Paweł Lańduch
Urząd Statystyczny w Poznaniu,
p.landuch@stat.gov.pl

Słowa kluczowe: kwestionariusz statystyczny, model dla błędu klasyfikacji, błąd pomiaru.

Wykorzystanie modeli statystycznych jako jednej z metod testowania i oceny kwestionariuszy statystycznych

Wstęp

Błąd pomiaru jest zdefiniowany jako różnica między prawdziwą wartością badanej cechy i wartością cechy otrzymanej przez pomiar. W klasyfikacji na skali nominalnej błąd klasyfikacji można przedstawić jako dychotomiczną różnicę osób które zostały błędnie przypisane do kategorii X, chociaż faktycznie nie należą do tej kategorii (false positive) i osoby, które zostały przypisane do innej kategorii, chociaż faktycznie należą do kategorii X (false negative). W badaniach z użyciem kwestionariusza oraz prowadzonymi poprzez ankietera przyczyny błędu mogą być powodowane zarówno poprzez konstrukcję kwestionariusza jak również sposób prowadzenia wywiadu przez ankietera, tzn. wywiad prowadzony metodą bezpośredniego kontaktu lub przeprowadzany telefonicznie. Trzecią przyczyną może być sam respondent, dla przykładu jego motywacja. Lohr (1999) zaznacza osiem przyczyn, które mogą prowadzić do błędu klasyfikacji bazując na trzech podstawach przeprowadzania wywiadu: ankieter, kwestionariusz, respondent. Te przyczyny to:

- respondenci czasami nie mówią prawdy,
- respondenci nie zawsze rozumieją pytania zawarte w kwestionariuszu,
- pamięć respondenta bywa ograniczona np. przez czas,
- różne odpowiedzi mogą być udzielone przez respondenta różnym ankieterom,
- respondenci mogą udzielić odpowiedzi, które uznają za oczekiwane przez ankietera,
- ankieter nie zawsze wiernie odczyta pytanie, w ten sposób wpływając na jakość odpowiedzi lub zarejestruje pomyłko inną odpowiedź niż podana,
- znaczenie słów może być różnie rozumiane przez różnych respondentów,
- sformułowanie pytań oraz kolejność ich zadawania mogą mieć wpływ na udzielane odpowiedzi.

Bradburn i Sudmann (2011) wskazują na szereg cech budowy kwestionariusza, które mogą mieć wpływ na błąd pomiaru. Podają przykłady jak duże znaczenie może mieć sformułowanie pytań (questions wording). Elementy takie jak długość pytań, użycie nieprecyzyjnych lub trudnych definicji prowadzą do trudności w rozumieniu pytań. Z jednej strony testowanie kwestionariuszy np. z użyciem metod kognitywnych, z drugiej większa rola ankietera, np.. przy zadawaniu dodatkowych pytań (probing) może być jednym z elementów prowadzących do poprawy jakości uzyskiwanych odpowiedzi. Innymi sposobami na lokalizację trudności lub ocenę pytań, które mogą być wpłynąć na błąd pomiaru są metody statystyczne. Podstawą do oceny tymi metodami są dane uzyskane przez powtórny pomiar tej samej cechy w badaniu poprzez np. powtórny pomiar dokonany na losowej podpróbie jednostek. Powtórny pomiar może być dokonany poprzez:

- powtórny wywiad dotyczący danej cechy lub cech poprzez ponowny kontakt z podpróbą jednostek uczestniczących w wywiadzie pierwotnym,
- pojedynczy pomiar, w którym badana cecha jest zawarta w kwestionariuszu statystycznym dwukrotnie poprzez różne sformułowania pytań,
- wielokrotny pomiar, gdzie jedna grupa danych uzyskana jest za pomocą kwestionariusza, a druga z innych źródeł, np. z danych administracyjnych,
- poprzez badanie panelowe odnoszące się do tych samych jednostek w różnych punktach czasu.

