Zastosowanie filtru Kalmana do szacowania poziomu szarej gospodarki w Polsce

Urząd Statystyczny w Kielcach

Katarzyna Brzozowska-Rup
Urząd Statystyczny w Kielcach
Politechnika Świętokrzyska

Działalność w szarej strefie definiowana jest jako działania produkcyjne w sensie ekonomicznym, całkowicie legalne pod względem spełniania norm i regulacji prawnych, ale ukrywane przed władzami publicznymi (ESA 2010).

Działania te ukrywane są z chęci uniknięcia:

- płacenia podatku dochodowego, podatku od wartości dodanej (VAT) i pozostałych podatków;
- płacenia składek na ubezpieczenie społeczne;
- stosowania wymogów prawa, takich jak płaca minimalna, maksymalny czas pracy czy warunki bezpieczeństwa pracy;
- stosowania procedur administracyjnych, takich jak wypełnianie kwestionariuszy statystycznych i innych formularzy.

Problem badawczy

Trudność związana z szacowaniem rozmiaru szarej gospodarki polega na tym, że podmioty gospodarcze działające w gospodarce nieobserwowanej dążą do pozostawania poza systemami ewidencjonowania i z tego względu nie jest możliwy bezpośredni pomiar tego zjawiska.

Jest to sytuacja, w której wielkość interesującego nas zjawiska potrafimy jedynie opisać za pomocą pewnych zmiennych obserwowalnych. Zagadnienie dodatkowo komplikuje fakt, że zarówno zmienne obserwowalne jak i zmienna nieobserwowalna (ukryta) mogą podlegać losowym zakłóceniom.

Filtr Kalmana

Model - założenia:

$$X = \tilde{x} + \varepsilon_x \qquad \text{(wektor stanu)}$$

$$Y = \tilde{y} + \varepsilon_y \qquad \text{(wektor pomiarów)}$$

Model, dla którego Rudolf Kalman podał algorytm rekurencyjny zadany jest równaniem

$$B\tilde{y} = A\tilde{x} + \zeta$$

gdzie B, A są odpowiednio $n_y \times n_y$ i $n_y \times n_x$ macierzmi współczynników, $\tilde{y}$, $\tilde{x}$ są n_y-wymiarowymi i n_x-wymiarowymi wektorami, których wartości są ukryte.

$\zeta, \varepsilon_x, \varepsilon_y$ - składniki losowe, o których zakłada się, że nie są wzajemnie skorelowane i są białymi szumami.

Model przestrzeni stanów (dynamiczny liniowy model gaussowski)

Wzorcem omawianej klasy modeli jest dynamiczny liniowy model gaussowski definiowany jako układ dwóch równań:

równanie obserwacji (sygnał pomiarowy)

$$y_t = \tilde{y}_t + \varepsilon_t = H_t x_t + G_t z_t^O + \varepsilon_t \qquad (1)$$

równanie stanu (równanie przejścia)

$$x_t = F_t x_{t-1} + \Gamma_t z_{t-1}^S + \eta_t \qquad (2)$$

gdzie $Z_t = \{ Z_t^O, Z_t^S \} \in \mathbb{R}^{n_z}$ jest wektorem obserwowalnych zmiennych egzogenicznych,

zakłócenia losowe ε_t , η_t są wzajemnie niezależne.

$$\begin{bmatrix} \eta_t \\ \varepsilon_t \end{bmatrix} \sim N\left(0, \begin{bmatrix} \Sigma_\eta & 0 \\ 0 & \Sigma_\varepsilon \end{bmatrix}\right)$$

o macierzach $H_t \in \mathbb{R}^{n_y \times n_x}$; $G_t \in \mathbb{R}^{n_y \times n_z}$; $F_t \in \mathbb{R}^{n_x \times n_x}$; $\Gamma_t \in \mathbb{R}^{n_x \times n_z}$ zakładamy, że są niezależne od składników losowych.

