DOI: 10.37317/biul-2011-0053

MARCIN KOZAK
Katedra Doświadczalnictwa i Bioinformatyki
Wydział Rolnictwa i Biologii
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie

Analiza związków przyczynowo-skutkowych w agronomii i hodowli roślin*

Analysis of cause-and-effect relationships in agronomy and plant breeding

Celem pracy jest dyskusja o analizie związków w systemach przyczynowo-skutkowych w agronomii i hodowli roślin. Systemy te są zawężone do takich, które obejmują związki liniowe między zmiennymi losowymi. Omówione są następujące metody statystyczne i zagadnienia: analiza ścieżek, sekwencyjna analiza plonu, dwukierunkowy podział zmienności plonu, oraz wnioskowanie o związkach przyczynowo-skutkowych dla populacji genotypów. Dyskusja teoretyczna opiera się na biologicznych aspektach rozwoju roślin uprawnych i ich współdziałaniu ze środowiskiem. W pracy ukazane są też praktyczne aspekty analizy związków przyczynowo-skutkowych, wskazując możliwości interpretacyjne konkretnych metod, a także ich ograniczenia.

Słowa kluczowe: analiza ścieżek, dwukierunkowy podział zmienności plonu, sekwencyjna analiza plonu

The paper discusses the cause-and-effect relationship systems in agronomy and plant breeding systems. The systems are limited to those which contain linear relationships between random variables. The following methods and problems are discussed: path analysis, sequential yield analysis, two-dimensional partitioning of yield variation, and analysis of cause-and-effect relationships for a population of genotypes. A theoretic discussion is based on biological aspects of plant development and their interaction with environment. The paper deals also with practical aspects of analyzing cause-and-effect relationships, pointing out interpretational possibilities of the methods discussed, but also their limitations.

Key words: path analysis, sequential yield analysis, two-dimensional partitioning of yield variation

1. WPROWADZENIE

W naukach rolniczych informacja o tym, co w badanym procesie jest przyczyną, a co skutkiem, jest niezwykle cenna: pozwala zrozumieć ów proces, oczywiście zakładając, że

* Praca była prezentowana w ramach I Warsztatów Biometrycznych, które odbyły się w IHAR-PIB w Radzikowie w dniach 14-15 września 2010 r

przynajmniej w części jego elementem jest zjawisko przyczynowości. Błędne określenie przyczyny i skutku może nie tylko prowadzić do błędnej interpretacji tego procesu, ale też czasami być wręcz komiczne. Wyobraźmy sobie, że badacz zastanawia się, co jest przyczyną, a co skutkiem w związku między dawką nawożenia azotem a plonowaniem pszenicy ozimej w danym sezonie wegetacyjnym. Zauważmy następstwo czasowe obu zjawisk, czyli nawożenia i plonowania: w danym sezonie najpierw nawozimy, a dopiero potem zbieramy plon. Pomyłka w tym przypadku, czyli uznanie, iż to poziom plonowania w danym sezonie jest przyczyną dawki nawożenia azotem w tym samym sezonie byłaby fatalna w skutkach.

Wiedza na temat charakteru związków w danym procesie (czyli rozpoznanie przyczyn i skutków) to jedno, a umiejętność ich statystycznego opisu i interpretacji to drugie. Jest to jeden z podstawowych celów statystyki: opis i interpretacja związków przyczynowo-skutkowych. Większość metod statystycznych opisuje właśnie związki przyczynowo-skutkowe, albo przynajmniej sposób ich działania można przedstawić w odniesieniu do tych związków.

Na przykład jednoczynnikowa analiza wariancji służy do porównania kilku populacji pod względem średniej wartości cechy, co jest niczym innym jak badaniem wpływu czynnika na cechę; jest więc to metoda badania związku przyczynowo-skutkowego między zmienną jakościową (czynnik) a ilościową zmienną losową (ewentualnie więcej zmiennych w wielowymiarowej analizie wariancji). Podobny związek z przyczynowością mają wieloczynnikowa analiza wariancji czy uogólnione modele liniowe, a nawet porównanie dwóch średnich (pod kątem wartości średniej, mediany, wariancji czy rozkładu zmiennej w populacjach). Analiza regresji jest powszechnie uznawaną metodą badania związków przyczynowo-skutkowych (zarówno liniowych, jak i nieliniowych). Należy zwrócić uwagę na to, że analiza regresji stosowana jest również wtedy, gdy wcale przyczynowości między zmiennymi nie ma (np. gdy celem analizy jest określenie najlepszej funkcji regresji między dwiema zmiennymi, choćby po to, aby wartości tej zmiennej, która jest obserwowana większym kosztem, można było szacować na podstawie drugiej zmiennej).

Są to wybrane metody statystyczne, te powszechnie znane, które badają związki przyczynowo-skutkowe i są stosowane w agronomii. Nie można jednak nie wspomnieć o jednej ogólnej tematyce, czy też grupie metod, której związek z przyczynowością jest oczywisty, a którą nazywa się „badaniem związków przyczynowo-skutkowych między losowymi zmiennymi ilościowymi", lub w skrócie „badaniem związków przyczynowo-skutkowych". Świadomie przedstawiłem powiązanie wcześniej wymienionych metod z przyczynowością, aby uniknąć błędu, w którym tylko i wyłącznie metody, o których będziemy mówić od tej chwili, zostałyby uznane za metody badania związków przyczynowo-skutkowych.

Tematyka ta obejmuje kilka metod czy podejść. Wszystkie one polegają na badaniu związków między losowymi zmiennymi ilościowymi, które tworzą system przyczynowo-skutkowy. Przez system przyczynowo-skutkowy będziemy rozumieć zespół cech, które biorą udział w badanym procesie biologicznym. System taki badany jest przy pomocy modelu związków przyczynowo-skutkowych; na podstawie analizy badacz decyduje, które zmienne mają się w tym modelu znaleźć, a decyzja ta zależy również od tego, jaki wycinek

procesu (czyli które cechy) jest badany. Ostateczna postać modelu zależy więc od tego, które zmienne znajdują się w kręgu zainteresowań badacza, oraz od tego, które z nich okażą się w modelu istotne.

Przykładem takiego modelowania może być badanie kształtowania plonu ziarna zbóż. Nie ma możliwości badania tego procesu ze wszystkich punktów widzenia naraz, gdyż wymagałoby to obserwacji zbyt wielu cech, robi się to więc na różne sposoby. Na przykład kształtowanie plonu ziarna bada się pod kątem uwarunkowania przez klasyczne składowe plonu, tj. liczbę kłosów na jednostce powierzchni, liczbę ziaren w kłosie oraz średnią masę ziarniaka (standardowo przeliczaną na masę tysiąca ziaren np. Rasmusson i Cannell, 1970; Dofing i Knight, 1992; Rozbicki, 1997; Mądry i in., 2003), albo przez inne, różnorodne cechy łanu i roślin niebędące składowymi plonu (np. Samonte i in., 1998, Mohammadi i in., 2003, Lorencetti i in., 2006, Samborski i in., 2006). Moglibyśmy mnożyć podobne przykłady, a każdy z nich byłby wycinkiem ogólnego procesu kształtowania plonu ziarna zbóż przez cechy łanu i roślin. Tego typu podejścia można łączyć w procesy ogólniejsze, każdy z nich można dzielić na procesy jeszcze bardziej szczegółowe, a które podejście, czy to ogólniejsze, czy to bardziej szczegółowe jest lepsze, nie sposób powiedzieć. Badanie dużej liczby zmiennych sprawia, iż patrzymy na badany proces bardziej ogólnie, uwzględniając współdziałania między zmiennymi, co z kolei jest pomijane, gdy zmienne nie są analizowane w jednym systemie. Z drugiej jednak strony, zbyt dużo zmiennych (których, nawiasem mówiąc, jednoczesna obserwacja może być bardzo kosztowna i pracochłonna) może wprowadzić pewien chaos do danych, co może z kolei sprawić, że pewne ważne zależności czy zmienne nie zostaną uznane przez analizę za istotne. Na przykład w analizie regresji wielokrotnej taka sytuacja występuje, gdy pewne zmienne przyczynowe są silnie skorelowane. Należy pamiętać, że odrzucenie zmiennej przez analizę nie musi oznaczać, że jest ona w badanym procesie nieważna w sensie biologicznym.

