Article citation info:
NOWAKOWSKA M. Spatial and temporal aspects of prior and likelihood data choices for Bayesian models in road traffic safety analyses. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2017; 19 (1): 68–75, http://dx.doi.org/10.17531/ein.2017.1.10.

Marzena NOWAKOWSKA

SPATIAL AND TEMPORAL ASPECTS OF PRIOR AND LIKELIHOOD DATA CHOICES FOR BAYESIAN MODELS IN ROAD TRAFFIC SAFETY ANALYSES

PRZESTRZENNY I CZASOWY ASPEKT WYBORU ROZKŁADÓW APRIORYCZNYCH I DANYCH DLA FUNKCJI WIARYGODNOŚCI DLA MODELI BAYESOWSKICH W ANALIZACH BEZPIECZEŃSTWA RUCHU DROGOWEGO*

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.

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

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.

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

1. Introduction

Traffic road safety, as an element of a *human–vehicle–road* system, has been the subject of scientific and research works for many years. There are many researchers and specialists in a wide range of fields or disciplines who are involved in the process of recognizing and understanding mechanisms related to a road crash. Many theories and models have been elaborated in order to evaluate the level of road traffic threats, as well as to identify circumstances, and cause and effect relationships of road accidents. The research area is extensive and covers: simulation and behavioural research (e.g. [8, 9]), elaboration of entropy models (e.g. [1, 12]), investigations of road polygons including road surroundings, and traffic and weather conditions (speed in particular) (e.g. [3, 10]), as well as exploration and mining of real road accident data (e.g. [15, 19]).

Statistical methods belong to the most important research techniques utilised in analysing real data. There are two approaches in such an analysis. The first one is a frequentist (also known as classical) approach, in which a random event's probability is assumed to be represented by the frequency of the event occurrence in a very large number of identical samples. The other one is a Bayesian (also known as non-classical) approach, according to which a prior (unconditional) probability of a random event is a measure of a rational belief that the event will occur. Then, the belief is modified using data from experiments or from observations of circumstances connected with the event. Prior knowledge is transformed into posterior

(*) Tekst artykułu w polskiej wersji językowej dostępny w elektronicznym wydaniu kwartalnika na stronie ***www.ein.org.pl*** knowledge, which is a resultant probability and a measure of a rational expectation of the event occurrence after getting information from the data. Bayesian thinking, supported by the development of numerical sampling techniques, has created modern statistics fundamentals, which enables formulating and solving problems not available in classical statistics.

Bayesian regression modelling is a non-classical methodology which becomes widespread in road traffic safety analyses, mainly because it allows eliminating various weaknesses of classical models. Bayesian regression models are difficult from both conceptual and computational points of view. Nevertheless, they bring a new quality to the development of scientific research methods, and they enable a flexible, though non-standard, approach to modelling issues. The models are used in order to develop safety performance functions (e.g. [6, 7, 13, 16]), including a before-after analysis (e.g. [17]), and also to classify descriptive road accident features, such as driver's behaviour, accident type, or accident severity (e.g. [2, 5, 16]).

The non-classical method of statistical inference was used in the study in order to develop logistic regression models, in which road accident severity is a response variable and selected features describing accident circumstances are input variables. A certain methodology of defining two basic sources of information for the Bayesian model was elaborated. The research is directed towards establishing informative priors as a general background for the model, and then towards choosing likelihood data in order to obtain posterior knowledge. Both elements would reflect various aspects of road safety research interests.

2. A Bayesian road accident severity classifier

The subject of the analysis is a statistical classifier – a logistic regression model that classifies road accident severity *AcSrv* into one of two values (categories): *LA* – light accident (assumed to be a failure) and *FSA* – fatal or serious accident (assumed to be a success). Input variables represent the description of a road accident location, at-fault driver's characteristics and accident features.

