Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 390 · 2019

Informatyka i Ekonometria 15

Krzysztof Ćwikliński

Ministerstwo Finansów[1]
Departament Analiz
Uniwersytet Ekonomiczny we Wrocławiu
Wydział Ekonomii i Finansów
Katedra Ekonometrii i Badań Operacyjnych
krzysztof.cwiklinski@ue.wroc.pl

MODELOWANIE I DEKOMPOZYCJA SZEREGÓW CZASOWYCH AKTUALIZACJI JEDNOLITYCH PLIKÓW KONTROLNYCH

Streszczenie: Modelowanie szeregów czasowych stało się niezbędne w procesie kontrolowania procesów zachodzących w systemach informatycznych Ministerstwa Finansów RP. Wymierne w sensie finansowym są problemy braku lub niepełnej aktualizacji relacyjnej bazy danych JPK_VAT w akceptowalnym przez prawo terminie. W tym przypadku niezwykle ważna okazuje się umiejętność zastosowania nie tylko klasycznych modeli uwzględniających składniki sezonowe (np. SARIMA), ale także złożone składniki systematyczne (BATS/TBATS). Dokonano analizy szeregów czasowych pod kątem występowania składników systematycznych, estymowano parametry strukturalne modeli, otrzymano i zestawiono wyniki testów wskazujące na konieczność zastosowania modelu TBATS.

Słowa kluczowe: Jednolity Plik Kontrolny, analiza szeregów czasowych, dekompozycja, prognozowanie, BATS/TBATS, SARIMA.

JEL Classification: C32, C53.

Wprowadzenie

Jeden z najważniejszych obowiązków płatnika podatku od towarów i usług stanowi ewidencjonowanie danych niezbędnych do prawidłowego sporządzenia deklaracji podatkowej [*Ewidencja dla podatku od towarów i usług*…, 2018, s. 1].

[1] Wszystkie zawarte w niniejszym artykule fakty, badania i wnioski nie reprezentują stanowiska Ministerstwa Finansów lub mojego jako pracownika Ministerstwa Finansów, a jedynie stanowisko osoby prywatnej (autora artykułu).

Zaniedbanie powinności zagrożone jest sankcjami karnymi o charakterze finansowym. Artykuł drugi Ustawy z dnia 11 marca 2004 r. o podatku od towarów i usług stanowi, że [2004, art. 109, ust. 2]: „w przypadku stwierdzenia, że podatnik nie prowadzi ewidencji, o której mowa w ust. 1, albo prowadzi ją w sposób nierzetelny, a na podstawie dokumentacji nie jest możliwe ustalenie wartości sprzedaży, organ podatkowy określi, w drodze oszacowania, wartość sprzedaży opodatkowanej i ustali od niej kwotę podatku należnego".

Podmioty będące płatnikami podatku od towarów i usług dostarczają drogą elektroniczną pliki JPK_VAT do Ministerstwa Finansów, które umieszczane są w relacyjnej bazie danych w tabelach: NAGLOWEK, PODMIOT, SUMA_ KONTROLNA, SPRZEDAZ, ZAKUP. Informacje w tabelach JPK_VAT powinny być zgodne z wymogami określonymi w Ustawie o podatku od towarów i usług [2004]. Dla przykładu, w tabeli NAGLOWEK znajdują się dane dotyczące m.in. celu złożenia, daty i czasu wytworzenia przesyłanego przez płatnika pliku kontrolnego [*Ewidencja dla podatku od towarów i usług*…, 2018, s. 4].

Podstawową zaletą systemu JPK_VAT jest możliwość kontrolowania szerokiej grupy instytucjonalnych podatników, którzy są płatnikami podatku od towarów i usług [*Ewidencja dla podatku od towarów i usług*…, 2018, s. 5].

1. Uzasadnienie wyboru celu i tematu artykułu

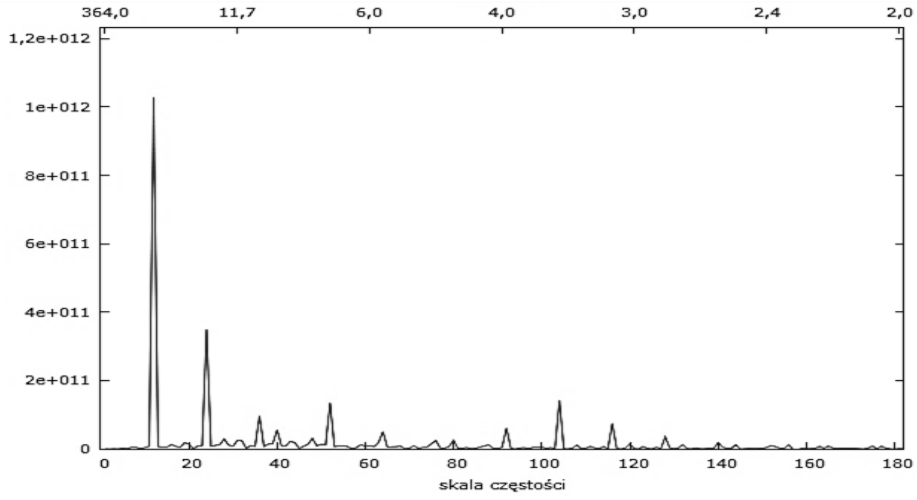
Sprawne działanie systemu ewidencjonowania służy m.in. poprawie ściągalności podatku od towarów i usług, sprzyja występowaniu dodatniego salda budżetu państwa, generuje przychody, zmniejsza tendencję do zadłużania, stanowi stymulantę poziomu życia społeczeństwa. Wymierne w sensie finansowym dla polskiego Ministerstwa Finansów są problemy braku lub niepełnej aktualizacji bazy danych JPK_VAT w akceptowalnym przez prawo terminie. Prowadzenie ewidencji w tej postaci wiąże się ze zdalnym aktualizowaniem informacji w relacyjnej bazie danych, przez którą rozumiemy uporządkowane (oparte na relacyjnym modelu danych) zbiory informacji przechowywane w pamięci komputera [Mazur, Mazur, 2004, s. 113]. Pojęcie relacji skojarzone jest z określeniem zmiennej jako obiektu, który może zmieniać swą wartość [Mazur, Mazur, 2004, s. 14, 51]. Sposób graficznej prezentacji relacji stanowi **tabela dwuwymiarowa**, której kolumny odpowiadają atrybutom relacji, wiersze zaś krotkom lub rekordom [Mazur, Mazur, 2004, s. 14, 51].