W niniejszym opracowaniu będziemy zajmować się następującymi wybranymi problemami i metodami badania związków przyczynowo-skutkowych w agronomii:

— analiza ścieżek,
— sekwencyjna analiza plonu,
— dwukierunkowy podział zmienności plonu,
— wnioskowanie o związkach przyczynowo-skutkowych dla populacji genotypów.

Ograniczymy się do liniowych związków między zmiennymi, które są bardzo częste w badaniach agronomicznych. Nie oznacza to, że związki nieliniowe nie występują albo są nieważne. Będziemy ponadto zakładać, że charakter związków przyczynowo-skutkowych między zmiennymi ilościowymi w badanym systemie jest badaczowi znany, a celem analizy jest ilościowy opis tego systemu. Oznacza to, że dany związek może istnieć (zakładamy jedynie konkretny kierunek wpływu w danym związku przyczynowo-skutkowym); stwierdzenie, czy jest on istotny i silny, jest celem analizy. W związku z tym możemy uznać, że celem analizy systemu związków przyczynowo-skutkowych między zmiennymi jest odnalezienie modelu jak najlepiej opisującego ten system. Z końcowego modelu można wnioskować o tym, które z postulowanych zależności okazały się istotne i, już tylko w odniesieniu do tych zależności, jaki jest ich charakter, czy są one dodatnie czy

ujemne, silne czy słabe. O systemie można się wypowiedzieć jeszcze bardziej wnikliwie, poprzez analizę oszacowanego modelu, która wykorzystuje pojęcia różnych efektów: bezpośrednich, pośrednich, całkowitych czy niezależnych.

Wynika z tego, że nie będziemy zajmować się badaniem kierunku wpływu w związkach między zmiennymi w systemie (czyli badaniem, która ze zmiennych jest przyczyną której), co nie stanowi zubożenia analizy, gdyż w większości badań agronomicznych jesteśmy w stanie stwierdzić, które z cech mogą być przyczyną lub skutkiem innych cech. Nie oznacza to, że nie istnieją metody, których celem jest odnalezienie przyczyny (przyczyn) i skutku (skutków) w danym procesie. Na przykład w modelowaniu równań strukturalnych (Shipley, 2002) możemy dobierać model, uwzględniając również różne kierunki wpływu między zmiennymi. Tematyka ta może zostać uznana za kontrowersyjną, bowiem czy dane liczbowe, bez żadnej dodatkowej informacji o badanym procesie, mogą dać odpowiedź na takie pytanie? Jedni mówią, że mogą, inni, że nie. Są to zagadnienia zarówno podejmowane przez filozofów nauki (np. Dowe, 1992; Steel, 2005), jak i statystyków (Shipley, 2002), my jednak nie będziemy zajmować się nimi zajmować. Tak jak m.in. Reynolds i in. (2007), uznajemy, że możliwe kierunku wpływu między zmiennymi powinny być ustalone przed analizą.

W dyskusji o metodach badania liniowych związków przyczynowo-skutkowych nie będziemy prezentować ich szczegółów; można się z nimi zapoznać w cytowanych pracach, a w przypadku analizy ścieżek w sporej liczbie innych opracowań.

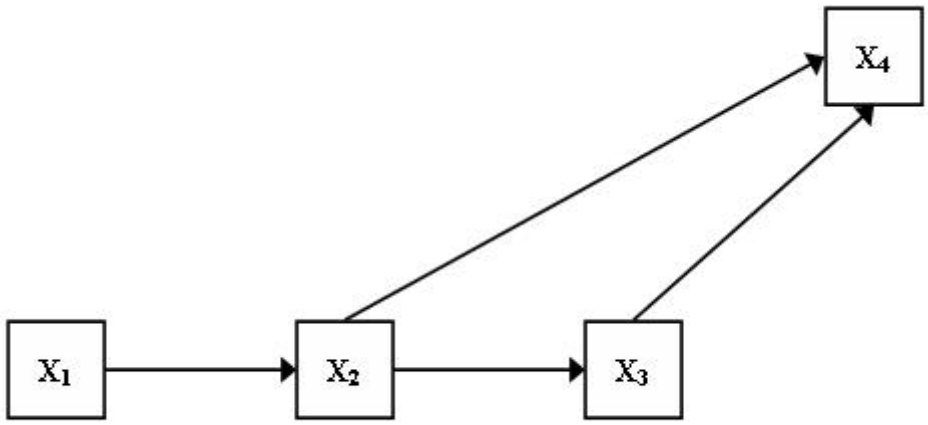
2. ANALIZA ŚCIEŻEK

Analiza ścieżek jest podstawową metodą analizy liniowych związków przyczynowo-skutkowych między zmiennymi. Jej twórcą jest Sewall Wright (1921, 1934); Wolfe (1999) przedstawił obszerny komentarz na temat bibliografii autorstwa Sewalla Wrighta dotyczącej analizy ścieżek. O analizie ścieżek napisano już wiele prac, zastosowano ją w wielu problemach badawczych. W naukach rolniczych za jeden z pionierskich artykułów, w których przedstawiono zastosowanie analizy ścieżek, uznaje się pracę Dewey i Lu (1957). Metoda ta w tej lub zbliżonej formie wciąż jest stosowana i ma wielu zwolenników, choć od pewnego czasu klasyczna analiza ścieżek (Przez klasyczną analizę ścieżek będę rozumiał tę opartą na metodologii Sewalla Wrighta, włączając w to również dalsze jej modyfikacje) znalazła konkurencyjne podejście, które stanowi element modelowania równań strukturalnych (ang. Structural Equation Modeling — SEM), w którym estymację i testowanie modelu związków przyczynowo-skutkowych prowadzi się na podstawie metody największej wiarygodności. Kozak i Kang (2006) podkreślili potrzebę stosowania nowego podejścia do estymacji w analizie ścieżek w naukach rolniczych, w których takich zastosowań do tej pory było niewiele (np. Guillen-Portal i in., 2006; Dhungana i in., 2007; Vargas i in., 2007; Kozak i in., 2007 c, 2008).

Interpretacja w analizie ścieżek opiera się na trzech podstawowych rodzajach efektów między zmiennymi: efektu bezpośredniego, pośredniego i całkowitego. Wciąż panuje opinia, że efekt bezpośredni jest podstawowym źródłem informacji o wpływie danej zmiennej przyczynowej na zmienną skutkową; nie jest to słuszna opinia, o czym

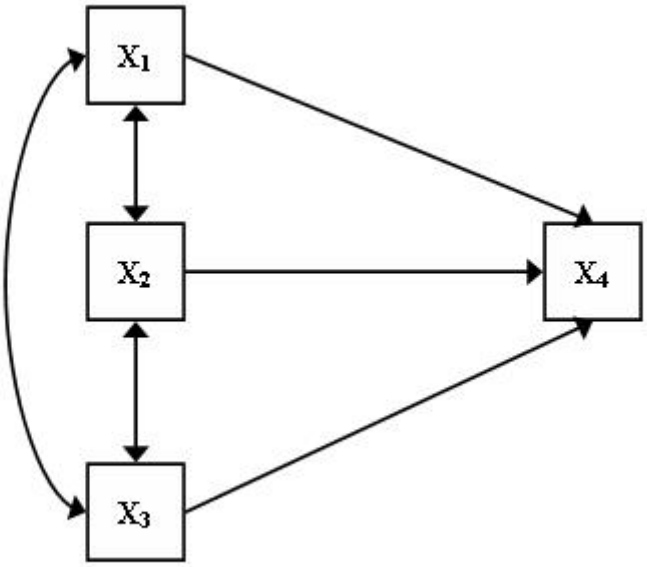
wspomnimy w tym opracowaniu. Efekt bezpośredni opisuje zmianę jednej zmiennej w reakcji na jednostkowy wzrost wartości drugiej zmiennej, gdy wartość pozostałych zmiennych nie zmienia się. Choć powszechnie uznaje się, iż zmiany te mierzone są w jednostkach odchylenia standardowego zmiennych (czyli efekt bezpośredni mierzy zmianę jednej zmiennej w jednostkach jej odchylenia standardowego, gdy wartość zmiennej na nią wpływającej zwiększy się o jej odchylenie standardowe, a wartości pozostałych zmiennych nie zmieniają się), to można je mierzyć w jednostkach oryginalnych zmiennych (Shipley, 2002). Na diagramie przyczynowo-skutkowym efekt bezpośredni jest oznaczany przy pomocy strzałki jednostronnej (→), która opisuje wpływ zmiennej, od której strzałka wychodzi, na zmienną, do której strzałka jest skierowana. Strzałkę tę nazwiemy ścieżką pojedynczą.