Logit is a link function in a logistic regression model. Conditional probability P(*AcSrv* = *FSA* | X_1, …, X_k) that an accident which occurred under circumstances described by a set of input variables values is fatal or serious constitutes the argument of the link function:

$$\operatorname{logit}\left(\mathrm{P}\left(AcSvr=FSA|X_1,\dots,X_k\right)\right)$$
$$=\ln\left(\frac{\mathrm{P}\left(AcSvr=FSA|X_1,\dots,X_k\right)}{1-\mathrm{P}\left(AcSvr=FSA|X_1,\dots,X_k\right)}\right)=\beta_0+\beta_1X_1+\dots+\beta_kX_k \quad (1)$$

The assumed model is relatively simple since the main purpose of the research is not to analyse the influence of the chosen features on the response variable, but to discuss the methodology that helps in developing a Bayesian regression model.

Contrary to the classical approach, it is assumed that Bayesian regression model parameters are not constants, but random variables. Therefore, each parameter is described by a certain posterior distribution that results from previous (prior) knowledge about the parameter and from the knowledge update using empirical data (Bayesian likelihood data) [18]:

$$\begin{aligned}&\mathrm{P}(\boldsymbol{\beta}\,|\,Y,\boldsymbol{X})=\mathrm{P}(\beta_0,\dots,\beta_k\,|\,Y,\boldsymbol{X})=\mathrm{P}(\beta_0,\dots,\beta_k)\cdot\mathrm{P}(Y,\boldsymbol{X}\,|\\&\beta_0,\dots,\beta_k)\,/\,\mathrm{P}\left(Y,\boldsymbol{X}\right)\propto\mathrm{P}(\beta_0,\dots,\beta_k)\cdot\mathrm{L}(Y,\boldsymbol{X}\,|\,\beta_0,\dots,\beta_k)\\&=\mathrm{P}(\boldsymbol{\beta})\cdot\mathrm{L}(Y,\boldsymbol{X}\,|\,\boldsymbol{\beta})\end{aligned} \quad (2)$$

The posterior distribution mean of the parameter $\boldsymbol{\beta}_i$ accompanying the variable X_i is the measure used to assess the magnitude and the direction of the variable influence on the response.

According to Bayes' rule, posterior distributions P($\boldsymbol{\beta}$ | Y, $\boldsymbol{X}$) contain information from two sources: prior distributions P($\boldsymbol{\beta}$) and likelihood functions L(Y, $\boldsymbol{X}$ | $\boldsymbol{\beta}$). A variety of posterior distributions for a regression parameter β_i is possible (Fig. 1), which is the consequence of the assumptions made about previous knowledge and likelihood data choices. Whenever one of the sources changes, the posterior changes as well.

Marcov Chains Monte Carlo (MCMC) sampling methodology [4, 18] is used in order to obtain posterior distributions P($\boldsymbol{\beta}$ |Y, $\boldsymbol{X}$). Each distribution is calculated from the series of numbers meeting the Marcov chain criteria. The Mertopolis-Hastings algorithm belongs to the most popular generators of the series. The Gibbs sampler is also frequently used. The results of the MCMC method depend on: the number of iterations in the chain, the number of burn-in values and the thinning rate. Converging the Marcov chain to stationarity is a significant issue in the generation process. It gives rise to an output sample from the stationary posterior distribution. Diagnostic tests (e.g. Gelman-Rubic, Geweke, Heidelberger-Welch), as well as trace diagnostic and correlations plots are used in order to assess the Marcov chain quality.

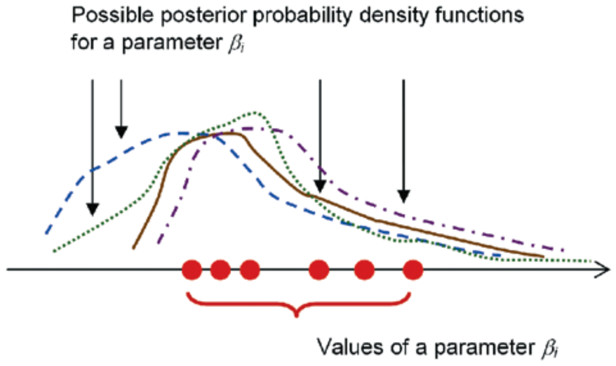

Fig. 1. A graphical interpretation of a Bayesian regression model parameter

3. Building a Bayesian road accident severity classifier

A Bayesian regression classifier (1) is created from a two-step Bayesian modelling procedure in which selected aspects of a road accident data investigation are adopted. The proposed approach and its results are strongly data-dependent: a several-year accident data registration period for a network of the same category roads in a given country region is needed (in particular roads supervised by a specific road administration unit). The data are selected in order to focus on either spatial or temporal aspect of the model estimation. The whole procedure extends and develops the concept presented in the investigation by Yu and Abdel-Aty [20] on the selection of informative priors for Bayesian models of a safety performance function.