Przykład (na podstawie literatury)

Równanie stanu

$$SE_t = \alpha_1 SE_{t-1} + \alpha_2 Tax_t + \alpha_3 GE_t + \alpha_4 TM_t + \alpha_5 SB_t + \eta_t$$

równanie obserwacji

$$PKB_t^P = \beta_1 PKB_{t-1}^P + \beta_2 SE_t + \beta_3 CK_t + \beta_4 SB_t + \varepsilon_t$$

$$= PG_t - ZP_t + PT_t - DP_t + SE_t + \varepsilon_t$$

gdzie

SE - poziom szarej gospodarki

Tax - obciążenia podatkowe

GE - jakość instytucji publicznych (w tym możliwość wykrycia działalności nierejestrowanej, poziom biurokracji),

TM - moralność obywatelska (tax morality,mierzona np. liczbą przestępstw karno-skarbowych)

SB - stopa bezrobocia,

PKB^P - produkt krajowy brutto wyznaczany od strony produkcji

CK - cykl koniunkturalny, PG - produkcja globalna, ZP - zużycie pośrednie

PT - podatki od produktów , DP - dotacje do produktów

Filtr Kalmana

$y_{i:j} = (y_i, y_{i+1}, \ldots, y_j)$ - realizacje procesu Y_t od chwili i do j, $(i \leq j)$.

Zagadnienie filtracji polega na wyznaczeniu optymalnego estymatora wartości zmiennych ukrytych (x_t) w oparciu o zaobserwowane wartości $y_{1:t}$

$$\hat{x}_{t|t} = E(x_t|y_{1:t})$$

macierz kowariancji błędu

$$\Sigma_{t|t} = Var(x_t|y_{1:t}) = E([x_t - \hat{x}_{t|t}][x_t - \hat{x}_{t|t}]^T|y_{1:t})$$

Zakłada się, że wartość początkowa x_0 jest generowana z rozkładu gaussowskiego o wartości średniej μ_0 i wariancji Σ_0.

FILTR KALMANA - FILTRACJA

$$x_t = F_t x_{t-1} + \Gamma_t z_{t-1}^S + \eta_t$$

$$y_t = H_t x_t + G_t z_t^O + \varepsilon_t$$

Algorytm szacuje stan układu w dwóch krokach:

prognozowania

$$\hat{x}_{t|t-1} = E(x_t|y_{1:t-1}) = F_t \hat{x}_{t-1|t-1} + \Gamma_t z_t^S \quad (3)$$

$$\hat{y}_{t|t-1} = E(y_t|y_{1:t-1}) = H_t \hat{x}_{t|t-1} + G_t z_t^O \quad (4)$$

$$\Sigma_{t|t-1} = Var(x_t|y_{1:t-1}) = F_t \Sigma_{t|t} F_t^T + \Sigma_\eta \quad (5)$$

$$v_t = y_t - \hat{y}_{t|t-1}$$

$$P_{t|t-1} = Var(y_t - \hat{y}_{t|t-1}) = H_t^T \Sigma_{t|t-1} H_t + \Sigma_\varepsilon \quad (6)$$

oraz **korekty (uaktualnienia)**

$$\hat{x}_{t|t} = E(x_t|y_{1:t}) = \hat{x}_{t|t-1} + K_t(y_t - \hat{y}_{t|t-1}) \quad (7)$$

$$\Sigma_{t|t} = \Sigma_{t|t-1} - K_t(H_t \Sigma_{t|t-1} H_t^T + \Sigma_\varepsilon) K_t^T \quad (8)$$

$K_t = \Sigma_{t|t-1} F_t (P_{t|t-1})^{-1}$ - tzw. korekta Kalmana

Filtr Kalmana - wygładzanie

Wygładzanie polega na tzw. wstecznym wyznaczaniu warunkowych estymatorów wektora stanu oraz macierzy kowariancji, gdy znany jest pełny zbiór obserwacji $y_{1:T}$. Przy danych warunkach początkowych $\hat{x}_{T|T}$, $\Sigma_{T|T}$.

