Andrzej Sokołowski
(Uniwersytet Ekonomiczny w Krakowie)
Marek Sobolewski
(Politechnika Rzeszowska)

Jak nie należy analizować danych regionalnych, czyli o błędnym stosowaniu współczynnika zmienności

Dwie zasady doboru zmiennych – podawane w niemal każdym artykule zawierającym analizę taksonomiczną

1. **Do analizy dokonujemy wyboru cech ważnych z merytorycznego punktu widzenia.**
2. **Przed wykonaniem właściwej analizy należy/warto dokonać zmniejszenia liczby cech na podstawie kryteriów statystycznych (a więc, bez wnikania w merytorykę!).**

Zasada z punktu 2. przeczy zasadzie z punktu 1.

Usuwanie zmiennych o niskim poziomie zmienności

Jedno z kryteriów doboru zmiennych – najczęściej stosowane w pierwszej kolejności – polega na usuwaniu cech o zbyt niskim poziomie zmienności – najczęściej jest to oceniane na podstawie wartości klasycznego współczynnika zmienności (V_s).

Zwykle przyjmuje się arbitralnie, iż odrzuca się zmienne, jeśli:

$$V_s < 20\% \text{ (ew. } V_s < 10\%).$$

Nie wiadomo, dlaczego przyjęto 20 lub 10%, a nie na przykład 5, 30 czy 50%.

Współczynnik zmienności – uwagi techniczne

W populacji	$\gamma_V = \sigma/\mu$
Zastosowanie do danych w skali	**ilorazowej**
Estymator obciążony	$C_V = S/\bar{x}$
Estymator nieobciążony	$\hat{C}_V = (1 + \frac{1}{4n})C_V$
Zakres zmienności	**[0;** $\sqrt{n-1}$ **] - przy obciążonym estymatorze odchylenia standardowego [0;** $\sqrt{n}$ **] - przy nieobciążonym estymatorze odchylenia standardowego**
Błąd standardowy	$\sigma_{C_V} = \frac{\gamma_V}{\sqrt{2n}}$ **(rozkład normalny)**

NISKI współczynnik zmienności… i co z tego nie powinno wynikać

Przyjęcie kryterium współczynnika zmienności, może prowadzić do zupełnie absurdalnych decyzji, jeśli chodzi o dobór zmiennych diagnostycznych do właściwej analizy.

Oto wartości statystyk opisowych *oczekiwanego czasu trwania życia mężczyzn* w latach 1960-2010 w krajach europejskich.

Rok	Oczekiwany czas trwania życia mężczyzn w państwach europejskich [lata]					
	Średnia	Mediana	s	min	max	V_s
1960	65,7	66,0	3,4	58,0	71,5	5,1%
1970	67,2	66,7	2,4	61,5	72,2	3,5%
1980	68,3	68,5	3,1	61,4	73,5	4,5%
1990	69,8	70,6	3,5	63,8	75,5	5,1%
2000	71,4	71,9	4,8	59,0	77,8	6,7%
2010	74,3	76,1	4,8	63,1	80,1	6,5%

NISKI współczynnik zmienności… i co z tego nie powinno wynikać

Jak widać, oczekiwany czas trwania życia mężczyzn nie spełnia kryterium współczynnika zmienności i nie powinien być uwzględniany jako miernik jakości życia w krajach europejskich.

