Zeszyty Naukowe
Wyższej Szkoły Bankowej we Wrocławiu
Nr 7(45)/2014

Tadeusz Kufel, Marcin Błażejowski, Paweł Kufel

Wyższa Szkoła Bankowa w Toruniu

Obszary modelowania niefinansowych danych transakcyjnych (tickowych)

Streszczenie. W artykule zostały przedstawione obszary modelowania niefinansowych danych transakcyjnych (tickowych). Niefinansowe dane transakcyjne dotyczące danych sprzedażowych – paragonowych – wykorzystywane są w analizie koszykowej. Analiza czasów trwania obsługi klienta, czyli obsługi transakcji, a także przerw w obsłudze klienta, pozwala na optymalizację pracy kanałów obsługi klientów. Analiza cen transakcyjnych sprzedaży nieruchomości i pojazdów pozwala na ocenę roli poszczególnych cech nieruchomości (pojazdów) i wyznaczanie cen ofertowych. Normowanie danych transakcyjnych poprzez agregowanie do danych o jednakowych jednostkach czasu (dane minutowe, godzinowe, dobowe, tygodniowe) pozwala na ocenę natężenia zjawiska w czasie, co umożliwia ocenę cykliczności zjawisk.

Słowa kluczowe: dane transakcyjne, dane tickowe, analiza koszykowa, analiza cen transakcyjnych, analiza czasów trwania, agregacja danych

Wstęp

Celem artykułu jest przedstawienie obszarów badawczych statystyczno-ekonometrycznego modelowania niefinansowych danych transakcyjnych (tickowych).

Dane o ultrawysokiej częstotliwości obserwowania definiowane są jako pełne rekordy transakcyjne wraz z towarzyszącymi im charakterystykami (moment realizacji, składniki i wartość transakcji). Przykładami danych są pełne rekordy

transakcyjne wraz z towarzyszącymi im charakterystykami, dotyczącymi informacji sprzedażowych. Kojarzone są one z transakcjami zakupu i sprzedaży akcji na giełdach papierów wartościowych[1].

Niefinansowe dane transakcyjne, o cechach danych tickowych, dotyczą rejestracji danych sprzedażowych – paragonowych (np. w supermarketach), faktur sprzedaży, danych rejestrujących operacje bankomatowe, a także danych dotyczących transakcji sprzedaży nieruchomości i pojazdów.

1. Modelowanie danych transakcyjnych – paragonowych Analiza koszykowa

Dane transakcyjne – paragonowe, to baza rekordów zawierająca zbiór danych sprzedażowych, zrealizowanych przez poszczególnych klientów. Dane sprzedażowe – paragonowe – dla pojedynczego paragonu obejmują informacje o momencie sprzedaży (data, godzinna, minuta i sekunda), nazwie towaru (kod towaru), ilości sprzedaży (rodzaj jednostki ilości: liczba sztuk lub waga produktu), cenie jednostkowej, wartości sprzedaży, udziale podatku VAT. Baza danych paragonowych jest podstawą wykonania analizy koszykowej. „Celem analizy koszykowej jest wykrywanie zależności ukrytych w bazach danych i przedstawienie ich w postaci prostych reguł. Reguły te mogą dotyczyć na przykład zwyczajów zakupowych klientów lub prawidłowości w korzystaniu z usług wybranego typu. Otrzymane reguły noszą nazwę reguł asocjacyjnych”[2]. Analiza koszykowa najczęściej dotyczy analizy danych transakcyjnych pochodzących z supermarketów[3]. Odkrywanie reguł zakupu produktów w supermarkecie, czyli zachowania się konsumentów, w tym ich reakcji na promocje czy bonusy, jest także celem

[1] Por. R.F. Engle, *The Econometrics of Ultra-High Frequency Data*, „Econometrica” 2000, Econometric Society, t. 68(1), s. 1-22; T.G. Anderson, T. Bollerslev, F.X. Diebold, P. Labys, *Modeling and Forecasting Realized Volatility*, „Econometrica” 2003, t. 71, nr 2, s. 579-625; M. Sokalska, *Modelowanie zmienności stóp zwrotu danych finansowych o wysokiej częstotliwości*, Dynamiczne Modele Ekonometryczne, Wyd. UMK, Toruń 2005, s. 137-142.

