Tsaya w pracy [Tsay 2002]. Wykorzystuje się tam funkcje kwadratowe i funkcje charakterystyczne zbiorów (zob. [Tsay 2002, s. 225]). Autor proponuje, aby multiplikatywny czynnik wewnątrzdziennej sezonowości wyznaczyć następująco:

$$\phi(t_i) = \exp(d(t_i)), \qquad d(t_i) = \beta_0 + \sum_{j=1}^{7} \beta_j f_j(t_i),$$

gdzie: $f_1(t_i), f_2(t_i), f_3(t_i), f_4(t_i)$ – proste funkcje kwadratowe określone przez badacza po wstępnej analizie danych,

$f_5(t_i)$ – funkcja charakterystyczna zbioru dla pierwszych 5 minut po otwarciu sesji ($f_5(\cdot) = 1$ wtedy i tylko wtedy, gdy t_i jest w pierwszych 5 minutach po otwarciu sesji),

$f_6(t_i)$ – funkcja charakterystyczna zbioru dla drugich 5 minut po otwarciu sesji,

$f_7(t_i)$ – funkcja charakterystyczna zbioru dla ostatnich 30 minut trwania sesji.

Współczynniki $\beta_0, \beta_1, \ldots, \beta_7$ są estymowane metodą najmniejszych kwadratów dla postaci liniowej

$$\ln(\Delta t_i) = \beta_0 + \sum_{j=1}^{7} \beta_j f_j(t_i) + \varepsilon_i.$$

Widzimy więc, że jest kilka propozycji estymacji odwzorowania wewnątrzdziennej sezonowości. Jeśli chodzi o podejścia szczegółowo omówione w opracowaniu, to w przypadku zastosowania funkcji sklejanych pewnym mankamentem może być początkowe uśrednianie czasów trwania pomiędzy zdarzeniami dla kolejnych godzin lub połówek godzin. Stopień agregacji danych oraz elastyczność estymowanej funkcji można minimalizować przez zwiększenie liczby węzłów dla splajnu. Z kolei w przypadku estymacji jądrowej funkcji regresji problemem może być odpowiedni dobór funkcji jądra, które wpływa na stopień wygładzenia czasów trwania, oraz wyznaczenie optymalnej wartości współczynnika wygładzania. Wydaje się, że naturalnym krokiem jest włączenie modeli wewnątrzdziennej sezonowości (zwłaszcza parametrycznych) do modeli podstawowych, zamiast uprzedniego filtrowania danych.

Poniższa praca ma charakter teoretyczny. Empiryczna weryfikacja prezentowanych metod na podstawie danych z polskiego rynku giełdowego będzie prezentowana w kolejnych pracach i badaniach autora. W części empirycznej oba prezentowane podejścia będą porównane pod względem skuteczności eliminacji wpływu zjawiska sezonowości na autokorelację czasów trwania oraz z punktu widzenia złożoności obliczeniowej.

Literatura

Bauwens L., Giot P., *Asymmetric ACD Models: Introducing Price Information in ACD Models*, CORE Discussion Paper 2002 no 9844.

Bauwens L., Giot P., *The Logarithmic ACD Model: an Application to the Bid-Ask Quote Process of Three NYSE Stocks*, „Annales d'Économie et de Statistique" 2000 nr 60, s. 117-149.

Bauwens L., Veredas D., *The Stochastic Conditional Duration Model: a Latent Variable Model for the Analysis of Financial Durations*, „Journal of Econometrics" 2004 no 119, s. 381-412.

Bień K., *Model ACD – podstawowa specyfikacja i przykład zastosowania*, „Przegląd Statystyczny" 2006 t. 53, nr 3.

Ćwik J., Koronacki J., *Statystyczne systemy uczące się*, Wydawnictwa Naukowo-Techniczne, Warszawa 2005.

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

Doman M., Doman R., *Ekonometryczne modelowanie dynamiki polskiego rynku finansowego*, AE, Poznań, 2004.

Domański C., Pruska K., *Nieklasyczne metody statystyczne,* Polskie Wydawnictwo Ekonomiczne, Warszawa 2000.

Engle R.F., Russell J.R., *Autoregressive Conditional Duration: a New Model for Irregularly Spaced Transaction Data*, „Econometrica" 1997 no 66, s.1127-1162.

Engle R.F., *The Econometrics of Ultra-High-Frequency Data*, „Econometrica" 2000 no 68, s. 1-22.

Fantazzini D., *Financial Markets Microstructure and High Frequency Data: Theoretical Issues, Stylized Facts and Econometric Tools*, D.U. Press, Bologna 2004.

Gatnar E., Walesiak M., *Statystyczna analiza danych z wykorzystaniem programu R*, Wydawnictwo Naukowe PWN, Warszawa 2009.

Giot P., *Market Risk Models for Intraday Data*, „European Journal of Finance" 2005 no 11, s. 187-212.

Hastie T., Tibshirani R., Friedman J.H., *The Elements of Statistical Learning*, Springer-Verlag, New York 2001.

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

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

Osińska M., *Ekonometria finansowa*, Polskie Wydawnictwo Ekonomiczne, Warszawa 2006.

Roll R., *A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market*, „The Journal of Finance" 1984 t. 39, no 4.

Stoer J., *Wstęp do metod numerycznych*, Polskie Wydawnictwo Naukowe, Warszawa 1979.

Strickland C.M., Forbes C.S., Martin G.M., *Bayesian Analysis of the Stochastic Conditional Duration Model*, Working Paper 14/2003, Monash University, Australia 2003.

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

ESTIMATION METHODS OF INTRADAY SEASONALITY IN TRANSACTION FINANCIAL DATA ANALYSIS

Summary: The aim of this article is to present special characteristics of the ultra-high-frequency financial data and the estimation methods of intraday seasonality of trading activity. Ultra-high-frequency financial data (transactions data or tick-by-tick data) are defined as a full record of transactions and their associated characteristics. In this case we study the transaction arrival times and accompanying measures.

Ultra-high-frequency data have some unique characteristics that do not appear in lower frequencies. In particular, the author discusses unequally spaced time intervals, discrete price changes and the existence of intraday seasonality of time duration between transactions. The author presents two main estimation methods of intraday seasonality: natural cubic splines and kernel estimation.

In case spline smoothing diurnal factor is determined by first averaging the durations between transactions over 30 minute or 60 minute intervals for each day and then fitting a cubic splines with nodes at each half-hour or an hour. In the second case a Nadaraya-Watson kernel estimator of regression of the raw duration on the time of day is used.