Celem głównym niniejszego artykułu jest wybór spośród trzech modeli (BATS, TBATS, SARIMA) najlepszej metody dekompozycji szeregów czasowych aktualizacji JPK_VAT i postawienie prognoz.

2. Modelowanie i dekompozycja szeregów czasowych aktualizacji bazy JPK_VAT

Weryfikacja hipotez wymaga przebadania szeregów czasowych liczby aktualizacji w okresach 18 maja 2018 r. – 17 maja 2019 r. (próba ucząca) oraz 18 maja 2019 r. – 22 lipca 2019 r. (próba testowa). Próbę uczącą stanowi 365 wartości empirycznych o dziennej częstotliwości w siedmiodniowym tygodniu. Modelowanie, dekompozycję oraz prognozowanie szeregów czasowych aktualizacji przeprowadzono w języku R, używając pakietu '*forecast*'. Przeanalizowano szeregi czasowe aktualizacji i stwierdzono, że „badane zjawisko może podlegać rozmaitym wahaniom (o różnych okresach) jednocześnie" [Dittmann, 2003, s. 83].

Wahania sezonowe wykazuje analiza spektralna (1) dla zróżnicowanych wartości szeregu czasowego aktualizacji tabeli SUMA_KONTROLNA.



Rys. 1. Periodogram zróżnicowanych wartości aktualizacji SUMA_KONTROLNA

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

Na rysunku 1 można zauważyć występowanie dwóch istotnych częstotliwości wahań sezonowych (tygodniowych oraz miesięcznych). Miesięczna sezonowość charakteryzuje się wysoką amplitudą występującą ok. 25. dnia miesiąca. W szeregach czasowych aktualizacji tabel JPK_VAT odzwierciedlony jest me-

chanizm wahań zgodny z Ustawą o podatku od wartości dodanej, która mówi, że płatnicy wysyłają pliki JPK_VAT do 25. dnia każdego miesiąca [Ustawa o podatku od towarów i usług, 2004, art. 109, ust. 2]. Wahania mają charakter złożonych składników sezonowości, których uwzględnienie w modelu wpływa na poprawienie jakości prognoz z otrzymaniem składników resztowych o charakterze losowym w szeregach pozbawionych autokorelacji [De Livera, Hyndman, Snyder, 2010, s. 6]. Przykładem nowoczesnego podejścia modelowego uwzględniającego niestacjonarność procesu stochastycznego, a także sezonowości, jest BATS (*Box-Cox Transform, ARMA Errors Trend, Seasonal Components*) [De Livera, Hyndman, Snyder, 2010, s. 9-10]:

$$y_t^{(\omega)} = \begin{cases} \dfrac{y_t^{(\omega)} - 1}{\omega} \; dla \; \omega \neq 0 \\ log y_t \end{cases} \tag{2.1}$$

$$y_t^{(\omega)} = l_{t-1} + \Phi b_{t-1} + \sum_{i=1}^{T} S_{t-m_i}^{(i)} + d_t \tag{2.2}$$

$$l_t = l_{t-1} + \Phi b_{t-1} + \alpha d_t \tag{2.3}$$

$$b_t = (1 - \Phi) b + \Phi b_{t-1} + \beta d_t \tag{2.4}$$

$$s_t^{(i)} = s_{t-m_i}^{(i)} + \gamma_i d_t \tag{2.5}$$

$$d_t = \sum_{i=1}^{p} \varphi_{t-i} d_{t-i} + \sum_{i=1}^{q} \theta_i \, \varepsilon_{t-i} + \varepsilon_t \tag{2.6}$$

gdzie: l_t jest lokalnym poziomem badanego zjawiska w okresie lub momencie *t*; *b* – trendem długookresowym; b_t – trendem krótkookresowym w okresie lub momencie *t*; $s_t^{(i)}$ – wartością składnika sezonowego w okresie *t*; α, β i γ – parametrami wygładzania; m_i – okresem sezonowym; d_t – wartościami teoretycznymi z modelu ARMA(*p*,*q*) [De Livera, Hyndman, Snyder, 2010, s. 10]. Model w formie pozwalającej na identyfikację rodzaju przekształceń i liczby parametrów przedstawia się w sposób następujący:

$$\text{BATS}(\omega, \Phi, p, q, m1, m2, \ldots, mk,) \tag{2.7}$$

gdzie: *ω* – parametr transformacji Boxa-Coxa; Φ – parametr tłumienia (*damping parameter* – odpowiadający za wpływ trendów krótko- i długookresowego na lokalny poziom badanego zjawiska); *p* oraz *q* – liczba parametrów autoregresyj-

nych (AR) i średniej ruchomej (MA) w modelu ARMA; m – liczba okresów sezonowych [De Livera, Hyndman, Snyder, 2010, s. 10].