[2] A. Pasztyła, *Analiza koszykowa danych transakcyjnych – cele i metody*, „Systemy IT” 2005, s. 51-54, www.statsoft.pl/pdf/artykuly/basket.pdf [12.08.2013].

[3] Zob. A. Pasztyła, *Analiza koszykowa...*; A. Pasztyła, *Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej*, StatSoft Polska Sp. z o.o., Kraków 2005, s. 55-66; M. Kukliński, M. Śniegocka-Łusiewicz, *Miary asocjacji w analizie koszykowej – przykład empiryczny*, „Acta Universitatis Nicolai Copernicum, Ekonomia”, zeszyt specjalny *Dynamiczne modele ekonometryczne*, Toruń 2009, s. 307-316; M. Kukliński, *Studium porównawcze wybranych algorytmów analizy koszykowej*, w: *Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Wyd. UE we Wrocławiu, Wrocław 2011, s. 603-611.

analizy koszykowej[4]. Opis metod analizy koszykowej, najczęściej sprowadzający się do wyznaczania reguł asocjacji na bazie milionowych zbiorów rekordów, został przedstawiony w pracach Larose'a[5]. Analiza koszykowa jest dzisiaj już samodzielnym obszarem badawczym.

Badanie preferencji zakupowych klientów dotyczy dwóch obszarów badawczych. Pierwszy obszar to odpowiedź na pytanie, co klienci wkładają do koszyka w sklepie, czyli analiza struktury zakupów. Drugi obszar badawczy analizy koszykowej pozwala uzyskać reguły zakupów typu „jeśli klient kupuje piwo i kupuje kabanosy, to często kupi musztardę", co umożliwia np. zoptymalizowanie układu towarów w sklepie, a także organizowanie akcji promocyjnych i bonusowych.

2. Modelowanie czasów trwania transakcji

Cechą danych tickowych jest rejestracja momentu realizacji transakcji oraz innych składników tej transakcji. W wielu bazach transakcyjnych rejestrowany jest moment rozpoczęcia i zakończenia transakcji, z dokładnością co do sekundy. Dotyczy to między innymi transakcji obsługi klienta przy kasie w supermarketach czy transakcji wypłat w bankomatach. Analiza czasów trwania transakcji pozwala na ocenę płynności obsługi klienta oraz czasów przerw pomiędzy kolejnymi transakcjami. Wyszukiwanie okresów z brakiem przerw pomiędzy transakcjami jest identyfikowane jako okresy o wzmożonych zakupach przez klientów, co pozwala na właściwe ustalenie liczby kanałów obsługi klienta.

W pracy Błażejowskiego[6] przedstawiono szerokie analizy czasów trwania operacji bankomatowych, a właściwie czasów trwania odstępów pomiędzy transakcjami, co pozwoliło odpowiedzieć na pytanie: kiedy występują okresy natężenia wypłat i tworzenia się kolejek w obsłudze klienta. Wykorzystanie baz danych transakcyjnych z kilkuletniego okresu pozwala ocenić złożoność występowania okresów natężenia operacji.

---

[4] F. Hansen, *Consumer Choice Behavior. A Cognitive Theory*, The Free Press, New York 1972; L. Rudnicki, *Zachowania konsumentów na rynku*, PWE, Warszawa 2000.

[5] D.T. Larose, *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, WN PWN, Warszawa 2006; idem, *Metody i modele eksploracji danych*, WN PWN, Warszawa 2008.

[6] M. Błażejowski, *Symulacyjna ocena mocy testu Dickeya na stacjonarność procesów o wysokiej częstotliwości obserwowania – przykłady dla danych godzinowych*, „Roczniki Naukowe, Seria B – Nauki Ekonomiczne i Informatyka", nr 1-2, Wyd. Uczelniane WSIiE TWP, Olsztyn 2010.