Efekt pośredni jest na diagramie reprezentowany przez tzw. ścieżkę złożoną, czyli taką, która biegnie wzdłuż co najmniej dwóch ścieżek pojedynczych. Jest on obliczany jako iloczyn efektów bezpośrednich odpowiadających poszczególnym ścieżkom pojedynczym, składającym się na tę ścieżkę złożoną. Efektów pośrednich między dwoma zmiennymi w systemie może więc być więcej niż jeden. Na przykład między zmienną X_1 i X_4 na rys. 1 możemy wyróżnić dwa efekty pośrednie (i dwie odpowiadające im ścieżki złożone): $X_1 \rightarrow X_2 \rightarrow X_4$ oraz $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ (brak jest za to efektu bezpośredniego między X_1 i X_4, więc zmienna X_1 wpływa na X_4 tylko pośrednio). Z kolei zmienna X_2 wywiera na X_4 wpływ bezpośredni ($X_2 \rightarrow X_4$) oraz pośredni, wzdłuż ścieżki: $X_2 \rightarrow X_3 \rightarrow X_4$. Efekt pośredni $X_2 \rightarrow X_3 \rightarrow X_4$ należy rozumieć jako efekt następującego procesu: zmienia się zmienna X_2, co, ze względu na jej wpływ na X_3, generuje zmianę X_3, ta zaś zmiana powoduje zmianę X_4.



Rys. 1. Przykładowy model analizy ścieżek złożonych
Fig. 3. The example of a complex path model

Jest jeszcze drugi sposób definiowania efektów pośrednich, popularny w przypadku badania modeli ścieżek pojedynczych. Termin ten został użyty np. przez Jończyk (2002) oraz Mądrego i in. (2003)). W modelu takim wszystkie zmienne przyczynowe znajdują się na tym samym poziomie ontogenetycznym i są skorelowane, reprezentuje więc on model wielokrotnej analizy regresji (Shipley, 2002; rozdz. 4.3); por. rys. 2). Efekt pośredni jest tu rozumiany jako efekt jednej zmiennej przyczynowej na zmienną skutkową poprzez inną zmienną przyczynową. Kozak i in. (2007 a) pokazali, że nie jest to prawidłowe podejście ze względu na brak związku przyczynowo-skutkowego między zmiennymi przyczynowymi, oraz zaproponowali, jak można interpretować tego typu efekty.



Rys. 2. Model analizy ścieżek pojedynczych dla trzech zmiennych przyczynowych i jednej skutkowej
Fig. 2. A model of single paths for the three dependent traits and one independent trait

Dla modelu ścieżek pojedynczych efekt całkowity równy jest współczynnikowi korelacji między zmienną przyczynową a zmienną skutkową. Jest to bardzo pożądana właściwość, jako że współczynnik korelacji opisuje ogólną (ostateczną, przy zignorowaniu pozostałych przyczyn i skutków i zmiennych) zależność dwóch zmiennych, którą to powinien opisywać efekt całkowity. W związku z tym dla tego modelu często wnioskuje się na podstawie tzw. podziału współczynnika korelacji między daną parą zmiennych (przyczynową i skutkową) na efekt bezpośredni oraz sumę efektów pośrednich poprzez inne zmienne.

Dla modelu ścieżek złożonych efekt całkowity ma nieco inne znaczenie niż dla ścieżek pojedynczych. Jest on sumą wszystkich efektów jednej zmiennej na drugą zmienną. Na przykład efekt całkowity wpływu X_1 na X_4 to suma następujących efektów pośrednich: $X_1 \rightarrow X_2 \rightarrow X_4$ oraz $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$, zaś efekt całkowity wpływu X_2 na X_4 to suma efektu bezpośredniego $X_2 \rightarrow X_4$ oraz efektu pośredniego $X_2 \rightarrow X_3 \rightarrow X_4$. Opisuje więc on ostateczną zmianę zmiennej przyczynowej w reakcji na zmianę zmiennej skutkowej. Dla zmiennych standaryzowanych jest to informacja o całkowitym (ostatecznym) wpływie zmiennych przyczynowych na zmienną skutkową, więc to właśnie efekt całkowity, a nie bezpośredni, niesie najwięcej informacji o wpływie jednej zmiennej na drugą, natomiast jego podział na efekty cząstkowe (bezpośredni i pośrednie) pomaga zrozumieć ten wpływ. Z przykładową interpretacją tego typu modeli opartej na efektach bezpośrednich, pośrednich i całkowitych można zapoznać się w pracach Kozaka i in. (2007 c, 2008). Niestety, efekt całkowity między dwoma zmiennymi nie jest równy współczynnikowi korelacji między nimi; o skutkach tej właściwości będziemy mówić później.

Estymacja w analizie ścieżek

Klasyczna estymacja w analizie ścieżek jest prowadzona metodą najmniejszych kwadratów, czyli, najprościej mówiąc, wykorzystuje metodologię analizy liniowej regresji wielokrotnej dla zmiennych standaryzowanych. Od pewnego jednak czasu analiza ścieżek to dużo więcej niż tylko podejście regresyjne dla zmiennych standaryzowanych; to już odrębne podejście, która pozwala spojrzeć na model ścieżek złożonych nie tylko jako na zbitek modeli regresyjnych (a tak właśnie patrzeć należałoby w przypadku „klasycznej" analizy ścieżek), lecz jako na jeden ogólny model, w którym znajdują się zmienne tzw. egzogeniczne (ang. exogenous variables; są to zmienne, które są przyczynami innych zmiennych w modelu, lecz które w modelu nie mają przyczyny), endogeniczne (ang. endogenous variables; są to zmienne, które mają przyczyny wśród zmiennych w modelu, ale same nie są przyczynami; zmienne te często nazywane też są zmiennymi skutkowymi) i interweniujące (ang. intervening variables; są to zmienne, które są zarówno przyczynami, jak i skutkami innych zmiennych w modelu). Ponadto każda ze zmiennych jest determinowana przez swoją zmienną resztową, wyjaśniającą tę część zmienności tej zmiennej, która nie jest tłumaczona przez jej przyczyny (zmienna resztowa jest też uwzględniana w klasycznej analizie ścieżek).

Obecnie estymacja w analizie ścieżek prowadzona jest przy pomocy metody największej wiarygodności, w której jednocześnie estymowane są wszystkie parametry (współczynniki) modelu (w przeciwieństwie do klasycznej analizy ścieżek, w której estymacja współczynników prowadzona jest niezależnie dla poszczególnych „pod-modeli", czyli modeli częściowych, składających się na pełny model). Taki model częściowy jest tworzony przez zmienną skutkową lub interweniującą i jej zmienne przyczynowe, więc jest tyle „pod-modeli", ile w modelu zmiennych endogenicznych i interweniujących). Szczegóły estymacji przedstawione są np. w książce Shipleya (2002). Należy zwrócić uwagę na pewien element testowania (które jest prowadzone przy pomocy testu χ^2), mianowicie na to, że polega ono na tym, że model jest odrzucany (przy określonym poziomie istotności) lub nie, czyli ze statystycznego punktu widzenia ten model jest najlepszy (spośród testowanych), dla którego osiągnięto największą wartość *p*.

Biorąc pod uwagę założoną wartość współczynnika istotności testu α, może się okazać, że kilka lub nawet wiele modeli nie zostało przez test odrzuconych, a tym samym możemy uznać je za istotne. W takim przypadku nie zaleca się wyboru tego modelu, dla którego uzyskano największą wartość *p*, lecz tego spośród modeli istotnych, który ma największe przesłanki merytoryczne (Shipley, 2002).

Co ciekawe, mimo że nowa metodologia analizy ścieżek nie jest bardzo młoda, nawet z punktu widzenia teorii naukowych, to jednak wciąż w naukach rolniczych jej zastosowania są sporadyczne. Co więcej, te kilka czy kilkanaście prac, które choćby częściowo sięgają po nową metodologię, to tylko kropla w morzu zastosowań analizy ścieżek, zwłaszcza w porównaniu do innych, wybranych dziedzin nauki (choćby psychologii czy nawet ekologii roślin, którą reprezentuje profesor Bill Shipley, autor jednej z ważniejszych książek dotyczących nowej metodologii analizy ścieżek).