The algorithm of building the Bayesian road accident severity classifier is presented hereafter.

Bayesian Modelling Step 1; defining the priors – the BM-S1 model

There are three general types of prior distributions used in Bayesian regression models: non-informative, semi-informative, and informative. The first one is utilized in road traffic safety analyses more often than the others, although it is dominated by likelihood data in the final output, and mean values of Bayesian model parameters are very close to parameter estimators of a classical regression model. Better results can be obtained when, instead of diffuse non-informative prior distributions, well-defined informative prior ones are used, because they reflect knowledge on an investigated subject. In order to generate such distributions, suitable data processing is proposed. It is the first step of the above-mentioned procedure, thanks to which the Bayesian BM-S1 model is obtained.

There are the following sources of information for the BM-S1 model:

- priors – non-informative, normal distributions with zero mean and a very big standard deviation (1E+06),
- Bayesian likelihood (likelihood function) – road accident data selected according to the chosen aspect of the analysis: spatial or temporal one.

The Bayesian likelihood for the BM-S1 model is defined in the following way:

- for the spatial aspect: all accident data registered on the same category roads in a given country region for an assumed period of time,
- for the temporal aspect: all historical accident data registered on the same category roads in a given country region, excluding the data from the latest (most recent) registration period covering the whole season cycle (a calendar year).

Means and standard deviations of posterior distributions obtained for the BM-S1 model become means and standard deviations of prior normal distributions for the parameters of the Bayesian regression model created in the second step.

Bayesian Modelling Step 2; defining the likelihood – the final BM-S2 model

Since normal distributions derived in the first step are not diffuse, they generate informative prior knowledge constituting a basic background (a generalisation) for the final BM-S2 Bayesian model which follows the chosen aspect of the analysis. The likelihood data for the BM-S2 model define a training data set and they are treated as a factor emphasising and clarifying the research context:

- for the spatial aspect: accident data for a given road that are selected from the whole data set modify priors related to the road,
- for the temporal aspect: the latest (most recent) accident data update historical knowledge related to the whole area.

Fatal accident observations are extremely rare in road accident data, which usually results in a weak classification quality of the accident fatality. Therefore, in order to overcome such a negative phenomenon and to strengthen the rare values influence on final modelling results, balancing [1, 14, 15] is applied to the likelihood data in the BM-S2 model forcing smaller differences in the proportions of the values of the response variable *AcSrv*. Firstly, the primary data set is split into three subsets according to the accident severity *AcSrv*: light, serious, and fatal. Then, all fatal accident observations are taken to create a 20% stratum in a new training data set. Next, serious and light accident observations are selected at random from the remaining subsets in order to constitute, in the newly created data set, 30% and 50% strata respectively. Finally, the data modification is carried out so as to receive the binary-valued response variable *AcSrv* which defines a failure by the light accident severity category and a success by combining the serious and fatal accident severity categories. In such a balanced likelihood data set, the fatal accident observations grow considerably and, at the same time, the relatively rare success category does not exceed 50% of the data set size.

The research experiment has been carried out utilising the balancing scheme in each aspect of the data definition for the likelihood function in the BM-S2 model.

4. Data description

The road accident data used in the study, acquired from the SEWiK police database system, were provided by the Police Headquarters of the Świętokrzyskie province, Poland. The accidents registered during the time period from 2008 to 2014 on all of the nine national roads in the province are analysed in the study. The roads are supervised by a national road administration unit (a division of the General Directorate for National Roads and Motorways) because they serve interregional connections.

The observations which meet the following criteria were selected for the research:

- accidents were registered outside towns with civic rights on two-lane single carriageways (national roads have the highest technical parameters among all the roads with such a profile),
- only one adult driver caused the accident (in Poland, adult relates to a person who is at least 18 years old),
- only motor vehicles were involved in the accidents,
- no pedestrians participated in the accidents.

