Dr hab. Marzena Nowakowska

Wydział Zarządzania i Modelowania Komputerowego
Politechnika Świętokrzyska
Al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Polska
E-mail: spimn@tu.kielce.pl

Przestrzenny i czasowy aspekt wyboru rozkładów apriorycznych i danych dla funkcji wiarygodności dla modeli bayesowskich w analizach bezpieczeństwa ruchu drogowego

***Słowa kluczowe**: model regresji bayesowskiej, informatywne rozkłady aprioryczne parametrów modelu, wiarygodność bayesowska, klasyfikator statystyczny, status wypadku drogowego, cechy wypadku drogowego*

Streszczenie: Parametry bayesowskiego modelu regresji nie są wartościami stałymi tylko zmiennymi losowymi opisanymi przez pewne rozkłady aposterioryczne. W celu zdefiniowania takiego rozkładu łączy się dwa źródła informacji: (1) rozkład aprioryczny, który reprezentuje wcześniejszą wiedzę o parametrze modelu oraz (2) funkcję wiarygodności (wiarygodność bayesowską), która uaktualnia wiedzę a'priori. Oba te elementy są przedmiotem badań w kontekście wykorzystania podejścia bayesowskiego w analizach bezpieczeństwa ruchu drogowego.
Badaniom podlega model wielokrotnej regresji logistycznej, który klasyfikuje status zdarzenia drogowego. W modelu uwzględniono trzy grupy zmiennych objaśniających: charakterystyki miejsca lokalizacji wypadku, cechy kierującego sprawcy oraz atrybuty wypadku. Ponieważ wypadki drogowe są rozproszone w czasie i przestrzeni, zaproponowano i poddano dyskusji dwa aspekty wyboru źródeł informacji w procedurze modelowania bayesowskiego: czasowy i przestrzenny. W obu podejściach rozkłady aprioryczne są definiowane na podstawie danych wybranych jako te, które generują uogólnioną wiedzę o parametrach modelu, tworząc tło podlegające modyfikacji – w ten sposób wiedza aprioryczna ma cechę informatywności. Wiarygodność bayesowska, modyfikująca rozkłady a'priori, jest definiowana za pomocą danych wprowadzających: (1) informację specyficzną dla wybranej drogi – w przypadku aspektu przestrzennego lub (2) informację najnowszą – w przypadku aspektu czasowego. Zaproponowane podejście zilustrowano w eksperymentach badawczych i przedstawiono wynikające z nich wnioski.

Spatial and temporal aspects of priors and likelihood data choices for Bayesian models in road traffic safety analysis

***Keywords:** Bayesian regression model, informative prior distributions for model parameters, likelihood data, statistical classifier, road accident severity, road accident features*

Abstract: In a Bayesian regression model, parameters are not constants, but random variables described by some posterior distributions. In order to define such a distribution, two pieces of information are combined: (1) a prior distribution that represents previous knowledge about a model parameter and (2) a likelihood function that updates prior knowledge. Both elements are analysed in terms of implementing the Bayesian approach in road safety analyses.
A Bayesian multiple logistic regression model that classifies road accident severity is investigated. Three groups of input variables have been considered in the model: accident location characteristics, at fault driver's features and accident attributes. Since road accidents are scattered in space and time, two aspects of information source choices in the Bayesian modelling procedure are proposed and

discussed: spatial and temporal ones. In both aspects, priors are based on selected data that generate background knowledge about model parameters – thus, prior knowledge has an informative property. Bayesian likelihoods which modify priors are data that deliver: (1) information specific to a road – in the spatial aspect or (2) the latest information – in the temporal aspect. The research experiments were conducted to illustrate the approach and some conclusions have been drawn.

1. Wprowadzenie

Bezpieczeństwo ruchu drogowego (brd) jako jeden z elementów systemu *człowiek-pojazd-droga* jest przedmiotem prac naukowych i badawczych od dziesiątków lat. Procesem poznania i zrozumienia mechanizmów związanych z wypadkiem drogowym zajmuje się wielu badaczy i specjalistów z różnych dziedzin i dyscyplin naukowych. W dążeniu do ustalenia stopnia zagrożenia w ruchu drogowym, określenia okoliczności powstawania wypadku, jego możliwych przyczyn i skutków, opracowano wiele teorii i modeli uwzględniających różne aspekty zjawiska. Obszar badań jest bardzo szeroki, obejmując m.in. badania symulacyjne i behawioralne (np. [8, 9]), opracowywanie modeli entropijnych (np. [1, 12]), badania poligonów drogowych z uwzględnieniem opisu warunków ruchu (szczególnie prędkości) oraz otoczenia drogi (np. [3, 10]) jak również eksplorację i drążenie rzeczywistych danych o zdarzeniach drogowych (np. [15, 19]).

W zbiorze różnych technik badawczych stosowanych w analizach danych rzeczywistych znaczące miejsce zajmują metody statystyczne, w których badacze stosują dwa podejścia: klasyczne i nieklasyczne. W podejściu klasycznym zakłada się, że prawdopodobieństwo zdarzenia losowego jest reprezentowane przez częstość wystąpienia tego zdarzenia w bardzo dużej liczbie takich samych obserwacji. Zgodnie z podejściem nieklasycznym, zwanym również bayesowskim, pierwotne (bezwarunkowe) prawdopodobieństwo zdarzenia jest miarą racjonalnego przekonania o zajściu zdarzenia. Przekonanie to jest następnie modyfikowane w drodze eksperymentów lub rejestracji danych o zjawiskach związanych ze zdarzeniem. Wiedzę aprioryczną przekształca się w wiedzę aposterioryczną, która jest prawdopodobieństwem wynikowym i miarą przewidywania zajścia zdarzenia po otrzymaniu informacji z zarejestrowanych danych. Myślenie bayesowskie, dzięki rozwojowi numerycznych technik próbkowania, stworzyło podstawy nowoczesnej statystyki, dzięki czemu możliwe było sformułowanie i rozwiązanie problemów niedostępnych dla statystyki klasycznej.

Nieklasyczną metodą statystyczną, która staje się coraz popularniejsza w analizach bezpieczeństwa ruchu drogowego, jest regresyjne modelowanie bayesowskie, zwłaszcza, że umożliwia ono wyeliminowanie różnych słabości modeli klasycznych. Bayesowskie modele regresji są trudne konceptualnie i obliczeniowo. Jednak, stwarzają nową jakość w rozwoju naukowych metod badawczych i umożliwiają elastyczne, chociaż niestandardowe, podejście do zagadnień modelowania. Są wykorzystywane w ocenie zagrożeń (np. [6, 7, 13, 16]), w tym w analizach „przed i po" (np. [17]), oraz w klasyfikacji cech jakościowych zdarzenia drogowego, takich jak zachowanie kierującego, rodzaj czy status tego zdarzenia (np. [2, 5, 16]).