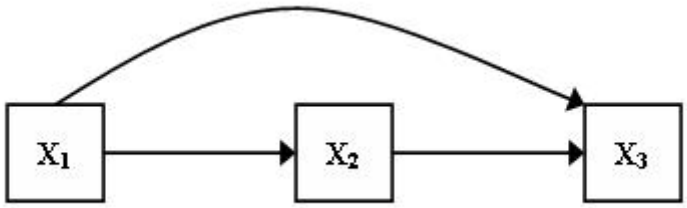
Jak podkreślili Kozak i Kang (2006), są sytuacje badawcze, w których nowe podejście do analizy ścieżek nie może zostać zastosowane. Może na przykład się okazać, że żaden model nie jest możliwy do zaakceptowania (czyli, w języku statystycznym, każdy model jest odrzucany, a dokładniej mówiąc, odrzucana jest hipoteza zerowa, że model jest prawidłowy) — z taką sytuacją przyszło się zmierzyć np. Gozdowskiemu i in. (2007). Autorzy uznali, że ich problemy z doborem próby wynikały prawdopodobnie z dwóch przyczyn. Po pierwsze, operowali na bardzo licznych próbach, które według Shipleya (2002) powodują, że test ma bardzo dużą moc, co z kolei sprawia, że hipoteza zerowa o tym, że model jest prawidłowy, jest łatwo odrzucana (Shipley, 2002; rozdz. 6.7). Po drugie, w modelu autorzy uwzględniali między innymi plon ziarna z kłosa oraz jego dwie składowe, liczbę ziaren w kłosie i średnią masę ziarniaka. Zależność plonu od składowych jest opisywana modelem multiplikatywnych (Kozak i Mądry, 2006), a nie liniowym; w związku z tym wszystkie modele zawierające te trzy zmienne były odrzucane, jako niespełniające warunku liniowości między zmiennymi. Możemy dodać jeszcze jedną możliwą przyczynę problemów z doborem modelu w analizie Gozdowskiego i in. (2007): łączenie danych z poziomów czynników (trzy lata, dwa terminy siewu i cztery dawki nawożenia azotem). Choć podejście to jest w agronomii bardzo częste (np. Sinebo, 2002; García del Moral i in., 2003; Mohammadi i in., 2003, Guillen-Portal i in., 2006), nie zawsze jest ono właściwe, na co wskazuje Shipley (2002, rozdział 7.2), co jednak nie oznacza, że zawsze jest nieprawidłowe.

Może się też zdarzyć, że badany model nie ma wystarczającej liczby stopni swobody, taki model nie może być estymowany przy pomocy metody największej wiarygodności (Shipley, 2002; rozdział 4.3). Taka sytuacja występuje na przykład wtedy, gdy model regresji wielokrotnej jest przedstawiany jako model ścieżkowy (Shipley, 2002; rozdział 4.3, Kozak i Kang, 2006), czyli dla modelu ścieżek pojedynczych, a także w przypadku sekwencyjnej analizy plonu, o której będziemy mówić w następnym rozdziale.

3. ANALIZA ZWIĄZKÓW PRZYCZYNOWO-SKUTKOWYCH MIĘDZY ZMIENNYMI W ŁAŃCUCHU ONTOGENETYCZNYM

3.1. Sekwencyjna analiza plonu

W agronomii bardzo ważnym elementem badań związków przyczynowo-skutkowych jest tzw. podejście ontogenetyczne (Kozak, 2007). Nie jest to metodologia statystyczna sensu stricte, a raczej określone podejście do analizy systemów przyczynowo-skutkowych o specyficznej postaci tzw. łańcucha ontogenetycznego. Ogólna definicja łańcucha ontogenetycznego jest taka, że zmienne go tworzące rozwijają się w określonej kolejności ontogenetycznej, co może mieć wpływ na postulowaną postać związków między nimi. Łańcuch ontogenetyczny jest jednym z wielu możliwych systemów przyczynowo-skutkowych, często spotykanym w agronomii. Kozak (2007) przeprowadził dyskusję na temat modelowania zależności między zmiennymi w łańcuchu ontogenetycznym, wskazując na to, że powinno być ono prowadzone przy pomocy modelu, w którym zakłada się, że kolejna zmienna w modelu może (ale nie musi) być skutkiem zmiennych ją poprzedzających w rozwoju ontogenetycznym, jak również może (ale nie musi) być przyczyną kolejnych zmiennych w modelu; model tego typu (dla wygody dalej nazywany modelem sekwencyjnej analizy plonu), dla trzech zmiennych, przedstawiony jest na rys. 3. W innym, alternatywnym podejściu do analizy tego typu zależności wykorzystuje się warunek Markova, w którym kolejna zmienna przyczynowa jest efektem tylko i wyłącznie jednej zmiennej, tej, która ją poprzedza w łańcuchu, oraz jest skutkiem tylko i wyłącznie jednej zmiennej, tej, którą poprzedza. W agronomii podejście to zwykle nie ma uzasadnienia merytorycznego (biologicznego) (Kozak, 2007).



Rys. 3. Model sekwencyjnej analizy plonu dla trzech zmiennych w łańcuchu ontogenetycznym
Fig. 3. A sequential yield analysis model for three traits in a ontogenetic chain

Podejście ontogenetyczne polega na uwzględnieniu w analizie kolejności pojawiania się zmiennych w przyrodzie, co z kolei przekłada się na to, która ze zmiennych może być przyczyną której; zakłada się, że zmienna, która kończy swój rozwój wcześniej niż inna zmienna, nie może być jej skutkiem, lecz może być jej przyczyną (może, bowiem nie jest to warunek wystarczający, o tym decyduje już sam biologiczny charakter procesu). Podejście ontogenetyczne wprowadzono już kilka dziesiątków lat temu (np. Grafius 1969, Rasmusson i Cannell, 1970; Thomas i in., 1971), choć dość długo trwała dyskusja nad jego słusznością (przykład takiej dyskusji w odniesieniu do analizy multiplikatywnych składowych plonu można znaleźć w pracach Spaarnaija i Bosa (1993), Piepho (1995) oraz

Kozaka i Mądrego (2006)). Obecnie uznaje się, że podejście to jest jak najzupełniej słuszne, gdyż uwzględnia biologiczny charakter systemów przyczynowo-skutkowych (oczywiście tylko i wyłącznie wtedy, gdy ów charakter w procesie występuje).

Jedną z metod analizy liniowych związków przyczynowo-skutkowych między zmiennymi w łańcuchu ontogenetycznym jest sekwencyjna analiza plonu (Mądry i inni 2005). W metodzie tej dokonuje się podziału współczynnika determinacji zmiennej skutkowej (ostatniej zmiennej w łańcuchu) na współczynniki niosące informację o niezależnym udziale zmiennych przyczynowych (czyli egzogenicznych i interweniujących) w determinacji zmiennej skutkowej (dalej opiszemy, na czym polega owa „niezależność"). Jest to metoda zbliżona do analizy ścieżek, choć często prowadząca do uzyskania innych wyników, pomimo tego że obie metody wykorzystują ten sam diagram zależności do badania związków między zmiennymi w łańcuchu ontogenetycznym. Diagram zależności (przyczynowo-skutkowych, diagram ścieżkowy) jest to graficzna prezentacja modelu oparta na metodzie grafów. Jest to nieodłączny sposób prezentacji modelu analizy ścieżek oraz równań strukturalnych. Szczegóły można znaleźć np. w książce Shipleya (2002). Należy też podkreślić, że wyniki obu metod nie są bezpośrednio porównywalne, bowiem analiza ścieżek nie bada niezależnego wpływu zmiennych przyczynowych na zmienną skutkową. Do tej pory w literaturze porównano sekwencyjną analizę plonu jedynie z analizą ścieżek pojedynczych (np. teoretyczne porównanie w pracy Mądrego i in., 2005), lecz brakuje jej porównania z analizą ścieżek złożonych, czyli z uwzględnieniem podejścia ontogenetycznego, a dopiero w takim przypadku obie metody analizowałyby ten sam model. Na podstawie dostępnej literatury nie sposób więc stwierdzić, która z tych metod jest lepsza czy bardziej prawidłowa.