Prior to the analysis, the data were cleaned and the records with outliers, missing or extremely rare values that couldn't be aggregated (considering the physical meaning of the values) were removed. The resultant data set includes 1329 observations and it consists of the following variables chosen for the investigation:

- the group of accident location characteristics (input variables):
    - ◦ *ArTp* – area type with the following values: *Bt* – built-up area (39.2%), *NBt* – non-built-up area (60.8%),
    - ◦ *LgCnd* – road lighting conditions with the following values: *NgDrk* – night darkness, i.e. no lighting at night (16.6%), *PrLg* – poor lighting, e.g. dawn, dusk or artificial lighting (usually poor on non-urban roads) at night (14.7%), *Dlg* – daylight (68.6%),
    - ◦ *RdSrf* – roadway surface conditions with the following values: *NDr* – not dry, i.e. wet, snow-covered or ice-covered (38.5%), *Dr* – dry (61.5%),
- the group of at-fault driver's features (input variables):
    - ◦ *VhTp* – vehicle type with the following values: *HvVh* – heavy vehicle (15.6%), *Mtr* – motorcycle, scooter, moped, i.e. single-track motor vehicle (3.2%), *Cr* – car (81.3%),
    - ◦ *Gndr* – at-fault driver's gender with the following values: *F* – female (12,5%), *M* – male (87,5%),
    - ◦ *AgGrp* – at-fault driver's age group with the following values: 02 – <18; 25) (25.1%), 03 – <25; 35) (27.5%), 04 – <35; 50) (25.9%), 05 – <50; 65) (16.3%), 06 – at least 65 (5.1%),
    - ◦ *Alh* – at-fault driver under the influence of alcohol or other toxic substances with the following values: *N* – no (89.8%), *Y* – yes (10.2%),
- the group of road accident attributes (input variables):
    - ◦ *NrVhIn* – number of vehicles involved with the following values: *Sng* – single vehicle accident (31.2%), *Mlt* – multiple vehicle accident (68.8%),
    - ◦ *Bhv* – at-fault driver's behaviour with the following values: *DrWrSdRd* – driving wrong side of a roadway (5.2%), *InSpPrCn* – inappropriate speed for prevailing traffic and weather conditions (44.2%), *NGvWy* – not giving right of way (10.3%), *InTrUTr* – incorrect turning or U-turning (4.1%), *InPs* – incorrect passing by (1.6%), *InOvBp* – incorrect overtaking or bypassing (12.9%), *PrPsCn* – poor psychophysical condition (8.3%), *FlCl* – following too close (13.5%),
- *AcSvr* – the response variable; accident severity defined by the status of a road crash according to the highest level of injuries experienced by a human casualty as follows [14, 15, 21]: *LA* – light accident (57%), *SA* – serious accident (29.4%), *FA* – fatal accident (13.5%).

5. Results

The Bayesian regression models were obtained from the 10000-element Marcov chains generated using the Metropolis algorithm for the following settings: the number of burn-out samples = 50000, the number of final chain iterations = 300000, the thinning indicator = 30. All the

Marcov chains reached the stationarity, which was verified by the autocorrelation and trace plots, as well as by the Geweke and Heilderberger-Welch tests. The resultant posterior distributions were unimodal.

The research experiments were conducted using the SAS® software: the in-built MCMC procedure and the author's own SAS 4GL and SAS macro language computer programs.

The data were prepared taking into account:

- for the spatial (S) aspect:
  - BM-S1(S): all the national roads in the Świętokrzyskie province, for the time period 2008-2014 (the data set length is equal to 1329 records),
  - BM-S2(S): the DK74 and DK7 roads for two independent models, for the period 2008-2014 (after balancing, the data set length is equal to 220 and 196 for the DK74 and DK7 roads respectively); the main difference between the roads is that the DK7 road, being the part of the European road network, additionally serves international traffic,
- for the temporal (T) aspect:
  - BM-S1(T): all the national roads in the Świętokrzyskie province, for the time period 2008-2013 (the data set length is equal to 1221 records),
  - BM-S2(T): all the national roads in the Świętokrzyskie province, for the year 2014 (after balancing, the data set length is equal to 60 records).