Nieklasyczna metoda wnioskowania statystycznego została wykorzystana w pracy do opracowania logistycznych modeli regresji, w których zmienną objaśnianą jest status wypadku drogowego a zmiennymi objaśniającymi są wybrane cechy opisujące okoliczności wypadku. Przedstawiono sposób określenia podstawowych źródeł informacji wymaganych przez model bayesowski, proponując taką jego interpretację, w której informatywna wiedza aprioryczna jest informacją bazową (tłem) dla modelu a dane do uaktualnienia wiedzy apriorycznej (dane wiarygodności) odzwierciedlają ukierunkowanie modelu zgodnie z aspektem i zakresem szczegółowości opisu zjawiska.

2. Bayesowski klasyfikator statusu wypadku drogowego

Analizie podlega klasyfikator statystyczny – model regresji logistycznej, w którym status wypadku drogowego *AcSvr* jest klasyfikowany do jednej z dwóch wartości (kategorii): *LA* – wypadek lekki (traktowany w modelu jako porażka) oraz *FSA* – wypadek ciężki lub śmiertelny (traktowany w modelu jako po sukces). Zmienne objaśniające reprezentują miejsce lokalizacji wypadku, charakterystykę kierującego sprawcy oraz cechy wypadku.

Funkcją łączącą w modelu regresji logistycznej jest logit. Argumentem funkcji jest prawdopodobieństwo warunkowe $P(AcSrv = FSA| X_1, ..., X_k)$, że wypadek, który zdarzył się w okolicznościach zdefiniowanych przez zbiór wartości zmiennych objaśniających $(X_1, ..., X_k)$ jest śmiertelny lub ciężki ($AcSrv = FSA$):

(1)

Przyjęty model jest stosunkowo prosty, ponieważ celem głównym pracy nie jest dyskusja wpływu wybranych cech na zmienną objaśnianą lecz prezentacja metody konstruowania logistycznego modelu bayesowskiego.

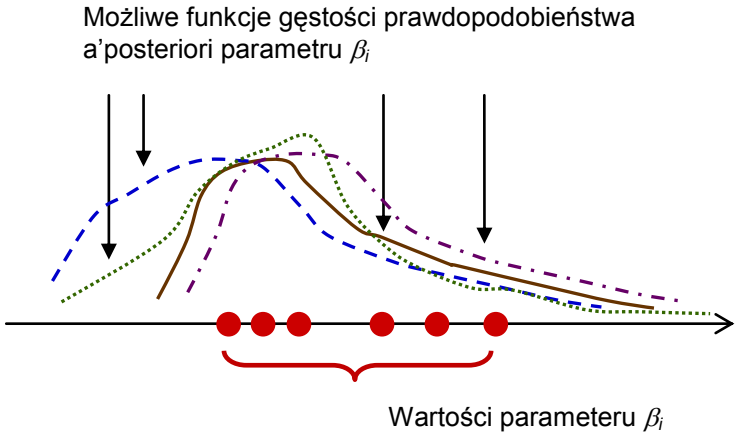
W przeciwieństwie do klasycznej, w regresji bayesowskiej zakłada się, że parametry modelu nie są stałymi tylko zmiennymi losowymi. W konsekwencji każdy parametr podlega pewnemu rozkładowi aposteriorycznemu będącemu wynikiem wiedzy pierwotnej o nim i uaktualnienia tej wiedzy poprzez wykorzystanie danych empirycznych (tworzących wiarygodność bayesowską) [18]:

$$P(\boldsymbol{\beta}|Y, \boldsymbol{X}) = P(\beta_0, ..., \beta_k \mid Y, \boldsymbol{X}) = P(\beta_0, ..., \beta_k) \cdot P(Y, \boldsymbol{X} \mid \beta_0, ..., \beta_k) / P(Y, \boldsymbol{X}) \propto P(\beta_0, \ldots, \beta_k) \cdot L(Y, \boldsymbol{X} \mid \beta_0, ..., \beta_k) = P(\boldsymbol{\beta}) \cdot L(Y, \boldsymbol{X} \mid \boldsymbol{\beta}) \quad (2)$$

W modelu bayesowskim podstawą oceny wielkości i kierunku wpływu zmiennej X_i na zmienną objaśnianą jest wartość średnia rozkładu aposteriorycznego parametru β_i stojącego przy tej zmiennej.

Zgodnie z regułą Bayesa rozkłady aposteriori $P(\boldsymbol{\beta} \mid Y, \boldsymbol{X})$ zawierają informację z dwóch źródeł: rozkładów a'priori $P(\boldsymbol{\beta})$ oraz funkcji wiarygodności $L(Y, \boldsymbol{X} \mid \boldsymbol{\beta})$. Konsekwencją założeń odnoszących się do wiedzy apriorycznej jak również doboru danych do funkcji wiarygodności są różne funkcje gęstości mogące opisywać rozkład aposterioryczny parametru strukturalnego β_i (rys. 1). Za każdym razem, gdy zmienia się któreś z tych źródeł rozkład aposterioryczny również ulega zmianie.

W celu wyznaczenia rozkładów aposteriorycznych $P(\boldsymbol{\beta} \mid Y, \boldsymbol{X})$ wykorzystano metodę Monte Carlo próbkowania łańcuchów Markowa – MCMC [4, 18]. Każdy rozkład jest wyznaczony na podstawie ciągu wartości numerycznych spełniających kryteria łańcucha Markowa. Najpopularniejszymi generatorami takich szeregów liczbowych są algorytm Metropolisa i jego uogólnienie – algorytm Metropolisa-Hastingsa. Często stosowany jest również próbnik Gibbsa. Wyniki metody MCMC zależą od liczby iteracji w łańcuchu, liczby wartości wypalonych (początkowych wartości odrzuconych z łańcucha) oraz wskaźnika przerzedzenia w łańcuchu docelowym. Kluczowym zagadnieniem w procesie generacji jest osiągnięcie stacjonarności łańcucha, dzięki czemu uzyskana próba pochodzi ze stacjonarnego rozkładu aposteriorycznego. Do oceny jakości wynikowego łańcucha Markowa wykorzystuje się testy diagnostyczne (np. Gelmana-Rubika, Geweke'a, Heidelbergera-Welcha) oraz wykresy diagnostyczne i korelacji.



Rys. 1. Graficzna interpretacja parametru bayesowskiego modelu regresji

3. Budowanie bayesowskiego klasyfikatora statusu wypadku drogowego

Proces budowania klasyfikatora bayesowskiego danego równaniem (1) oparto na dwukrokowej procedurze modelowania, w której zaadaptowano wybrane aspekty profilowania danych o wypadkach drogowych. Dane takie mają więc istotny wpływ na ostateczne wyniki zastosowania zaproponowanego podejścia: obejmują kilkuletni okres rejestracji i dotyczą wypadków drogowych z sieci dróg tej samej kategorii wybranego obszaru kraju (w szczególności mogą to być drogi pod zarządem określonej jednostki administracyjnej). Dane są dobrane tak, aby uwzględnić przestrzenny lub czasowy aspekt estymacji modelu. Prezentowane podejście jest rozszerzeniem i rozwinięciem koncepcji zaprezentowanej przez Yu i Adgel-Aty'ego [20], w której autorzy poddali dyskusji dobór rozkładów apriorycznych dla bayesowskiego prognostycznego modelu oceny bezpieczeństwa ruchu drogowego (*safety preformacje function*).