Sam model sekwencyjnej analizy plonu ma swe korzenie w pracach Grafiusa (1969), Rasmussona i Cannella (1970) i Thomasa i in. (1971), lecz w obecnej formie sekwencyjna analiza plonu opiera się na metodologii zaproponowanej prawdopodobnie przez Eatona i Kyte'a (1978) i Eatona i MacPhersona (1978). W pracach tych autorzy badali zależność plonu od jego multiplikatywnych składowych plonu rozwijających się w łańcuchu ontogenetycznym; model multiplikatywny przekształcano przy pomocy transformacji logarytmicznej, co prowadzi do modelu addytywnego ze składowymi i plonem w skali logarytmicznej. Metoda ta została również przystosowana do analizy zmiennych w łańcuchu ontogenetycznym, które nie są składowymi plonu; źródłem tego podejścia była prawdopodobnie propozycja Gołaszewskiego (1996), według którego analizie można poddać zmienne nietransformowane logarytmicznie. Na podstawie tego pomysłu analizę tę można stosować dla wszelkich zmiennych, dla których związki liniowe mają postać łańcucha ontogenetycznego. Dla odróżnienia od sekwencyjnej analizy składowych plonu metodę tę nazwano sekwencyjną analizą plonu, w skrócie SAP (Mądry i Kozak, 2000).

Dokładny opis omawianej metody znajduje się np. w pracy Mądrego i in. (2005). Warto zwrócić uwagę na pewien element tej metody, który może czynić ją atrakcyjną w porównaniu np. do analizy ścieżek. W metodzie SAP wnioskowanie opiera się na tzw. wpływie niezależnym danej zmiennej przyczynowej na zmienną skutkową w łańcuchu ontogenetycznym. Dla danej, j-tej zmiennej przyczynowej (na razie załóżmy, że $1 < j < k$, gdzie k jest liczbą zmiennych przyczynowych w łańcuchu; łańcuch więc składa się z $k + 1$

zmiennych — k przyczynowych i jedna skutkowa) niezależny wpływ na zmienną skutkową oznacza wpływ niezależny od zmiennych, które ją poprzedzają w łańcuchu (czyli wyodrębniamy z tego efektu ten wpływ j-tej zmiennej, który wynika z jej zależności od zmiennych poprzedzających ją w łańcuchu); we wpływie tym są natomiast zawarte efekty pośrednie poprzez zmienne, które występują w łańcuchu po j-tej zmiennej, gdyż mają one swe źródło w tej (j-tej) zmiennej, która generuje (a dokładniej: może generować) zmiany kolejnych po niej zmiennych.

Interpretacja w sekwencyjnej analizie plonu opiera się przede wszystkim na dwóch współczynnikach. Pierwszym z nich jest p_i, który jest w pewnym stopniu odpowiednikiem efektu całkowitego (w wersji standaryzowanej), ale opisuje nie wpływ bezpośredni, lecz wpływ niezależny i-tej zmiennej przyczynowej na zmienną skutkową. Informuje on o zmianie zmiennej skutkowej (w jednostkach odchylenia standardowego) na skutek wzrostu i-tej zmiennej przyczynowej (o wartość odchylenia standardowego), ale nie oryginalnej, lecz tej jej części, która jest niezależna (dokładniej mówiąc, niezależna liniowo, gdyż zajmujemy się wyłącznie związkami liniowymi) od zmiennych poprzedzających ją w łańcuchu; bierze się przy tym pod uwagę zarówno efekt bezpośredni, jak i efekty pośrednie i-tej zmiennej. Taką zmienną niezależną od poprzedzających ją zmiennych nazywamy zmienną ortogonalną. Drugim przydatnym współczynnikiem jest kwadrat współczynnika p_i, czyli p_i^2, który ma tę ważną właściwość, że suma tych współczynników po wszystkich $i = 1,...,k$ równa jest wartości współczynnika determinacji dla modelu liniowego ze zmienną skutkową i k zmiennymi przyczynowymi (zarówno oryginalnymi, jak i ortogonalnymi). W związku z tym p_i^2 informują o niezależnym udziale zmiennych przyczynowych w determinacji zmiennej skutkowej.

Więcej szczegółów dotyczących sekwencyjnej analizy plonu można znaleźć w pracach cytowanych w tym rozdziale, np. u Mądrego i in. (2005). Metoda ta, zwłaszcza aby mogła znaleźć szerszy zakres zastosowań, powinna być rozwinięta w kierunku analizy bardziej rozbudowanych systemów przyczynowo-skutkowych niż łańcuch ontogenetyczny. Kozak i in. (2006 a) zaproponowali podejście do analizy systemu, w którym zmienne przyczynowe są ułożone w łańcuch ontogenetyczny, na końcu którego znajduje się kilka zmiennych skutkowych (autorzy przedstawili przykład dla dwóch zmiennych skutkowych, plonu ziarna jęczmienia jarego oraz zawartości białka w ziarnie [zmienna jakościowa plonu], analizowanych jako cechy wynikowe procesu kształtowania składowych plonu). Podejście to polega na zastosowaniu ortogonalizacji Grama-Schmidta dla zmiennych przyczynowych (przez co stają się one liniowo niezależne — jest to istotny element sekwencyjnej analizy plonu), po czym transformowane dane poddawane są analizie korelacji kanonicznych, w której jednym zbiorem zmiennych są ortogonalne zmienne przyczynowe, zaś drugim — zmienne skutkowe. Dzięki temu badacz uzyskuje informację o niezależnym wpływie zmiennych przyczynowych na zmienne skutkowe, przy czym brany jest pod uwagę wielowymiarowy charakter zbioru zmiennych skutkowych.

Podejście to, niewątpliwie atrakcyjne z tego względu, że bada niezależny wpływ zmiennych przyczynowych w łańcuchu ontogenetycznym na kilka zmiennych skutkowych, jest dopiero pierwsza próbą uogólnienia sekwencyjnej analizy plonu dla bardziej zaawansowanych systemów przyczynowo-skutkowych. Udostępnia jednak ono

inne, uboższe możliwości interpretacyjne niż sekwencyjna analiza plonu, bowiem nie rozpatruje się w nim podziału współczynnika determinacji zmiennej skutkowej, co wydaje się być atrakcyjnym elementem interpretacji w sekwencyjnej analizie plonu.

3.2. Analiza związków między zmiennymi w łańcuchu ontogenetycznym: sekwencyjna analiza plonu a analiza ścieżek

Porównanie sekwencyjnej analizy plonu oraz analizy ścieżek ma sens jedynie dla analizy łańcucha ontogenetycznego, bowiem pierwsza z tych metod może być zastosowana wyłącznie w analizie systemu przyczynowo-skutkowego między zmiennymi w łańcuchu ontogenetycznym. W związku z tym cały niniejszy podrozdział dotyczyć będzie tylko badania związków między zmiennymi w łańcuchu ontogenetycznym.

Interpretacja w analizie ścieżek i sekwencyjnej analizie plonu opiera się na różnych efektach. Wpływ bezpośredni opisuje zmianę zmiennej skutkowej w reakcji na jednostkową zmianę *j*-tej zmiennej przyczynowej przy założeniu, że wartość pozostałych zmiennych nie ulega zmianie. Jest to podejście nieco sztuczne, gdyż w przyrodzie takie zjawisko (czyli wpływ o takim charakterze) występuje wyłącznie w przypadku, gdy *j*-ta zmienna przyczynowa jest niezależna od wszystkich pozostałych zmiennych przyczynowych w łańcuchu. W sytuacji przeciwnej, zdecydowanie częstszej w biologii, możemy uznać, iż efekt bezpośredni opisuje pewien sztuczny „twór" interpretacyjny. Efekt bezpośredni dla danej zmiennej nie bierze pod uwagę efektu zmiennych ją poprzedzających, co więcej, efektu tego nie ujmuje się z efektu całkowitego dla tej zmiennej. Jest to o tyle „niesprawiedliwe", że pewna część *j*-tej zmiennej jest generowana przez zmienne ją poprzedzające, w związku z czym część jej efektu na zmienną skutkową powinna być zaliczona na poczet tych właśnie zmiennych. W analizie ścieżek jest to robione poprzez efekty pośrednie tych zmiennych, ale nie jest to ujmowane z efektu *j*-tej zmiennej. W tym właśnie aspekcie leży podstawowa różnica między analizą ścieżek a sekwencyjną analizą plonu; może to sprawić, że ich wyniki będą się znacznie różnić.