The results of Bayesian modelling for the spatial aspect are presented in Table 1, and for the temporal aspect in Table 2. The BM-S1models obtained in the first step are called prior models since they deliver informative prior knowledge for the second step. The BM-S2 models obtained in the second step are called posterior models because they are the final classifiers of the whole modelling procedure. Both tables have a similar structure:

- mean, and standard deviation values (*Mean (S.D.)*) of parameter distributions for the prior models (*BM-S1 – prior*) and for the posterior models (*BM-S2 – posterior*),
- reference of each posterior model to its corresponding prior model by determining the index that, for any parameter, compares the posterior distribution mean with its corresponding prior distribution mean. The index is calculated by the expression ($mean_{posterior}$ – $mean_{prior}$)/|$mean_{prior}$|. The index values are given in the *Comparison* columns for: *DK74 vs. prior*, *DK7 vs. prior*, and *2014 vs. prior*,
- comparison of two posterior models for the spatial aspect (for the DK74 and DK7 roads) by showing the difference between the dis-

Table 1. Results of Bayesian accident severity classifiers for the spatial aspect

Model	BM-S1(S) – prior	BM-S2(S) – posterior for DK74		BM-S2(S) – posterior for DK7		Posteriors comparison
Specification	Mean (S.D.)	Mean (S.D.)	Comparison: DK74 vs. prior	Mean (S.D.)	Comparison: DK7 vs. prior	DK74–DK7
Constant	−1.396 (0.378)	−1.224 (0.235)	12.3%	−1.192 (0.249)	14.6%	−0.032
The group of accident location characteristics						
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)		
The group of at–fault driver's features						
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)		
The group of road accident attributes						
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		
Sensitivity	38.9%	59.3%		57.9%		
Specificity	82.6%	67.8%		65.3%		
HMSS	52.9%	63.3%		61.4%		

tribution means of the corresponding model parameters calculated by the expression: $(mean_{posterior(DK74)} - mean_{posterior(DK7)})$. The difference values are given in the *Posterior comparison* column in Table 1,

- Deviance Information Criterion (*DIC*) measure calculated from the training data sets: the unbalanced one for the BM-S1 model and the balanced one for the BM-S2 model,
- classification quality assessment measures: sensitivity (the percentage of correctly classified *FSA* cases), specificity (the percentage of correctly classified *LA* cases), and the harmonic mean of sensitivity and specificity *HMSS* (which balances the two measures). All the indices were calculated from the primary likelihood data set for the BM-S1 model and from the primary (nor balanced) likelihood data set for the BM-S2 model.

For each parameter of a Bayesian model, the highest probability density HPD interval can be constructed unambiguously, provided that the parameter distribution is not uniform. To some extent, the HPD interval corresponds to a credible interval in classical statistics – if it contains zero, values of its parameter cannot be clearly interpreted. The uncertainty is also indicated when the absolute value of the parameter coefficient of variation exceeds 50%. Such statistically insignificant parameters are highlighted in red in Tables 1 and 2. The HPD intervals for the statistically significant parameters of the final models (the BM-S2 models obtained in the second step) are illustrated in Figures 2 and 3.

In Tables 1 and 2, and in Figures 2 and 3, all the input variables are grouped according to their substantial meaning, i.e. accident location characteristics, at-fault driver's features, and accident features.

