TOMASZ KULPA
dr inż., Politechnika Krakowska,
ul. Warszawska 24
31–155 Kraków,
tel. +48 604 093 965,
email: tkulpa@pk.edu.pl

METODY UZUPEŁNIANIA BRAKUJĄCYCH DANYCH NA PRZYKŁADZIE LICZBY ZAREJESTROWANYCH POJAZDÓW[1,2]

Streszczenie. W artykule podjęto problematykę braków w bazach danych wykorzystywanych w analizach transportowych. Braki występują zarówno w wynikach badań ankietowych, jak i innych bazach danych. Ich źródłem może być odmowa udzielenia odpowiedzi na pytanie, błędny pomiar lub po prostu dane nie są zbierane dla wszystkich elementów próby. W artykule wyróżniono trzy typy brakujących danych oraz dwie grupy metod ich uzupełniania: proste i złożone. Występowanie brakujących danych może mieć charakter losowy lub może być uzależnione od pewnych cech charakteryzujących populację. W pierwszym przypadku możliwe jest zastosowanie pełnej gamy metod uzupełniania brakujących danych. Pomimo tego częstą praktyką jest usuwanie wybrakowanych rekordów. Przy dużej próbie jest to metoda dopuszczalna, jednak w małych próbach powoduje dodatkowe zmniejszenie liczebności próby. Stąd konieczne jest poszukiwanie innych metod, np. uzupełnianie na podstawie podobnych rekordów, regresji liniowej lub metoda k-najbliższych sąsiadów. Różne metody uzupełniania brakujących danych zostały zilustrowane na fikcyjnych przykładach pokazujących ich istotę. Następnie wybrane metody wykorzystano do szacowania liczby zarejestrowanych samochodów ciężarowych w powiatach. Dokonana ocena poszczególnych metod pokazała, że najgorsze rezultaty uzyskano przy uzupełnianiu wartością średnią, natomiast najlepsze przy wykorzystaniu regresji liniowej. Zadowalające wyniki uzyskano również w przypadku metod złożonych. W podsumowaniu sformułowano wnioski dotyczące zastosowania technik uzupełniania brakujących danych, między innymi stosowanie usuwania brakujących rekordów tylko dla dużych prób oraz rezygnacji z uzupełniania wartością średnią na rzecz innych metod.

Słowa kluczowe: modelowanie podróży, badania ruchu, bazy danych

Wprowadzenie

Wiarygodne i kompletne bazy danych są podstawą do tworzenia modeli podróży. W analizach transportowych głównymi źródłami danych są badania ankietowe (np. w gospodarstwach domowych w ramach KBR, typu źródło-cel w przekrojach dróg) oraz pomiary ruchu (np. natężenia ruchu, ruchu tranzytowego). Inną grupę danych stanowią ogólnodostępne bazy Głównego Urzędu Statystycznego (GUS).

Przyczyny występowania braków w bazach danych mogą być bardzo różne. Najczęściej wynikają z odmowy odpowiedzi na pytanie zadane w ankiecie wykonywanej w gospodarstwie domowym lub przedsiębiorstwie wykorzystującym samochody ciężarowe. Przyczyną braków może być również niestaranność osoby wykonującej ankietę lub pomiar (np. pominięcie pytania, błędne zanotowanie odpowiedzi, nieprecyzyjny pomiar). W niektórych przypadkach dane mogą być niedostępne, ponieważ nie są zbierane, tak jak w przedstawionym w niniejszym referacie przykładzie dotyczącym uzupełniania liczby zarejestrowanych samochodów ciężarowych w powiatach.

Wśród pozostałych przyczyn można wymienić niepewność co do poprawności danych (pomimo ich uzyskania), uzyskanie informacji w formie przedziału lub opisowej zamiast konkretnej wartości lub przypisaniu jednemu obiektowi różnych wartości tego samego parametru.

Rozróżniamy trzy typy brakujących danych ([1], [2]):

- braki w pełni losowe, MCAR (ang. *Missing Completely at Random*) – brakujące wartości rozłożone są losowo wśród wszystkich obserwacji,
- braki częściowo losowe, MAR (ang. *Missing at Random*) – występowanie brakujących wartości może zależeć od innej zmiennej;
- braki nielosowe, NMAR (ang. *Not Missing at Random*).