BATS umożliwia modelowanie szeregów czasowych zawierających wiele okresów sezonowych i jest najbardziej oczywistym uogólnieniem tradycyjnych modeli sezonowych [De Livera, Hyndman, Snyder, 2010, s. 10]. Nie może jednak uwzględniać tzw. sezonowości ułamkowej i może mieć bardzo dużą liczbę stanów – początkowy składnik sezonowy zawiera wiele wartości niezerowych [por. De Livera, Hyndman, Snyder, 2010, s. 11]. W ulepszonym, trygonometrycznym BATS składniki sezonowe podlegają aproksymacji trygonometrycznymi szeregami Fouriera:

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} \tag{2.8}$$

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t \tag{2.9}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1} sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t \tag{2.10}$$

gdzie: $\lambda_1^{(i)}$ i $\lambda_2^{(i)}$ są parametrami wygładzania; $\lambda_j^{(i)} = 2\pi j/m_i$ [por. De Livera, Hyndman, Snyder, 2010, s. 11]. We wzorze (2.9) zawarty jest stochastyczny poziom *i*-tego składnika sezonowego ($s_{j,t}^{(i)}$), którego zmiany wpływają na wartości komponentu w czasie ($s_{j,t}^{*(i)}$) [por. De Livera, Hyndman, Snyder, 2010, s. 11]. Modelowanie złożonych składników sezonowych wymaga określonej liczby par szeregów Fouriera dla i-tego składnika sezonowego i oznaczona jest przez *k* (oczekiwana jest niewielka ich liczba) [por. De Livera, Hyndman, Snyder, 2010, s. 11].

Model TBATS również często jest przedstawiany w formie pozwalającej na łatwą identyfikację postaci strukturalnej:

$$\text{TBATS}(\omega, \Phi, p, q, \{m1,k1\}, \{m2,k2\}, \ldots, \{mT, kT\}) \tag{2.12}$$

gdzie: m – okresy sezonowe; k – liczba par szeregów Fouriera [por. De Livera, Hyndman, Snyder, 2010, s. 12]. Parametry modelu oszacowano metodą największej wiarygodności, wybór najlepszej postaci odbywa się w wyniku porównania wartości kryteriów informacyjnych Akaike'a dla różnych kombinacji parametrów [De Livera, Hyndman, Snyder, 2010, s. 16-22].

Rezultatem estymacji wynikających z trygonometrycznej postaci sezonowych składników w modelu jest znacznie mniejsza liczba początkowych wartości, które trzeba zastosować do modelowania częstotliwości niecałkowitych [De Livera,

Hyndman, Snyder, 2010, s. 16-22]. Model uwzględnia nieliniowości występujące w szeregach czasowych, a także wymaga prostszej procedury szacowania, umożliwia uwzględnienie autokorelacji w szeregach reszt, czego rezultatem jest uzyskanie trafnych prognoz [De Livera, Hyndman, Snyder, 2010, s. 12].

Przyjęte częstotliwości wraz z liczbą składników sezonowych są zgodne z wnioskami płynącymi z wykresu spektrum, a także znajomości wiedzy w zakresie wymogów ustawowych dotyczących dnia ewidencji podatku od towarów i usług [por. Ustawa o podatku od towarów i usług, 2004, art. 109, ust. 2].

W tabeli 1 zawartość kolumny Struktura ogólna składa się ze wzorów określających postać modeli BATS dla tabel JPK_VAT estymowanych w okresie próby uczącej. Wartości w nawiasach klamrowych {7, 30} świadczą o dekompozycji szeregów z uwzględnieniem sezonowości tygodniowej oraz miesięcznej.

Tabela 1. Ogólne postaci modelu BATS dla aktualizacji tabel bazy JPK_VAT

Lp.	Tabela JPK_VAT	Struktura ogólna
1.	NAGLOWEK	BATS(1, {0,0}, 1, {7,30})
2.	PODMIOT	BATS(1, {0,0}, 1, {7,30})
3.	SUMA_KONTROLNA	BATS(1, {0,0}, 1, {7,30})
4.	ZAKUP	BATS(1, {0,0}, 1, {7,30})
5.	SPRZEDAZ	BATS(0.483, {0,0}, 1, {7,30})

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

Z uwagi na zerowe wartości parametry ARMA (co ukazano w pierwszym nawiasie klamrowym) nie występuje konieczność korygowania wartości teoretycznych równań BATS w sposób bardziej złożony niż średnią z szeregu czasowego. Precedensem były aktualizacje SPRZEDAZ, gdzie konieczna była transformacja danych wejściowych przed właściwą estymacją; sugeruje to mniejsza od 1 wartość parametru ω (0,483). Parametr tłumienia (Φ) równy jedności oznacza, że trend krótkookresowy w okresie t zależy wprost proporcjonalnie od trendu z poprzedniego okresu (2.4).

Lokalny poziom zjawiska (l_t) zależy również od poziomu z okresu poprzedniego (l_{t-1}) lub wprost proporcjonalnie od trendu krótkookresowego *ceteris paribus* z okresu poprzedzającego obecny. Szczegółową prezentację wartości parametrów strukturalnych modelu BATS po estymacji przedstawia tabela 2. Aktualizacje SPRZEDAZ charakteryzują się innym w porównaniu do pozostałych rozkładem, co pokazuje wartość parametru ω równa 0,483. Różna od jedności wartość tego parametru oznacza, że wystąpiła transformacja Boxa–Coxa dla szeregu czasowego aktualizacji.

