Wiadomości Statystyczne. The Polish Statistician, 2021, vol. 66, 5, 7–25 DOI: 10.5604/01.3001.0014.8781
Studia metodologiczne / Methodological studies
Otrzymano/received: 10.11.2020, zaakceptowano/accepted: 15.04.2021

Kilka uwag na temat pomiaru zależności pomiędzy zmiennymi o panelowej strukturze danych

Mieczysław Kowerski[a], Jarosław Bielak[b]

Streszczenie. W licznych artykułach modelowanie na podstawie danych panelowych rozpoczyna się od przedstawienia macierzy współczynników korelacji liniowej Pearsona pomiędzy zmiennymi przyjętymi do badania. Celem niniejszego artykułu jest pokazanie nieprzydatności takiego podejścia w analizie zależności w przypadku danych panelowych oraz próba zaproponowania bardziej adekwatnej miary – współczynnika korelacji pomiędzy wartościami empirycznymi i teoretycznymi zmiennej objaśnianej oszacowanego modelu panelowego (z efektami stałymi lub losowymi) względem zmiennej, której zależność w stosunku do zmiennej objaśnianej jest obliczana.

Współczynnik korelacji liniowej Pearsona nie uwidacznia podstawowej zalety danych panelowych, jaką jest dostarczanie informacji o zależnościach badanych zjawisk jednocześnie w czasie i przestrzeni. Dla obliczenia tego współczynnika nie jest bowiem istotne, że pewna obserwacja dotyczy obiektu *i* w okresie *t*, a inna – obiektu *j* w okresie $t + 1$. Można go natomiast wykorzystać w analizach danych panelowych do obliczeń cząstkowych.

Prowadzone rozważania zilustrowano obliczeniami zależności pomiędzy strukturą kapitału oraz rentowością i wielkością 17 spółek budowlanych notowanych na Giełdzie Papierów Wartościowych w Warszawie w latach 2009–2018 (170 obserwacji), tworzących panel zbilansowany. Obliczenia te pozwoliły na sformułowanie zalet i wad zaproponowanego rozwiązania.

Słowa kluczowe: dane o strukturze panelowej, współczynnik korelacji liniowej Pearsona, model panelowy

JEL: C18, C23

A selection of remarks on the measurement of correlations between variables of a panel data structure

Abstract. Many articles featuring panel data modelling tend to begin their considerations with an introduction of the Pearson linear correlation coefficients matrix between the analysed variables. The aim of the article is to prove such an approach unsuitable in the analysis of panel data dependencies. Instead, an attempt has been made to propose a more appropriate measure – a correlation coefficient between the empirical and fitted values of the dependent variable of the estimated panel model (with fixed or random effects) in relation to the variable whose dependency towards the dependent variable is being studied.

[a] Uczelnia Państwowa im. Szymona Szymonowica w Zamościu, Instytut Społeczno-Ekonomiczny, Polska / State Higher School of Vocational Education in Zamość, Department of Social and Economic Sciences, Poland. ORCID: https://orcid.org/0000-0002-2147-2037. Autor korespondencyjny / Corresponding author, e-mail: mieczyslaw.kowerski@upz.edu.pl.

[b] Uczelnia Państwowa im. Szymona Szymonowica w Zamościu, Instytut Społeczno-Ekonomiczny, Polska / State Higher School of Vocational Education in Zamość, Department of Social and Economic Sciences, Poland. ORCID: https://orcid.org/0000-0001-8537-8624. E-mail: jaroslaw.bielak@upz.edu.pl.

Pearson's linear correlation coefficient does not reflect the basic advantage of panel data, which is the ability to provide information about the dependencies of the studied phenomena simultaneously in time and space. The fact that one observation relates to object *i* during period *t* and another to object *j* during period $t + 1$ is irrelevant for the calculation of the coefficient. Pearson's coefficient, however, can be used when conducting sub-calculations in panel data analysis.

The presented considerations have been illustrated by the calculations of the relationships between the structure of capital and the profitability and size of 17 construction companies listed on the Warsaw Stock Exchange in the years 2009–2018 (170 observations) which created a balanced panel. A specification of the advantages and disadvantages of the proposed solution was formulated on the basis of the calculations.

Keywords: panel data structure, Pearson linear correlation coefficient, panel model

1. Wprowadzenie

Terminem *dane panelowe* określa się zbiór informacji o jednoznacznie identyfikowalnych obiektach (jednostkach) obserwowanych w czasie. Metody analizy i modelowania zjawisk przy użyciu takich danych cieszą się w ostatnich latach rosnącym zainteresowaniem ekonomistów. Wynika to z wielu zalet danych panelowych i przewagi modelowania zjawisk ekonomicznych na podstawie tego rodzaju danych nad oddzielnym modelowaniem zjawisk opisanych przez szeregi czasowe czy przekrojowe; wykorzystanie danych panelowych umożliwia pełniejszy opis analizowanego zjawiska, większą liczbę dostępnych informacji, a także eliminację lub redukcję obciążenia estymatorów (Witkowski, 2012, s. 268–269).

Jak wskazuje literatura przedmiotu (zob. przykłady omówione poniżej), pierwszym krokiem w modelowaniu na podstawie danych panelowych dość często jest badanie korelacji liniowej między zmiennymi za pomocą współczynnika Pearsona.

Na podstawie zgromadzonych danych panelowych opisujących sytuację finansową badanych spółek w latach 1998–2016 Herman (2019, s. 45–46) obliczył wartości współczynnika korelacji liniowej Pearsona, a następnie wykorzystał je do przeprowadzenia wnioskowania o sile i kierunku zależności pomiędzy potencjalnymi zmiennymi objaśniającymi a zmiennymi objaśnianymi. Stwierdził przy tym, że właśnie taka analiza korelacji powinna poprzedzać podejście wielowymiarowe.

Borsuk i Kostrzewa (2020) poprzedzili prezentację wyników estymacji modeli ryzyk systemowych oszacowanych na podstawie kwartalnych danych panelowych obejmujących działalność banków komercyjnych na polskim rynku w latach 2007–2017 przedstawieniem macierzy współczynników korelacji liniowej Pearsona analizowanych zmiennych, ale nie podjęli się ich interpretacji.

Driver i in. (2020) zaprezentowali macierz współczynników korelacji liniowej Pearsona zmiennych przyjętych do badania przed wykonaniem estymacji panelowych modeli Lintnera opisujących politykę dywidendową spółek notowanych na

giełdzie londyńskiej w latach 1997–2012. Za pomocą analizy poziomu istotności współczynników korelacji autorzy próbowali odpowiedzieć na pytanie, które zmienne powinny znaleźć się w modelach panelowych jako zmienne objaśniające. Dociekali także, czy może wystąpić zjawisko przybliżonej współliniowości pomiędzy potencjalnymi zmiennymi objaśniającymi.

Karkowska (2019) obliczyła macierz współczynników korelacji Pearsona przyjętych zmiennych i dokonała oceny ich istotności oraz kierunku zależności pomiędzy potencjalnymi zmiennymi objaśniającymi i zmiennymi objaśnianymi, aby później wykonać estymację – na podstawie danych panelowych z 4678 banków w 31 krajach europejskich w latach 1996–2011 – modeli źródeł ryzyka systematycznego sektora bankowego.

Pluskota (2020) obliczyła macierz korelacji przyjętych zmiennych i oceniła istotność współczynników korelacji, a następnie, na podstawie danych panelowych pochodzących z rozwiniętych krajów europejskich w latach 1996–2017, przeprowadziła estymację modeli opisujących wpływ korupcji na tempo wzrostu gospodarczego i innowacje. Nieistotnie skorelowana ze zmiennymi objaśnianymi zmienna *corruption* okazała się istotna w oszacowanych modelach panelowych (s. 82–84), co może być sygnałem, że współczynnik korelacji liniowej Pearsona nie mierzy zależności w przypadku danych panelowych.