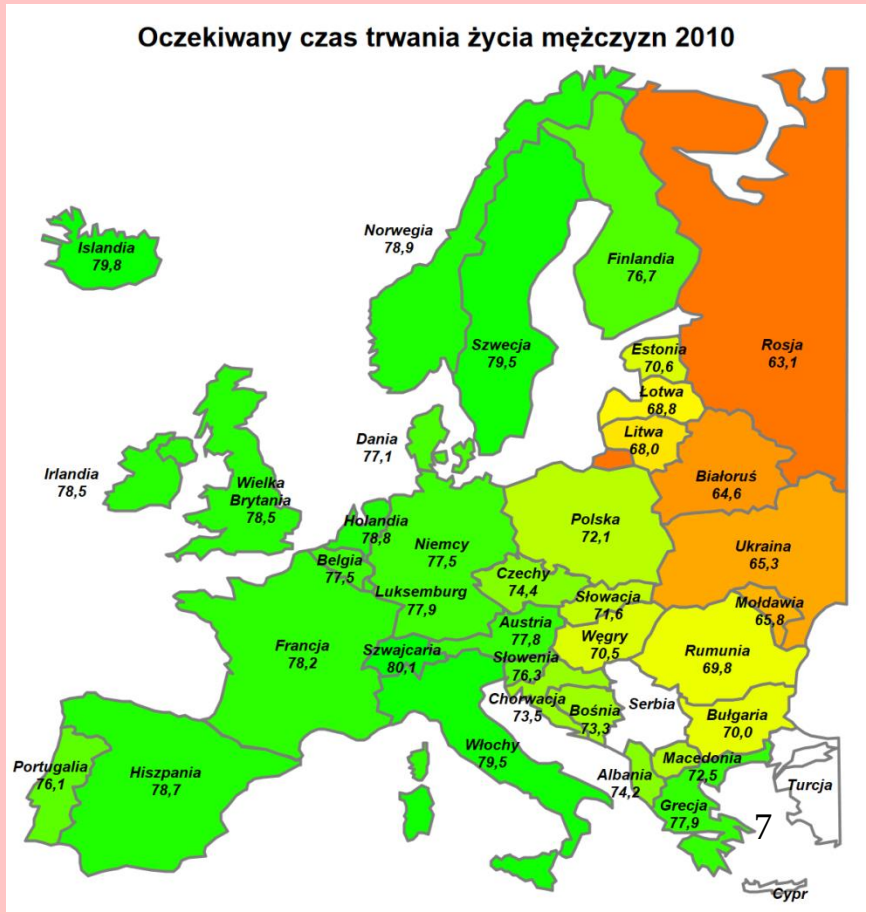
A tymczasem:

- **w roku 2010 LE dla mężczyzn w Rosji wynosił 63 lata, a w Szwajcarii 80 lat.**
- **LE dla mężczyzn w Rosji w 2010 roku był niższy niż w Szwajcarii w roku 1960 – 69 lat.**

Niestety, w wielu analizach danych regionalnych czas trwania życia „pada ofiarą” bezkrytycznie stosowanego kryterium wysokiej zmienności.

NISKI współczynnik zmienności… i co z tego nie powinno wynikać

Wizualizacja danych z 1960 i 2010 roku nie pozostawia wątpliwości, że zróżnicowanie LE mężczyzn w 2010 roku, w kontekście zmian tej wielkości w czasie jest znaczne.



NISKI współczynnik zmienności… i co z tego nie powinno wynikać

A skoro tak, to może lepiej nie stosować kryterium, które daje tak bezsensowne wyniki lub przeformułować je tak, by nie dotyczyło jednego, wyrwanego z kontekstu okresu obserwacji.

Może kryterium przydatności danej cechy, powinna być na przykład miara zdefiniowana następująco: *„Ile lat wstecz wartość dla najgorszego obiektu z danego roku byłaby najlepsza”*.

Oczywiście kwestią dyskusyjną jest potrzeba przyjmowania wartości progowej dla takiej miary – i pytania, czy ma być to 10 lat (bo okrągła liczba), czy może 25 lat (jedno pokolenie).

WYSOKI współczynnik zmienności… i co z tego nie musi wynikać

Powodem wysokiej zmienności cechy statystycznej może być jej losowość, niestabilność w czasie.

Wtedy wysoki poziom zmienności powinien stanowić raczej przesłankę wykluczenia (!!!) cechy statystycznej z analizy, a nie jej pozostawienia.

Przykład dotyczy bezpieczeństwa ruchu drogowego w miastach na prawach powiatu, a rozważaną cechą jest wskaźnik ofiar wypadków drogowych (liczba ofiar w przeliczeniu na 100 tys. mieszk.).

WYSOKI współczynnik zmienności… i co z tego nie musi wynikać

Wartości współczynnika zmienności wielokrotnie przekraczają próg minimalnej zmienności (niezależnie od tego, czy wynosiłby on 10 czy 20%).

Rok	Liczba ofiar wypadków na 100 tys. mieszk.					
	Średnia	Mediana	*s*	min	max	V_s
2011	4,6	4,2	2,4	0,0	11,1	50,8%
2012	3,6	3,4	2,4	0,0	9,0	65,6%
2013	3,8	3,6	2,0	0,0	8,5	53,8%
2014	3,3	3,0	1,9	0,0	7,8	57,7%
2015	3,1	2,7	2,2	0,0	10,6	70,2%

Ale, jak to zostanie pokazane za chwilę, w tym przypadku powinien być to raczej argument za wykluczeniem jej z analizy.