Algorytm budowania bayesowskiego klasyfikatora statusu wypadku drogowego składa się z dwóch kroków.

Modelowanie bayesowskie 1; definiowanie rozkładów a'priori – model BM-S1

W regresyjnych modelach bayesowskich można wyodrębnić następujące główne typy rozkładów apriorycznych: nieinformatywne, semi-informatywne oraz informatywne. Pierwszy z nich jest wykorzystywany w analizach bezpieczeństwa ruchu drogowego dużo częściej niż dwa pozostałe, mimo że w ostatecznym wyniku rozkład taki jest mocno zdominowany przez funkcję wiarygodności a wartości średnie rozkładów parametrów modelu bayesowskiego niewiele się różnią od wartości parametrów klasycznego modelu regresji. Lepsze wyniki można uzyskać stosując zamiast nieprecyzyjnych rozkładów nieinformatywnych dobrze zdefiniowane rozkłady informatywne, które odzwierciedlają wiedzę o przedmiocie badań. Aby uzyskać takie rozkłady zaproponowano odpowiednie przetwarzanie danych, definiując pierwszy krok wspomnianej procedury, której wynikiem jest bayesowski model regresji BM-S1.

Źródła informacji do modelu BM-S1 są następujące:

- rozkłady apriori parametrów – nieinformatywne: rozkłady normalne ze średnią równą zero i bardzo dużymi odchyleniami standardowymi (1E+06),
- wiarygodność bayesowska (funkcja wiarygodności) – zdefiniowana przez dane o wypadkach drogowych zgodnie z wybranym aspektem analiz: czasowym lub przestrzennym.

Wiarygodność bayesowska dla modelu BM-S1 jest określana na podstawie:

- dla aspektu przestrzennego: wszystkich danych o wypadkach zarejestrowanych dla dróg tej samej kategorii wybranego regionu kraju dla zadanego okresu analizy,
- dla aspektu czasowego: wszystkich danych historycznych o wypadkach zarejestrowanych dla dróg tej samej kategorii wybranego regionu kraju z wyłączeniem ostatniego (najnowszego) okresu rejestracji o pełnym cyklu sezonowym (rok).

Przyjęto, że wartości średnie i odchylenia standardowe rozkładów aposteriorycznych otrzymanych dla modelu BM-S1 stają się średnimi i odchyleniami standardowymi apriorycznych rozkładów normalnych dla parametrów bayesowskiego modelu regresji budowanego w kroku drugim.

Modelowanie bayesowskie 2; definiowanie funkcji wiarygodności – model BM-S2

Ponieważ, wyprowadzone w kroku pierwszym, rozkłady normalne nie są rozmyte, tworzą informatywną wiedzę aprioryczną, stanowiąc bazowe tło (uogólnienie), dla docelowego bayesowskiego modelu regresji BM-S2, w którym kontynuuje się wybrany aspekt analiz. Dane dla funkcji wiarygodności w tym modelu definiują zbiór uczący i są interpretowane jako czynnik, który uwypukla i doprecyzowuje kontekst badań:

- dla aspektu przestrzennego: dane o wypadkach dla wskazanej drogi wybrane z całego analizowanego zbioru modyfikują wiedzą aprioryczną w odniesieniu do tej drogi,
- dla aspektu czasowego: dane z ostatniego okresu uaktualniają wiedzę historyczną w odniesieniu do całego badanego zbioru dróg.

Obserwacje z wypadkami śmiertelnymi występują bardzo rzadko w zbiorze danych, co zazwyczaj skutkuje słabą jakością klasyfikacji tej kategorii zdarzeń. Dlatego, aby zniwelować to negatywne zjawisko i wzmocnić wpływ kategorii rzadkich na ostateczny wynik modelowania wprowadzono balansowanie danych [1, 14, 15] dla funkcji wiarygodności w modelu BM-S2, forsując bardziej zrównoważony rozkład zmiennej objaśnianej *AcSvr*. W procesie balansowania pierwotny zbiór danych jest dzielony na trzy podzbiory zgodnie z wartościami statusu wypadku: lekki, ciężki, śmiertelny. Następnie wszystkie obserwacje dotyczące wypadków śmiertelnych są pobierane do zbioru uczącego jako warstwa o liczebności 20%. Z pozostałych podzbiorów są pobierane w losowaniu prostym obserwacje tak, aby utworzyć w zbiorze uczącym 30% warstwę obserwacji dla wypadków ciężkich i 50% warstwę dla wypadków lekkich. Na końcu, powstaje binarna zmienna objaśniana *AcSrv*, w której kategoria wypadku lekkiego definiuje porażkę, a dwie pozostałe kategorie, wypadku ciężkiego oraz śmiertelnego, są agregowane definiując sukces. W tak zbalansowanym zbiorze wiarygodności znacząco wzrasta liczebność kategorii wypadku śmiertelnego, przy czym stosunkowo rzadka kategoria sukcesu nie przekracza 50% liczebności całego zbioru uczącego.

W eksperymencie badawczym balansowanie zastosowano w każdym aspekcie definiowania danych dla funkcji wiarygodności modelu BM-S2.

4. Opis danych do badań

Dane wykorzystane w eksperymencie badawczym pochodzą z policyjnego Systemu Ewidencji Wypadków i Kolizji (SEWiK) z wojewódzkiej komendy policji w Kielcach. Przedmiotem analiz są wypadki drogowe zarejestrowane w okresie 2008-2014 na drogach krajowych województwa świętokrzyskiego; tych dróg jest dziewięć. Obsługują one połączenia międzyregionalne i są zarządzane przez jednostkę szczebla krajowego (Generalna Dyrekcja Dróg Krajowych i Autostrad Oddział w Kielcach).

Do badań wybrano obserwacje spełniające następujące kryteria:

- wypadki zarejestrowano na drogach zamiejskich jednojezdniowych, dwupasowych, dwukierunkowych (spośród wszystkich kategorii dróg tego rodzaju, drogi krajowe charakteryzują się najwyższymi parametrami technicznymi),
- jeden pełnoletni kierujący był sprawcą wypadku,
- kierujący uczestnicy wypadków prowadzili pojazdy silnikowe,
- w wypadkach nie uczestniczyli piesi.