Z kolei Urbanek (2017), zanim przystąpił do estymacji modeli panelowych zależności pomiędzy wynikami finansowymi i siłą marki spółek notowanych na Giełdzie Papierów Wartościowych (GPW) w Warszawie w latach 2008–2014, obliczył macierz korelacji przyjętych zmiennych i już na podstawie oceny istotności współczynników korelacji liniowej Pearsona doszedł do wniosku, że ich wartości w pełni potwierdzają postawione hipotezy[1].

Wymienione wyżej artykuły zawierają prawidłowo zastosowane metody estymacji modeli panelowych uwzględniające zróżnicowanie obserwacji ze względu na przestrzeń i czas. Zdaniem autorów niniejszej pracy obliczanie i prezentacja wartości współczynnika korelacji liniowej Pearsona pomiędzy zmiennymi o panelowej strukturze danych nie są więc potrzebne, a niekiedy wręcz wprowadzają badaczy w błąd. Nie oznacza to, że nie należy obliczać zależności pomiędzy zmiennymi opisanymi za pomocą danych panelowych, powinno się jednak robić to z uwzględnieniem ich panelowego charakteru.

Celem artykułu jest pokazanie nieprzydatności współczynnika korelacji liniowej Pearsona do analizy zależności pomiędzy zmiennymi o panelowej strukturze danych oraz próba zaproponowania bardziej adekwatnej miary. Prowadzony wywód został

[1] W związku z czym można zapytać o celowość szacowania modeli panelowych w dalszej części artykułu.

zilustrowany obliczeniami zależności pomiędzy strukturą kapitału oraz rentownością i wielkością 17 spółek budowlanych notowanych na GPW w Warszawie w latach 2009–2018 (170 obserwacji), tworzących panel zbilansowany.

2. Podstawowe pojęcia

Dane panelowe mają jednocześnie cechy danych przekrojowych (opisujących zbiorowość w jednym momencie) i szeregów czasowych (opisujących jednostkę w różnych okresach). Jeżeli we wszystkich okresach obserwowane są te same jednostki, to mamy do czynienia z panelem zbilansowanym, a jeżeli dla niektórych okresów brakuje danych o wszystkich jednostkach – z panelem niezbilansowanym.

Podstawową zaletą danych panelowych, w porównaniu np. z danymi przekrojowymi, jest większa liczba informacji o tych samych obiektach. Dane panelowe umożliwiają jednoczesną obserwację zróżnicowania analizowanych obiektów oraz ich ewolucję w czasie, a zatem lepsze rozpoznanie badanego zjawiska. Pozwalają także na kontrolowanie i/lub identyfikację nieobserwowalnych efektów indywidualnych w modelach regresji. W związku z tym dzięki wykorzystaniu panelu możliwe staje się usunięcie obciążenia estymatora z powodu pominięcia ważnego czynnika, jakim jest stały, indywidualny (specyficzny) efekt charakterystyczny dla każdej jednostki (Witkowski, 2012, s. 268–269). Dane panelowe zapewniają znacznie większą liczbę obserwacji, co poprawia precyzję wnioskowania oraz umożliwia oszacowanie dynamiki zjawisk, nawet gdy liczba okresów jest niewielka.

Współczynnik korelacji liniowej Pearsona[2] pozwala zmierzyć siłę i kierunek zależności pomiędzy dwiema jednowymiarowymi (w szeregu czasowym lub przekrojowym) zmiennymi, przede wszystkim gdy zakłada się, że jest to zależność liniowa. Stąd też, jeżeli mamy panel zbilansowany N jednostek w T okresach, to w formule współczynnika korelacji liniowej każdą z $N \cdot T$ obserwacji traktuje się niezależnie od tego, jakiego obiektu i jakiego okresu dotyczy. Współczynnik korelacji liniowej Pearsona jest więc interpretowany jako miara zależności pomiędzy dwiema zmiennymi w jednym obiekcie w $N \cdot T$ okresach (np. latach) lub jako miara zależności w $N \cdot T$ jednostkach w jednym okresie (np. roku). Przykładowo jeżeli mamy panel 100 obiektów w ciągu 10 lat (a więc 1000 obserwacji), to współczynnik korelacji możemy traktować jako miarę zależności dwóch zmiennych w 1000 obiektów w jednym roku (co w praktyce zachodzi dość często) lub w jednym obiekcie w ciągu 1000 lat (co jest w zasadzie niemożliwe).

Współczynnik korelacji liniowej Pearsona nie uwzględnia podstawowej zalety danych panelowych, jaką jest dostarczanie pełniejszej informacji o badanym zjawi-

[2] W 1895 r. Karl Pearson podał ostateczną matematyczną formułę współczynnika korelacji liniowej, który odtąd nazywany jest współczynnikiem Pearsona.

sku. Dla jego obliczenia nie jest istotne, że pewna obserwacja dotyczy obiektu i w okresie t, a inna – obiektu j w okresie $t + 1$, a to ma zasadnicze znaczenie w analizie danych panelowych. Tak więc obliczanie i wnioskowanie o zależnościach pomiędzy zmiennymi o panelowej strukturze danych na podstawie macierzy liniowych współczynników korelacji Pearsona jest niewłaściwe.

3. Zastosowanie współczynnika korelacji liniowej Pearsona do badania zależności w przypadku danych panelowych

Nieprzydatność współczynnika korelacji liniowej Pearsona do obliczania zależności pomiędzy zmiennymi o panelowej strukturze danych nie oznacza, że nie znajduje on zastosowania w analizach panelowych. Można go wykorzystać do pewnych obliczeń cząstkowych.

Do unaocznienia problemu niech posłuży następujący przykład. Analizie poddano zależności finansowe w zbilansowanym panelu 17 spółek budowlanych notowanych na GPW w Warszawie w latach 2009–2018 (170 obserwacji). Badano zależności struktury kapitału spółki mierzonej relacją zadłużenia ogółem do aktywów ogółem (CS) względem wielkości tej spółki mierzonej logarytmem naturalnym z wartości aktywów ogółem tego podmiotu (SIZE) oraz rentowności spółki mierzonej wskaźnikiem ROA.

Tabl. 1. Wartości współczynnika korelacji liniowej Pearsona (r) pomiędzy CS oraz SIZE i ROA (dla 170 obserwacji, bez uwzględnienia struktury panelowej danych)