Tabela 2. Oszacowane wartości parametrów BATS dla tabel JPK_VAT

Parametr	NAGLOWEK	PODMIOT	SUMA_KONTROLNA	ZAKUP	SPRZEDAZ
ω	1	1	1	1	0,483
Φ	1	1	1	1	1
α	1,41583	1,41583	1,31995	1,22552	1,12595
β	–0,00500	–0,00500	0,00067	–0,00661	0,06346
$\gamma 1$	–0,00701	–0,00701	–0,01173	0,09091	–0,00013
$\gamma 2$	–0,03002	–0,03002	–0,04776	–0,04026	–0,09378

Źródło: Obliczenia własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

Cechą charakterystyczną wyników estymacji jest także istotność parametrów α i β, które odzwierciedlają wpływ wartości teoretycznych modelu ARMA na poziom lokalny zjawiska (2.3) oraz trend krótkookresowy. Przy tym parametry α są dodatnie i większe od jedności, natomiast β są w trzech przypadkach ujemne, istotne statystycznie, choć bliskie zeru. Zaobserwowano charakterystyczną właściwość: dwa identyczne zestawy wartości parametrów dla danych NAGLOWEK oraz PODMIOT.

Szeregi czasowe aktualizacji dla tych tabel powinny być identyczne ze względu na relacje, które wynikają z treści danych i z umocowania prawnego, jakie zajmują w ewidencji podatku od towarów i usług. Dla przykładu, NAGLOWEK zawiera początkową i końcową datę dotyczącą danych zawartych w tabeli PODMIOT, która również posiada informacje o numerach NIP, nazwę i adres e-mail podatnika rozliczającego się za podany okres. Wszelkie rozbieżności w aktualizacjach dwóch tabel można wytłumaczyć jedynie poważnymi błędami systemu związanymi z mechanizmem ładowania danych do bazy JPK_VAT – najczęściej występują duplikaty danych powstałe w wyniku błędnego działania systemu. W tabeli 2 zestawiono parametry wygładzania γ, które odpowiadają za modelowanie składnika sezonowego, a występują parami: jeden z nich odpowiada za modelowanie sezonowości tygodniowej ($\gamma 1$), a drugi miesięcznej ($\gamma 2$). Dziewięć na dziesięć parametrów wygładzania γ jest niższych od zera, tylko dla aktualizacji ZAKUP parametr ($\gamma 1$) jest dodatni. Nieco inaczej przedstawiają się wyniki modelowania TBATS, gdy również szacuje się parametry wygładzania odpowiadające za składniki sezonowe.

W tabeli 3 zestawiono informacje: czy wystąpiła konieczność przeprowadzenia dla szeregów czasowych transformacji Boxa–Coxa, jaki jest wpływ trendów długookresowych na poziom badanego zjawiska, a także liczbę parametrów w modelowaniu ARMA oraz składnikach sezonowych.

Tabela 3. Ogólne postaci modelu TBATS dla aktualizacji tabel bazy JPK_VAT

Lp.	Tabela JPK_VAT	Struktura ogólna
1.	NAGLOWEK	TBATS(1, {2,2}, 1, {<7,3>, <30.42,6>})
2.	PODMIOT	TBATS(1, {2,2}, 1, {<7,3>, <30.42,6>})
3.	SUMA_KONTROLNA	TBATS(1, {2,2}, 1, {<7,3>, <30.42,5>})
4.	ZAKUP	TBATS(1, {2,2}, 1, {<7,3>, <30.42,5>})
5.	SPRZEDAZ	TBATS(1, {0,0}, 1, {<7,3>, <30.42,5>})

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

Pierwszy argument dotyczy parametrów ω, które są równe jedności dla wszystkich tabel – algorytm nie poddał danych transformacji Boxa–Coxa [Brozyna i in., 2018, s. 238]. Wartość parametru tłumienia wynosi $\Phi = 1$ dla wszystkich tabel, co oznacza, że trend krótkookresowy w okresie t zależy jedynie od wartości trendu krótkookresowego z poprzedniego okresu $(t - 1)$, a nie od trendu długookresowego b w równaniu (2.4).

W tabeli 3 wartości znajdujące się w nawiasach klamrowych (po parametrze ω) odnoszą się do konfiguracji parametrów modelu, np. {2, 2} oznacza, że model miał po dwa parametry ARMA (dla części autoregresyjnej i średniej ruchomej).

ARMA nie został zastosowany do szeregu aktualizacji tabeli SPRZEDAZ (korekta równań modelu teoretycznymi wartościami z modelu ARMA nie wystąpiła). Model TBATS szacuje wahania sezonowe w sposób bardziej skomplikowany niż BATS, stosuje przybliżenia wieloma szeregami trygonometrycznymi Fouriera. Wartości w drugim nawiasie (klamrowym) dotyczą sezonowości tygodniowej, którą przybliżono trzema parami szeregów trygonometrycznych Fouriera. Sezonowość miesięczna dla dwóch tabel (NAGLOWEK i PODMIOT) modelowana była sześcioma, dla pozostałych pięcioma parami szeregów trygonometrycznych Fouriera. W rezultacie daje nam to ok. kilkudziesięciu odmiennych, co do wartości parametrów wygładzania (sezonowość tygodniowa i miesięczna). Wartości głównych parametrów γ odpowiadających za wygładzanie składników sezonowych zestawiono w następnej tabeli 4. W pierwszym wierszu tabeli 4 zawarte są dane dotyczące transformacji Boxa–Coxa (ω) – nie występowała w stosunku do tabel JPK_VAT.

Wartość parametrów tłumienia (*damping parameter* – Φ) równa była jedności – istnieje liniowy wpływ trendu krótkookresowego z okresu poprzedniego na wartość bieżącą zmiennej prognozowanej.