Powtórny pomiar i dane uzyskiwane powyższymi metodami ogólnie nazwać można metodami wielokrotnego pomiaru. Dane uzyskane w ten sposób następnie mogą być poddane obróbce metod statystycznych, jako jedna z metod w strategii badań nad jakością gromadzonych danych, której celem jest:

- analiza danych w celu identyfikacji pytań, które potencjalnie generują odpowiedzi o małej wiarygodności,
- dokonanie dalszej analizy, która prowadzi do identyfikacji źródeł błędów w pytaniach, które zostały uznane jako mające małą wiarygodność,
- weryfikacji źródeł błędów poprzez dodatkowo uzyskane dane z eksperymentów, testów pilotażowych, metod kognitywnych, itd.,
- wybór strategii, które mają wyeliminować lub zmniejszyć błąd pomiaru poprzez jej implementację.

Metody statystyczne wykorzystywane są przede wszystkim w realizacji pierwszych dwóch punktów.

Współczynnik niezgodności

Punktem wyjścia może być *współczynnik zgodności* κ Cohen'a (1960), który zaproponował ocenę jakości danych poprzez podwójny pomiar dla kategorii klasyfikującej badaną cechę na skali nominalnej. Założenia dla pomiarów są następujące:

- jednostki są niezależne,
- kategorie na skali są niezależne, wzajemnie wykluczające się i wyczerpujące,
- oceniający działają niezależnie.

Wówczas w tablicy (tabl. 1) kontyngencji mamy:

Tabl. 1 Przykładowa tablica kontyngencji dla dwóch pomiarów

WYSZCZE-GÓLNIENIE-	Pomiar A			
	kategoria	1	2	P_{iB}
Pomiar B	1	a (a')	b (b')	P_{1+}
	2	c (c')	d (d')	P_{2+}
	P_{iA}	P_{+1}	P_{+2}	$\sum P_i$

Wartości a, b, c, d są częstościami obserwowanymi.

Wartości a',b',c',d' są częstościami oczekiwanymi, zakładając przypadkową zbieżność liczoną na podstawie częstości brzegowych. Wtedy:

p_0= a+d, częstość, w której oceny są zgodne,

$p_c=a'+d'$, częstość, w której zgodność jest przypadkowa.

Współczynnik κ wyraża się wzorem:

$$\kappa = \frac{p_0 - p_c}{1 - p_c}.$$

Hansen i in. (1964) zaproponowali prosty model na interpretację niezgodności między wynikami pomiaru i powtórnego pomiaru bazując na tablicy kontyngencji wyników z obu pomiarów. Na tej podstawie został zbudowany *indeks niezgodności I,* który można przedstawić jako miara proporcji rozbieżności między odpowiedziami na pytanie, która jest spowodowana rozbieżnością pomiaru. Hess i in., (1999) pokazali zależność współczynnika κ oraz współczynnika 1-*I*. Dla przykładu mając tablicę zmiennej dychotomicznej, której pomiar został wykonany dwukrotnie można obliczyć dla niej wymienione współczynniki. Podwójny pomiar rozumiany jest jako taki, który spełnia założenia niezależności między kategoriami fałszywie pozytywnych i fałszywie negatywnych.

Indeks *I* ma postać:

$$I = \frac{g}{p_1 q_2 + p_2 q_2},$$

gdzie: $p_1=(a+c)/n$, czyli ilość osób która wybrała wariant 1 w pierwszym pomiarze (A), $p_2=(a+b)/n$, liczba osób którzy wybrali kategorię 1 w powtórnym pomiarze, a $q_1=1-p_1$, $q_2=1-p_2$, tzn. q_1 i q_2 są wyborem kategorii 2 zarówno w pierwszym jak i drugim pomiarze, g jest wskaźnikiem niezgodności (gross difference rate), który wynosi $g=(b+c)/n$, a liczba wszystkich jednostek wynosi $n=a+b+c+d$.

Zakłada się, że powtórny pomiar jest odzwierciedleniem pierwszego, co do warunków przeprowadzania wywiadu i w związku z tym można traktować te dwa pomiary jako równoległe i niezależne. Innym wskaźnikiem może być współczynnik liczony jako p_1-p_2 oznaczany jako NDR (net difference rate), co przy założeniu wysokiej wiarygodności drugiego pomiaru lub przyjmując drugi pomiar jako *złoty standard* może być wskaźnikiem błędu pomiaru (bias) w pierwszym wywiadzie. Natomiast jeśli przyjmujemy, że drugi wywiad jest odwzorowaniem pierwszego nie można wówczas robić założenia o złotym standardzie. Powyższe wskaźniki mogą być użyteczną miarą jako punkt wyjścia do oceny danych gromadzonych za pomocą kwestionariuszy.