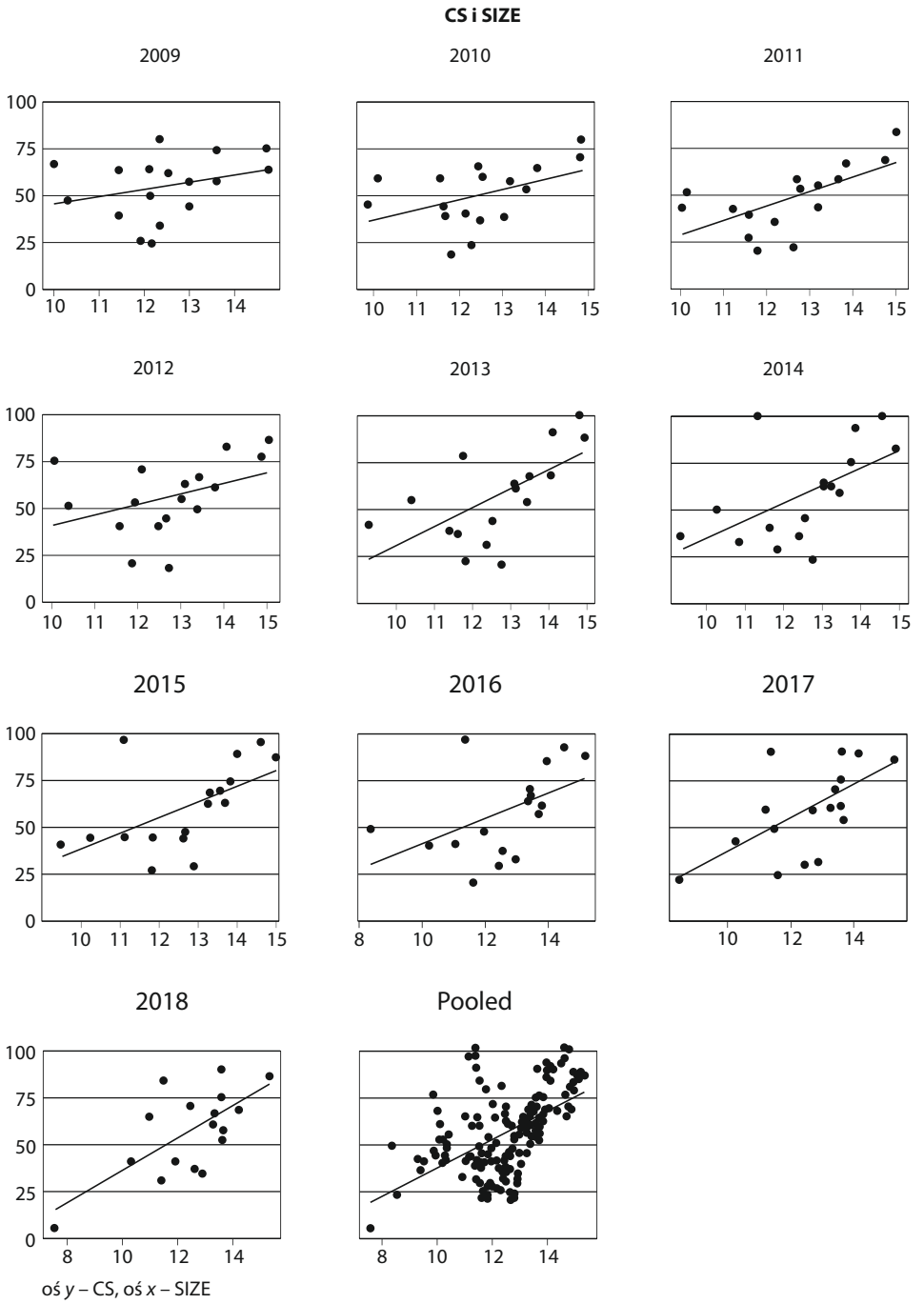
Zmienne	r	p
SIZE	0,5363	<0,0001
ROA	–0,0737	0,3384

Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.).

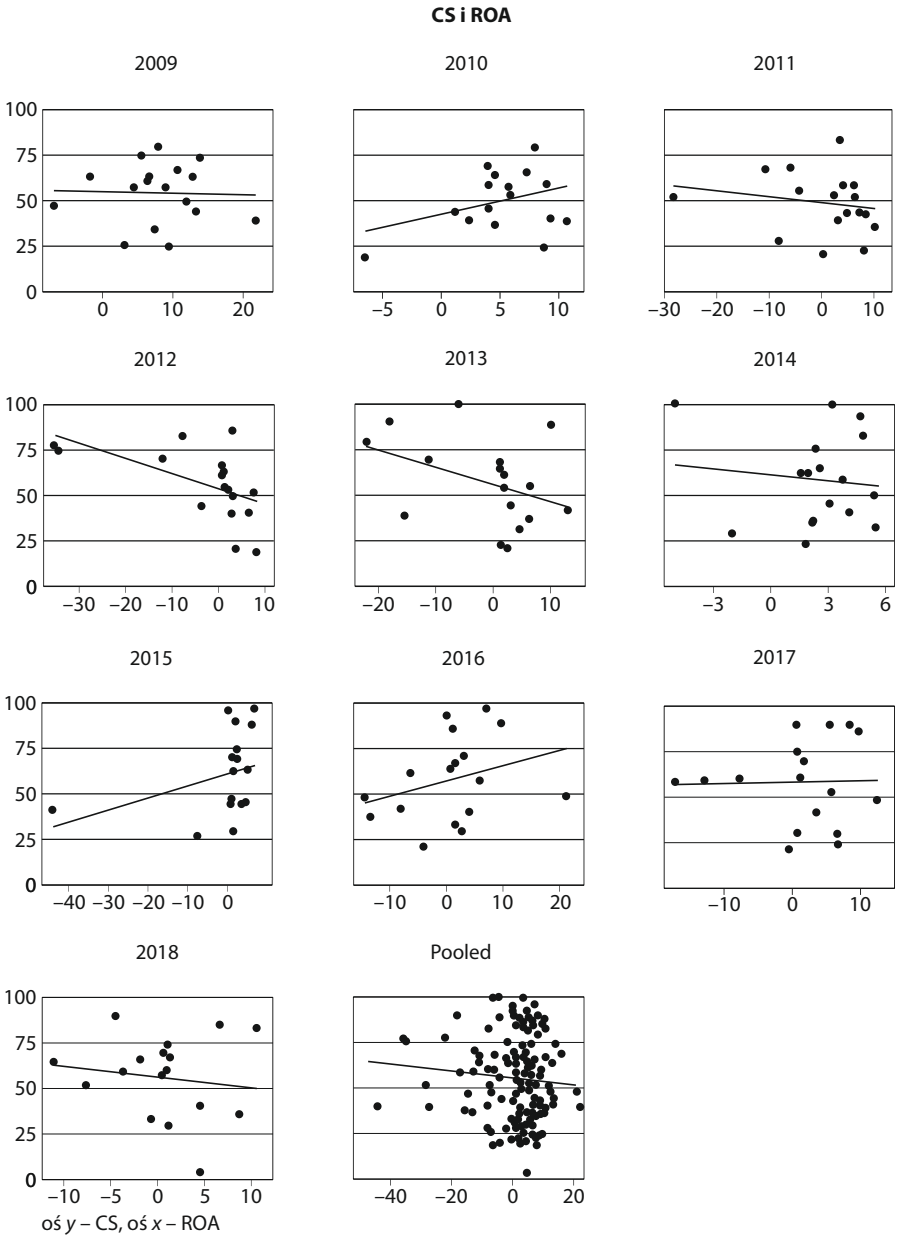
Obliczone wartości współczynnika korelacji (tabl. 1), zgodnie z wcześniejszym wywodem, odzwierciedlają zależności występujące dla jednej hipotetycznej spółki w ciągu 170 lat lub 170 hipotetycznych spółek w jednym roku. Z formalnego punktu widzenia warto odnotować, że zależność pomiędzy CS i SIZE jest istotna na poziomie <0,0001, natomiast pomiędzy CS i ROA – nieistotna[3]. Zbiór danych to jednak panel zbilansowany 17 spółek w ciągu 10 lat, a zaprezentowane współczynniki nie uwzględniają tego faktu.

[3] W artykule do oceny istotności przyjęto poziom $p = 0{,}05$.

Wykr. 1. Wykresy rozrzutu z uwzględnieniem struktury panelowej wraz z ilustracją liniowej korelacji pomiędzy zmiennymi CS i SIZE oraz CS i ROA dla spółek w poszczególnych latach



Wykr. 1. Wykresy rozrzutu z uwzględnieniem struktury panelowej wraz z ilustracją liniowej korelacji pomiędzy zmiennymi CS i SIZE oraz CS i ROA dla spółek w poszczególnych latach (dok.)



Uwaga. Wykresy zatytułowane „Pooled" dotyczą wszystkich obserwacji bez uwzględnienia struktury panelowej.
Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.) z wykorzystaniem pakietu oprogramowania R.