Bayesian models for the spatial aspect

1. The sets of statistically significant input variables are roughly the same in the BM-S1(S), as well as in both BM-S2(S) models. The driver's age group proved significant in the BM-S2(S) model for the DK74 road only due to the significance of the coded variable *AgGrp_05* (50-65 years old).
2. The directions of the influence of the individual statistically significant variables on the accident severity are the same in the BM-S1(S) model and in both BM-S2(S) models.
3. The nature (magnitude and direction) of the change in the values of the statistically significant posterior parameters (the BM-S2(S) models) in relation to the values of the corresponding prior parameters (the BM-S1(S) model) is road-dependent:
   - the positive influence of the accident location characteristics on the accident severity is greater by more than 20% in the BM-S2(S) model for the DK74 road, whereas the change of the influence in the BM-S2(S) model for the DK7 road is different – there is a rise by 5% in the parameter mean for built-up area *ArTp_Bt* and a drop by 6% in the parameter mean for night darkness *LgCnd_NgDrk*,
   - the positive influence of single-track motor vehicle (motorcycle, scooter, and moped) *VhTp_Mtr* and the negative influence of female driver's gender *Gndr_F* on the accident severity identified in the prior parameter distributions become smaller by nearly 10% in the posterior distributions for the DK74 road, whereas they remain at almost the same level for the DK7 road,
   - the modification of the parameter prior distribution for the single vehicle accident variable *NrVhIn_Sng* by using the likelihood data taken from different roads caused different results in the posterior distributions: the parameter mean value rose by 7% for the DK74 road and dropped by 20% for the DK7 road,
   - the range of the change in the parameter posterior distributions for driver's behaviour is different for the DK74 and DK7 roads, which is particularly evident for not giving right of way *Bhv_NGvWy* (a drop in the mean value by 6.8% and a rise by 11.7% respectively), for incorrect turning or U-turning *Bhv_InTrUTr* (almost without a change and a drop by 9.1% respectively), and for poor psychophysical condition *Bhv_PrPsCn* (a rise by 5.2% and a drop by 7.7% respectively).

Bayesian models for the temporal aspect

Table 2. Results of Bayesian accident severity classifiers for the temporal aspect

Model	BM-S1(T) – prior	BM-S2(T) – posterior for 2014	
Specification	Mean (S.D.)	Mean (S.D.)	Comparison: 2014 vs. prior
Constant	−1.192 (0.393)	−1.169 (0.289)	1.9%
The group of accident location characteristics			
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)	
The group of at−fault driver's features			
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)	
The group of road accident attributes			
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	
Sensitivity	36.9%	61.9%	
Specificity	82.8%	74.2%	
HMSS	51.0%	67.5%	

1. The sets of statistically significant input variables in the BM-S1(T) and BM-S2(T) models differ in two variables: (1) night lighting condition *LgCnd_NgDrk* is insignificant in the BM-S1(T) model,

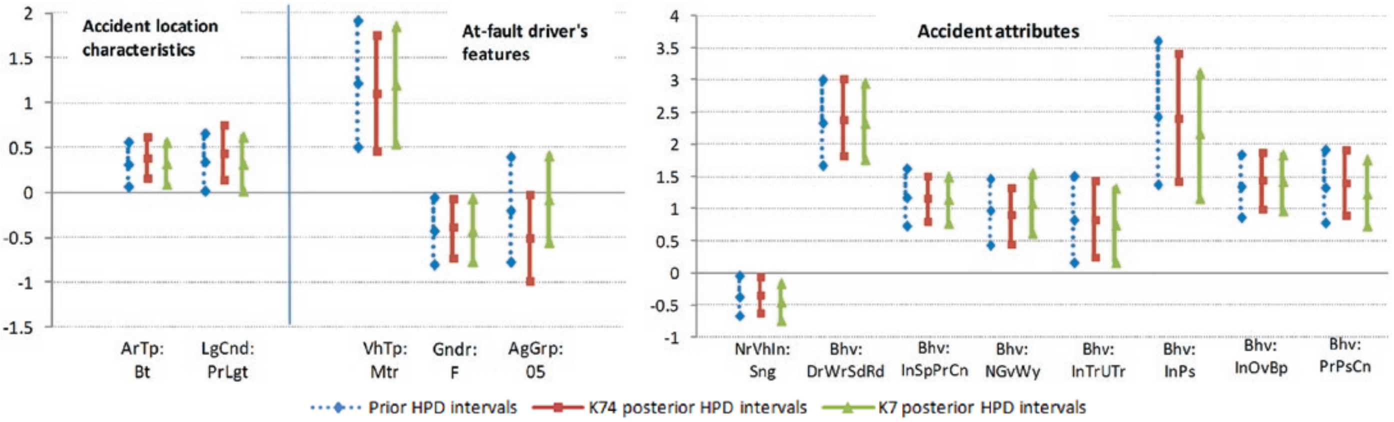

Fig. 2. HPD intervals for statistically significant parameters of Bayesian models for the spatial aspect