Klasa ukryta

Inną próbą oceny jakości danych uzyskanych z dwóch pomiarów może być wykorzystanie analizy klasy ukrytej (LCA). W modelowaniu z wykorzystaniem klasy ukrytej (LCA) zakłada się, że prawdziwa cecha klasyfikacji danej osoby jest ukryta, a cechy zmierzone za pomocą pytań w kwestionariusza są tylko wskaźnikiem (manifest variable) tej cechy. Analiza klasy ukrytej początek bierze w literaturze psychometrycznej, która wykorzystała i rozwinęła tą metodę do kategoryzacji populacji na podstawie zmiennych zaobserwowanych. Przykładowo, tworząc klasyfikacje statusu społeczno-ekonomicznego respondenta, można posłużyć się zmiennymi dotyczącymi poziomu dochodu, wykształcenia i płci. Odnosząc się do kwestii podwójnego pomiaru oznaczamy S jako próbę liczącą n jednostek z dużej populacji, X_i jako prawdziwą charakterystykę jednostki i, natomiast A_i i B_i oznaczają dwa pomiary jednostki i. Dwa pomiary uzyskane jako dane z wywiadu i ponownego wywiadu lub dane uzyskane z dwóch pytań charakteryzujących tą samą cechę w jednym wywiadzie. Pomijając dla uproszenia indeks i, niech zmienne A, B i X przyjmują wartości 1 dla wyboru kategorii

1 odpowiedzi i 0 dla kategorii 2. Oznaczając π_z jako prawdopodobieństwo zdarzenia Z, mamy dla przykładu $\pi_{a|x}$ co oznacza P(A=a | X=x). Dla błędów klasyfikacji $\pi_{a=0|x=1}$ oznacza prawdopodobieństwo znalezienia się respondenta poza daną kategorią, chociaż do niej należy. Szacowanie parametrów prawdopodobieństwa modelu następuję metodą maksymalizacji funkcji prawdopodobieństwa (maximum likehood estimation). Hui and Walter (1980) zauważyli, że dla R pomiarów zastosowanych dla S populacji liczba parametrów do oszacowania wynosi (2R + 1)S dla (2R-1)S stopni swobody. Dla dwukrotnego pomiaru z jednej populacji dla dychotomicznej zmiennej liczba parametrów do oszacowania wynosi 5 przy 3 stopniach swobody. W związku z tym model jest nieidentyfikowalny i potrzebne są dodatkowe założenia. Dla przykładu można założyć rolę jednego z pomiarów jako złotego standardu. Następnie, stosując statystykę chi-kwadrat można oszacować stopień dopasowania modelu. W modelu niezależności równoległego pomiaru potrzebne są jeszcze dodatkowe założenia 1. $\pi_{a|x}=\pi_{b|x}$ oraz 2. lokalnej niezależności $\pi_{a\,b|x} = \pi_{a|x}\pi_{b|x}$.