W przypadku danych typu MCAR brakujące wartości są rozłożone losowo w zbiorze obserwacji. Oznacza to, że występowanie brakujących danych nie zależy od zmiennej, dla której występują braki ani od żadnej innej zmiennej znajdującą sie w zbiorze obserwacji [1]. Jeśli w badaniu ankietowym braki w odpowiedziach na pytanie o zarobki nie zależą od płci, miejsca zamieszkanie, wieku lub innych cech osoby ankietowanej będą one typu MCAR. Dane typu MAR charakteryzują się tym, że występowanie brakujących wartości może zależeć od jakiejś zmiennej objaśniającej, jednak nie zależą od zmiennej dla której występują braki [2]. Przykładowo jeśli w badaniu ankietowym osoby starsze lub mieszkające w dużych miastach będą unikać odpowiedzi na pytanie o zarobki, to brakujące dane będą typu MAR. Ostatni typ danych, NMAR, to takie, w których braki nie występują losowo i mogą być zależne również od zmiennej, dla której występują braki. Kontynuując przykład, jeśli brakujące odpowiedzi na pytanie o zarobki zależą od poziomu dochodów, to dane są typu NMAR. W zależności od typu brakujących danych możliwe jest zastosowanie określonej metody uzupełniania brakujących danych.

Można wyróżnić następujące metody uzupełniania brakujących danych [3]:

[2] Artykuł opracowano na podstawie referatu wygłoszonego na IX konferencji naukowo-technicznej „Problemy komunikacyjne miast w warunkach zatłoczenia komunikacyjnego", Poznań–Rosnówek, 19-–21 VI 2013 r.

- proste
  - usuwanie niekompletnych rekordów (przez rekord rozumiany jest pojedynczy wpis do bazy danych),
  - wstawienie wartości średniej lub mediany w miejsce brakujących danych,
  - zastępowanie brakującej wartości pochodzącą z rekordu o podobnym profilu (ang. *Pattern matching* lub *hotdecking*),
  - uzupełnianie na podstawie regresji liniowej (pojedynczej lub wielorakiej), jeśli zmienna z brakującymi wartościami jest skorelowana z innymi zmiennymi; w takim przypadku na podstawie formuły regresji przewiduje się brakujące wartości;
- złożone
  - wielokrotne wstawianie (ang. *multiple imputation,* MI) z wykorzystaniem metod największej wiarygodności (ang. *maximum likelihood,* ML).

Obecnie metody uzupełniania brakujących danych stosowane są bardzo często w medycynie i naukach społecznych [4]. Można znaleźć również pojedyncze zastosowania w dziedzinie transportu ([5], [6]), które w przytoczonych pozycjach bibliograficznych dotyczą uzupełniania brakujących danych wejściowych w systemach ITS.

Proste metody uzupełniania brakujących danych

Usuwanie niekompletnych rekordów (*Listwise Deletion*)

Jest to metoda najprostsza i posiada dwie podstawowe zalety: może być stosowana w każdym rodzaju analizy statystycznej oraz nie wymaga żadnych zaawansowanych metod obliczeniowych. Polega ona na usunięciu rekordów, które zawierają braki. W przypadku danych typu MCAR statystyki dla zredukowanej próbki będą odpowiadać wartościom w próbce głównej bez braków i nie będą obarczone błędem [4].

Zastępowanie wartością średnią (*Mean Imputation*)

Metoda polega na zastąpieniu brakujących wartości wartością średnią, obliczoną na podstawie znanych wartości. Zależność można opisać wzorem:

$$Y_B = \frac{\sum_{i=1}^{n_Z} Y_{Z_i}}{n_Z} \quad (1)$$

gdzie:

Y_B – brakująca wartość,

Y_{Zi} – znane wartości,

n_Z – liczba znanych wartości w całej próbie n.

W tabeli 1 pokazano fikcyjny przykład wyników badań ankietowych, w których brakującą informacją jest ruchliwość piątej osoby. Poszczególne wiersze są kolejnymi rekordami w bazie danych.