Przed rozpoczęciem prac, dane zostały wyczyszczone. Odrzucono obserwacje z bardzo rzadkimi kategoriami, które (uwzględniając ich fizyczne znaczenie) nie mogły być zagregowane oraz obserwacje z wartościami brakującymi lub odstającymi. Otrzymano zbiór o liczności 1329 obserwacji, opisany przez następujące zmienne wybrane do analiz:

- grupa charakterystyk miejsca wypadku drogowego (zmienne objaśniające):
  - *ArTp* (*area type*) – rodzaj obszaru z wartościami: *Bt* (*built-up*) – obszar zabudowany (39,2%), *NBt* (*no built-up*) – obszar niezabudowany (60,8%),
  - *LgCnd* (*lighting conditions*) – oświetlenie drogi z wartościami: *NgDrk* (*night darkness*) – brak oświetlenia w nocy (16,6%), *PrLg* (*poor lighting*) – niedostateczne oświetlenie, takie jak zmierzch, świt, sztuczne oświetlenie w nocy (14,7%), *Dlg* (*daylight*) – światło dzienne (68,6%),
  - *RdSrf* (*road surface*) – stan nawierzchni z wartościami: *NDr* (*not dry*) – niesucha, tzn. mokra, oblodzona lub ośnieżona (38,5%), *Dr* (*dry*) – sucha (61,5%),
- grupa cech kierującego sprawcy wypadku drogowego (zmienne objaśniające):
  - *VhTp* (*vehicle type*) – rodzaj pojazdu z wartościami: *HvVh* (*heavy vehicle*) – pojazd ciężki (15,6%), *Mtr* (*moped, scooter, motorcycle*) – motorowy pojazd jednośladowy (3,2%), *Cr* (*car*) – samochód osobowy (81,3%),
  - *Gndr* (*gender*) – płeć kierującego z wartościami: *F* (*female*) – kobieta (12,5%), *M* (*male*) – mężczyzna (87,5%),
  - *AgGrp* (*age group*) – grupa wiekowa kierującego z wartościami: 02 – <18; 25) (25,1%), 03 – <25; 35) (27,5%), 04 – <35; 50) (25,9%), 05 – <50; 65) (16,3%), 06 – co najmniej 65 lat (5,1%),
  - *Alh* (*alcohol*) – obecność alkoholu lub innych środków odurzających we krwi kierującego sprawcy z wartościami: *N* (*no*) – nie (89,8%), *Y* (*yes*) – tak (10,2%),
- grupa atrybutów wypadku drogowego (zmienne objaśniające):
  - *NrVhIn* (*number of vehicles involved*) – typ wypadku ze względu na liczbę uczestniczących w nim pojazdów z wartościami: *Sng* (*single*) – z udziałem jednego pojazdu (31,2%), *Mlt* (*multiple*) – z udziałem co najmniej dwóch pojazdów (68,8%),
  - *Bhv* (*behaviour*) – zachowanie kierującego sprawcy wypadku: *DrWrSdRd* (*driving wrong side of a roadway*) – jazda po niewłaściwej stronie drogi (5,2%), *InSpPrCn* (*inappropriate speed for prevailing traffic and weather conditions*) – niedostosowanie prędkości do warunków ruchu (44,2%), *NGvWy* (*not giving right of way*) – nieudzielenie pierwszeństwa przejazdu (10,3%), *InTrUTr* (*incorrect turning or u-turning*) – nieprawidłowe skręcanie lub zawracanie (4,1%), *InPs* (*incorrect passing by*) – nieprawidłowe mijanie (1,6%), *InOvBp* (*incorrect overtaking or baypassing*) – nieprawidłowe wyprzedzanie lub omijanie (12,9%), *PrPsCn* (*poor psychophysical condition*) – ograniczenie sprawności psychomotorycznej (w tym zmęczenie lub zaśnięcie) (8,3%), *FlCl* (*following too close*) – niezachowanie bezpiecznej odległości między pojazdami (13,5%),
- *AcSvr* (*acident severity*) – zmienna objaśniana; status wypadku drogowego zdefiniowany wg największego stopnia poszkodowania wśród ofiar ludzkich [14, 15, 21]: *LA* (*light accident*) – wypadek lekki (57%), *SA* (*serious accident*) – wypadek ciężki (29,4%), *FA* (*fatal accident*) – wypadek śmiertelny (13,5%).

5. Wyniki

Bayesowskie modele regresji otrzymano z 10000-elementowych łańcuchów wygenerowanych za pomocą algorytmu Metropolisa dla następujących ustawień: liczba prób wypalonych = 50000, liczba iteracji docelowych = 300000, wskaźnik przerzedzenia = 30. Stacjonarność uzyskano dla wszystkich łańcuchów Markowa, co zostało zweryfikowane za pomocą wykresów śladu i autokorelacji oraz testów Geweke'a i Heidelbegera-Welcha. Wynikowe rozkłady aposterioryczne są unimodalne.

Eksperymenty badawcze zostały przeprowadzone z wykorzystaniem środowiska systemu SAS: wbudowanej procedury MCMC oraz autorskich programów komputerowych napisanych w języku SAS 4GL oraz języku makr.

Dane do badań przygotowano uwzględniając:

- dla aspektu przestrzennego (S – *spatial aspect*):
  - BM-S1(S): wszystkie drogi krajowe województwa świętokrzyskiego, przedział czasowy 2008-2014 (długość zbioru danych jest równa 1329 rekordów),
  - BM-S2(S): drogi DK74 oraz DK7 dla dwóch niezależnych modeli, przedział czasowy 2008-2014 (po zbalansowaniu długości zbiorów są równe odpowiednio 220 i 196 rekordów); najważniejsza różnica między tymi dwiema drogami polega na tym, że droga DK7 prowadzi dodatkowo ruch międzynarodowy będąc częścią europejskiego systemu dróg,
- dla aspektu czasowego (T – *temporal aspect*):
  - BM-S1(T): wszystkie drogi krajowe województwa świętokrzyskiego, przedział czasowy 2008-2013 (długość zbioru danych jest równa 1221 rekordów),
  - BM-S2(T): wszystkie drogi krajowe województwa świętokrzyskiego, rok 2014 (po zbalansowaniu długość zbioru danych jest równa 60 rekordów),

Wyniki modelowania bayesowskiego przedstawiono: dla aspektu przestrzennego – w tabeli 1, dla aspektu czasowego – w tabeli 2. Modele BM-S1 otrzymane w pierwszym kroku nazwano apriorycznymi, ponieważ dostarczają apriorycznej wiedzy dla kroku BM-S2. Modele BM-S2 otrzymane w drugim kroku nazwano aposteriorycznymi, ponieważ są ostatecznymi klasyfikatorami całego procesu modelowania. Zestawienia obejmują informacje wg układu:

- wartości średniej i odchylenia standardowego (*Średnia (O.S.)*) rozkładów parametrów modeli: apriorycznych (*BM-S1 – a'priori*) i aposteriorycznych (*BM-S2 – a'posteriori*),
- odniesienie każdego modelu aposteriorycznego do odpowiadającego mu modelu apriorycznego poprzez wyznaczenie wartości wskaźnika porównującego średnią rozkładu parametru w modelu aposteriorycznym ze średnią rozkładu odpowiadającego mu parametru w modelu apriorycznym. Wskaźnik jest zdefiniowany za pomocą wyrażenia $(średnia_{a'posteriori} - średnia_{a'priori})/|średnia_{a'priori}|$ a jego wartości są zawarte w kolumnach *Porównanie* dla: *DK74 vs. a'priori, DK7 vs. a'priori* oraz *2014 vs. a'priori*,
- porównanie dwóch modeli aposteriorycznych dla aspektu przestrzennego (dla dróg: DK74 i DK7) poprzez wyznaczenie różnic między średnimi rozkładów odpowiadających sobie parametrów tych modeli wg zależności: $(średnia_{a'posteriori\ (DK74)} - średnia_{a'posteriori\ (DK7)})$. Obliczone wartości są zawarte w kolumnie *Porównanie modeli BM-S2(S)* w tabeli 1,
- wskaźnik jakości modelu *DIC* policzony w odniesieniu do zbiorów treningowych: niezbalansowanego dla kroku BM-S1 oraz zbalansowanego dla kroku BM-S2,
- miary oceny jakości klasyfikacji; czułość (*sensitivity*, procent prawidłowych klasyfikacji wartości *FSA*), swoistość (*specificity*, odsetek prawidłowo sklasyfikowanych wartości *LA*), średnia harmoniczna czułości i swoistości *HMSS* (która balansuje obie te miary) –

miary policzono na podstawie pierwotnego zbioru wiarygodności dla modelu BM-S1 oraz na podstawie pierwotnego (przed zbalansowaniem) zbioru wiarygodności dla modelu BM-S2.