Metoda Hui-Waltera

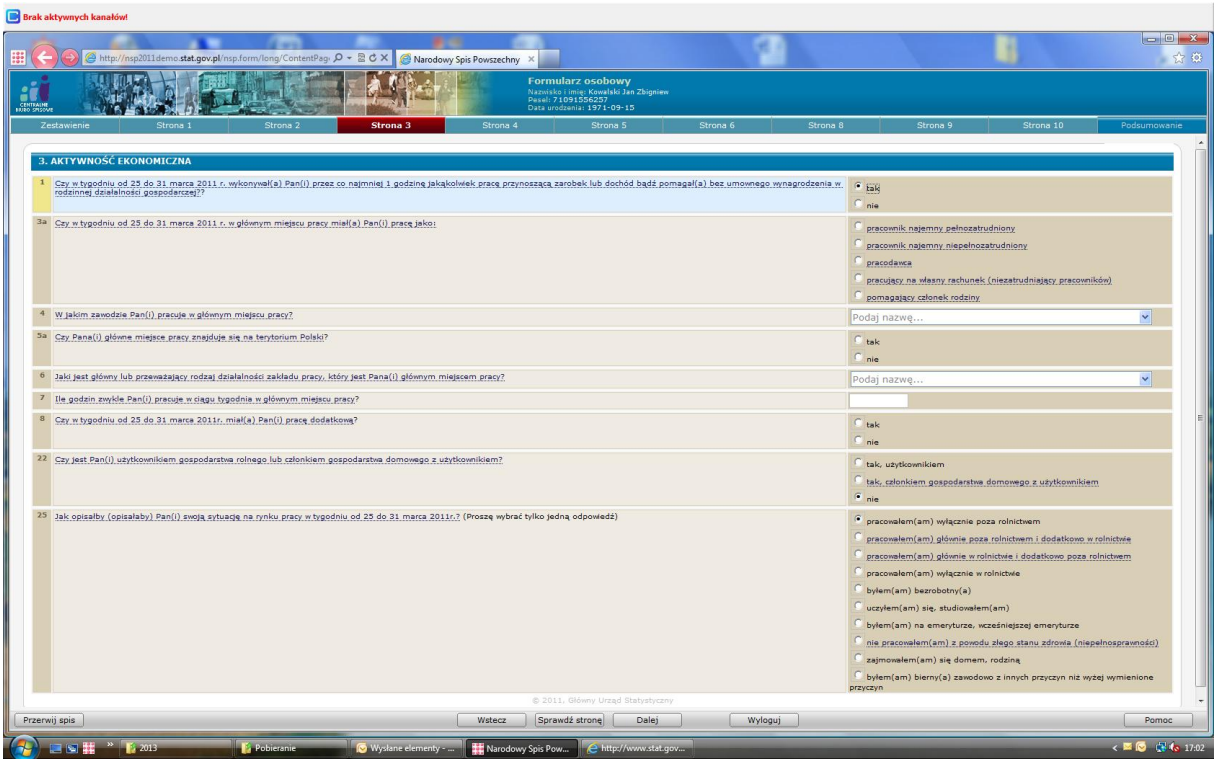
W celu uzyskania identyfikowalnego modelu dla dwukrotnego pomiaru Hui i Walter (1980) zastosowali użyteczne usprawnienie. Metoda ta wprowadza zmienna grupującą G oraz przyjmuje ograniczenie, że błędy pomiaru klasyfikacji są takie same w grupach. Prawdziwy wskaźnik cechy natomiast jest różny w przyjętych grupach. Dla przykładu przyjmując cechę grupującą G, wskutek której dla pomiaru i powtórnego pomiaru powstają dwie grupy, liczba parametrów do oszacowania dla dychotomicznej zmiennej wynosi osiem wraz tabelą klasyfikacyjną zawierającą osiem liczb. Metoda nie zakłada potrzeby, by którykolwiek pomiar odgrywał rolę złotego standardu.

Przykład

Jako przykład wykorzystania modelu Hui-Waltera, można przyjąć pomiar statusu osoby na rynku pracy mierzony za pomocą kwestionariusza w Narodowym Spisie Powszechnym Ludności i Mieszkań przeprowadzonego w 2011 r. W pytaniach mierzono status osoby na rynku pracy przyjmując definicję Międzynarodowej Organizacji Pracy (ILO). Klasyfikacja wyróżnia trzy kategorie osoby na rynku pracy: Pracujący, Bezrobotny i Bierny zawodowo. Nie wchodząc w szczegóły definicji ILO, kiedy osoba spełnia określone warunki, aby zostać przypisana do określonej kategorii, w Spisie Powszechnym, przypisanie osoby do danej kategorii następowało na podstawie zadania szeregu następujących po sobie pytań badających

status osoby na rynku pracy, których efektem była zmienna klasyfikująca respondentów pomiędzy kategorie wspomniane powyżej. Sekcja kwestionariusza zawierająca te pytania nazwała się Aktywność Ekonomiczna. Na załączonym rysunku (Rys. 1) przedstawiony jest fragment elektronicznego formularza internetowego, który zawierał automatyczne przekierowania i za pomocą szeregu pytań klasyfikował respondenta.

Rys. 1 Ekran przedstawiający stronę osobowego formularza internetowego użytego w Narodowym Spisie Powszechnym Ludności i Mieszkań przeprowadzonym w 2011 roku



Drugim źródłem danych do porówania danych klasyfikujących jest pytanie zawarte w tym samym kwastionariuszu (Rys. 1), które brzmiało:

Jak opisłaby (opisałaby) Pan(i) swoją sytaucję na rynku pracy w tygodniu od 25 do 31 marca 2011 r. ? (Proszę wybrać jedną odpowiedź) ?