Przypomnijmy, że w analizie ścieżek pojedynczych popularną metodą interpretacji jest podział współczynnika korelacji między zmienną przyczynową i zmienną skutkową na sumę efektu bezpośredniego tej pierwszej oraz jej efektów pośrednich. Można w takim przypadku dokonać podziału współczynnika determinacji zmiennej skutkowej przez zmienne przyczynowe na udziały związane z efektami bezpośrednimi i pośrednimi poszczególnych zmiennych (Kozak, 2007 a). Niestety w przypadku ścieżek złożonych tego typu zależności nie występują, co oznacza, że w tych efektach panuje pewien „bałagan": efekt całkowity danej zmiennej na zmienną skutkową zwykle nie jest równy współczynnikowi korelacji między tymi zmiennymi. Porównajmy diagramy z rysunków 1 i 2: w przypadku tego drugiego (czyli dla modelu ścieżek pojedynczych) właściwości te, czyli podział współczynnika korelacji i determinacji, występują, natomiast w przypadku tego drugiego (czyli dla modelu ścieżek złożonych) — już nie, gdyż na przykład efekt całkowity X_1 na X_4 nie jest równy współczynnikowi korelacji między tym zmiennymi — w takim przypadku mówi się o tzw. fałszywym efekcie (ang. spurious effect). Efekt ten jest fałszywy, bowiem nie przedstawia prawdziwego (rzeczywistego) efektu dla danego systemu, ale jest reprezentowany w danych, a właściwie w wynikach analizy. Nie jest

ponadto możliwy podział współczynnika determinacji zmiennej skutkowej, jak to się robi dla modelu ścieżek pojedynczych, bez udziału fałszywych efektów. W praktyce fałszywe efekty po prostu się pomija i prowadzi się interpretację modelu na podstawie efektów bezpośrednich, pośrednich i całkowitych.

Sekwencyjna analiza plonu ma kilka zalet w porównaniu do analizy ścieżek. Efekty niezależne oferują ciekawą informację o wpływie zmiennych przyczynowych na zmienną skutkową. Ponadto brak fałszywych efektów sprawia, że możemy dokonać podziału współczynnika determinacji na niezależne udziały zmiennych przyczynowych. Za jej wadę niektórzy uznaliby to, że w modelowaniu należy uwzględniać wszystkie zależności między zmiennymi (zarówno przyczynowymi, jak i zmienną skutkową), nawet te nieistotne — w przeciwnym wypadku zmienne nie będą ściśle ortogonalne, co sprawiłoby utratę możliwości podziału współczynnika determinacji. Z aspektem tym wiąże się brak możliwości doboru modelu, który można przeprowadzić w analizie ścieżek. Z drugiej jednak strony, w przypadku badania łańcucha ontogenetycznego w procesach agronomicznych badanie istotności w sensie statystycznym może zostać uznane za sprawę drugorzędną, a wnioskowanie o ważności zmiennych przyczynowych opiera się na ich udziale w determinacji zmiennej skutkowej. Zmienna o udziale 30% może w przypadku niewielkiej próby zostać uznana za statystycznie nieistotną w danym modelu, zaś w przypadku dużej próby zmienna o udziale 10% może okazać się statystycznie istotna. Tego typu problemy niejednokrotnie były podkreślane (np. Webster, 2001; Reese, 2004): choćby i istotny w sensie statystycznym, współczynnik (w sensie ogólnym, w naszym przypadku mówimy o wpływie zmiennej przyczynowej na zmienną skutkową) jest ważny wtedy, gdy jest znaczący (istotny) w sensie biologicznym. Sama istotność statystyczna nie jest wystarczającym elementem do uznania danej zmiennej przyczynowej za ważną w procesie kształtowania zmiennej skutkowej.

Z tego względu raczej nie należy uznać braku możliwości doboru modelu w sekwencyjnej analizie plonu za wadę. Co więcej, czy rzeczywiście jesteśmy zainteresowani doborem modelu w tego typu systemie przyczynowo-skutkowym? Naszym celem nie jest ani znalezienie modelu optymalnego, ani redukcja liczby zmiennych, lecz opisanie procesu w kategoriach wpływu zmiennych przyczynowych na zmienną skutkową; do tego dobór modelu nie jest potrzebny. Pamiętajmy, że cały czas mówimy o zmiennych w łańcuchu ontogenetycznym. Gdy badamy procesy bardziej złożone, w których bierze udział kilkanaście czy kilkadziesiąt zmiennych o różnych możliwych związkach między nimi, dobór modelu może okazać się niezbędny. Przykłady takich problemów przedstawione są w pracach Kozaka i in. (2007 c) i Kozaka i in. (2008).

Warto zwrócić uwagę na jeszcze jeden element. Dla łańcucha ontogenetycznego nie ma możliwości zastosowania nowego podejścia do analizy ścieżek, bowiem rozpatrywany model nie ma wystarczającej liczby stopni swobody. Dlatego albo należy zastosować klasyczną estymację w analizie ścieżek (godząc się na obecność efektów fałszywych), albo też przeprowadzić dobór modelu przy pomocy nowej analizy ścieżek, np. startując od modelu z wybranymi trzema zmiennymi i dwoma ścieżkami pojedynczymi między nimi. To ostatnie podejście nie byłoby jednak prawidłowe w przypadku, gdy wszystkie efekty

bezpośrednie rozpatrywane w modelu (np. na rys. 3 wszystkie ścieżki pojedyncze) są ważne w sensie biologicznym i istotne w sensie statystycznym.

Przede wszystkim jednak należy podjąć dyskusję nad znaczeniem i interpretacją, a przede wszystkim prawidłowością (od strony statystycznej i biologicznej) efektów w obu metodach, tj. efektu bezpośredniego, pośredniego i całkowitego w analizie ścieżek oraz niezależnego w sekwencyjnej analizie plonu. Ten aspekt ma prawdopodobnie największe znaczenie w porównaniu tych dwóch metod.

3.3. TDP — dwukierunkowy podział zmienności plonu

W oryginalnym wydaniu (Eaton i inni 1976) dwukierunkowy podział zmienności plonu (ang. Two-Dimensional Partitioning of Yield Variation —TDP) polegał na połączeniu sekwencyjnej analizy składowych plonu (SYCA) z analizą wariancji, dzięki czemu plon analizowany był pod kątem determinacji przez dwa kierunki zmienności: jeden pochodzący z wpływu składowych plonu, zaś drugi — z wpływu czynników doświadczenia. Gołaszewski (1996) uznał, że metodę tę można również stosować dla cech plonotwórczych z łańcucha ontogenetycznego niebędących składowymi plonu, w którym to przypadku pojawia się dodatkowa zmienność resztowa, również uwzględniana w analizie. Twórcami metody byli Eaton i inni (1976), zaś wnikliwy jej opis można znaleźć w pracy Gołaszewskiego (1996), Gołaszewskiego i in. (1998) lub Kozaka (2006); metoda TDP znalazła pewne uznanie w kręgach agronomii, hodowli i fizjologii roślin (Akwilin Tarimo, 1997) i została zastosowana w co najmniej kilkudziesięciu analizach w uprawie i hodowli roślin polowych i ogrodniczych (obszerny wybór cytowań takich prac znajduje się w pracy Kozaka (2006)).

Sama idea metody TDP jest bardzo ciekawa, gdyż łączy dwie inne metody badania związków przyczynowo-skutkowych: analizę wariancji i analizę związków przyczynowo-skutkowych w łańcuchu ontogenetycznym. Metoda tego typu dostarczyłaby ciekawych wniosków dotyczących procesu kształtowania plonu, gdyż determinacja plonu przez czynniki doświadczenia oraz przez zmienne w łańcuchu ontogenetycznym jest zazwyczaj badana niezależnie.

Pomimo powyższego, metoda TDP nie jest niestety pozbawiona wad, które przedstawił Kozak (2006). Sprawiają one, że interpretacja analizy i wnioski wyciągnięte na tej podstawie mogą być, i zazwyczaj są, zafałszowane. Dwoma podstawowymi problemami jest wykorzystanie sum kwadratów z analizy wariancji jako wskaźników udziału poszczególnych czynników i zmiennych w zmienności plonu oraz obecność tzw. „cross-products", które pojawiają się w tabeli TDP, a które mają, lub mogą mieć, znaczący wpływ na interpretację. Ostateczny wniosek z pracy Kozaka (2006) jest taki, że problem postawiony przy konstrukcji metody jest ciekawy, ale jej podstawy statystyczne są w pewnych aspektach błędne. Warto zwrócić uwagę na to, że metoda ta zakłada, że zależności między plonem (lub inną zmienną skutkową) a zmiennymi przyczynowymi w łańcuchu ontogenetycznym są takie same; zwykle jest to założenie bardzo nierealistyczne.