Korzystając z danych panelowych, można natomiast obliczyć wartości współczynnika korelacji liniowej Pearsona pomiędzy wyspecyfikowanymi zmiennymi, które są:

- „przekrojowe" – w każdym roku ($T = 10$) dla wszystkich spółek ($N = 17$);
- „czasowe" (dynamiczne) – dla każdej spółki w ciągu 10 lat.

Tabl. 2. Wartości współczynnika korelacji pomiędzy CS oraz SIZE i ROA w poszczególnych latach

Lata	CS i SIZE		CS i ROA	
	r	p	r	p
2009	0,2968	0,2474	–0,0250	0,9241
2010	0,4496	0,0702	0,3643	0,1506
2011	0,6379	0,0059	–0,1938	0,4560
2012	0,3868	0,1251	–0,5581	0,0199
2013	0,6176	0,0082	–0,3934	0,1182
2014	0,5518	0,0216	–0,1177	0,6528
2015	0,5606	0,0192	0,3401	0,1817
2016	0,4825	0,0498	0,3034	0,2365
2017	0,6091	0,0095	0,0327	0,9010
2018	0,6759	0,0029	–0,1362	0,6023

Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.).

W poszczególnych latach wartości „przekrojowego" współczynnika korelacji są zróżnicowane, co oznacza, że okres analizy ma wpływ na wyniki. W przypadku korelacji pomiędzy CS i SIZE we wszystkich analizowanych okresach zależności były dodatnie i w siedmiu latach istotne na poziomie <0,05. Natomiast w przypadku korelacji pomiędzy CS i ROA w sześciu okresach wartości te były ujemne, w tym w jednym roku współczynnik okazał się istotny, podczas gdy w czterech latach zaobserwowano dodatnie, ale nieistotne statystycznie wartości tego współczynnika (tabl. 2, wykr. 1).

Tabl. 3. Wartości współczynnika korelacji pomiędzy CS oraz SIZE i ROA w latach 2009–2018 według spółek

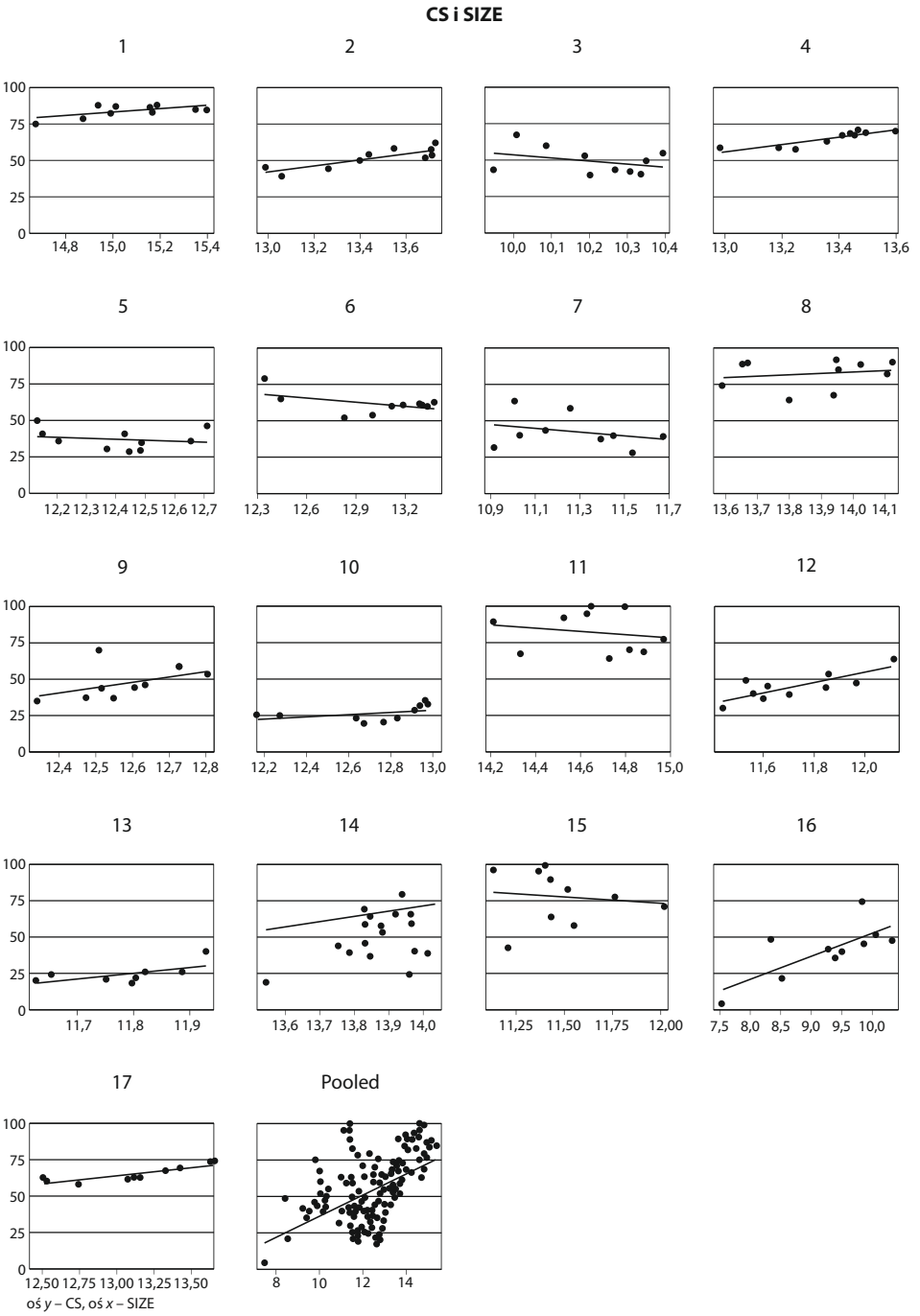
Lp.	Spółki	CS i SIZE		CS i ROA	
		r	p	r	p
1	Budimex	0,6588	0,0383	0,3130	0,3785
2	Elbudowa	0,8459	0,0020	–0,4080	0,2419
3	Enap	–0,3610	0,3054	0,9552	<0,0001
4	Erbud	0,8757	0,0009	–0,2013	0,5771
5	Instalkrk	–0,1981	0,5832	0,3427	0,3323
6	Mirbud	–0,5376	0,1090	0,5117	0,1306
7	Mostalplc	–0,3129	0,3788	–0,3397	0,3368
8	Mostalwar	0,2084	0,5634	–0,1734	0,6319
9	Mostalzab	0,4370	0,2066	–0,2449	0,4952
10	Panova	0,4487	0,1934	–0,5959	0,0691
11	Polimexms	–0,1784	0,6220	–0,0585	0,8725
12	Prochem	0,8014	0,0053	–0,1913	0,5965
13	Projprzem	0,6758	0,0320	0,3332	0,3469
14	Rafako	0,7082	0,0219	–0,2073	0,5655
15	Remak	–0,1276	0,7255	0,0104	0,9774
16	Resbud	0,7331	0,0158	–0,4069	0,2432
17	Unibep	0,9046	0,0003	–0,5848	0,0758

Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.).