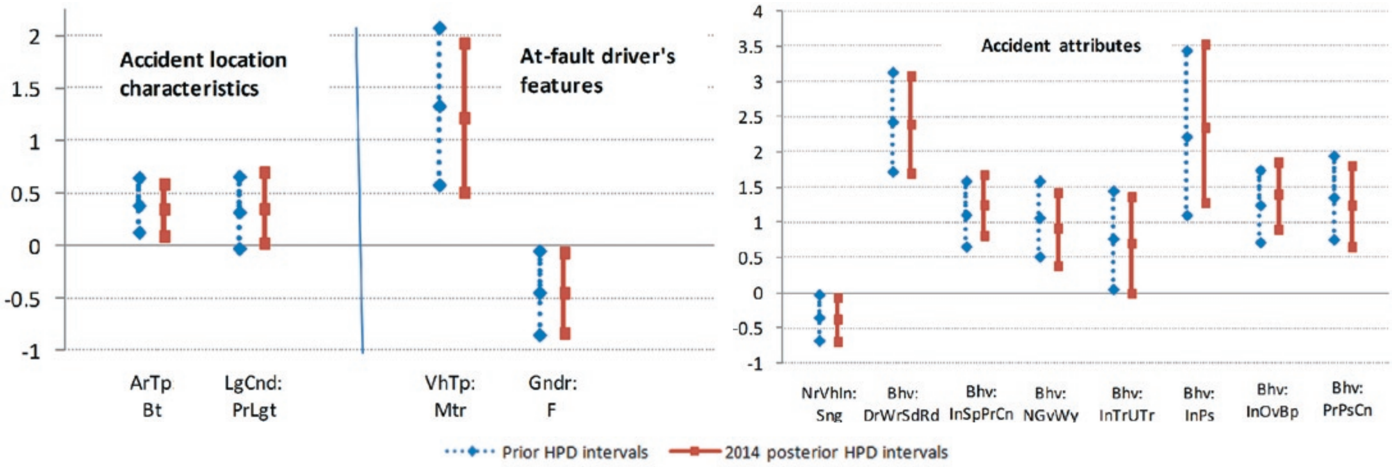

Fig. 3. HPD intervals for statistically significant parameters of Bayesian models for the temporal aspect

but significant in the BM-S2(T) model, (2) incorrect turning or U-turning *Bhv_InTrUTr* is significant in the BM-S1(T) model, but insignificant in the BM-S2(T) model.

2. Similarly to the spatial models, the influence directions of the corresponding statistically significant variables are the same in the BM-S1(T) model in the first modeling step and in the BM-S2(T) model in the second modeling step.
3. The latest information modified the up-till-now (prior) knowledge about the importance of the individual input variables in the posterior model, and in particular it caused strengthening the following:
   - the positive influence on the fatal or serious accident status of the factors: night lighting condition *LgCnd_NgDrk* (increase by 10.6%), inappropriate speed for the prevailing traffic and weather conditions *Bhv_InSpPrCn* (an increase by 12.9%), incorrect overtaking or bypassing *Bhv_InOvBp* (an increase by 12.5%),
   - the negative influence on the fatal or serious accident status of the single-vehicle accident variable *NrVhIn_Sng* (a decrease by 5.9%).

Balancing the likelihood data in the second modelling step, both in spatial and temporal aspects, improves the classification quality of all the final Bayesian models. The values of the quality assessment measures are satisfactory:

- sensitivity is greater than 57%,
- specificity is greater than 65%,
- the HMSS coefficient is greater than 61%.

A general picture of the coefficients of variation for the statistically significant parameters of the models is presented in Fig. 4 in the form of a bubble plot, where the centres represent mean values and the radii are standard deviations of the coefficients. The standard deviation values are similar, irrespective of the step (prior or posterior models) and the aspect (spatial or temporal) of modelling. A slightly greater difference can be noticed for the mean values – they are smaller for the parameters of the second step models, which indicates the better estimation precision of the final posterior models.