Tabela 4. Oszacowane wartości parametrów TBATS dla tabel JPK_VAT

Parametr	NAGLOWEK	PODMIOT	SUMA_KONTROLNA	ZAKUP	SPRZEDAZ
ω	1	1	1	1	1
Φ	1	1	1	1	1
α	0,07105	0,07105	0,06102	0,08761	1,05527
β	0,00195	0,00195	0,00166	0,00137	0,02379
$\gamma 1$	–0,00001	–0,00001	0,00004	0,00017	0,00086
$\gamma 2$	–0,00007	–0,00007	–0,00008	–0,00015	–0,01010
$\gamma 3$	0,00006	0,00006	0,00004	0,00019	0,00258
$\gamma 4$	–0,00002	–0,00002	–0,00009	–0,00016	–0,00690
$\theta 1$	0,65454	0,65454	0,66820	0,66733	–
$\theta 2$	0,07168	0,07168	0,07719	0,02027	–
$\phi 1$	0,45915	0,45915	0,47536	0,34947	–
$\phi 2$	0,22200	0,22200	0,22408	0,27919	–

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

Różnicą w stosunku do „nietrygonometrycznego" BATS było wystąpienie niezerowych konfiguracji parametrów modelu ARMA dla wszystkich (poza jedną) tabelą (SPRZEDAZ). W tym przypadku wartości teoretyczne i parametry modelu ARMA używane są do korekty wartości teoretycznych generowanych przez równania składające się na model TBATS, którego koncepcja powstała w 2010 r. Znacznie wcześniej niż w latach dwutysięcznych, bo już od lat 70. XX w. „czyste" modele ARMA oraz ich modyfikacje (ARIMA, SARIMA, ARFIMA) są regularnie wykorzystywane do dekompozycji i prognozowania szeregów czasowych (także finansowych). Szczególnym przypadkiem jest SARIMA (*Seasonal Autoregressive Integrated Moving Average*), gdy występuje różnicowanie niestacjonarnych szeregów czasowych w zakresie składnika sezonowego.

W swojej budowie SARIMA zawiera zarówno niesezonowe, jak i sezonowe pierwiastki jednostkowe, a także opóźnienia zmiennych w strukturze sezonowej oraz czasowej procesu [Osińska, 2006, s. 66-67]. „Użyteczność modelu SARIMA polega na założeniu, że cykle sezonowe nie muszą odtwarzać dokładnie tego samego przebiegu co roku, jak zakładano w modelu sezonowości deterministycznej" [Osińska, 2006, s. 67]. W procesie estymacji modelu SARIMA najpierw usuwa się pierwiastki sezonowe, a następnie niesezonowe za pomocą odpowiedniego filtru różnicującego – dwa pierwiastki odpowiadają niesezonowej częstości, jeden z nich usuwa trend, a pozostałe likwidują sezonową strukturę [Osińska, 2006, s. 66]. Postać strukturalną modelu należy rozpatrywać jako kombinację parametrów oraz ich wartości, które obrazują wystąpienie w estymacji odpowiednich przekształceń związanych z modelowaniem składników: sezonowego i niesezonowego. M. Osińska zauważa, że: „Model (SARIMA) ma

dość skomplikowaną strukturę, ponieważ zakłada zarówno zwykłe, jak i sezonowe pierwiastki jednostkowe, a ponadto opóźnienia AR i MA w strukturze czasowej procesu oraz strukturze sezonowej" [2006, s. 66-67].

Przyjmując, że *d* i *D* są nieujemnymi liczbami całkowitymi, a proces *Yt* jest sezonowym procesem ARIMA o okresie *s*, przedstawia się kombinację parametrów SARIMA w postaci: SARIMA(*p, d, q*)(*P, D, Q*), gdzie: *p* – liczba parametrów autoregresyjnych (AR – niesezonowe); *q* – liczba parametrów średniej ruchomej (MA – niesezonowe); *d* – stopień integracji niesezonowej części szeregu czasowego; *P* – liczba parametrów autoregresyjnych (składnik sezonowy); *D* – stopień integracji sezonowej części szeregu czasowego; *Q* – liczba parametrów średniej ruchomej (składnik sezonowy) [Osińska, 2006, s. 67].

Powyższy sposób zapisu modelu SARIMA ujęto w tabeli 5. Pozwala to na odtworzenie kombinacji parametrów modelu oraz stopnia zintegrowania szeregów czasowych zarówno pod względem rozkładu sezonowej (*D*), jak i niesezonowej (*d*) części procesu.

Tabela 5. Ogólne postaci modelu SARIMA dla aktualizacji tabel bazy JPK_VAT

Tabela JPK_VAT	Struktura ogólna
NAGLOWEK	SARIMA(1,1,3)(1,0,1)
PODMIOT	SARIMA(1,1,3)(1,0,1)
SUMA_KONTROLNA	SARIMA(1,1,3)(1,0,2)
ZAKUP	SARIMA(1,1,3)(1,0,1)
SPRZEDAZ	SARIMA(0,1,4)

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

Wyniki modelowania tabel (NAGLOWEK i PODMIOT) mają tę samą postać, co świadczy o prawidłowych procesach aktualizacyjnych odbywających się wewnątrz systemu ewidencji. Najczęściej występująca postać modelu (SARIMA(1,1,3)) w zakresie procesu stochastycznego o charakterze niesezonowym wynika z istotnego wpływu na obecną wartość okresów poprzednich i koniecznością zastosowania stosunkowo dużego opóźnienia szeregu średniej ruchomej w celu uzyskania nieskorelowanego procesu resztowego. Wartości oszacowań poszczególnych parametrów niosą ze sobą informacje, które przedstawiono w tabeli 6. Na uwagę zasługuje wartość stopnia integracji szeregu symbolizowanego przez literę „D". Przy stopniu integracji sezonowej równej zero ($D = 0$) nie zachodzi różnicowanie szeregu czasowego w części dotyczącej składnika sezonowego.