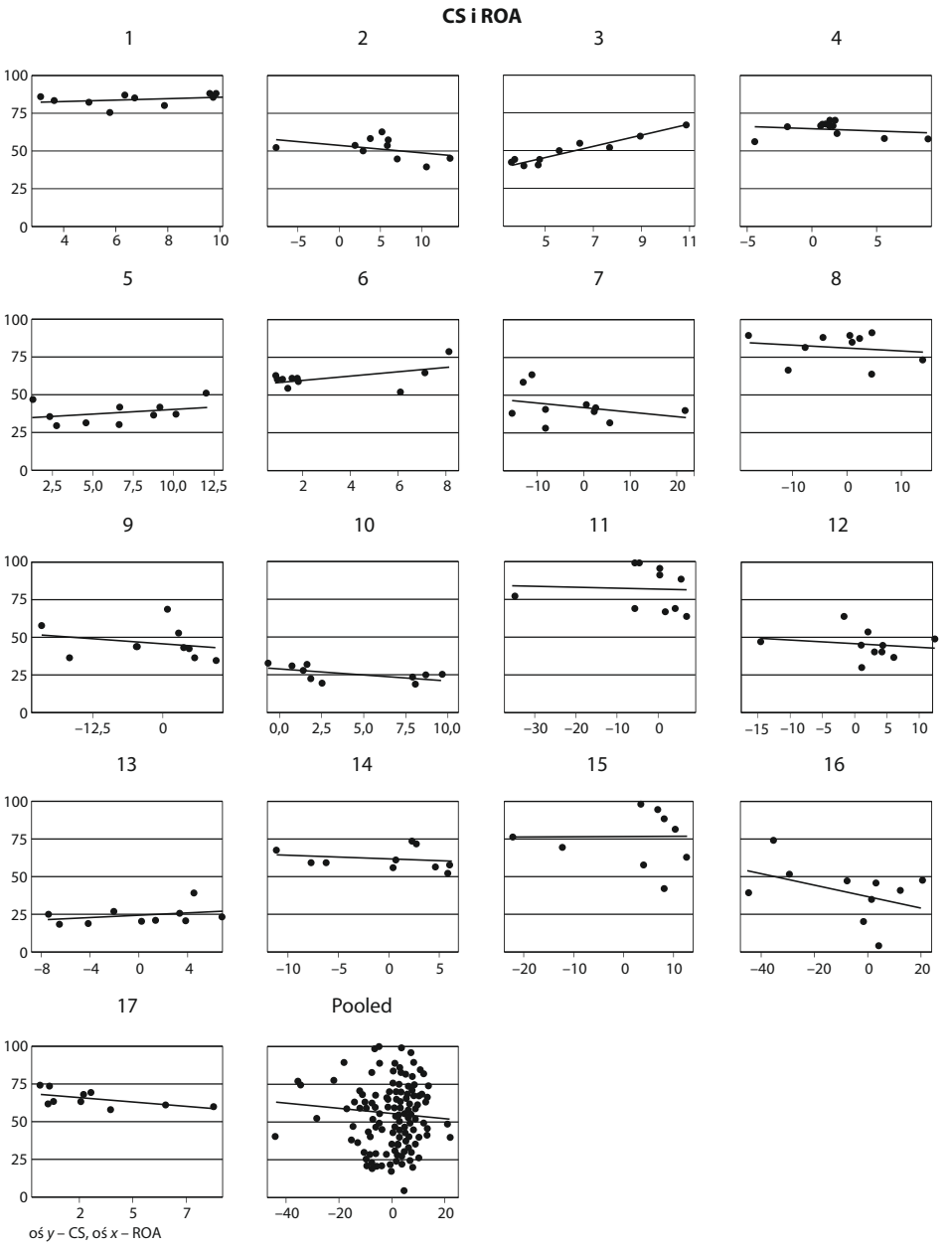
Zróżnicowane są również wartości „czasowego" współczynnika korelacji. W przypadku korelacji pomiędzy CS i SIZE w latach 2009–2018 dla sześciu spółek wartości te były ujemne, ale nieistotne, a dla 11 spółek – dodatnie, w tym dla ośmiu istotne na poziomie 0,05. Natomiast zależności pomiędzy CS i ROA dla 11 spółek były ujemne, ale nieistotne, a dla sześciu spółek – dodatnie, w tym dla jednej współczynnik korelacji był istotny (tabl. 3, wykr. 2).

Powyższa analiza nie daje syntetycznej odpowiedzi na pytanie, jakie są kierunek i siła zależności pomiędzy analizowanymi zmiennymi w badanym panelu.

Wykr. 2. Wykresy rozrzutu z uwzględnieniem struktury panelowej wraz z ilustracją liniowej korelacji pomiędzy zmiennymi CS i SIZE oraz CS i ROA według spółek



Wykr. 2. Wykresy rozrzutu z uwzględnieniem struktury panelowej wraz z ilustracją liniowej korelacji pomiędzy zmiennymi CS i SIZE oraz CS i ROA według spółek (dok.)



Uwaga. Numery od 1 do 17 oznaczają kolejne spółki (zob. tabl. 3). Wykresy zatytułowane „Pooled" dotyczą wszystkich obserwacji bez uwzględnienia struktury panelowej.
Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.) z wykorzystaniem pakietu oprogramowania R.

4. Propozycje pomiaru zależności pomiędzy zmiennymi o panelowej strukturze danych

Przedstawiona argumentacja nie oznacza, że nie należy badać zależności pomiędzy poszczególnymi zmiennymi, które są przewidziane do modelowania panelowego. Badanie zależności daje możliwość wyboru właściwego zestawu zmiennych objaśniających do modelu ekonometrycznego, jak również zbadania koincydencji parametrów (Hellwig, 1976). Jednak narzędzia oceny zależności muszą być adekwatne do panelowego charakteru danych. Nie ma panelowego współczynnika korelacji analogicznego do współczynnika Pearsona, ale podejmowane są próby rozwiązania tego problemu.

4.1. Metoda Famy-MacBetha

Jako pierwszą warto omówić metodę Famy-MacBetha (Fama i MacBeth, 1973), polegającą na szacowaniu w każdym roku parametrów (w tym wypadku są to wartości współczynnika korelacji liniowej Pearsona) na podstawie danych przekrojowych, a następnie – badaniu za pomocą testu t-Studenta istotności średnich wartości parametrów z całego analizowanego okresu. Fama i French (2002, s. 11–12) zaproponowali dodatkowo, aby w procesie wnioskowania wartość krytyczną statystyki t powiększyć 2,5 razy ze względu na autokorelację w czasie.

Już w latach 70. XX w. znane i stosowane były metody estymacji modeli panelowych z efektami ustalonymi (*least square with dummy variables* – LSDV) albo losowymi (*generalized least squares* – GLS) (Maddala, 2006, s. 644–648). Na początku lat 80. Anderson i Hsiao (1981) opracowali metodę estymacji dynamicznych modeli panelowych, a 10 lat później Arellano i Bond (1991) przedstawili uogólnioną metodę momentów (*general moments method*). Ale finansiści jeszcze w pierwszej dekadzie XXI w. stosowali metodę Famy-MacBetha do oceny istotności parametrów modeli panelowych[4].

Tabl. 4. Test t-Studenta istotności średniego współczynnika korelacji liniowej Pearsona w latach 2009–2018

Wyszczególnienie	CS i SIZE	CS i ROA
Średni r	0,5269	–0,0384
Statystyka t (9)	13,7686	0,3918

Uwaga. Przy poziomie istotności 0,05 dla dziewięciu stopni swobody wartość statystyki t wynosi 2,2622. Ze względu na autokorelację wartość krytyczna statystyki t wynosi $2{,}2622 \cdot 2{,}5 = 5{,}6555$.

Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.).

Na podstawie kryterium zaproponowanego przez Famę i Frencha można stwierdzić, że tylko zależność pomiędzy CS i SIZE okazała się istotna (w rzeczywistości współczynnik korelacji jest istotny na poziomie 0,001 – zob. tabl. 4).

[4] Jako przykłady można podać prace: Booth i Zhou (2008), Chay i Suh (2009), Denis i Osobov (2008), Fama i French (2001).