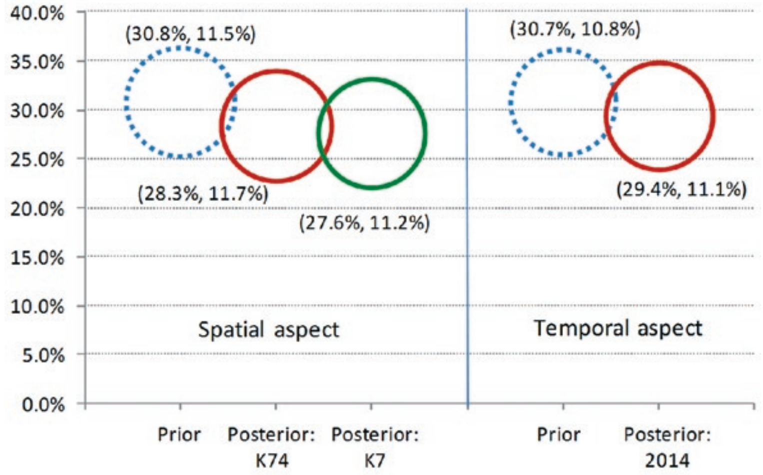

Fig. 4. Bubble plots of mean and standard deviation values of the coefficient of variation for statistically significant parameters of Bayesian models

6. Conclusions

Parameters are random variables in Bayesian regression models. Their so-called posterior distributions are obtained by combining systematic (prior) knowledge about the parameters with Bayesian likelihood – the knowledge derived from data. Some issues concerning the methodology of such models development for road traffic safety analyses is presented in the study. A logistic regression model that classifies road accident severity is analysed.

Road accident data are treated as a potential source of both information types for the Bayesian model: prior knowledge and Bayesian likelihood. Some researchers apply such an approach in their road safety investigations. In the study, however, a specific interpretation of both sources has been proposed and consequently their special application in the modelling process in which an additional task to obtain the best possible final classifiers was considered as well.

Prior knowledge about regression parameters can be obtained from data the range of which depends on the subject of a research. If the investigation focuses on the spatial aspect, all accident data recorded on the same technical class roads in a given country region are a possible source of informative priors, creating a reference background for being updated by Bayesian likelihood originating from accident data recorded on a chosen road. Thus, a model related to the road is obtained. If the investigation focuses on the temporal aspect, historical road accident data create informative prior background, and new accident data from the latest registration period update the priors, providing a new general picture of the region road network safety.

Balancing likelihood data, in both spatial and temporal aspects, positively affects the classification quality of the final Bayesian logistic regression models. The result is particularly important since the level of correct classification of rare success categories, i.e. serious or fatal road accident severity, is crucial.

Bayesian regression models work well when a quasi-complete or complete separation of data points appears [15] in a short data set with qualitative input variables. Classical regression models estimated on the basis of such data are not credible due to some constrains of the maximum likelihood method used in the estimation process. To solve the problem, enlarging the data set (not always efficient for some specific data structures) or a suitable aggregation of categories within chosen qualitative variables (which causes the reduction of information delivered to the model) is recommended. In the research, the quasi-complete separation was detected in the training data set for the time aspect. However, no interference into the data was necessary owing to the Bayesian approach to the modelling tasks.

Notwithstanding their complex nature, Bayesian models become more and more widely used in road traffic safety analyses. As it was shown in the study, they can provide great possibilities in interpreting and utilizing real data. Further studies are recommended to confirm the obtained findings and to widen possible implementations of the discussed technologies.

References

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, http://dx.doi.org/10.1016/j.aap.2014.08.014.
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, http://dx.doi.org/10.1016/j.aap.2007.04.002.
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, http://dx.doi.org/10.1016/j.aap.2013.11.001.
7. Huang H., Abdel-Aty M. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis and Prevention 2010; 42: 1556-1565, http://dx.doi.org/10.1016/j.aap.2010.03.013.
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, http://dx.doi.org/10.1016/j.aap.2006.08.002.
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, http://dx.doi.org/10.1016/j.aap.2010.12.026.
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, http://dx.doi.org/10.1016/j.aap.2009.06.028.
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, http://dx.doi.org/10.1016/j.aap.2011.03.025.
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, http://dx.doi.org/10.1016/j.aap.2013.03.023.
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.

Marzena NOWAKOWSKA
Faculty of Management and Computer Modelling
Kielce University of Technology
Al. Tysiąclecia Państwa Polskiego 7, 25-314 Kielce, Poland
E-mail: spimn@tu.kielce.pl