Tabela 6. Oszacowane wartości parametrów SARIMA dla tabel JPK_VAT

Tabela JPK_VAT	*p*(1)	*d*	*q*(1)	*q*(2)	*q*(3)	*q*(4)	*P*(1)	*D*	*Q*(1)	*Q*(2)
NAGLOWEK	0,5129	1	0,1973	–0,005	0,1336	–	–0,5769	0	0,3960	–
PODMIOT	0,5129	1	0,1973	–0,005	0,1336	–	–0,5769	0	0,3960	–
SUMA_KONTROLNA	0,1488	1	0,4165	0,021	0,0889	–	0,7335	0	–0,7867	0,2805
ZAKUP	0,3796	1	0,1897	–0,009	0,1833	–	–0,5395	0	0,3789	–
SPRZEDAZ	–	1	0,3285	0,177	0,2168	0,119	–	–	–	–

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

Stopień integracji niesezonowej części szeregu czasowego *d* równy jest jedności. Gdy szereg jest zintegrowany w stopniu pierwszym, wystarczy jednokrotne różnicowanie do uzyskania stacjonarności. Sezonowe parametry strukturalne nie występują dla modelu powstałego na podstawie szeregów aktualizacji SPRZEDAZ. Odmienna jest postać modelu, którego kombinacja parametrów reprezentuje model średniej ruchomej po jednokrotnym różnicowaniu: SARIMA(0,1,4)(0,0,0). W przypadkach czterech szeregów aktualizacji tabel (NAGLOWEK, PODMIOT, SUMA_KONTROLNA oraz ZAKUP) oprócz parametru autoregresyjnego AR(1) występowały trzy parametry MA na trzech poziomach opóźnień ($t-1$ do $t-3$).

Analizując wyniki modelowania NAGLOWEK, PODMIOT, ZAKUP i SPRZEDAZ), można stwierdzić, że sezonowy parametr MA na drugim opóźnieniu nie był obecny w modelu, co świadczy o stosunkowo niewielkim wpływie starszych zaburzeń składnika losowego pochodzącego ze składnika sezonowego na całkowitą dynamikę procesu. Prezentacja wyników modelowania w sensie kombinacji i wartości parametrów strukturalnych modeli nie kończy procedur analiz, pojawiają się pytania: czy modele po estymacji są prawidłowe w sensie niezależności reszt; czy wartości empiryczne z próby testowej zawierają się w przedziałach prognoz przy założeniu odpowiednich poziomów ufności (0,80 i 0,95); czy ekstrapolacja wartości miałaby charakter prognoz trafnych?

Testowanie autokorelacji w szeregach reszt porównywanych do siebie modeli jest niezbędne z punktu widzenia oceny *ex ante*, który model może dać trafne prognozy. Autokorelacja w szeregach reszt modelu oznacza, że pominięto w strukturze zmiennych te, które mają istotny wpływ na jakość oraz horyzont predykcji [Maddala, 2008, s. 46]. Uznanymi metodami do badania autokorelacji szeregów czasowych są testy Boxa–Pierce'a i Ljunga–Boxa.

W teście Ljunga–Boxa jest zmodyfikowana statystyka znaną z testu Boxa tak, aby uzyskać większą moc testu dla skończonych prób [Doman, Doman, 2009, s. 46]. Efekt autokorelacji był testowany na wielu maksymalnych opóźnieniach, w tym na zalecanym przez R.S. Tsaya w *Analysis of Financial Time*

Series logarytmie naturalnym z wielkości próby uczącej [2002, s. 25]. W tabeli 7 zawarte są *p-value* wyrażające prawdopodobieństwo odrzucenia hipotezy o braku autokorelacji na zakładanych poziomach istotności. Postawiono hipotezę zerową o braku autokorelacji przeciwko hipotezie o braku niezależności składnika losowego.

Tabela 7. Wyniki testowania braku autokorelacji reszt dla wybranych modeli

Model	***p-value* dla…**	**Autokorelacja reszt**									
	Tabela JPK_VAT	Box–Pierce					Ljung–Box				
	Opóźnienia:	1	5	14	21	25	1	5	14	21	25
SARIMA	NAGLOWEK	0,922	0,896	0,425	0,168	0,259	0,922	0,892	0,392	0,132	0,208
BATS	NAGLOWEK	0,629	0,148	**0,051**	0,073	0,158	0,628	0,141	**0,043**	0,056	0,126
TBATS	NAGLOWEK	1,000	0,839	0,499	0,568	0,699	1,000	0,833	0,465	0,516	0,680
SARIMA	PODMIOT	0,922	0,896	0,425	0,168	0,259	0,922	0,892	0,392	0,132	0,208
BATS	PODMIOT	0,629	0,148	**0,051**	0,073	0,158	0,628	0,141	**0,043**	0,056	0,126
TBATS	PODMIOT	1,000	0,839	0,499	0,568	0,699	1,000	0,833	0,465	0,516	0,646
SARIMA	SUMA	0,972	0,782	0,588	0,427	0,584	0,972	0,774	0,558	0,375	0,526
BATS	SUMA	**0,033**	**0,020**	**0,005**	**0,009**	**0,026**	**0,032**	**0,019**	**0,004**	**0,007**	**0,019**
TBATS	SUMA	0,996	0,843	0,456	0,512	0,665	0,996	0,839	0,421	0,459	0,610
SARIMA	ZAKUP	0,961	0,991	0,273	**0,042**	*0,085*	0,961	0,990	0,243	**0,029**	*0,060*
BATS	ZAKUP	0,271	0,093	**0,058**	**0,061**	0,101	0,269	0,088	**0,048**	**0,047**	0,076
TBATS	ZAKUP	0,966	0,952	0,542	0,649	0,708	0,966	0,951	0,601	0,507	0,654
SARIMA	SPRZEDAZ	0,894	0,971	0,246	*0,089*	0,128	0,894	0,970	0,220	*0,067*	*0,095*
BATS	SPRZEDAZ	0,112	0,282	0,342	0,440	0,349	0,110	0,274	0,318	0,397	0,294
TBATS	SPRZEDAZ	0,856	0,919	0,817	0,564	0,676	0,856	0,917	0,795	0,507	0,616