Aktualizacja zmiennej stanu

$$\hat{x}_{t|T} = E[x_t|y_{1:T}]$$

$$= \hat{x}_{t|t} + \Sigma_{t|t} F_t^T \Sigma_{t+1|t}^{-1} (\hat{x}_{t+1|T} - \hat{x}_{t+1|t})$$

Aktualizacja macierzy kowariancji

$$\Sigma_{t+1|T} = \Sigma_{t|t} - \Sigma_{t|t} F_t^T \Sigma_{t+1|t}^{-1} (\Sigma_{t+1|T} - \Sigma_{t+1|t}) \Sigma_{t+1|t}^{-1} F_t \Sigma_{t|t}^T$$

Funkcja wiarygodności

Rozważany model przestrzeni stanów (1-2) zależy od nieznanych parametrów

$\boldsymbol{\theta_t} = [\boldsymbol{F_t}, \boldsymbol{\Gamma_t}, \boldsymbol{H_t}, \boldsymbol{G_t}, \boldsymbol{\Sigma_\eta}, \boldsymbol{\Sigma_\varepsilon}]$.

Warto zauważyć, że wybrane składowe wektora parametrów $\boldsymbol{\theta}$ są zależne od czasu, więc ich oszacowania mogą być aktualizowane za każdym razem, gdy pojawią się nowe wartości zmiennych obserwowalnych w chwili $\boldsymbol{t}$.

Zakładając, że znany jest rozkład stanu początkowego $\boldsymbol{N(\mu_0;\ \Sigma_0)}$, logarytm funkcji wiarygodności, z dokładnością do stałej zadany jest wzorem

$$\ell(\boldsymbol{\theta}) = \boldsymbol{lnL}\left(\boldsymbol{\theta}\middle|\boldsymbol{y_{1:T}}\right) = \boldsymbol{p}(\boldsymbol{y_1}|\boldsymbol{\theta}) \prod_{t=2}^{T} \boldsymbol{p}(\boldsymbol{y_t}|\boldsymbol{y_{1:t-1}}, \boldsymbol{\theta})$$

$$\propto -\frac{1}{2}\left(\sum_{t=1}^{T} \boldsymbol{ln}(\boldsymbol{det}(\boldsymbol{P_{t|t-1}})) + \sum_{t=1}^{T} \boldsymbol{v_t^T}\ \ \boldsymbol{P_{t|t-1}^{-1}} \boldsymbol{v_t}\right)$$

gdzie $\boldsymbol{P_{t|t-1}} = \boldsymbol{Var}\left(\boldsymbol{y_t} - \boldsymbol{\hat{y}_{t|t-1}}\right)$,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{arg}\ \boldsymbol{\max_{\theta \in \Theta}}\ \boldsymbol{\ell}(\boldsymbol{\theta})$$

Metoda Newtona-Raphsona

Zakłada się, że

- istnieją drugie pochodne funkcji $lnL(\theta|y_{1:T})$
- funkcja wiarygodności jest wypukła, tzn. hesjan $I(\theta; y_{1:T}) = \frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}$ jest macierzą ujemnie określoną na całej przestrzeni parametrów.

W wyniku odpowiednich przekształceń wartość skorygowaną parametrów modelu określa wzór

$$\theta_k = \theta_{k-1} - (I(\theta_{k-1}; y_{1:T}))^{-1} \frac{\partial \ell(\theta)}{\partial \theta}\Big|_{\theta=\theta_{k-1}}$$

Najczęściej za kryterium zatrzymania algorytmu przyjmujemy jeden z poniższych warunków.

- $\frac{\|\theta_k - \theta_{k-1}\|}{\|\theta_{k-1}\|} \leq \Delta$ dla pewnego arbitralnie zadanego poziomu krytycznego Δ;
- odgórnie ustala się ilość iteracji, po jej zrealizowaniu algorytm powtarzany jest dla nowych warunków początkowych (nowe wartości parametru).

Idea Proponowanego Podejścia

Konstruowany model opiera się na szeregu założeń:

- Porodukcja globalna jest sumą produkcji wytworzonej w sektorze oficjalnym (rejestrowanym) oraz w szarej strefie

$$PG = PG_r + PG_{SE}$$

- Przedsiębiorstwa działające w szarej strefie nie informują władz o odziałalności gospodarczej oraz mają możliwość uchylania się od części opodatkowania.
- W każdym momencie przedsiębiorstwo może być poddane kontroli, wykrycie działalności nierejestrowanej następuje z pewnym prawdopodobieństwem i wiąże się z nałożeniem kary oraz zobligowaniem do zapłacenia podatku od dochodów z produkcji ukrytej.
- Informacja o działalności w szarej strefie jest ukryta w przychodach przedsiębiorstw.