3. Modelowanie cen transakcyjnych

Kolejnym przykładem danych tickowych są transakcje sprzedaży nieruchomości i pojazdów. Cechy transakcji zakupu nieruchomości to: moment zakupu, wartość transakcji wraz cechami charakterystycznymi nieruchomości, np. przeznaczenie (mieszkaniowe, usługowe, produkcyjne), wielkość (m.kw.), powierzchnia działki, położenie itp. Podobne cechy transakcji można wskazać przy zakupie pojazdów, to jest: cena, marka pojazdu, rok produkcji, typ silnika, pojemność silnika, przeznaczenie (osobowy/ciężarowy) itp.

Modelowanie cen nieruchomości lub pojazdów ma na celu wyznaczenie ceny oczekiwanej ze względu na szeroki zestaw cech charakteryzujących nabywaną nieruchomość lub samochód, która mogłaby być ceną ofertową[7]. Modelowanie danych o zrealizowanych transakcjach pozwala na ocenę roli charakterystycznych cech nieruchomości, takich jak: lokalizacja (dzielnica), piętro, garaż itp.[8] lub przy sprzedaży samochodów cech takich jak: rocznik, przebieg czy marka.

Modelowanie ceny nieruchomości mieszkalnej w zależności od powierzchni lokalu mieszkalnego można wykonać, testując postać analityczną liniową lub kwadratową, zakładając, że wraz ze wzrostem powierzchni mieszkania cena jednostkowa ulega zmniejszeniu.

Powyższy typ analizy cen nieruchomości można także wykonać na bazie zrealizowanych transakcji zakupu/sprzedaży samochodów, a wyznaczone przez modele ceny można wykorzystać do oceny wartości pojazdu dla celów ubezpieczeniowych.

4. Agregowanie danych transakcyjnych

Kolejną sferą wykorzystania danych tickowych jest analiza natężenia zjawiska w czasie. Cechy charakterystyczne danych tickowych to:

– precyzyjnie określony moment realizacji (data, godzina, minuta, sekunda),

– różne odstępy czasu pomiędzy kolejnymi transakcjami.

Druga cecha – różne odstępy czasu pomiędzy kolejnymi transakcjami – powoduje, że analiza natężenia zjawiska w jednej jednostce czasu jest możliwa po agregowaniu transakcji odbywających się w danej jednostce czasu.

---

[7] T. Kotrasiński, *Dane transakcyjne i ofertowe w analizie rynku i wycenie nieruchomości*, Realexperts Sp. z o.o., Warszawa 2011, www.realexperts.pl [12.08.2013].

[8] Por. K. Kiczek, *Wpływ atrybutów nieruchomości na cenę rynkowa nieruchomości mieszkaniowych w Płocku*, praca magisterska, promotor: T. Kufel, Wyższa Szkoła Bankowa w Toruniu, Toruń 2011.

Dane tikowe dla cechy typu „ilość" mają charakter strumienia, co powoduje, że sumujemy ilości transakcji. Dla kategorii ceny jednostkowej – zmiennej momentów – należy określić metodę agregacji, to jest wyboru: pierwszej czy ostatniej ceny transakcji w okresie agregacji albo ceny średniej ważonej z transakcji w okresie agregacji.

Agregacja, czyli normowanie natężenia zjawiska na jednakową jednostkę czasu, odbywa się najczęściej na następujące okresy:

– minutowe,
– 5-, 10- lub 15-minutowe,
– półgodzinne,
– godzinne,
– dobowe,
– tygodniowe.

4.1. Modelowanie złożenia cykliczności

Ekonometryczne modelowanie natężenia zjawiska w czasie dla danych o wysokiej częstotliwości przedstawiono w pracy Kufla[9]. Bardzo ważnym elementem w modelowaniu natężenia zjawiska jest modelowanie złożenia cykliczności kalendarzowych, które dla danych godzinowych lub krótszych okresów (1-, 5-, 10-, 15- lub 30-minutowych) umożliwia wyznaczenie złożenia cykliczności: dobowej, tygodniowej, miesięcznej oraz rocznej. We wspomnianej pracy przedstawiono też szerokie zestawy metod wyznaczania amplitud cykliczności zjawisk o stałych i zmiennych amplitudach wraz z ilustracją dla danych dziennych, dotyczących sprzedaży detalicznej w supermarkecie, operacji wypłat i wpłat bankowych oraz dla danych godzinowych procesów wypłat bankomatowych.