WYSOKI współczynnik zmienności… i co z tego nie musi wynikać

Jeżeli cecha ma mieć jakąś wartość merytoryczną, to powinna charakteryzować się pewną stabilnością – inaczej analiza tych samych danych nawet z dwóch sąsiednich okresów da zupełnie inne wyniki.

Tymczasem, wartości rozważanego wskaźnika BRD z poszczególnych lat są ze sobą niemal zupełnie nieskorelowane.

Dla przykładu, korelacja pomiędzy wartościami wskaźnika z 2014 i 2015 roku to -0,06.

Współczynnik Spearmana	2011	2012	2013	2014	2015
2011		0,13	0,19	0,07	0,07
2012	0,13		0,23	0,34*	0,04
2013	0,19	0,23		0,13	0,05
2014	0,07	0,34*	0,13		-0,06
2015	0,07	0,04	0,05	-0,06	

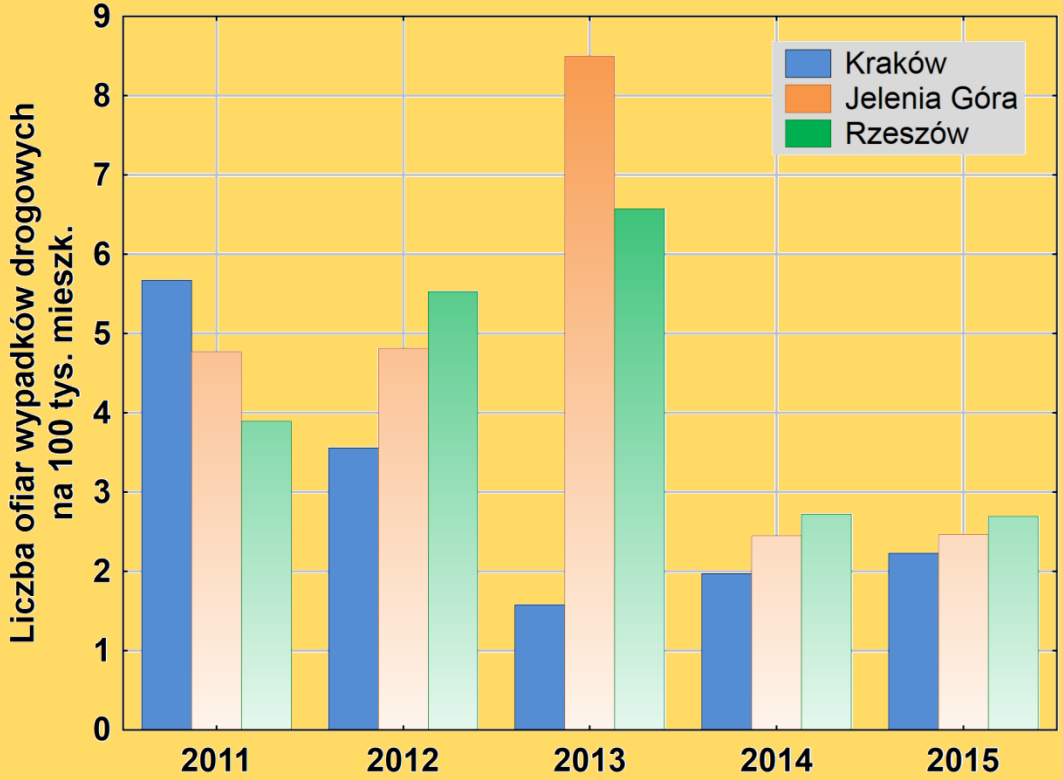
WYSOKI współczynnik zmienności… i co z tego nie musi wynikać

A oto wyjaśnienie otrzymanych wyników:

1. **Na terenie miast na prawach powiatu liczba wypadków ze skutkiem śmiertelnych jest na tyle niewielka, że wartości wskaźnika charakteryzują się dużą losowością.**
2. **Współczynnik zmienności jest tak wysoki głównie z powodu dużej losowości w przekroju powiatów.**
3. **Tak więc, wysoka zmienność może oznaczać, że cecha statystyczna jest niestabilna, a wtedy jej wartość diagnostyczna jest żadna – czyli słuszne byłoby odwrócone kryterium, nie wysokiej a niskiej, zmienności (!!!).**
4. **Kryterium zmienności nie powinno być stosowane zwłaszcza dla relatywnie małych jednostek terytorialnych, gdzie zmienność cech statystycznych ma genezę w losowości.**

WYSOKI współczynnik zmienności… i co z tego nie musi wynikać

I jeszcze ilustracja wartości tej tak „przydatnej” w świetle kryterium wysokiej zmienności cechy. Dla wybranych trzech miast widać niestabilność wartości w czasie i efemeryczność tworzonych w ten sposób pod-rankingów.