Dla każdego parametru strukturalnego modelu bayesowskiego można w sposób jednoznaczny wyznaczyć przedział największej gęstości rozkładu a'posteriori (HPD), pod warunkiem, że rozkład ten nie jest równomierny. Do pewnego stopnia przedział HPD odpowiada przedziałowi ufności w statystyce klasycznej – jeżeli zawiera zero, to nie można jednoznacznie interpretować wartości odpowiadającego mu parametru. Ta niepewność jest również sygnalizowana przez współczynnik zmienności parametru, większy (co do modułu) od 50%. Takie nieistotne statystycznie parametry są wyróżnione w obu tabelach czerwoną czcionką. Przedziały największej gęstości rozkładu a'posteriori (HPD) dla istotnych statystycznie parametrów modeli wynikowych (modeli BM-S2 otrzymanych w kroku drugim) zilustrowano na rysunkach 2 i 3.

W wynikach przedstawionych w tabelach 1 i 2 oraz na rysunkach 2 i 3 zmienne objaśniające są pogrupowane zgodnie z ich znaczeniem merytorycznym: charakterystyki miejsca wypadku drogowego, cechy kierującego sprawcy, atrybuty wypadku.

Modele bayesowskie dla aspektu przestrzennego

Tabela 1. Charakterystyka modeli bayesowskich dla aspektu przestrzennego klasyfikujących status wypadku drogowego

Model	**BM-S1(S) – a'priori**	**BM-S2(S) – a'posteriori dla DK74**		**BM-S2(S) – a'posteriori dla DK7**		**Porównanie modeli BM-S2(S)**
Specyfikacja	**Średnia (O.S.)**	**Średnia (O.S.)**	**Porównanie: DK74 vs. a'priori**	**Średnia (O.S.)**	**Porównanie: DK7 vs. a'priori**	**DK74-DK7**
Stała	-1,396 (0,378)	-1,224 (0,235)	12,3%	-1,192 (0,249)	14,6%	-0,032
Grupa charakterystyk miejsca wypadku drogowego						
ArTp_Bt	0,311 (0,127)	0,381 (0,117)	22,5%	0,326 (0,119)	4,9%	0,055
LgCnd_NgDrk	0,341 (0,165)	0,434 (0,156)	27,1%	0,321 (0,153)	-5,8%	0,112
LgCnd_PrLg	-0,090 (0,174)	-0,103 (0,159)		0,020 (0,166)		
RdSrf_NDr	0,011 (0,126)	-0,070 (0,116)		0,009 (0,118)		
Grupa cech kierującego sprawcy wypadku drogowego						
VhTp_HvVh	-0,082 (0,172)	-0,039 (0,159)		-0,062 (0,159)		
VhTp_Mtr	1,217 (0,361)	1,101 (0,333)	-9,5%	1,203 (0,333)	-1,1%	-0,102
Gndr_F	-0,428 (0,191)	-0,386 (0,172)	9,8%	-0,422 (0,181)	1,5%	0,036
AgGrp_02	-0,043 (0,289)	0,202 (0,226)		0,023 (0,234)		
AgGrp_03	-0,156 (0,288)	0,003 (0,215)		-0,159 (0,229)		
AgGrp_04	-0,112 (0,288)	-0,026 (0,224)		-0,142 (0,224)		
AgGrp_05	-0,201 (0,300)	-0,509 (0,245)	153,2%	-0,078 (0,246)		0,432
Alh_N	0,008 (0,204)	0,062 (0,176)		-0,099 (0,184)		
Grupa atrybutów wypadku drogowego						
AcTp_Sng	-0,366 (0,158)	-0,339 (0,143)	7,3%	-0,440 (0,146)	-20,2%	0,101
Bhv_DrWrSdRd	2,342 (0,343)	2,390 (0,308)	2,0%	2,340 (0,304)	-0,1%	0,050
Bhv_InSpPrCn	1,175 (0,229)	1,161 (0,181)	-1,1%	1,149 (0,187)	-2,2%	0,013
Bhv_NGvWy	0,975 (0,263)	0,908 (0,225)	-6,8%	1,089 (0,237)	11,7%	-0,181
Bhv_InTrUTr	0,829 (0,345)	0,832 (0,307)	0,3%	0,753 (0,290)	-9,1%	0,079
Bhv_InPs	2,439 (0,569)	2,410 (0,511)	-1,2%	2,171 (0,500)	-11,0%	0,238
Bhv_InOvBp	1,354 (0,250)	1,450 (0,226)	7,1%	1,435 (0,222)	6,0%	0,016
Bhv_PrPsCn	1,336 (0,295)	1,405 (0,258)	5,2%	1,233 (0,262)	-7,7%	0,173
DIC	1168,6	249,5		231,5		
Czułość	38,9%	59,3%		57,9%		

Swoistość HMSS	82,6% 52,9%	67,8% 63,3%		65,3% 61,4%		

1. Zbiory istotnych statystycznie zmiennych objaśniających są w modelu BM-S1(S) i obu modelach BM-S2(S) prawie takie same. Grupa wiekowa kierującego sprawcy wypadku drogowego okazała się istotna w modelu BM-S2(S) wyznaczonym dla drogi DK 74 tylko dzięki istotności zmiennej kodowanej *AgGrp_05* (kierujący w wieku 50-65 lat).
2. Kierunki wpływu poszczególnych zmiennych istotnych statystycznie są takie same zarówno w modelu pierwszego kroku jak i w obu modelach kroku drugiego.
3. Charakter (wielkość i kierunek) zmiany wartości istotnych statystycznie parametrów aposteriorycznych (modele BM-S2(S)) w odniesieniu do wartości odpowiadających im parametrów apriorycznych (model BM-S1(S)) zależą od drogi:
   - dodatni wpływ charakterystyk miejsca wypadku jest o ponad 20% większy w modelu BM-S2(S) dla drogi DK74 podczas gdy zmiana tego wpływu w modelu BM-S2(S) dla drogi DK7 jest inna – średnia parametru dla obszaru zabudowanego *ArTp_Bt* jest większa o 5% a dla braku oświetlenia w nocy *LgCnd_NgDrk* – mniejsza o 6%,
   - dodatni wpływ jednośladowych pojazdów motorowych *VhTp_Mtr* i ujemny płci żeńskiej kierującego sprawcy *Gndr_F* zidentyfikowany w apriorycznych rozkładach parametrów zmniejszył się w rozkładach aposteriorycznym dla drogi DK74 o blisko 10%, ale w rozkładach aposteriorycznym dla drogi DK7 pozostał na prawie tym samym poziomie,
   - modyfikacja rozkładu apriorycznego parametru dla zmiennej identyfikującej wypadek z udziałem jednego pojazdu *NrVhIn_Sng* przez dane z różnych dróg spowodowała różne skutki w rozkładach aposteriorycznych: dla drogi DK74 wartość średnia parametru wzrosła o 7% ale dla drogi DK7 spadła o 20%,
   - zakres zmian w rozkładach aposteriorycznych parametrów odnoszących się do zachowania kierującego jest różny dla dróg DK74 i DK7, co jest szczególnie widoczne w przypadku nieustąpienia pierwszeństwa przejazdu *Bhv_NGvWy* (odpowiednio spadek wartości średniej o 6,8% i wzrost o 11,7%), nieprawidłowego skręcania lub zawracania *Bhv_TrUTr* (odpowiednio prawie bez zmian i spadek o 9,1%) oraz ograniczenia sprawności psychomotorycznej kierującego *Bhv_PrPsCn* (odpowiednio wzrost o 5,2% i spadek o 7,7%).