Metoda o silnych podstawach statystycznych, która pozwalałaby na interpretację i wnioskowanie postulowane dla metody dwukierunkowego podziału zmienności plonu,

mogłaby znaleźć zastosowanie w analizie wielu doświadczeń agronomicznych i hodowlanych.

4. WNIOSKOWANIE O ZWIĄZKACH PRZYCZYNOWO-SKUTKOWYCH DLA POPULACJI GENOTYPÓW

Dysponując wynikami z doświadczenia przeprowadzonego dla wystarczająco dużej populacji genotypów, przy pomocy omawianych metod można wnioskować o procesach przyczynowo-skutkowych zachodzących w obrębie tej populacji. Wnioskowanie to można wzbogacać przy pomocy innych metod. Przykładowo, Kozak i inni (2007c) połączyli wyniki analizy ścieżek z wynikami przedstawionymi przez Singha i in. (2006), którzy przy pomocy analizy skupień pogrupowali genotypy wykorzystane do badania związków przyczynowo-skutkowych między plonem ziarna i jakością mielenia oraz cechami roślin badanymi przez Kozaka i in. (2007 c); grupowanie to, jak i analiza ścieżek, zostało przeprowadzone na podstawie średnich wartości cech dla genotypów. W ten sposób Kozak i in. (2007 c) wskazali te grupy genotypów, które charakteryzowały się średnimi wartościami badanych cech zbliżonymi do optymalnych, ale biorąc pod uwagę tylko te zmienne, które okazały się ważne (istotne) w procesie kształtowania ostatecznego plonu ryżu z roślin oraz jakości mielenia. Można więc uznać, że Kozak i in. (2007 c) wzbogacili wnioskowanie o badanym procesie poprzez analizę podobieństw między genotypami, a wspólne wnioski z tych dwóch analiz wykorzystali do wskazania genotypów, dla których osiągnięto pożądane średnie wartości cech istotnych w procesie tworzenia plonu ziarna i jakości mielenia.

Kozak i in. (2008) wzbogacili powyższe podejście. Zauważyli, że choć ostateczne wnioski wyciągane są na podstawie obu metod, tj. analizy ścieżek i analizy skupień, to wyniki tych metod nie są łączone na etapie analizy. Stąd też na przykład w analizie skupień grupuje się genotypy na podstawie wszystkich zmiennych, choć analiza ścieżek może wykazać, że niektóre z nich w ogóle nie są istotne w procesie kształtowania zmiennej skutkowej (zmiennych skutkowych). Takie podejście może sprawić, że uzyskane skupienia genotypów będą wynikać z podobieństwa genotypów i różnic między nimi w odniesieniu do tych zmiennych, które dla zmiennych skutkowych nie mają znaczenia. Aby więc pogrupować genotypy pod kątem badanego procesu, a nie wszystkich zmiennych, które były obserwowane, Kozak i in. (2008) zaproponowali odpowiednie ważenie zmiennych w analizie skupień. Ważenie to oparte jest na wynikach analizy ścieżek, które informują o ważności poszczególnych cech w procesie kształtowania zmiennej przyczynowej. Podejście zaproponowane przez Kozaka i in. (2008) może zostać zastosowane dla systemów przyczynowo-skutkowych z jedną zmienną przyczynową. Jak sugerują autorzy, następnym krokiem jest opracowanie systemu ważenia dla modeli o większej liczbie zmiennych skutkowych.

5. PODSUMOWANIE

W niniejszym opracowaniu dyskutowaliśmy o analizie liniowych związków między zmiennymi losowymi w agronomicznych systemach przyczynowo-skutkowych. Jest to tematyka obszerna i bogata w modele, metody, podejścia i problemy, spośród których zajmowaliśmy się tylko wybranymi: analizą ścieżek, sekwencyjną analizą plonu, dwukierunkowym podziałem zmienności plonu oraz wnioskowaniem o związkach przyczynowo-skutkowych w populacji genotypów.

W przypadku analizy ścieżek wskazaliśmy potrzebę stosowania podejścia opartego na metodzie największej wiarygodności, stanowiącego alternatywę do klasycznej analizy ścieżek i dostarczającego nowe możliwości estymacji i interpretacji. Dyskutowaliśmy również o problemach związanych z tą metodą: obecności fałszywych efektów oraz pomijaniu ich w interpretacji, a także ograniczonych możliwościach wnioskowania opartego na efektach bezpośrednich, pośrednich i całkowitych, co jest związane ze „sztucznym" podejściem do związków przyczynowo-skutkowych, jakie efekty te reprezentują.

O sekwencyjnej analizie plonu powiedzieliśmy, że jej główną zaletą jest wykorzystanie efektów niezależnych, które uznaliśmy za bardziej naturalne niż te, na których opiera się analiza ścieżek. Co więcej, w sekwencyjnej analizie plonu dokonuje się podziału współczynnika determinacji zmiennej skutkowej na niezależne udziały poszczególnych zmiennych przyczynowych w łańcuchu ontogenetycznym. Za wadę tej metody niektórzy uznaliby brak możliwości doboru optymalnego modelu, co sprawia, że wszystkie efekty, nawet te praktycznie nieistotne (zarówno z punktu widzenia biologii procesu, jak i statystyki), muszą być w analizie uwzględnione. Jak jednak uznaliśmy, wcale nie musi to zostać uznane za wadę w przypadku, gdy omawianym systemem przyczynowo-skutkowym jest łańcuch ontogenetyczny.

Warto zwrócić uwagę na perspektywy dalszej pracy nad omawianymi metodami i podejściami. Po pierwsze, interesujące może być uogólnienie sekwencyjnej analizy plonu na systemy przyczynowo-skutkowe o innej formie niż łańcuch ontogenetyczny. Zastosowanie analizy korelacji kanonicznych (o czym mówiliśmy wcześniej) jest prawdopodobnie jedną z możliwości, nie bez wad, przyszłe badania powinny mieć na celu znalezienie innych podejść, być może bardziej efektywnych, zwłaszcza w interpretacji. Kolejne pytanie to jaki system ważenia zmiennych należy zastosować w analizie skupień do łączenia jej wyników z wynikami analizy ścieżek, gdy w systemie związków przyczynowo-skutkowych jest więcej niż jedna zmienna skutkowa. Ponadto warto podjąć dyskusję o fałszywych efektach w analizie ścieżek. Są one standardowo pomijane w zastosowaniach i właściwie nie do końca wiadomo, jaki może być ich wpływ na interpretację, czyli w jakim stopniu mogą one fałszować wnioskowanie. Innym ciekawym kierunkiem badań nad analizą związków przyczynowo-skutkowych dla zmiennych w łańcuchu ontogenetycznym obserwowanych w doświadczeniu czynnikowym jest stworzenie metody, która realizowałaby cele stawiane metodzie dwukierunkowego podziałowi zmienności plonu TDP, będąc przy tym pozbawioną jej wad.

Ważnym pytaniem, wciąż pozostającym bez odpowiedzi, jest: które podejście do analizy związków przyczynowo-skutkowych jest bardziej prawidłowe, to oparte na analizie efektów bezpośrednich (oraz pośrednich i całkowitych, które wykorzystują tę samą filozofię i metodologię), czy to oparte na efektach niezależnych (czyli estymowanych w sekwencyjnej analizie plonu)? Są to podejścia alternatywne, a z dyskusji przedstawionej w niniejszym opracowaniu wynika, że ostateczne wnioskowanie nie powinno opierać się na łączonej interpretacji z obu tych podejść, w wielu bowiem aspektach filozoficznych nie zgadzają się one ze sobą, co wyklucza możliwość ich wspólnego stosowania. Nie pokusiliśmy się o jednoznaczną odpowiedź na pytanie, które podejście jest bardziej prawidłowe. Wymaga to dalszych wnikliwych badań, a te przedstawione tutaj są niejako ich początkiem. Pojęcie wpływu niezależnego wymaga dodatkowej uwagi, gdyż do tej pory znajdowało się w cieniu klasycznego podejścia do badania związków przyczynowo-skutkowych, czyli tego poprzez efekty bezpośrednie, pośrednie i całkowite. Idea efektu niezależnego jest ciekawa i są podstawy, by podjąć dalsze nad nim badania

LITERATURA

Dewey D. R., Lu K. H. 1959. A correlation and path-coefficient analysis of components of crested wheatgrass seed production. Agronomy Journal 51: 515 — 518.