Brakującą wartość możemy uzupełnić średnią w kategorii ruchliwość, która wyniesie 2,0. Możliwe jest również wstawienie wartości średniej, jednak obliczonej dla mężczyzn (2,2), osób w wieku 30 lat (2,15) lub osób posiadających samochód (1,9).

Tabela 1

Przykładowe wyniki badań ankietowych dla potrzeb uzupełniania brakujących danych				
Lp.	**Płeć**	**Ruchliwość**	**Dostępność samochodu**	**Wiek**
1	K	1,9	T	30
2	K	1,7	T	52
3	M	2,4	N	30
4	M	2,0	T	44
5	M	?	T	30

Uzupełnianie na podstawie podobnych rekordów (*Hot-Decking, Pattern Matching*)

Metoda ta jest bardzo prosta i polega na odnalezieniu w całej próbie pełnego rekordu najbardziej podobnego do rekordu z brakującymi danymi pod względem jednej lub wielu cech. Jeśli w zbiorze pełnych rekordów znajduje się więcej niż jeden pasujący, wtedy albo brakującą wartość uzupełnia się wartością z pierwszego znalezionego rekordu lub wybraną losowo z wszystkich pasujących rekordów.

W analizowanym przykładzie, biorąc pod uwagę płeć i wiek, najbardziej pasującym rekordem jest rekord 3 i jako brakującą wartość można wstawić 2,4. Natomiast biorąc pod uwagę dostępność samochodu i wiek, najbliższym rekordem jest rekord 1 i jako brakującą wartość można wstawić 1,9. W przypadku uwzględnienia wszystkich cech (płeć, dostępność samochodu i wiek) trudno określić najbliższy rekord do rekordu 5. W takich przypadkach możliwe jest po prostu wylosowanie wartości z cechy ruchliwość i wstawienie jej w miejsce brakującej wartości.

Jak widać na powyższym przykładzie uzupełniona wartość zależy od przyjętych cech, które będą brane pod uwagę przy określaniu podobieństwa rekordów. Jednocześnie uwzględnienie zbyt dużej liczby cech może utrudnić znalezienie podobnego rekordu.

Wstawianie ostatniej zaobserwowanej wartości (*Last Value/Observation Carried Forward*, LVFC/LOFC)

Metoda ma zastosowanie, gdy analizowana cecha jest zmienna w czasie. Metoda zakłada, że pomimo występującej zmienności w czasie uzupełniane wartości cechy pozostają stałe od momentu zaobserwowania ostatniej wartości. W analizowanym przypadku (tab. 2) brakująca wartość dla wiersza 1 wynosiłaby 2,6. W wierszu 3 wszystkie brakujące wartości zostałyby uzupełnione liczbą 1,8, natomiast w wierszu 5 liczbą 1,7.

Tabela 2

Przykładowe ruchliwości 5 osób w zależności od dnia tygodnia dla potrzeb uzupełniania brakujących danych					
Lp.	**Ruchliwość (pn.)**	**Ruchliwość (wt.)**	**Ruchliwość (śr.)**	**Ruchliwość (czw.)**	**Ruchliwość (pt.)**
1	2,3	2	2,5	2,6	?
2	1,9	2,2	2,5	1,9	1,7
3	1,6	1,8	?	?	?
4	1,7	1,9	2,5	2,4	1,7
5	1,8	2,5	1,7	?	?

Uzupełnianie brakujących danych na podstawie wzoru regresji (*Regression Imputation*)

Jeżeli zmienna Y, dla której występują brakujące wartości, jest zależna od innej zmiennej (lub zmiennych) X, w której nie ma braków, to możliwe jest uzupełnienie brakujących danych zmiennej Y na podstawie regresji liniowej. W pierwszej kolejności należy z całej próby usunąć rekordy, w których występują braki wartości. Następnie należy znaleźć zależności regresyjne pomiędzy zmienną Y i zmienną (zmiennymi X). Następnie obliczamy brakujące wartości zmiennej Y na podstawie stworzonego modelu regresji.