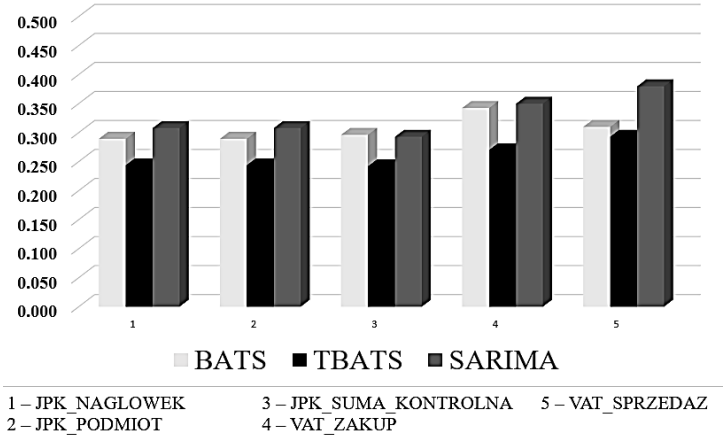
Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

W tabeli 7 przedstawiono rezultaty testowania autokorelacji reszt SARIMA, dla którego testy wykazały zanikającą autokorelację na poziomie istotności 0,1 dla modelowanych aktualizacji dwóch tabel (ZAKUP, SPRZEDAZ). Nastąpiło odrzucenie hipotezy zerowej, reszty są zależne, co ogranicza możliwości związane z wykorzystaniem modelu do dekompozycji szeregu czasowego i trafnego prognozowania. Przy tym zależność reszt jest związana z brakiem losowości, a także niesymetrycznym rozkładem, którego postać jest tak odmienna od rozkładu normalnego.

W przypadku modelu TBATS wszystkie wartości *p-value* były większe od zakładanych poziomów istotności (α = 0,05, α = 0,10), **nie było podstaw do odrzucenia hipotezy o braku autokorelacji składnika resztowego**. Reszty modelu BATS dla aktualizacji tabeli NAGLOWEK oraz PODMIOT są zależne. Testy wykazały, że występuje autokorelacja na czternastym opóźnieniu (wartość *p-value* jest niższa od zakładanego poziomu istotności równego 0,05). Reszty

modelu BATS na podstawie szeregów aktualizacji tabeli SUMA są zależne: wystąpiła wyraźna i niezanikająca autokorelacja.

Przy aktualizacjach tabeli ZAKUP wystąpiła autokorelacja w szeregach reszt dwóch opóźnień (14 i 21). Wybór modelu jest uzależniony od zdolności do generowania trafnych prognoz, a także stosunkowo niewielkiej rozpiętości przedziałów predykcji w okresie próby testowej. Istotne dla analityka jest oszacowanie postaci modelu do prognozowania przed wykonaniem ekstrapolacji i wybranie tej, która rokuje największe oczekiwania, jeśli chodzi o wysoką dokładność prognoz. Na rysunku 2 przedstawiono błędy MAPE, które pozwalają na wstępne rozeznanie problemu dokładności predykcji w okresie próby uczącej.



Rys. 2. Średni bezwzględny procentowy błąd prognozy (MAPE) aktualizacji tabel

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2018 r. – 17 maja 2019 r.

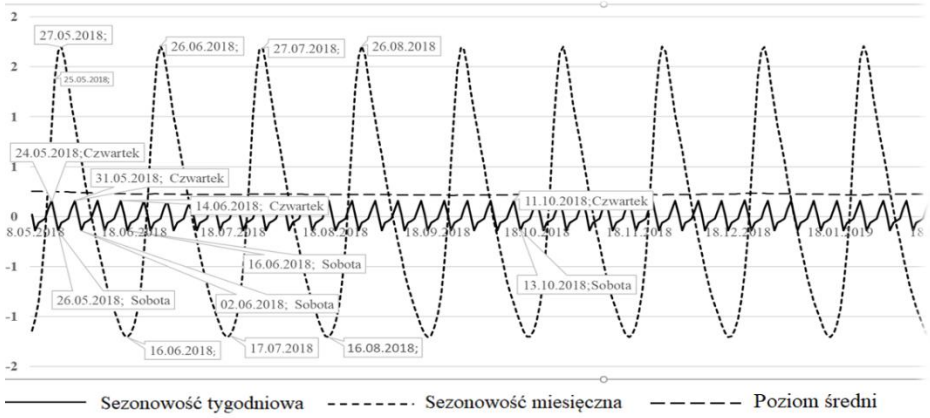
Błąd MAPE określa, o ile procent wartość prognozowania różni się od wartości empirycznej, pomijając przy tym informację o jej niedowartościowaniu lub przewartościowaniu. Rysunek 2 uwzględnia pięć szeregów aktualizacji bazy JPK_VAT. Dwie tabele (NAGLOWEK oraz PODMIOT) są takie same, jeśli chodzi o aktualizacje, w związku z tym tabele charakteryzują się identycznymi wartościami MAPE.