Modele bayesowskie dla aspektu czasowego

Tabela 2. Charakterystyka modeli bayesowskich dla aspektu czasowego klasyfikujących status wypadku drogowego

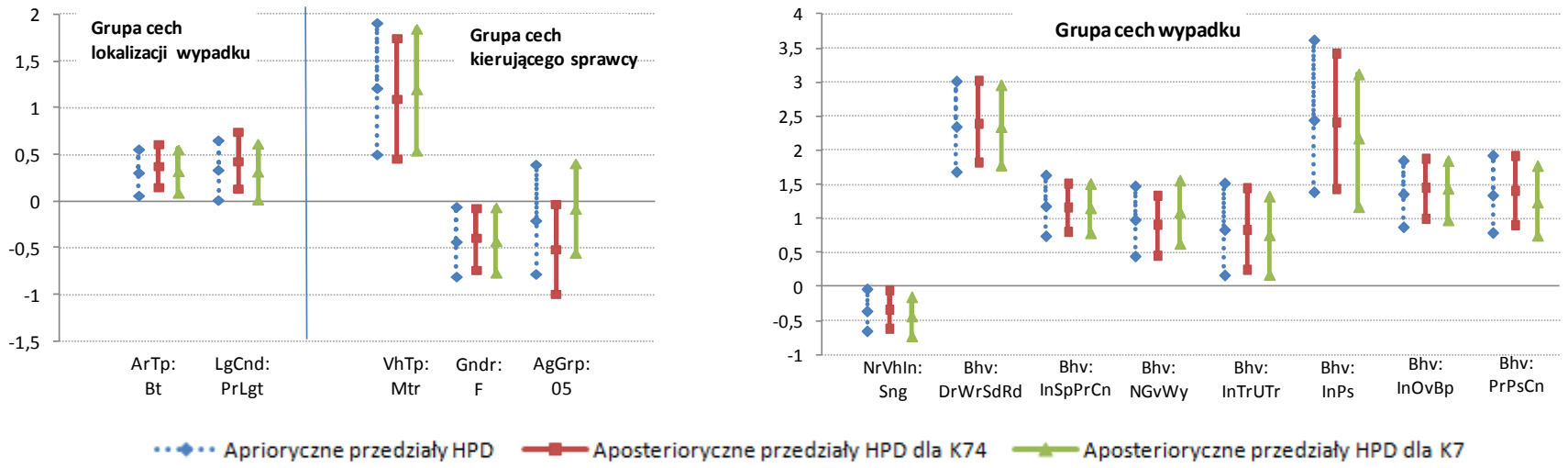
Model	**BM-S1(T) – a'priori**	**BM-S2(T) – a'posteriori dla 2014**	
Specyfikacja	**Średnia (O.S.)**	**Średnia (O.S.)**	**Porównanie: 2014 vs. a'priori**
Stała	-1,192 (0,393)	-1,169 (0,289)	1,9%
Grupa charakterystyk miejsca wypadku drogowego			
ArTp_Bt	0,383 (0,133)	0,351 (0,127)	-8,3%
LgCnd_NgDrk	0,319 (0,176)	0,353 (0,171)	10,6%
LgCnd_PrLg	-0,048 (0,178)	-0,040 (0,176)	
RdSrf_NDr	-0,042 (0,131)	0,008 (0,128)	
Grupa cech kierującego sprawcy wypadku drogowego			
VhTp_HvVh	-0,053 (0,177)	-0,077 (0,171)	
VhTp_Mtr	1,329 (0,387)	1,216 (0,362)	-8,6%
Gndr_F	-0,446 (0,202)	-0,448 (0,198)	-0,4%
AgGrp_02	-0,222 (0,304)	0,099 (0,272)	

AgGrp_03	-0,279 (0,300)	-0,388 (0,277)	
AgGrp_04	-0,258 (0,303)	-0,344 (0,280)	
AgGrp_05	-0,303 (0,314)	-0,283 (0,286)	
Alh_N	-0,051 (0,214)	-0,046 (0,198)	
Grupa atrybutów wypadku drogowego			
NrVhIn_Sng	-0,356 (0,166)	-0,377 (0,160)	-5,9%
Bhv_DrWrSdRd	2,423 (0,361)	2,389 (0,353)	-1,4%
Bhv_InSpPrCn	1,103 (0,239)	1,246 (0,220)	12,9%
Bhv_NGvWy	1,061 (0,272)	0,912 (0,263)	-14,0%
Bhv_InTrUTr	0,764 (0,357)	0,700 (0,352)	-8,3%
Bhv_InPs	2,211 (0,594)	2,344 (0,568)	6,1%
Bhv_InOvBp	1,240 (0,262)	1,395 (0,243)	12,5%
Bhv_PrPsCn	1,350 (0,306)	1,239 (0,295)	-8,2%
DIC	1092,4	74,8	
Czułość	36,9%	61,9%	
Swoistość	82,8%	74,2%	
HMSS	51,0%	67,5%	