Nawiązując do przykładu z tabeli 1, można stworzyć zależność regresyjną pomiędzy ruchliwością a wiekiem respondenta. Równanie regresji ma postać: $2{,}7 - 0{,}02 \cdot WIEK$ ($R2 = 0{,}47$). Stąd obliczona ruchliwość dla 5 osoby wynosi 2,1.

Powyższy przykład jest poglądowy i pokazuje procedurę postępowania w przypadku uzupełniania danych na podstawie wzoru regresji. Z punktu widzenia analizy regresji liniowej liczebność próby, jak i uzyskana wartość współczynnika determinacji są zbyt niskie, aby przyjąć taki model.

Złożone metody uzupełniania brakujących danych

Najczęściej stosowaną metodą złożoną jest metoda wielokrotnego wstawiania (ang. *Multiple Imputation*, MI). Po raz pierwszy metoda ta została zaproponowana przez Rubina w 1970 roku ([1]). Polega ona na wygenerowaniu od 3 do 10 różnych zestawów z uzupełnionymi danymi z wykorzystaniem wybranej metody (np. k-najbliższych sąsiadów lub oceny skłonności). Wygenerowanie kilku losowych zestawów brakujących danych odzwierciedla niepewność wartości, jaka powinna być wstawiona w miejsce brakującej. Następnie każdy zestaw uzupełnionych danych jest analizowany oddzielne, co daje m częściowych parametrów. W kolejnym kroku obliczane są wynikowe parametry (np. współczynniki regresji, wartości średnie, błędy).

Z uwagi na złożoność metod wielokrotnego wstawiania trudne jest przedstawienie przykładu opartego na konkretnej metodzie. Stąd w niniejszej pracy zawarto prosty przykład na podstawie danych z tabeli 1 z wykorzystaniem metody k-najbliższych sąsiadów pod względem wieku. Zakładając, że $k=3$, najbliższymi do brakującej są obserwacje: 1, 3 i 4. Następnie z trzech wartości ruchliwości (dla obserwacji 1, 3 i 4) losujemy (z jednakowym prawdopodobieństwem) jedną wartość. W ten sposób uzyskujemy pierwszy zestaw uzupełnionych danych. W celu uzyskania kolejnych zestawów uzupełnionych danych procedurę należy powtórzyć. Przykładowo, dla $m=2$ wstawień i $k=3$ najbliższych sąsiadów wyniki mogą wyglądać następująco: 1,9 i 2,0. Stąd ruchliwość 5 osoby będzie równa 1,95.

Zwykle liczba wstawień wynosi od 3 do 5. Powyżej 5 wstawień przy wzrastającym wysiłku obliczeniowym efektywność modelu wzrasta nieznacznie (tabela 3) zgodnie ze wzorem:

$$e = \left(1 + \frac{\gamma}{m}\right)^{-1} \tag{2}$$

gdzie:
e – procentowa efektywność,
γ – udział braków brakujących danych [-],
m – liczba wstawień.

Tabela 3

Procentowa efektywność e uzupełniania danych w zależności od udziału brakujących danych i liczby wstawień [1]

		Udział brakujących danych γ				
		0,1	0,3	0,5	0,7	0,9
Liczba wstawień m	3	97	91	86	81	77
	5	98	94	91	88	85
	10	99	97	95	93	92
	20	100	99	98	97	96

W odniesieniu do tabeli 3 można kwestionować, czy przy 70 lub 90% braków można stosować metody uzupełniania danych. W ogólności procedura wielokrotnego wstawiania wygląda następująco [1]:

- wygenerować brakujące wartości z wykorzystaniem odpowiedniego modelu, który uwzględnia zmienność losową;
- wykonać uzupełnianie danych m razy w celu uzyskania m kompletnych zestawów danych;
- obliczyć estymaty dla każdego uzupełnionego zestawu danych;
- uśrednić obliczone estymaty pomiędzy m zestawami uzupełnionych danych.