Najwyższe błędy były właściwe dla modelu SARIMA w przypadkach: NAGLOWEK, PODMIOT, SPRZEDAZ. Model BATS dawał gorsze wyniki (pod względem wielkości MAPE) od SARIMA oraz TBATS w przypadku szeregu aktualizacji SPRZEDAZ.

Najniższy możliwy błąd wynikający z porównania wartości teoretycznych z modelu do wartości empirycznych w okresie próby uczącej przypadał na mo-

del TBATS, co zachęca do przedstawienia wyników dekompozycji szeregu czasowego pochodzącej z estymacji tego modelu. Dekompozycja wykonana dzięki zastosowaniu modelu TBATS z uwzględnieniem sezonowości tygodniowej i miesięcznej umożliwia analizę składników sezonowych.

Na rysunku 3 pokazano fragment szeregu czasowego aktualizacji SUMA_ KONTROLNA (18 maja 2018 r. – 18 lutego 2019 r.) w postaci składników systematycznych.



Rys. 3. Wahania sezonowe w szeregu czasowym po dekompozycji modelem TBATS

Źródło: Opracowanie własne na podstawie danych z okresu 18 maja 2019 r. – 18 lutego 2019 r.

Na rysunku 3 przedstawiono momenty charakterystyczne dla maksymalnych i minimalnych aktualizacji tabeli SUMA_KONTROLNA.

Cechami miesięcznych wahań sezonowych aktualizacji jest występowanie opóźnień w ewidencji, które zarówno zależą od opóźnień spowodowanych przez podatników, jak i (w pewnych okresach) spowodowane są występowaniem dni wolnych (wówczas termin ulega przesunięciu). Przy tym minimalne stany aktualizacji zwykle występują w połowie miesiąca (16.-17. dzień).

Podatnicy zwykle czekają na zgromadzenie wystarczających informacji o przeprowadzonych transakcjach i często rozliczają się na ostatnią chwilę (przed upływem 25. dnia miesiąca). Jeśli chodzi wahania sezonowe o częstotliwości tygodniowej, najczęściej w czwartki występują maksymalne wartości załadowań ewidencji do bazy JPK_VAT, a minimalne aktualizacje mają miejsce w soboty. Na częstość aktualizacji w danych dniach wpływa harmonogram ustalony w podmiotach gospodarczych, które najczęściej rozliczają się z urzędem skarbowym pod koniec tygodnia. Sobota dla większości jest dniem wolnym i nie jest często wybierana jako najlepszy termin aktualizacji danych w ewidencji.

Podsumowanie

Modelowanie szeregów czasowych aktualizacji tabel JPK_VAT wykonywane jest nie tylko z potrzeb wynikających z poszerzenia wiedzy o rozkładach szeregów aktualizacji, ale przede wszystkim z chęci predykcji ich wartości. Pomiar podobieństwa próby uczącej do testowej urzeczywistni się w porównaniu do średnich błędów MAPE z próby uczącej do błędów z okresu próby testowej. Będzie to możliwe dopiero po wygaśnięciu prognoz i zestawieniu ich z wartościami empirycznymi – tego dotyczy artykuł autora pt. *Prognozowanie szeregów czasowych aktualizacji Jednolitych Plików Kontrolnych* w niniejszym zeszycie „Studiów Ekonomicznych. Zeszytów Naukowych Uniwersytetu Ekonomicznego w Katowicach", w którym również nastąpi podsumowanie badań w kontekście analiz ujętych w niniejszych rozważaniach.

Literatura

Brozyna J., Mentel G., Strielkowski W., Szetela B. (2018), *Multi-Seasonality in the BATS Model Using Demand for Electric Energy as a Case Study*, "Economic Computation & Economic Cybernetics Studies & Research", Vol. 52, https://pdfs.semanticscholar.org/385c/ (dostęp: 31.12.2019).

De Livera A.M., Hyndman R.J., Snyder R.D. (2010), *Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing*, Department of Econometrics and Business Statistics, Working Paper, No. 15/09, http:/www.buseco.monash.edu.au/depts/ebs/pubs/wpapers (dostęp: 6.05.2019).

Ewidencja dla podatku od towarów i usług JPK_VAT. Broszura informacyjna dot. struktury JPK_VAT (3) (2018), Ministerstwo Finansów, styczeń, Warszawa.

Dittmann P. (2003), *Prognozowanie w przedsiębiorstwie*, Oficyna Ekonomiczna, Kraków.

Doman M., Doman R. (2009), *Modelowanie zmienności i ryzyka*, Oficyna a Wolters Kluwer business, Kraków.

Maddala G.S. (2008), *Ekonometria*, WN PWN, Warszawa.

Mazur H., Mazur Z. (2004), *Projektowanie relacyjnych baz danych*, Politechnika Wrocławska, Wrocław.

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

Tsay R.S. (2002), *Analysis of Financial Time Series. Financial Econometrics*, Wiley & Sons, United States.

Ustawa z dnia 11 marca 2004 r. o podatku od towarów i usług, Dz.U. z 2004 r., nr 54, poz. 535.

MODELING AND DECOMPOSITION OF STANDARD AUDIT FILES FOR TAX (SAF-T) UPDATES

Summary: The modeling different time series became necessary process at the Ministry of Finance IT systems. The problems with lack of information and actual updates of Standard Audit Files for Tax are known. Capabilities to choosing right model of time series with complex seasonal patterns are crucial in some cases. In the article, author made the decomposition of time series with complex seasonal patterns. The results of modeling and testing indicated the best predicting (according to Mean Absolute Percentage Error) and time series decomposition method – TBATS.

Keywords: Standard Audit Files for Tax (SAF-T), analyze of time series, decomposition, predicting, BATS, TBATS, SARIMA.