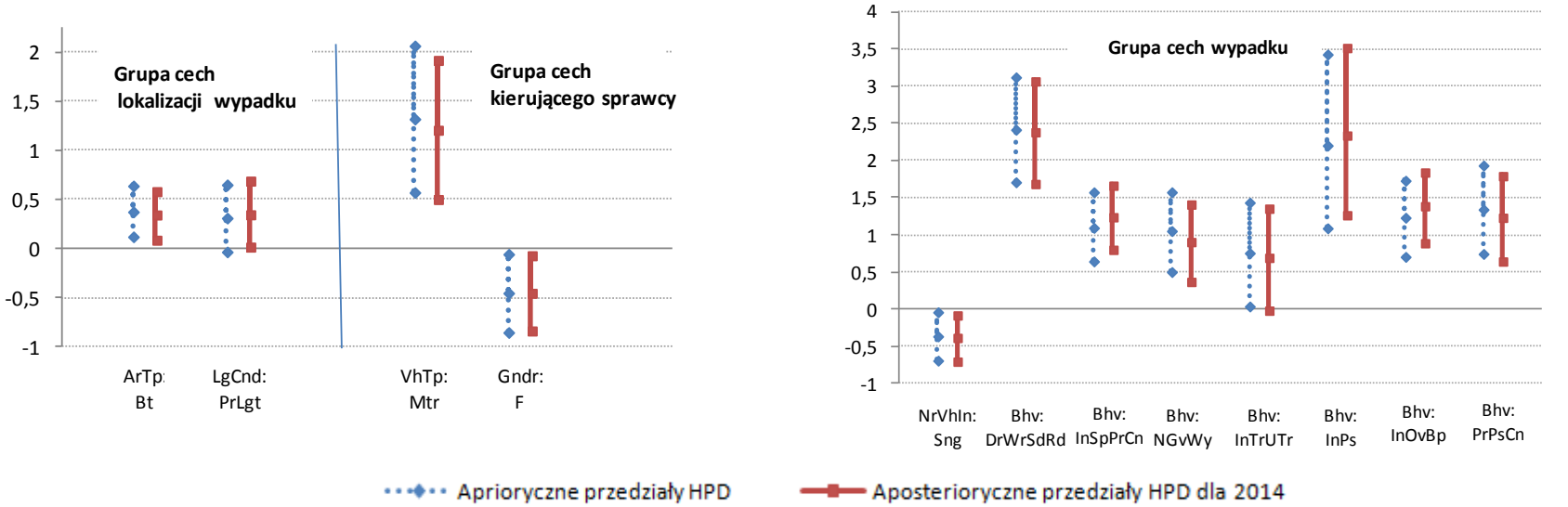
1. Zbiory istotnych statystycznie zmiennych objaśniających w modelach BM-S1(T) i BM-S2(T) różnią się w zakresie dwóch zmiennych: (1) brak oświetlenia w nocy *LgCnd_NgDrk* jest statystycznie nieistotny w modelu BM-S1(T) ale istotny w modelu BM-S2(T), (2) nieprawidłowe skręcanie lub zawracanie *Bhv_InTrUTr* jest istotne statystycznie w modelu BM-S1(T) ale nieistotne w modelu BM-S2(T).
2. Podobnie jak w modelach dla aspektu przestrzennego, kierunki wpływu odpowiadających sobie istotnych statystycznie zmiennych w modelu pierwszego kroku BM-S1(T) i w modelu kroku drugiego BM-S2(T) są takie same.
3. Nowa informacja zmodyfikowała dotychczasową (aprioryczną) wiedzę o znaczeniu poszczególnych zmiennych objaśniających w modelu aposteriorycznym, w szczególności przyczyniając się do wzmocnienia:
   - pozytywnego wpływu na ciężki lub śmiertelny status wypadku drogowego następujących cech: brak oświetlenia w nocy *LgCnd_NgDrk* (wzrost o 10,6%), niedostosowanie prędkości do warunków ruchu *Bhv_InSpPrCn* (wzrost o 12,5%), nieprawidłowe wyprzedzanie lub omijanie *Bhv_InOvBp* (wzrost o 12,5%) ,
   - negatywnego wpływu na ciężki lub śmiertelny status wypadku drogowego zmiennej identyfikującej wypadek z udziałem jednego pojazdu *NrVhIn_Sng* (spadek o 5,9%).

Zarówno w aspekcie przestrzennym jak i czasowym balansowanie zbioru danych wiarygodności w drugim kroku modelowania wpływa pozytywnie na jakość klasyfikacji bayesowskich modeli logistycznych. Satysfakcjonujące są wartości wszystkich miar: czułość jest większa niż 57%, swoistość jest większa niż 65%, wskaźnik HMSS jest większy niż 61%.

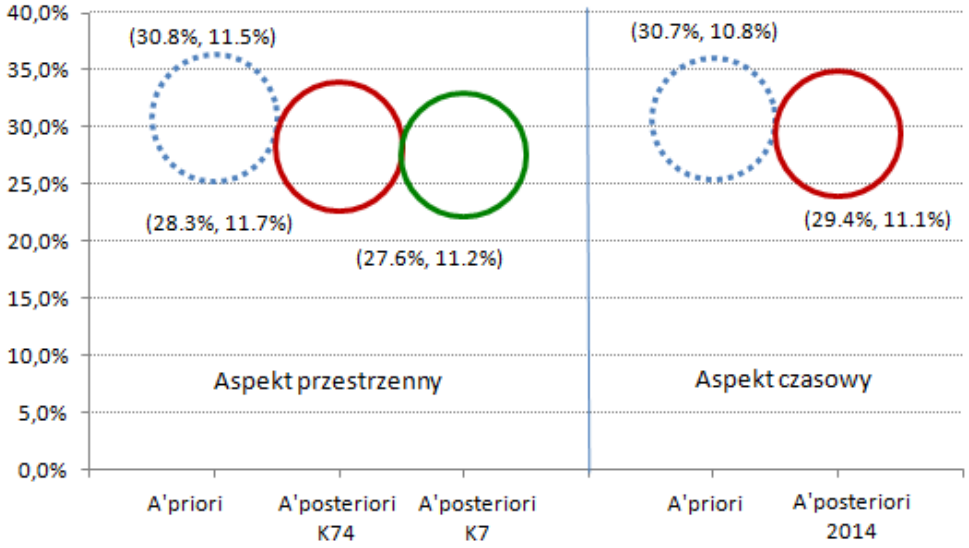
Uogólniony obraz wskaźników zmienności istotnych statystycznie parametrów modeli przedstawiono na rys. 4 w postaci wykresów bąbelkowych; środki i promienie okręgów reprezentują odpowiednio średnie i odchylenia standardowe wskaźników. Wartości odchyleń standardowych są bardzo podobne, niezależnie od kroku (modele aprioryczne lub aposterioryczne) oraz aspektu (przestrzenny, czasowy) modelowania. Nieco większą różnicę można zauważyć w przypadku wartości średnich dyskutowanych wskaźników – mniejszych dla parametrów modeli kroku drugiego, co wskazuje na lepszą precyzję oszacowania docelowych modeli aposteriorycznych.



Rys. 2. Przedziały HPD dla istotnych statystycznie parametrów modeli bayesowskich aspektu przestrzennego



Rys. 3. Przedziały HPD dla istotnych statystycznie parametrów modeli bayesowskich aspektu czasowego



Rys. 4. Wykresy bąbelkowe średnich i odchyleń standardowych współczynników zmienności istotnych statystycznie parametrów modeli bayesowskich

6. Wnioski końcowe

W modelu regresji bayesowskiej parametry są zmiennymi losowymi. Ich rozkłady, zwane rozkładami aposteriorycznymi, uzyskuje się poprzez łączenie wiedzy systematycznej (apriorycznej) o tych parametrach z wiarygodnością bayesowską – wiedzą pochodzącą z danych. Pewne zagadnienia związane z metodologią budowy takich modeli dla potrzeb analiz bezpieczeństwa ruchu drogowego przedstawiono w tej pracy. Badaniom został poddany model regresji logistycznej do klasyfikacji statusu wypadku drogowego.