Największą użyteczność posiadają ekonometryczne modele wykorzystujące zestawy zmiennych zero-jedynkowych. Dla szacowania cykli rocznych dla danych godzinowych, dziennych czy tygodniowych preferowane jest wykorzystanie składowych harmonicznych o długich okresach, to jest 12-, 6-, 4-, 3-, a także 2-miesięcznym[10]. Pozostałe cykle o krótszych okresach można modelować za pomocą zestawów zmiennych zero-jedynkowych.

Powyższe modele opisu natężenia zjawiska dla danych o wysokiej częstotliwości obserwowania mogą być podstawą prognozowania powyższych zjawisk,

---

[9] T. Kufel, *Ekonometryczna analiza cykliczności procesów o wysokiej częstotliwości obserwowania.* WN Uniwersytetu Mikołaja Kopernika, Toruń 2010.

[10] Por. T. Kufel, *Składowe harmoniczne w modelowaniu cykliczności zjawisk ekonomicznych dla danych o wysokiej częstotliwości*, w: *Prognozowanie w zarządzaniu firmą*, red. P. Dittmann, J. Szanduła, Wyd. AE we Wrocławiu, Wrocław 2008, s. 87-97.

a także budowy scenariuszy akcji promocyjnych (upustów cenowych) dla grupy asortymentów w supermarketach[11].

4.2. Modelowanie efektów kalendarzowych

Modelowanie efektów kalendarzowych związane jest z poszukiwaniem prawidłowości obserwowanych w procesach gospodarczych, a związanych z występowaniem nietypowych układów kalendarzowych, np. długich weekendów czy ruchomych dni wolnych lub świątecznych. Te nietypowe układy kalendarzowe skutkują zmianami w typowych zachowaniach klientów. Na przykład zakup pieczywa na długie weekendy zdeterminowany jest liczbą dni wolnych, momentem rozpoczęcia (np. środa, czwartek czy piątek), w tym także wstrzymaniem pracy hipermarketów (13 dni świątecznych w roku bez handlu), co skutkuje zmiennym popytem – trudnym do prognozowania[12].

Kolejnym obszarem badawczym w modelowaniu efektów kalendarzowych jest analiza nieterminowych spłat należności za faktury w firmach. Najczęściej nieterminowe spłaty należności mają charakter krótkoterminowego opóźnienia (kilkudniowego) wynikającego z przypadania spłaty należności na piątek, sobotę lub niedzielę, co skutkuje wzrostem wpłat w poniedziałki i wtorki, ale opóźnienie spłaty może uleć wydłużeniu, ponieważ kumuluje się z innymi ważnymi przyczynami spłaty zobowiązań w firmie dłużnika, takimi jak: wpłata do ZUS, wpłata należności za VAT, podatku dochodowego czy wypłata wynagrodzeń. Wymienione powyżej efekty kalendarzowe powodujące nieterminowe spłaty należności są możliwe do ekonometrycznego modelowania za pomocą zestawów zmiennych zero-jedynkowych, a zarazem do ich prognozowania. Oszacowane efekty kalendarzowe można interpretować jako efekt poniedziałku (zawsze dodatni), efekt VAT, ZUS czy wypłaty wynagrodzeń (zawsze ujemny) w spłacie należności badanego kontrahenta, które korygują oczekiwaną kwotę spłaty wyliczoną przez system finansowo-księgowy w firmie.

Wysoka zmienność wartości średniej procesów niefinansowych ujawnia się w modelowaniu występowaniem zmienności wariancji procesu resztowego (heteroskadastycznością reszt), którą można modelować za pomocą modeli klasy GARCH, co przedstawiono w pracy Kufla[13].

---

[11] Por. M. Błażejowski, *Ekonometryczne modelowanie popytu konsumpcyjnego na podstawie danych dziennych,* WN Uniwersytetu Mikołaja Kopernika, Toruń 2009.

[12] Por. T. Kufel, *Dane dzienne w modelowaniu ekonometrycznym (przykład empiryczny)*, w: *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, red. A. Zeliaś, Wyd. AE w Krakowie, Kraków 1998, s. 141-156.