Metoda obliczania współczynnika korelacji dla danych panelowych według Famy-MacBetha ma dość poważną wadę. W przypadku gdy w poszczególnych grupach panelu występują korelacje dodatnie i ujemne, podczas obliczania średniej ich wartości mogą się wzajemnie „znosić", czego rezultatem jest ostateczny wynik bliski 0 (por. tabl. 4 dla korelacji CS i ROA), pomimo dość wysokich wartości dla poszczególnych grup. Rozwiązaniem tego problemu może być obliczanie średniej na podstawie wartości bezwzględnych współczynników korelacji uzyskanych na pierwszym etapie. Dzięki temu wyeliminowane zostaje „znoszenie się" wartości o przeciwnych znakach, a ostateczny współczynnik znacznie lepiej informuje o sile zależności między dwiema cechami z uwzględnieniem ich panelowej struktury. Odbywa się to kosztem utraty informacji o kierunku zależności, niemniej jednak autorzy niniejszej pracy proponują, żeby metodę Famy-MacBetha skorygować w opisany wyżej sposób i obliczać średnią na podstawie wartości bezwzględnych otrzymanych współczynników korelacji. Końcowy współczynnik będzie miarą unormowaną w przedziale [0, 1], informującą o sile zależności liniowej pomiędzy dwiema zmiennymi, z uwzględnieniem panelowego charakteru danych, chociaż niepokazującą kierunku tej zależności.

Tabl. 5. Absolutna średnia wartość współczynnika korelacji według skorygowanej metody Famy-MacBetha w latach 2009–2018

Wyszczególnienie	CS i SIZE	CS i ROA
Średni r	0,5269	0,2465
Statystyka t (9)	13,7686	4,4928

Uwaga. Przy poziomie istotności 0,05 dla dziewięciu stopni swobody wartość statystyki t wynosi 2,2622. Ze względu na autokorelację wartość krytyczna statystyki t wynosi $2{,}2622 \cdot 2{,}5 = 5{,}6555$.

Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.).

W przypadku zależności pomiędzy CS i SIZE – dla których wszystkie wartości współczynnika korelacji były dodatnie – wartość statystyki empirycznej się nie zmieniła. W przypadku zależności CS i ROA wartość statystyki t znacznie wzrosła, ale po zastosowaniu zaproponowanego przez Famę i Frencha skorygowanego kryterium istotności zależność dla wskazanych zmiennych była nadal nieistotna (tabl. 5).

4.2. Dobór zmiennych do mierników taksonomicznych konstruowanych na podstawie danych panelowych

Młodak i in. (2016, s. 8–9) zaproponowali metodę doboru zmiennych do mierników taksonomicznych, która uwzględnia panelowy charakter danych. W tym celu dla każdego roku policzyli macierz współczynników korelacji liniowej Pearsona pomiędzy potencjalnymi zmiennymi objaśniającymi. Następnie skonstruowali kompleksową macierz z tych współczynników korelacji, które miały największą wartość bez-

względną. Dalej przyjęto postępowanie analogiczne do postępowania przy metodzie odwróconej macierzy korelacji (Malina i Zeliaś, 1998; Młodak, 2006; Położij, 1966).

Kompleksowa macierz zaproponowana w tej metodzie nie jest co prawda macierzą korelacji, ponieważ nie zostały spełnione kryteria przechodniości skorelowania (Hellwig, 1976) i z tego powodu wartości na głównej przekątnej macierzy odwrotnej były nawet ujemne, niemniej jednak przy „wniesionej dozie subiektywizmu" (Młodak i in., 2016, s. 9) umożliwiła wybranie właściwych zmiennych o strukturze panelowej potrzebnych do konstrukcji miernika.

4.3. Współczynnik korelacji pomiędzy wartościami empirycznymi i teoretycznymi zmiennej objaśnianej modelu panelowego

Autorzy niniejszego artykułu proponują rozwiązanie polegające na szacowaniu modeli panelowych z jedną zmienną objaśniającą, którą jest kolejno każda ze zmiennych przeznaczonych do tej roli (w przykładzie przedstawionym w artykule: SIZE i ROA), oraz ocenie siły zależności za pomocą współczynnika korelacji liniowej Pearsona pomiędzy wartościami empirycznymi i teoretycznymi zmiennej objaśnianej (w przedstawionym przykładzie: CS). Kwadrat tego współczynnika jest często stosowany jako miara dopasowania modelu do danych empirycznych. Warto podkreślić, że w przypadku nieuwzględniania panelowego charakteru danych, a więc zastosowania metody najmniejszych kwadratów, współczynnik ten jest równy współczynnikowi korelacji liniowej Pearsona pomiędzy obiema zmiennymi, a jego kwadrat – współczynnikowi determinacji. Kierunek zależności określa się na podstawie znaku parametru przy zmiennej objaśniającej.

Każdy z modeli należałoby oszacować dwukrotnie: jako model z ustalonymi oraz losowymi efektami indywidualnymi i czasowymi. W omawianym przykładzie szacowane więc będą modele zmiennej CS w zależności od SIZE i od ROA, każdy w dwóch alternatywnych wersjach (łącznie cztery modele). Oczywiście należy się spodziewać innych wyników w zależności od przyjętego założenia o charakterze efektów, które mogą być stałe lub losowe[5].

Obliczenia rozpoczyna się od określenia zależności pomiędzy CS i SIZE. Oszacowany model, ze stałymi efektami indywidualnymi i czasowymi CS względem SIZE, pokazuje, że efekty indywidualne dla spółek w sposób istotny wpływały na zależność pomiędzy tymi zmiennymi (co oznacza odrzucenie hipotezy zerowej o wspólnym wyrazie wolnym), natomiast wpływ efektów czasowych był nieistotny na poziomie 0,05. Współczynnik korelacji wartości empirycznej i teoretycznej zmiennej objaśnianej CS wynosi 0,9077 ($p < 0{,}0001$). Jest on więc znacznie większy niż współczynnik

[5] Stanowi to problem, jeżeli ta procedura miałaby zostać zastosowana do doboru zmiennych, ponieważ może się okazać, że niektóre potencjalne zmienne objaśniające są silnie powiązane ze zmienną objaśnianą w przypadku założenia stałych efektów, a inne – w przypadku efektów losowych.

korelacji obliczony dla obu zmiennych przy założeniu, że dane nie mają charakteru panelowego (np. pochodzą ze 170 spółek w jednym roku). Oszacowany model z losowymi efektami indywidualnymi i czasowymi CS względem SIZE pokazuje również, że tylko efekty indywidualne istotnie wpływały na zależność pomiędzy tymi zmiennymi (tzn. brakuje podstaw do odrzucenia hipotezy zerowej o zgodności estymatora, czyli do odrzucenia hipotezy o efektach losowych – test Hausmana). W tym przypadku współczynnik korelacji wartości empirycznej i teoretycznej zmiennej objaśnianej CS wynosi 0,8959 (p < 0,0001). Jest on tylko nieznacznie mniejszy niż współczynnik korelacji obliczony dla modelu z ustalonymi efektami.

Z oszacowanych modeli o ustalonych i losowych efektach wynika, że parametr przy zmiennej SIZE jest dodatni i istotny statystycznie, a więc zależność pomiędzy CS i SIZE jest dodatnia i istotna. Jednocześnie siła tej zależności znacznie wzrosła (w porównaniu z zależnością mierzoną współczynnikiem korelacji liniowej Pearsona) po uwzględnieniu panelowego charakteru danych (tabl. 6, wykr. 3).