Dane o wypadkach drogowych są traktowane jako potencjalne źródło obu rodzajów informacji do modelu bayerowskiego: wiedzy apriorycznej i wiarygodności bayesowskiej. Niektórzy badacze stosują już takie podejście, jednak zaproponowano szczególny sposób interpretacji obu źródeł i w konsekwencji ich wykorzystania, uwzględniając dodatkowo zadanie uzyskania jak najlepszych klasyfikatorów końcowych.

Wiedza aprioryczna o parametrach regresji może być uzyskana z danych, których zakres zależy od przedmiotu badań. W aspekcie przestrzennym są to dane o wypadkach z grupy dróg tej samej klasy technicznej określonego regionu kraju. Wtedy stają się źródłem informatywnej wiedzy apriorycznej tworząc bazę (rodzaj tła odniesienia), która jest kalibrowana (uaktualniana) przez wiarygodność bayesowską pochodzącą z danych o wypadkach drogowych zarejestrowanych na wybranej drodze. Dzięki temu uzyskuje się model specyficzny dla tej właśnie drogi. Jeżeli badania dotyczą aspektu czasowego, informatywne tło aprioryczne tworzą dane historyczne o wypadkach, a nowe dane (z ostatniego okresu rejestracji) uaktualniają tę wiedzę aprioryczną, dostarczając najnowszego, uogólnionego obrazu stanu brd sieci dróg regionu.

Wprowadzając, zarówno w aspekcie przestrzennym jak i czasowym, balansowanie zbioru danych dla wiarygodności bayesowskiej uzyskuje się dobrą jakość klasyfikacji ostatecznych bayesowskich modeli regresji logistycznej. Ten wynik jest szczególnie ważny, ponieważ poziom poprawnych klasyfikacji rzadkich kategorii sukcesu, tzn. ciężkiego lub śmiertelnego statusu wypadku drogowego, ma znaczenie kluczowe.

Modele bayesowskie bardzo dobrze sprawdzają się, gdy w krótkim zbiorze treningowym, w którym występują zmienne jakościowe, wystąpi pozornie całkowita lub całkowita separacja punktów w wielowymiarowej przestrzeni obserwacji [15]. Klasyczne

modele regresyjne tworzone na podstawie takich danych nie są wiarygodne, co wynika z pewnych ograniczeń w metodzie największej wiarygodności stosowanej do estymacji takich modeli. Rozwiązaniem jest wtedy zwiększenie zbioru danych (nie zawsze skuteczne przy specyficznej strukturze zbioru treningowego) albo odpowiednia agregacja wartości wybranych zmiennych jakościowych (co powoduje redukcję informacji dostarczanej do modelu). Pozornie całkowita separacja w danych pojawiła się w danych treningowych dla aspektu czasowego w tej pracy, ale dzięki wykorzystaniu podejścia bayesowskiego do modelowania nie było potrzeby ingerencji w te dane.

Mimo, że trudne koncepcyjnie i wymagające obliczeniowo, modele bayesowskie są coraz powszechniej stosowane w analizach bezpieczeństwa ruchu drogowego. Jak zaprezentowano w pracy, dają duże możliwości w obszarze interpretacji i zakresu wykorzystania danych rzeczywisty do badań. Istnieje potrzeba dalszych prac w celu potwierdzenie uzyskanych wyników oraz poszerzenie możliwości aplikacyjnych dyskutowanej technologii.

Bibliografia

1. Bąk J., Bąk-Gajda D. Psychological factors in road safety. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2007; 17(3): 22–29.

2. El-Basyouny K., Barua S., Islam M. T. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention 2014; 73: 91-99.

3. Gaca S. Badania prędkości pojazdów i jej wpływu na bezpieczeństwo ruchu drogowego (The investigation of vehicle speed and its influence of road traffic safety). In Polish. Zeszyty Naukowe Politechniki Krakowskiej, Inżynieria Lądowa nr 75, Kraków, 2002.

4. Häggström O. Finite Markov Chains and Algorithmic Applications. Cambridge University Press (Virtual Publishing), 2003.

5. Helai H., Chor C.H., Haque M.M. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 2008; 40: 45-54.

6. Heydari S., Miranda-Moreno L.F., Lord D., Fu L. Bayesian methodology to estimate and update safety performance functions under limited data conditions: A sensitivity analysis. Accident Analysis and Prevention 2014; 64: 41-51.

7. Huang H., Abdel-Aty M. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis and Prevention 2010; 42: 1556-1565.

8. Jurecki R., Jaśkiewicz M., Guzek Z., Lozia Z., Zdanowicz P. Driver's reaction time under emergency braking a car – Research in a driving simulator. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2012; 14 (4): 295–301.

9. Jurecki R.S., Stańczyk T.L. Test methods and the reaction time of drivers. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2011; 3 (51): 84–91.

10. Kieć M. Wpływ dostępności do dróg na warunki i bezpieczeństwo ruchu (The influence of road accessibility on road traffic conditionas and road traffic safety – PhD thesis). Rozprawa doktorska na Wydziale Inżynierii Lądowej Politechniki Krakowskiej, Kraków, 2009.

11. Larose D.T. Data Mining Methods and Models. John Wiley & Sons, Inc., 2006.

12. Mitas A.W., Czapla Z., Bugdol M., Ryguła A. Rejestracja i ocena parametrów biometrycznych kierowcy dla poprawy bezpieczeństwa ruchu drogowego. Zeszyty Naukowe Politechniki Śląskiej, seria Transport 2010; 6 (1825): 71-79.

13. Mitra S., Washington S. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 2007; 39: 459-468.

14. Nowakowska M. Logistic models in the crash severity classification on the basis of chosen road characteristics. Transportation Research Record, Journal of the Transportation Research Board, Highway Safety Data, Analysis, and Evaluation, Volume 2, Washington D.C. 2010; 2148: 16-26.

15. Nowakowska M. Modelowanie związków między cechami drogi a zagrożeniami w ruchu na drogach zamiejskich (Modelling the relationship between road features and traffic threats on national roads). In Polish. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa, 2013.

16. Pei X. Wong S.C., Sze N.N. A joint probability approach to crash prediction models. Accident Analysis and Prevention 2011; 43: 1160-1166.

17. Persaud B., Lan B., Lyon C., Bhim R. Comparison of empirical Bayes and full Bayes approach for before-after road safety evaluations. Accident Analysis and Prevention 2010; 42: 38-43.

18. SAS/STAT® 9.2 User's Guide. Introduction to Bayesian Analysis Procedures. Second Edition, SAS Institute Inc., Cary, NC, USA, 2009.

19. Savolainen P.T., Mannering F.L., Lord D., Quddus M.A. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 2011; 43: 1666−1686.

20. Yu R., Abdel-Aty M. Investigation different approaches to develop informative priors in hierarchical Bayesian safety performance functions. Accident Analysis and Prevention 2013; 56: 51-58.

21. Zarządzenie nr 653 Komendanta Głównego Policji z dnia 30 czerwca 2006 r. w sprawie metod i form prowadzenie przez Policję statystyki zdarzeń drogowych (The regulation No 635 by the Main Commanding Officer of the Polish Police Headquarters from the 30-th of June 2006 regarding the methods and the forms of processing road crash statistics by the police). In Polish. Warszawa, 2006.