Chcąc obliczyć wynikowe wartości dowolnej estymaty dla m zestawów uzupełnionych danych, należy uśrednić wartości tych estymat uzyskane dla poszczególnych zestawów uzupełnionych danych według poniższego wzoru:

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{3}$$

gdzie:
m – liczba wstawień (wygenerowanych zestawów danych),
$\bar{Q}$ – wartość estymaty dla i-tego zestawu uzupełnionych danych,
$\hat{Q}_i$ – wynikowa wartość estymaty dla m wstawień.

Szczegółowy opis poszczególnych metod uzupełniania danych podano w [10].

Uzupełnianie danych o liczbie zarejestrowanych samochodów ciężarowych

W ramach własnych badań [9] dokonano adaptacji metody Vomberga do warunków polskich dla szacowania międzygminnych potoków samochodów ciężarowych, wykorzystując wyniki badań ankietowych typu źródło-cel wykonanych na drogach krajowych i wojewódzkich w roku 2006 [7]. Poprzez analogię do oryginalnej metody [8] w adaptacji założono, że wielkość potoku ruchu samochodów ciężarowych będzie zależna od liczby samochodów ciężarowych zarejestrowanych w gminach (zamiast miast) i odległości między nimi. Dla potrzeb adaptacji metody Vomberga niezbędna była informacja o liczbie zarejestrowanych samochodów ciężarowych w poszczególnych gminach w roku 2006 (rok wykonania badań ankietowych). Przyjęto, że wskaźnik motoryzacji dla powiatu będzie obowiązywał dla wszystkich gmin tego powiatu. Dopiero

od roku 2009 GUS publikuje w Banku Danych Lokalnych (BDL) liczby zarejestrowanych pojazdów dla wszystkich powiatów, stąd niezbędne było uzupełnienie informacji o zarejestrowanych samochodach ciężarowych w powiatach w roku 2006. Dane dla roku 2011 zostały wykorzystane do oceny metod uzupełniania brakujących wartości.

Pozyskano dane dla wszystkich powiatów dla roku 2011. Dla roku 2006 na 379 powiatów dla 142 dostępna była informacja o liczbie zarejestrowanych samochodów ciężarowych. W celu zastosowania metod uzupełniania danych w roku 2011 usunięto te same rekordy (dane o liczbie zarejestrowanych samochodów ciężarowych), które były niedostępne w roku 2006. Wykorzystując pozostałe rekordy, uzupełniono dane o liczbie zarejestrowanych pojazdów, stosując pakiet SOLAS [10]. Uzyskane uzupełnione liczby zarejestrowanych samochodów ciężarowych porównano z rzeczywistymi poprzez obliczenie średniego względnego błędu procentowego (MAPE). Wyniki zamieszczono w tabeli 4.

W metodzie opartej na uzupełnieniu wartości średniej, obliczono średnią liczbę pojazdów zarejestrowanych w powiatach, dla których dostępne były dane. W uzupełnianiu na podstawie podobnych rekordów wśród kompletnych rekordów wyszukiwano takie, które byłyby najbardziej zbliżone do rekordów nie zawierających liczby zarejestrowanych samochodów ciężarowych pod względem liczby mieszkańców (LM) i liczby podmiotów REGON w transporcie (REGT).

Brakujące wartości zostały uzupełnione na podstawie równań regresji opracowanych dla dostępnych danych dla roku 2011.

W metodzie wielokrotnego wstawiania wykorzystywana jest również regresja liniowa. Różnicą w stosunku do metody pojedynczego wstawiania jest uwzględnienie losowości przy doborze parametrów modelu regresji. W metodzie „analizy skłonności" w zbiorze kompletnych rekordów poszukiwane były takie, których skłonność, rozumiana jako prawdopodobieństwo wystąpienia braków, były podobne jak w przypadku rekordu z brakującymi danymi. W metodzie „odległości Mahalanobisa" poszukiwane były kompletne rekordy, dla których odległość Mahalanobisa od rekordu z brakami była jak najmniejsza. W każdej metodzie wielokrotnego wstawiania generowano 5 zestawów uzupełnionych danych, dla których następnie wyliczono wartości średnie. Uzyskane liczby zarejestrowanych samochodów ciężarowych porównano z danymi GUS i obliczono średni błąd względny (MAPE) dla każdej z metod.