[13] T. Kufel, *Ekonometryczna analiza...*

Testowanie występowania pierwiastka jednostkowego dla danych o wysokiej częstotliwości (godzinowych, dobowych, tygodniowych i innych), czyli dla cykli składających się z okresów równych d = 5, 6, 7, 24, 48, 52, 365 przedstawiono w pracach Dickeya[14] oraz Błażejowskiego[15].

Podsumowanie

Przedstawione obszary modelowania niefinansowych danych transakcyjnych (tickowych) dotyczą:

– analizy struktury zakupów, w tym analizy koszykowej, na podstawie danych transakcyjnych z supermarketów (sprzedaży detalicznej i hurtowej),

– analizy czasów trwania obsługi klienta, a także przerw w obsłudze klienta, co pozwala na optymalizację pracy kanałów obsługi klientów,

– analizy cen transakcyjnych sprzedaży nieruchomości i pojazdów, co pozwala na ocenę roli poszczególnych cech nieruchomości (pojazdów) i wyznaczanie cen ofertowych,

– normowania danych transakcyjnych poprzez ich agregowanie do danych o jednakowych jednostkach czasu (dane minutowe, godzinowe, dobowe, tygodniowe), co pozwala na ocenę natężenia zjawiska w czasie i umożliwia ocenę złożoności cykliczności zjawisk poprzez wyznaczanie cykli: dobowych, tygodniowych, miesięcznych i rocznych,

– budowy scenariuszy akcji promocyjnych (upustów cenowych) dla sprzedaży detalicznej w supermarketach,

– analizy efektów kalendarzowych w sprzedaży towarów i usług, w szczególności efektów tzw. długich weekendów, wpływających na wielkość i strukturę zakupów,

– danych transakcyjnych zawartych na fakturach sprzedaży (m.in. termin spłaty należności) i pozwalają oszacować efekty nieterminowych spłat należności, tzw. efekt kumulacji wpłat w poniedziałki, efekt spadku spłat należności za towary wynikający z terminów wpłat należności do ZUS, za podatek dochodowy czy VAT.

Niefinansowe dane transakcyjne, jako dane o najwyższej szczegółowości rejestracji zdarzeń gospodarczych, umożliwiają tworzenie kolejnych obszarów badawczych zjawisk ekonomicznych.

---

[14] D. Dickey, *Stationarity Testing in High-Frequency Seasonal Time Series*, w: *SAS Global Forum 2009, Statistics and Data Analysis*, s. 233-244. Paper 235-2009.

[15] M. Błażejowski, *Symulacyjna ocena mocy testu Dickeya na stacjonarność procesów o wysokiej częstotliwości obserwowania – przykłady dla danych godzinowych*, „Roczniki Naukowe, Seria B – Nauki Ekonomiczne i Informatyka", nr 1-2, Wyd. Uczelniane WSIiE TWP, Olsztyn 2010; M. Błażejowski, *Symulacyjna ocena mocy testu Dickeya na stacjonarność procesów o wysokiej częstotliwości obserwowania*, „Zeszyty Naukowe WSB we Wrocławiu" 2011, nr 20.

Literatura

Anderson T.G., Bollerslev T., Diebold F.X., Labys P. , *Modeling and Forecasting Realized Volatility*, „Econometrica" 2003, t. 71, nr 2.

Błażejowski M., *Ekonometryczne modelowanie popytu konsumpcyjnego na podstawie danych dziennych,* WN Uniwersytetu Mikołaja Kopernika, Toruń 2009.

Błażejowski M., *Symulacyjna ocena mocy testu Dickeya na stacjonarność procesów o wysokiej częstotliwości obserwowania – przykłady dla danych godzinowych*, „Roczniki Naukowe, Seria B – Nauki ekonomiczne i Informatyka", nr 1-2, Wyd. Uczelniane WSIiE TWP, Olsztyn 2010.

Błażejowski M., *Wykorzystanie danych tickowych jako narzędzia oceny skuteczności obsługi konsumentów*, w: *Osiągnięcia i perspektywy modelowania i prognozowania zjawisk społeczno-gospodarczych*, red. B. Pawełek, Wyd. UE w Krakowie, Kraków 2011.