MODEL - PROPOZYCJA

- Na podstawie badania CBSG/01 (Badanie podmiotów o liczbie pracujących do 49 osób) uzyskiwana jest informacja o ponoszonych przez przedsiębiorców kosztach prowadzenia działalności, wysokości wynagrodzeń i dochodów.

- Na ich podstawie w Urzędzie Statystycznym w Kielcach dokonuje się szacunków **przychodów przedsiębiorstw**, które z kolei stanowią bazę do szacowania, pierwotnie, **produkcji globalnej**, a następnie **zużycia pośredniego i wartości dodanej brutto**.

- Szacunki, z podziałem na: grupy sekcji PKD, województwa i podregiony wykonywane są dla trzech grup podmiotów:
  1. mikroprzedsiębiorstw (do 9 pracujących) – osób fizycznych i spółek cywilnych,
  2. mikroprzedsiębiorstw (do 9 pracujących) – osób prawnych,
  3. podmiotów małych (od 10 do 49 pracujących) – sektora prywatnego z wyłączeniem spółdzielni.

Model – Propozycja Danych

PP	przychody przedsiębiorstw z działalności gospodarczej (przychody ze sprzedaży netto obejmują środki pieniężne ze sprzedaży wyrobów, usług oraz towarów po potrąceniu VAT) z naszego bada nia CBSG/01
SD	sprzedaż detaliczna
PI	przychody inne: przychody ze sprzedaży towarów i materiałów, tj. nabyte w celu odsprzedaży w stanie nieprzetworzonym, przychody operacyjne, przychody finansowe, tj. m.in.: kwoty należne z tytułu dywidend i udziałów w zysku
DP	dotacje przedmiotowe
WD	wartość dodana brutto
ZP	zużycie pośrednie
PG	Produkcja globalna
KWS	koszty wytworzenia świadczeń na własne potrzeby jednostki z naszego badnia CBSG/01
KP	koszty produkcji, na które składają się wydatki na materiały i energię (plus podatki)
AP	amortyzacja środków trwałych
WBP	fundusz wynagrodzeń
ZFP	inne zobowiązania finansowe

DYNAMICZNY MODEL LINIOWY - PROPOZYCJA

Równanie obserwacji

$$PP_t = \alpha_{11,t}PP_{t-1} + \alpha_{12,t}PSP_t + \alpha_{13,t}SD_t + \alpha_{14,t}DP_t + \alpha_{15,t}SE_t + \varepsilon_{1,t}$$

$$KWS_t = \alpha_{21,t}KP_t + \alpha_{22,t}AP_t + \alpha_{23,t}WBP_t + \alpha_{24,t}ZFP_t + \alpha_{25,t}SE_t + \varepsilon_{2,t}$$

$$PG_t = \alpha_{31,t}PG_{t-1} + \alpha_{32,t}PP_t + \alpha_{33,t}\frac{PP_{t-1}}{PG_{t-1}} + \alpha_{34,t}\frac{PG_{t-1}}{PG_{t-2}} + \alpha_{35,t}\frac{ZP_t}{ZP_{t-1}} + \alpha_{36,t}SE_t + \varepsilon_{3,t}$$

$$ZP_t = \alpha_{41,t}\frac{ZP_t}{ZP_{t-1}} + \alpha_{42,t}KWS_t + \alpha_{43,t}SE_t + \varepsilon_{4,t}$$

$$WD_t = PG_t - ZP_t$$

Równanie stanu

$$SE_t = \beta_{1t}SE_{t-1} + \beta_{2t}p_tG_t + \beta_{3t}SB_t + \beta_{4t}Tax_t + \eta_t$$

gdzie p_t - prawdopodobieństwo wykrycia prowadzenia działalności nierejestrowanej, G_t - grzywna za prowadzenie działalności nierejestrowanej, SB_t - stopa bezrobocia