Tabl. 6. Panelowe zależności pomiędzy CS (zmienna objaśniana) oraz SIZE i ROA (zmienne objaśniające); modele z jedną zmienną

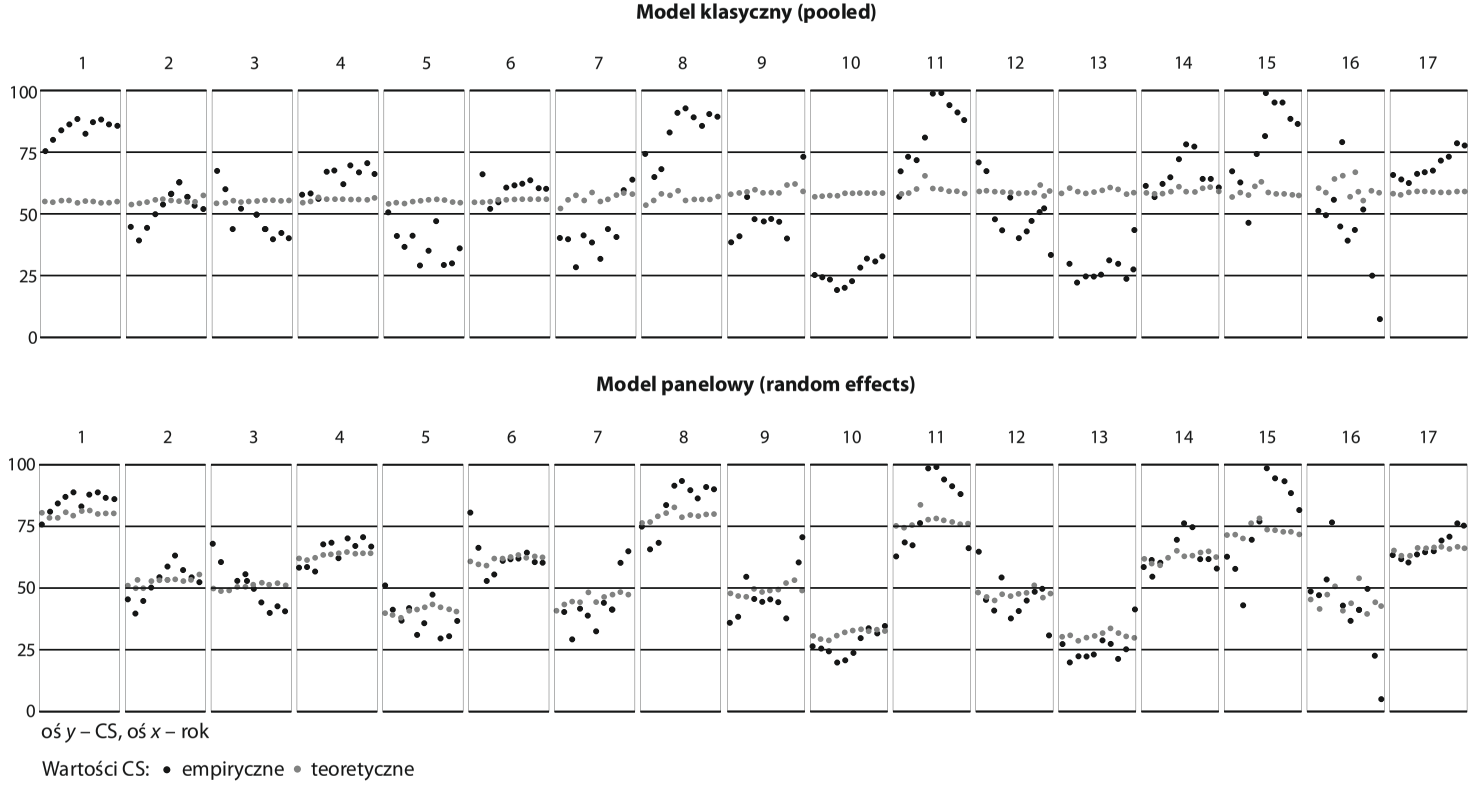
Modele	Wartość parametru przy zmiennej objaśniającej	Współczynnik korelacji Pearsona między wartościami teoretycznymi i empirycznymi CS[a]	Test Lagrange'a na istotność efektów. Hipoteza zerowa: efekty nieistotne. Statystyka testu LM[b]		Test Hausmana. Hipoteza zerowa: estymator GLS jest zgodny. Statystyka testu chi-kwadrat
			efekty indywidualne	efekty czasowe	
SIZE					
Pooled[c]	7,5195	0,5363	–	–	–
	(<0,0001)	(<0,0001)			
Panelowy: FE[d]	9,3440	0,9077	19,2000	–1,2388	0,1173
	(0,0391)	(<0,0001)	(<0,0001)	(0,8923)	(0,7320)
RE[e]	8,7319	0,8959	–	–	
	(<0,0001)	(<0,0001)			
ROA					
Pooled[c]	–0,1749	0,0737	–	–	–
	(0,3829)	(0,3394)			
Panelowy: FE[d]	–0,1989	0,9013	21,1840	–1,4453	0,0480
	(0,0297)	(<0,0001)	(<0,0001)	(0,9258)	(0,8266)
RE[e]	–0,2057	0,8962	–	–	
	(0,0063)	(<0,0001)			

a „Klasyczny" współczynnik korelacji liniowej Pearsona między dwiema zmiennymi. b Test Lagrange'a Multiplier z efektami two-way (Honda) dla paneli zbilansowanych. c Parametry szacowane klasyczną metodą najmniejszych kwadratów (KMNK), odporne błędy standardowe (HC1). d Model panelowy z ustalonymi efektami indywidualnymi i czasowymi, estymator KMNK na danych przekształconych, odporne błędy standardowe (HC1, Arellano, 1987). e Model panelowy z losowymi efektami two-way, estymator według metody Swamy-Arora, odporne błędy standardowe (HC1, Arellano, 1987).

Uwaga. W nawiasach wartości poziomu istotności p.

Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.).

Wykr. 3. Zależność pomiędzy wartościami empirycznymi i teoretycznymi zmiennej CS w modelach ze zmienną objaśniającą ROA według szeregów czasowych dla poszczególnych spółek



Uwaga. Numery od 1 do 17 oznaczają kolejne spółki (zob. tabl. 3).
Źródło: opracowanie własne na podstawie danych z Bankier.pl (b.r.) z wykorzystaniem pakietu oprogramowania R.

Podobne wyniki otrzymano w przypadku zależności pomiędzy CS i ROA. Warto jednak zauważyć, że współczynnik korelacji nieuwzględniający struktury panelowej okazał się ujemny i nieistotny statystycznie. Po uwzględnieniu panelowej struktury zmiennych zależność nadal była ujemna (ze względu na ujemną wartość oszacowanego parametru przy ROA), ale istotna statystycznie. A zatem także w tym przypadku uwzględnienie panelowej struktury zmiennych znacznie poprawiło siłę zależności. Gdyby jako kryterium doboru zmiennych przyjąć istotność współczynnika korelacji Pearsona, to wskaźnik ROA nie znalazłby się wśród zmiennych objaśniających. Natomiast ujemna i istotna zależność pomiędzy strukturą kapitału mierzoną udziałem zadłużenia w aktywach i rentownością została potwierdzona bardzo wieloma badaniami przeprowadzonymi na danych panelowych (Nehrebecka i in., 2016, s. 22–49).

Niedoskonałością zaproponowanej miary jest (w przeciwieństwie do współczynnika korelacji Pearsona) jej niesymetryczność. Wyznaczona w ten sposób zależność pomiędzy SIZE i CS nie będzie taka sama, jak pomiędzy CS i SIZE.

5. Podsumowanie

W artykule autorzy wykazują, że w przypadku zmiennych o panelowej strukturze danych nie należy obliczać wartości liniowego współczynnika korelacji Pearsona, ponieważ nie odzwierciedlają one panelowego charakteru danych. Wniosek ten dla większości badaczy jest oczywisty, ale w związku z tym, że wciąż publikowane są artykuły uwzględniające takie podejście, warto zabrać głos w tej sprawie.

Obliczanie wartości współczynnika korelacji Pearsona dla zmiennych o panelowej strukturze danych jest prostym przeniesieniem metodologii modelowania na podstawie szeregów czasowych lub szeregów przekrojowych, które rozpoczyna się od prezentacji statystyk opisowych i macierzy korelacji; w takim przypadku szacowanie wartości współczynnika Pearsona jest w pełni uzasadnione i zazwyczaj pomaga w wyborze zmiennych objaśniających do modeli.