Błażejowski M., *Symulacyjna ocena mocy testu Dickeya na stacjonarność procesów o wysokiej częstotliwości obserwowania*, „Zeszyty Naukowe WSB we Wrocławiu" 2011, nr 20.

Dickey D., *Stationarity Testing in High-Frequency Seasonal Time Series*, w: *SAS Global Forum 2009, Statistics and Data Analysis*, paper 235-2009.

Engle, R.F., *The Econometrics of Ultra-High Frequency Data*, „Econometrica", Econometric Society 2000, t. 68(1).

Hansen F., *Consumer Choice Behavior. A Cognitive Theory*, The Free Press, New York 1972.

Kiczek K., *Wpływ atrybutów nieruchomości na cenę rynkowa nieruchomości mieszkaniowych w Płocku*, praca magisterska, promotor: T. Kufel, Wyższa Szkoła Bankowa w Toruniu, Toruń 2011.

Kotrasiński T., *Dane transakcyjne i ofertowe w analizie rynku i wycenie nieruchomości*, Realexperts Sp. z o.o., Warszawa 2011, www.realexperts.pl [12.08.2013].

Kufel T., *Dane dzienne w modelowaniu ekonometrycznym (przykład empiryczny)*, w: *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych*, red. A. Zeliaś, Wyd. AE w Krakowie, Kraków 1998.

Kufel T., *Składowe harmoniczne w modelowaniu cykliczności zjawisk ekonomicznych dla danych o wysokiej częstotliwości*, w: *Prognozowanie w zarządzaniu firmą*, red. P. Dittmann, J. Szanduła, Wyd. AE we Wrocławiu, Wrocław 2008.

Kufel T., *Ekonometryczna analiza cykliczności procesów o wysokiej częstotliwości obserwowania*, WN Uniwersytetu Mikołaja Kopernika, Toruń 2010.

Kukliński M., *Studium porównawcze wybranych algorytmów analizy koszykowej*, w: *Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Wyd. UE we Wrocławiu, Wrocław 2011.

Kukliński M., Śniegocka-Łusiewicz M., *Miary asocjacji w analizie koszykowej – przykład empiryczny*, „Acta Universitatis Nicolai Copernicum, Ekonomia", zeszyt specjalny *Dynamiczne modele ekonometryczne*, Toruń 2009.

Larose D.T., *Metody i modele eksploracji danych*, WN PWN, Warszawa 2008.

Larose D.T., *Odkrywanie wiedzy z danych. Wprowadzenie do eksploracji danych*, WN PWN, Warszawa 2006.

Pasztyła A., *Analiza koszykowa danych transakcyjnych – cele i metody*, „Systemy IT" 2005, www.statsoft.pl/pdf/artykuly/basket.pdf [12.08.2013].

Pasztyła A., *Przykład badania wzorców zachowań klientów za pomocą analizy koszykowej*, StatSoft Polska Sp. z o.o., Kraków 2005.

Rudnicki L., *Zachowania konsumentów na rynku*, PWE, Warszawa 2000.

Sokalska M., *Modelowanie zmienności stóp zwrotu danych finansowych o wysokiej częstotliwości*, Dynamiczne Modele Ekonometryczne, Wyd. UMK, Toruń 2005.

Śniegocka-Łuniewska M., *Analiza koszykowa w badaniu cykliczności reguł asocjacji w handlu detalicznym*, w: *Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania*, red. K. Jajuga, M. Walesiak, Wyd. UE we Wrocławiu, Wrocław 2011.

Modelling nonfinancial transaction (tick) date

Summary. In a paper we present the areas of possible analysis of non financial transaction (tick) data. Non financial transaction data mostly concern sales data – single receipt records which can be used in affinity analysis. Duration analysis of customer service time (transaction time) and also breaks in customer service let optimize the work of customer service channels. Analysis of transaction prices on real estate and car markets enables to assess the role (weight) of particular real estate (car) price and to estimate bid prices. Normalising transaction data by aggregating to time-series (minute, hourly, daily, weekly) enables to assess the intensity of these phenomena in time and to estimate their cycles.

Key words: transaction data, tick data, affinity analysis, duration analysis, aggregation data