Model – propozycja II

$$PP_t = \alpha_{11,t}PP_{t-1} + \alpha_{12,t}PSP_t + \alpha_{13,t}SD_t + \alpha_{14,t}DP_t + \alpha_{15,t}SE_t + \varepsilon_{1,t}$$

$$KWS_t = \alpha_{21,t}KP_t + \alpha_{22,t}AP_t + \alpha_{23,t}WBP_t + \alpha_{24,t}ZFP_t + \alpha_{25,t}SE_t + \varepsilon_{2,t}$$

$$SE_t = \beta_t SE_{t-1} + \eta_t$$

Ogólny model przestrzeni stanów (SSM) znanych również jako dynamiczne modele liniowe (DLM)

$$Y_t = A_t\theta_t + \epsilon_t ,$$

$$\theta_t = T_t\theta_{t-1} + e_t$$

$t = 1, 2, \dots, T$

gdzie $Y_t = [PP_t \;\; KWS_t]$ jest wektorem obserwacji

θ_t - wektor zmiennych stanu,

A_t , T_t - macierze współczynników modelu.

R Pakiet dml

- Pakiet dlm koncentruje się na analizie bayesowskiej dynamicznych modeli liniowych (DLM) , zawiera również **filtr Kalmana**

Function	Model
`dlmModARMA`	ARMA process
`dlmModPoly`	nth order polynomial DLM
`dlmModReg`	Linear regression
`dlmModSeas`	Periodic – Seasonal factors
`dlmModTrig`	Periodic – Trigonometric form

- W pakiecie dostępne są funkcje do szacowania parametrów DLM metodą największej wiarygodności **dlmMLE**.

Podsumowanie

Idea proponowanego podejścia polega na szacowaniu poziomu szarej gospodarki w oparciu o zarejestrowane całkowite przychody z działalności gospodarczej oraz ponoszone koszty (z podziałem na grupy sekcji).

Zalety proponowanego podejścia:

- Modele przestrzeni stanów obejmują niezwykle ogólną klasę modeli dynamicznych.
- Proponowana metoda sprawdza się w szacowaniu modeli, w których parametry zmieniają się w sposób ciągły.
- Metody filtracji (w szczególności filtr Kalmana) umożliwiają szacowanie nieobserwowalnych zmiennych oraz parametrów modelu w przypadku gdy nie jest znana dokładna natura modelowanego układu.
- Szczególną zaletą proponowanej metody jest istnienie ogólnego rozwiązania, które można wyznaczyć w oparciu o zasady rachunku prawdopodobieństwa i równania filtru Kalmana.

Uzyskane wyniki zostaną porównane wynikami badania SP3 (Sprawozdanie o działalności gospodarczej przedsiębiorstw).

Dziękuję za uwagę

Katarzyna Brzozowska-Rup

K.Brzozowska-Rup@stat.gov.pl

Urząd Statystyczny
w Kielcach

Literatura

Anderson B. D. O. , Moore J. B. , (1984), *Filtracja optymalna*, Wydawnictwo Naukowo-Techniczne, Warszawa.

M. Doman, R. Doman, (2004), *Ekonometryczne modelowanie dynamiki polskiego rynku finansowego*, Prace habilitacyjne, Wydawnictwo Akademii Ekonomicznej w Poznaniu.

Kalman R. E., (1960), *A New Approach to Linear Filtering and Prediction Problems*, Journal of Basic Engineering 82 (Series D) 35-45

P. Malczewska, (2019), *Szara strefa. Determinanty i mechanizmy kształtowania*, Wydawnictwo Uniwersytetu Łódzkiego.

G. J. McLachlan, T. Krishnan, (1997), *The EM Algorithm and extensions*, John Wiley & Sons, INC.

Petris G, Petrone S, Campagnoli P (2009). Dynamic Linear Models with R. Springer-Verlag, New York.

Literatura

Tusell F., (2011), Kalman Filtering in R, Journal of Statistical Software

Petris G., (2009), dlm: an R package for Bayesian analysis of Dynamic Linear Models, https://cran.r-project.org/web/packages/dlm/vignettes/dlm.pdf

Rocznik Statystyczny Przemysłu, Statistical Yearbook of Industry – Poland, (2022), Główny Urząd Statystyczny, Statistics Poland, Warszawa.

OECD, (2002), Measuring the Non-Observed Economy A Handbook, https://www.oecd.org/sdd/na/1963116.pdf