Oczywiście nie oznacza to, że w ogóle nie należy liczyć zależności pomiędzy parami zmiennych o panelowej strukturze danych. Takie miary można wykorzystać przy opisie analizowanych zjawisk, jak również podczas wstępnego doboru zmiennych do modeli panelowych. Jednak miary te muszą uwzględniać panelowy charakter danych. Niekiedy w takiej sytuacji wykorzystuje się współczynnik Pearsona do obliczeń cząstkowych (pośrednich).

Zaproponowana w artykule miara wykorzystująca współczynnik korelacji pomiędzy wartościami empirycznymi i teoretycznymi modelu panelowego z jedną zmienną objaśniającą została skonstruowana przede wszystkim w celu pokazania braku przydatności współczynnika korelacji Pearsona do pomiaru zależności pomiędzy zmiennymi o strukturze panelowej. Ta miara, jakkolwiek bardziej adekwatna, jest niedoskonała. Jej mankamentem jest to, że w przeciwieństwie do współczynnika korelacji Pearsona jest ona niesymetryczna. Problemy mogą się pojawiać również przy próbach zastosowania tej miary do procedury doboru zmiennych. Na przykład

gdy jedna zmienna zostanie wybrana na podstawie modelu z ustalonymi efektami (ponieważ zastosowane testy pokazały, że ten model będzie właściwy), a druga – na podstawie modelu z efektami losowymi, to pojawia się wątpliwość, który model należy zastosować do oszacowania modelu panelowego zmiennej objaśnianej względem obu zmiennych.

Do pomiaru zależności pomiędzy zmiennymi o strukturze panelowej potrzebna jest miara o cechach współczynnika korelacji liniowej Pearsona (unormowanie na przedział [–1, 1], spełnienie zasady, zgodnie z którą im większa wartość bezwzględna, tym silniejsza zależność), a jednocześnie uwzględniająca panelowy charakter danych. Taką miarę można by nazwać „panelowym współczynnikiem korelacji".

Artykuł stanowi przyczynek do dyskusji nad badanym zagadnieniem. Nie przedstawiono w nim rozwiązań najodpowiedniejszych dla badania zależności między zmiennymi o panelowej strukturze danych, niemniej jednak zasugerowano pewne rozwiązania. Zwrócono przy tym uwagę na problem metodologiczny, który wymaga dalszych prac.

Bibliografia

Anderson, T. W., Hsiao, Ch. (1981). Estimation of Dynamic Models with Error Components. *Journal of the American Statistical Association*, *76*(375), 598–606. https://doi.org/10.2307/2287517.

Arellano, M. (1987). Computing Robust Standard Errors for Within-groups Estimators. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431–434. https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x.

Arellano, M., Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, *58*(2), 277–297. https://doi.org/10.2307/2297968.

Bankier.pl. (b.r.). *Notowania GPW – akcje*. Pobrane 20 lutego 2020 r. z https://www.bankier.pl/gielda/notowania/akcje.

Booth, L., Zhou, J. (2008). *Market Power and Dividend Policy: A Risk-Based Perspective.* https://dx.doi.org/10.2139/ssrn.1296940.

Borsuk, M., Kostrzewa, K. (2020). Miary ryzyka systemowego dla Polski. Jak ryzyko systemowe wpływa na akcję kredytową banków?. *Bank i Kredyt*, *51*(3), 211–238. https://bankikredyt.nbp.pl/content/2020/03/BIK_03_2020_01.pdf.

Chay, J. B., Suh, J. (2009). Payout Policy and Cash-Flow Uncertainty. *Journal of Financial Economics*, 93(1), 88–107. https://doi.org/10.1016/j.jfineco.2008.12.001.

Denis, D., Osobov, I. (2008). Why do firms pay dividends? International evidence on the determinants of dividend policy. *Journal of Financial Economics*, *89*(1), 62–82. https://doi.org/10.1016/j.jfineco.2007.06.006.

Driver, C., Grosman, A., Scaramozzino, P. (2020). Dividend policy and investor pressure. *Economic Modelling*, *89*, 559–576. https://doi.org/10.1016/j.econmod.2019.11.016.

Fama, E. F., French, K. R. (2001). Disappearing dividends: changing firm characteristics or lower propensity to pay?. *Journal of Financial Economics*, *60*(1), 3–43. https://doi.org/10.1111/j.1745-6622.2001.tb00321.x.

Fama, E. F., French, K. R. (2002). Testing Trade-Off and Pecking Order Predictions About Dividends and Debt. *The Review of Financial Studies*, *15*(1), 1–33. https://doi.org/10.1093/rfs/15.1.1.

Fama, E. F., MacBeth, J. D. (1973). Risk, Return, and Equilibrium: Empirical Tests. *Journal of Political Economy*, *81*(3), 607–636. https://doi.org/10.1086/260061.

Hellwig, Z. (1976). Przechodniość relacji skorelowania zmiennych losowych i płynące stąd wnioski ekonometryczne. *Przegląd Statystyczny, 23*(1), 3–20.

Herman, S. (2019). Impact of joint-stock companies' financial condition on real activities manipulation to manage earnings. *Wiadomości Statystyczne. The Polish Statistician*, *64*(10), 36–52. https://doi.org/10.5604/01.3001.0013.7588.

Karkowska, R. (2019). Systematic risk affected by country level development. The case of the European banking sector. *Argumenta Oeconomica*, (2), 255–282. https://doi.org/10.15611/aoe.2019.2.11.

Maddala, G. S. (2006). *Ekonometria*. Warszawa: Wydawnictwo Naukowe PWN.

Malina, A. (2002). Wielokryterialna taksonomia w analizie porównawczej struktur gospodarczych Polski. W: A. Zeliaś (red.), *Przestrzenno-czasowe modelowanie i prognozowanie zjawisk gospodarczych* (s. 305–312). Kraków: Wydawnictwo Akademii Ekonomicznej.

Malina, A., Zeliaś, A. (1998). On Building Taxonometric Measures on Living Conditions. *Statistics in Transition*, *3*(3), 523–544.

Młodak, A. (2006). *Analiza taksonomiczna w statystyce regionalnej*. Warszawa: Difin.

Młodak, A., Józefowski, T., Wawrowski, Ł. (2016). Zastosowanie metod taksonomicznych w estymacji wskaźników ubóstwa. *Wiadomości Statystyczne, 61*(2), 1–24. https://doi.org/10.5604/01.3001.0014.0900.

Nehrebecka, N., Białek-Jaworska, A., Dzik-Walczak, A. (2016). *Źródła finansowania przedsiębiorstw. Stan badań i ich metaanaliza*. Warszawa: Difin.

Pearson, K. (1895). Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, *58*(347–352), 240–242. https://doi.org/10.1098/rspl.1895.0041.

Pluskota, A. (2020). The Impact of Corruption on Economic Growth and Innovation in an Economy in Developed European Countries. *Annales Universitatis Mariae Curie-Skłodowska. Sectio H. Oeconomia*, *54*(2), 77–87. https://doi.org/10.17951/h.2020.54.2.77-87.

Położij, G. N. (1966). *Metody przybliżonych obliczeń*. Warszawa: Wydawnictwa Naukowo-Techniczne.

Urbanek, G. (2017). Analysing brand strength – corporate financial performance link for companies listed on the Warsaw Stock Exchange. *Ekonometria*, (2), 92–102. https://doi.org/10.15611/ekt.2017.2.06.

Witkowski, B. (2012). Modele danych panelowych. W: M. Gruszczyński (red.), *Mikroekonometria. Modele i metody analizy danych indywidualnych* (s. 267–308). Warszawa: Oficyna Wolters Kluwer Business.