Przypuszczenie

Pierwotnym powodem redukowania liczby zmiennych w badaniach taksonomicznych były zapewne obawy przed zbyt dużą złożonością obliczeniową.

Trudno stosować te same zasady w XXI wieku, kiedy dowolny zakres obliczeń nie powinien stanowić żadnego problemu.

Propozycje

We wszystkich poniższych propozycjach założono, że ocena przydatności cechy statystycznej w badaniach regionalnych wymaga odniesienia rozkładu wartości w poszczególnych latach do danych z dłuższego okresu.

1. **Porównanie odchylenia standardowego z danego roku (czyli zmienności pomiędzy obiektami) do średniej z odchyleń standardowych danych rocznych dla każdego obiektu.**
2. **Porównanie suma kwadratów analizy wariancji z powtarzanymi pomiarami (SSS) do (SST-SSE).**
3. **Dla danych wykazujących trend wzrostowy lub spadkowy – określnie liczby lat, dzielących dany okres, od okresu, w którym najlepszy obiekt miał wartość na poziomie obecnie najgorszego.**

Wyniki dla jednej z propozycji

A oto wyniki analizy zmienności z uwzględnieniem szerszego kontekstu dla obu rozważanych wcześniej przykładów.

Wyniki są oparte na analizie wariancji z powtarzanymi pomiarami, więc ostatni rok z serii pomiarów nie jest tu specjalnie wyróżniony – dlatego też dla V_S podano zakres wartości z poszczególnych lat.

„Nasze" kryterium daje zupełnie inny sygnał niż klasyczny współczynnik zmienności – i w świetle przedstawiono wcześniej argumentów, jest to lepsza metoda.

Miara zmienności	LE mężczyzn	Wskaźnik ofiar wypadków
V_S	3,5-6,7%	50,8-70,2%
Zmienność pomiędzy przypadkami (SSS)	2437	485
Zmienność całkowita (SST)	5010	1621
SSS/(SST-SSS)	0,95	0,43

Wnioski końcowe

1. **W analizie danych regionalnych – zwłaszcza porządkowaniu i grupowaniu – nie należy wstępnie eliminować zmiennych o małym współczynniku zmienności.**
2. **Kryterium wysokiej zmienności należy odrzucić.**
3. **Alternatywnie może korzystać z innych miar zmienności, uwzględniając szerszy zakres danych w aspekcie czasowym.**

P.S. Wnioski dodatkowe

Słabością wielu analiz regionalnych jest brak określenia celu analizy i perspektywy z jakiej są wykonywane:

- **po co grupujemy powiaty?**
- **po co porządkujemy regiony?**
- **kto ma być odbiorcą tych analiz?**

P.S. Usuwanie zmiennych skorelowanych (też dyskusyjne...)

Drugim „obowiązkowym” etapem analiz taksonomicznych danych regionalnych jest usuwanie cech skorelowanych, tak aby „nie powielać tych samych informacji”.

Jeżeli jednak z merytorycznego punktu widzenia cechy te były ważne, dlaczego usuwać je z analizy?

P.S. Usuwanie zmiennych skorelowanych (zapowiedź kolejnej „rewolucji”)

Jeżeli mamy przyjętą określoną perspektywę badania – na przykład atrakcyjności danego miasta na prawach powiatu jako miejsca zamieszkania dokonujemy wyboru pewnych kluczowych cech diagnostycznych. Najlepiej na podstawie ankiety.

Przyjęto orientację w 70% kulturalną (7 wskaźników określających dostępność kin, teatrów, muzeów, koncertów, etc.) oraz w 30% ekonomiczną (3 wskaźniki – wynagrodzenia, bezrobocie, cena mieszkania).

Jednak wskaźniki oferty kulturalnej okazały się być skorelowane i w efekcie w analizie pozostał jeden z nich i 3 miary ekonomiczne. W efekcie, otrzymamy ranking, w którym „wygra” miasto o najlepszych parametrach ekonomicznych, mimo, że ranking miał być tworzony głównie pod kątem oferty kulturalnej.