Tabela 4

Ocena metod uzupełniania danych wykorzystanych do uzupełnienia liczby zarejestrowanych samochodów ciężarowych w powiatach w roku 2011

Metoda		Średni błąd względny [%]
Zastępowanie wartością średnią (*Mean Imputation*)		48,0
Uzupełnianie na podstawie podobnych rekordów (*Hot-Decking, Pattern Matching*)		20,1
Uzupełnianie brakujących danych na podstawie regresji (*Regression Imputation*)	Zmienna objaśniająca LM	16,8
	Zmienna objaśniająca REGT	16,9
Wielokrotne wstawianie na podstawie regresji (*Predictive Model Based Method*)		17,5
Wielokrotne wstawianie na podstawie analizy skłonności (*Propensity Score Method*)		40,0
Wielokrotne wstawianie na podstawie odległości Mahalanobisa (*Mahalanobis Method*)		18,4

Najgorsze wyniki, tj. najwyższy błąd względny, uzyskano dla uzupełniania brakujących danych wartością średnią. Najlepsze wyniki uzyskano dla prostej metody uzupełniania na podstawie regresji liniowej, która została wykorzystana do uzupełniania liczby zarejestrowanych samochodów dla roku 2006. Porównywalne wyniki dały metody złożone, z wyjątkiem metody wykorzystującej analizę skłonności. Niewiele większy błąd uzyskano dla metody uzupełniania na podstawie podobnych rekordów.

Podsumowanie i wnioski

Niemal w każdej dziedzinie nauki na etapie tworzenia baz danych występują braki. Najczęściej wybrakowane rekordy są usuwane kosztem zmniejszenia liczebności próby. Niestety, w niektórych przypadkach może to prowadzić do znacznego zredukowania próby lub usunięcia rekordów, które mogą być istotne. Stąd stosowanie metod uzupełniania brakujących danych jest uzasadnione.

Często w przypadku brakujących danych uzupełniane są one wartością średnią. Wynika to z łatwości zastosowania tej metody i jednocześnie mylnego założenia, że przyjęcie wartości średniej jest „bezpiecznym" rozwiązaniem. W przeprowadzonych analizach pokazano, że uzupełnianie brakujących wartości średnią daje najgorsze rezultaty ze wszystkich analizowanych metod. Dlatego metoda ta jest niezalecana. Przy dużej próbie usuwanie wybrakowanych rekordów może okazać się lepszą metodą niż uzupełnianie wartością średnią.

W artykule zilustrowano przydatność metod uzupełniania danych do szacowania liczby zarejestrowanych samochodów ciężarowych. Niemniej zakres zastosowania tych metod jest bardzo szeroki i możliwe jest ich wykorzystanie w uzupełnianiu brakujących wyników badań ankietowych (por. przykład teoretyczny) lub pomiarów ruchu (np. brak ciągłości pomiaru).

Literatura

1. Rubin D.B., *Multiple Imputation for Nonresponse in Surveys*, J. Wiley & Sons, New York 1987.
2. Schafer J.L., *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London 1997.
3. Lynch S.M., *SOC504 Course Website, Missing Data Notes*, http://www.princeton.edu/~slynch/soc504/soc504index.html (odczyt z dn. 28 listopada 2012 r.).
4. Acock A.C., *Working with missing Values*, Journal of Marriage and Family 67 (November 2005): 10121028.
5. Conklin J.H., Scherer W.T., *Data Imputation Strategies for Transportation Management Systems*, Research Report No. UVACTS-13-0-80 May, 2003.
6. Nguyen L.N., Scherer W.T., *Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications*, Research Report No. UVACTS-13-0-78, May, 2003.
7. *Studium układu dróg szybkiego ruchu w Polsce*, Politechnika Warszawska, Instytut Dróg i Mostów, Warszawa 2007.
8. Suwara T., *Analiza ruchu zamiejskiego*, WKiŁ, Warszawa 1988.
9. Kulpa T., *Modelowanie potencjałów ruchotwórczych w drogowych przewozach ładunków w skali regionu*, Praca doktorska, Politechnika Krakowska, 2013.
10. *SOLAS Imputation Manual*, SOLAS™ Version 4.0, 2011.