Pytanie to zawierało następujące warianty odpowiedzi:

1 - pracowałem(am) wyłącznie poza rolnictwem

2 - pracowałem(am) głównie poza rolnictwem i dodatkowo w rolnictwie

3 - pracowałem(am) głównie w rolnictwie i dodatkowo poza rolnictwem

4 - pracowałem(am) wyłącznie w rolnictwie

5 - byłem(am) bezrobotny(a)

6 - uczyłem(am) się, studiowałem(am)

7 - byłem(am) na emeryturze, wcześniejszej emeryturze

8 - nie pracowałem(am) z powodu złego stanu zdrowia (niepełnosprawności)

9 - zajmowałem(am) się domem, rodziną

10 - byłem(am) bierny(a) zawodowo z innych przyczyn niż wyżej wymienione

99 – nieustalona

W celu otrzymania klasyfikacji analogicznej, dzielącej osoby między 3 kategorie na rynku pracy, połączyłem warianty odpowiedzi samooceny na rynku pracy w grupy w następujacy sposób:

Pracujący – wariant 1 + wariant 2 + wariant 3 +wariant 4,

Bezrobotni – kategeria 5,

Bierny zawodowo – wariant 6 + wariant 7 + wariant +8 + wariant 9 +wariant 10,

Przyjmując ten sposób wylicznenia do porówania dwóch zmiennych klasyfikacyjnych i wprowadzając zmienną grupującą według płci, otrzymałem tablicę liczności, przedstwioną w tablicy 2.

Tabl. 2 Tablica liczności osób na rynku pracy – wiersze zmienna samoocena, kolumny zmienna status ILO - na podstawie danych ze spisu reprezentacyjnego (Narodowy Spis Powszechny Ludności i Mieszkań 2011 r.)

Status - samoocena	Kobiety				Mężczyźni			
	Status – ILO				Status - ILO			
	pracujący	bezrobotni	bierni	brak danych	pracujący	bezrobotni	bierni	brak danych
Pracujący	1269252	8213	80104	0	1614995	10832	69822	3
Bezrobotni	9096	140452	82258	3	13421	178354	69981	1
Bierni	98221	58194	1612792	1	59002	34769	1026976	1
brak danych	144759	26818	73280	270042	194635	25668	48079	274044

Źródło: badanie reprezentacyjne NSP 2011.

Dane dotyczą osób ze spisu reprezentacyjnego, dla których wiek jest większy lub równy 15 lat. Liczba respondentów wynosi w tym wypadku 7494068 osób. Przygotowując w ten sposób dane dostajemy klasyfikację osób na rynku mierzonych za pomocą dwóch zmiennych, które mogą posłużyć jako przykład do porówania ich metodą klasy ukrytej. Ponieważ nie można tutaj założyć równoległości pomiaru, a w związku z tym, nie zakaładając żadnych restrykcji na spełnienie założeń do użycia tych danych do oceny samej klasyfikacji (tzn. nie zakładając

lokalnej niezależności i tej samej cechy ukrytej w obu zmiennych), można spojrzeć na wyliczenia jako możliwość zastosowania klasy ukrytej do każdej ze zmiennych z osobna. W wyliczeniach przyjęto płeć respondenta do podziału na dwie grupy danych G=1 dla kobiet i G=2 dla mężczyzn (zakładamy, że błąd pomiaru jest taki sam w obu grupach, wartość prawdziwej cechy jest natomiast w nich inna). Klasa ukryta X oznacza prawdziwą cechę X=1 jako pracujący, X=2 jako bezrobotni oraz X=3 jako bierni zawodowo. Zmienne A i B są danymi z klasyfikacji odpowiednio dla samooceny i statusu ILO, tzn. A=B=1 jako pracujący, A=B=2 jako bezrobotni, A=B=3 jako bierni zawodowo. W modelu szacujemy 18 parametrów, z 0 stopniami swobody, a w związku z tym nie można również ocenić stopnia dopasowania modelu.

Wyliczenia dla klasy ukrytej wykonałem w programie LEM Software (Vermunt,1997), który posiada analizę modeli klas ukrytych (wyniki wyliczeń przedstawione na Rys. 3).