Dhungana P., Eskridge K.M., Baenziger P.S., Campbellc B.T., Gilld K.S., Dweikat I. 2007. Analysis of genotype-by-environment interaction in wheat using a structural equation model and chromosome substitution lines. Crop Science 47: 477 — 484.

Dofing S. M., Knight C. W. 1992. Alternative model for path analysis of small-grain yield. Crop Science 32: 487 — 489.

Dowe P. 1992. Wesley Salmon's process theory of causality and the conserved quantity theory. Philosophy of Science 59 2: 195 — 216.

Eaton, G. W. Kyte T. R. 1978. Yield component analysis in strawberry. Journal of the American Society for Horticultural Science 103: 578 — 583.

Eaton G. W., MacPherson E. A. 1978. Morphological components of yield in cranberry. Horticultural Research 17: 73 — 82.

García del Moral L. F., Rharrabti Y., Villegas D., Royo C. 2003. Evaluation of grain yield and its components in durum wheat under Mediterranean conditions: An ontogenetic approach. Agronomy Journal 95: 266 — 274.

Gołaszewski J. 1996. A method of yield component analysis. Biometrical Letters 332: 79 — 88.

Grafius J. E. 1969. Stress: a necessary ingredient of genotype by environment interactions. International Barley Genet. II. Wash. State Univ. Press.: 346 — 355.

Guillen-Portal F. R., Stougaard R. N., Xue Q., Eskridge K. M. 2006. Compensatory mechanisms associated with the effect of spring wheat seed size on wild oat competition. Crop Science 46: 935 — 945.

Jolliffe P. A., Courtney W. H. 1984. Plant growth analysis: additive and multiplicative components of growth. Annals of Botany 54: 243 — 254.

Jończyk B. 2002. Statystyczna ocena uwarunkowania plonu pszenżyta ozimego przez jego składowe za pomocą analizy ścieżek pojedynczych i złożonych. Praca magisterska. Wydział Rolniczy SGGW, Warszawa.

Kozak M. 2002. Statystyczna analiza uwarunkowania plonu roślin przez jego składowe. Praca doktorska. Wydział Rolniczy. Szkoła Główna Gospodarstwa Wiejskiego, Warszawa.

Kozak M. 2006. Two-dimensional partitioning of yield variation: A critical note. Plant Breeding and Seed Science 53: 37 — 42.

Kozak M. 2007. Ontogenetic chain and the Markov condition in crop science. Nature and Science 53: 5 — 8.

Kozak M., Bocianowski J., Rybiński W. 2008. Selection of promising genotypes based on path and cluster analyses. Journal of Agricultural Science 146: 85 — 92.

Kozak M., Gozdowski D., Hossain S., Ahmed S. E., Laudański Z., Wyszyński Z. 2006 a. Canonical correlations in studying grain yield and protein content as affected by yield components: An ontogenetic approach. Plant Breeding and Seed Science 54: 17 — 27.

Kozak M., Gozdowski D., Wyszyński Z. 2006 b. An approach to analyzing a response variable as affected by its additive components: Example for spring barley grain yield. Cereal Research Communications 342-3: 981 — 988.

Kozak M., Kang M. S. 2006. Note on modern path analysis in application to crop science. Communications in Biometry Crop Science 11: 32 — 34.

Kozak M., Mądry W. 2006. Note on yield component analysis. Cereal Research Communications 34 2 — 3: 933 — 940.

Kozak M., Mądry W., Wyszyński Z. 2002. Metoda analizy wpływu masy korzeni z różnych frakcji na plon korzeni buraka cukrowego. Fragmenta Agronomica 2 74: 251 — 262.

Kozak M., Kang M.S., Stępień M. 2007a. Causal Pathways when Independent Variables are Co-Related: New Interpretational Possibilities. Plant, Soil and Environment 536: 267 — 275.

Kozak M., Samborski S., Rozbicki J., Mądry W. 2007b. Winter triticale grain yield, a comparative study of 15 genotypes. Acta Agriculturae Scandinavica. Section B – Plant & Soil Science 57: 263 — 270.

Kozak M., Singh P.K., Verma M.R., Hore D.K. 2007c. Causal mechanism for determination of grain yield and milling quality of lowland rice. Field Crops Research 102: 178 — 184.

Lorencetti C., de Carvalho F.I.F., de Oliveira A. C., Valério I.P., Hartwig I., Benin G., Schmidt D.A.M. 2006 Applicability of phenotypic and canonic correlations and path coefficients in the selection of oat genotypes. Scientia Agricola 63: 11 — 19.

Mądry W., Kozak M. 2000. Analiza ścieżek i sekwencyjna analiza plonu w badaniach zależności plonu od cech łanu. Cz. I. Opis metod. Roczniki Nauk Rolniczych. Seria A 115: 143 — 157.

Mądry W., Kozak M., Pluta S., Żurawicz E. 2005. A new approach to sequential yield component analysis SYCA: Application to fruit yield in blackcurrant *Ribes nigrum* L.. Journal of New Seeds 7 1: 85 — 107.

Mądry W., Laudański Z., Kozak M., Rozbicki J. 2003. Empiryczne porównanie sekwencyjnej analizy składowych plonu i analizy ścieżek pojedynczych dla plonu ziarna pszenżyta ozimego i jego składowych. Biuletyn IHAR 230: 147 — 156.

Mohammadi S. A., Prasanna B. M., Singh N.N. 2003. Sequential path model for determining interrelationships among grain yield and related characters in maize. Crop Science 43: 1690 — 1697.

Piepho H. P. 1995. A simple procedure for yield component analysis. Euphytica 84: 43 — 48.

Rasmusson D. C., Cannell R. Q. 1970. Selection for grain yield and components of yield in barley. Crop Science 10: 51 — 54.

Reese R. A. 2004. Does significance matter? Significance 1: 39 — 40.

Rencher A. C. 1998. Multivariate statistical inference and applications. John Wiley and Sons, New York.

Reynolds M., Calderini D., Condon A., Vargas M. 2007. Association of source/sink traits with yield biomass and radiation use efficiency among random sister lines from three wheat crosses in a high-yield environment. The Journal of Agricultural Science 1451: 3 — 16.

Rozbicki J. 1997. Agrotechniczne uwarunkowania wzrostu, rozwoju i plonowania pszenżyta ozimego. Rozprawa habilitacyjna. Fundacja Rozwój SGGW, Warszawa.

Samborski S., Kozak M., Rozbicki J. 2006. The usefulness of chlorophyll meter SPAD-502 for winter triticale grain yield estimation. Folia Universitas Agriculturae Stetinensis. Agricultura 247 100: 157 — 162.

Samonte S.O.P.B., Wilson L.T., McClung A.M. 1998. Path analyses of yield and yield-related traits of fifteen diverse rice genotypes. Crop Science 38: 1130 — 1136.

Shipley B. 2002. Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference. Cambridge University Press, Cambridge.

Sinebo W. 2002. Yield relationships of barleys in a tropical highland environment. Crop Science 42: 428 — 437.

Singh P. K., Mishra M. N., Hore D. K., Verma M.R. 2006. Genetic divergence in lowland rice of north eastern region of India. Communications in Biometry and Crop Science 11: 35 — 40.

Steel D. 2005. Indeterminism and the causal Markov condition. The British Journal for the Philosophy of Science 56 1: 3 — 26.

Thomas R.L., Grafius J.E., Hahn S.K. 1971. Transformation of sequential quantitative characters. Heredity 26: 189 — 193.

Vargas M., Crossa J. Reynolds M.P., Dhungana P., Eskridge K.M. 2007. Structural equation modeling for studying genotype × environment interactions of physiological traits affecting yield in wheat. The Journal of Agricultural Science 1452: 151 — 161.

Webster R. 2001. Statistics to support soil research and their presentation. European Journal of Soil Science 52: 331 — 340.

Wolfe L.M. 1999. Sewall Wright on the method of path coefficients: An annotated biography. Structural Equations Modeling 63: 280 — 291.

Wright S. 1921. Correlation and causation. Journal of Agricultural Research Wash., D.C 20: 557 — 585.

Wright S. 1934. The method of path coefficients. Annals of Mathematical Statistics 5: 161 — 215.