Rys. 3. Wyniki obliczeń w programie LEM Software (Vermunt, 1997) na podstawie wejściowych danych z tablicy 2

```
*** INPUT ***
      lat 1
      man 3
      dim 3 2 3 3
      lab X G A B
      mod X G|X A|X B|X
      dat [1269252 8213 80104 9096 140452 82258 98221 58194 1612792 1614995 10832
69822 13421 178354 69981 59002 34769 1026976]
*** LATENT CLASS OUTPUT ***
          X 1    X 2     X 3
        0.4573  0.0650  0.4776
 G 1    0.4398  0.4343  0.6122
 G 2    0.5602  0.5657  0.3878
 A 1    1.0000  0.0489  0.0290
 A 2    0.0000  0.9511  0.0310
 A 3    0.0000  0.0000  0.9400
 B 1    0.9778  0.0442  0.0543
 B 2    0.0000  0.7930  0.0322
 B 3    0.0222  0.1628  0.9135
```

Zmienna A	Prawdziwy status		
Obserwowany Status	Pracujący	Bezrobotni	Bierni
Pracujący	1,0	0,0	0,0
Bezrobotni	0,0489	0,9511	0,0
Bierni	0,0290	0,0310	0,9400

Zmienna B	Prawdziwy status		
Obserwowany Status	Pracujący	Bezrobotni	Bierni
Pracujący	0,9778	0,0	0,0222
Bezrobotni	0,0442	0,7930	0,1628
Bierni	0.0543	0,0322	0,9135

W interpretacji wyników przyjmujemy oszacowanie poprawnego statusu jako wynik dla P(A=s|X=s), P(B=s|X=s) dla statusu s, s=1,2,3. Dla klasyfikacji A (samoocena), wynik dla wszystkich kategorii jest wysoki i wynosi dla pracujących 100 %, bezrobotnych 95,1 %, a dla biernych 94 %. Natomiast dla zmiennej B (status ILO) dużo niższy wynik jest dla kategorii bezrobotnych, który wynosi 79,4 % z liczbą 16,3 % jako niepoprawnie sklasyfikowanych do kategorii biernych zawodowo. Wynik jest wyłącznie pochodną przyjętego modelu, dlatego wyliczenia należy traktować tylko jako przykład zastosowania modelu analizy klasy ukrytej. Przykład ten może jednak stanowić punkt wyjścia do oceny pomiaru klasyfikacji statusu osoby na rynku pracy za pomocą kwestionariusza i oceny jej uzyskiwanej jakości. Z drugiej jednak strony badania nad jakością pomiaru takiej klasyfikacji, za pomocą równoległego pomiaru, które przeprowadzane są w U.S. Census Bureau w miesięcznym badaniu U.S Current Population Survey (CPS), wskazują, że kategoria bezrobotnych (unemployment) jest najtrudniejsza do uchwycenia i najmniej wiarygodna.

Literatura:

Berzofsky, M., Biemer, P., Kalsbeek, W., 2008, *A Brief History of Classification Models*, "Proceedings of the Survey Research Methods Section, American Statistical Association", online edition

Biemer, P., 2004, Modeling Measurement Error to Identify Flawed Questions In: Presser, S., et al. (eds.). *Methods for Testing and Evaluating Survey Questionnaires*, Chapter 12, Wiley, New York,

Biemer, P., Groves R., Lyberg, L., Mathiowetz, N., Sudman, S., 2004, *Measurement errors in surveys*, John Wiley & Sons, Inc, Hoboken, New Jersey,

Bradburn, N.M., Sudman S., (2011) The Current Status of Questionnaire Design In: Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., & Sudman, S. (Eds.). (2011). *Measurement errors in surveys* (Vol. 173). Wiley.com

Cohen, J., 1960, *A coefficient of agreement for nominal scales*, "Educational and Psychological Measurements", 20: 37–46.

Hansen, M., Hurwitz, W., and Pritzker, L., 1964, *The estimation and interpretation of gross differences and the simple response variance*, in C. Rao (ed.), "Contributions to Statistics.", Calcutta: Pergamon Press, pp. 111–136.

Hui, S., Walter, S., 1980, *Estimating the error rates of diagnostic tests*, "Biometrics", 36: 167–171.

Lohr, S.L., 1999, *Sampling: Design and analysis*, New York, Duxbury Press,

Vermunt, J., 1997, *REM: A General Program for the Analysis of Categorical Data*. 1997 Tilburg, The Netherlands: Tilburg University.