Pesquisa Operacional (2020) 40: e230462 p.1-20
doi: 10.1590/0101-7438.2020.040.00230462
© 2020 Brazilian Operations Research Society
Printed version ISSN 0101-7438 / Online version ISSN 1678-5142
www.scielo.br/pope
ARTICLES

COMPOSITIONAL STATISTICAL MODELS UNDER A BAYESIAN APPROACH: AN APPLICATION TO TRAFFIC ACCIDENT DATA IN FEDERAL HIGHWAYS IN BRAZIL

Ricardo Puziol de Oliveira[1*] and Jorge Alberto Achcar[2]

Received October 28, 2019 / Accepted August 19, 2020

ABSTRACT. This study considers the use of a composicional statistical model under a Bayesian approach using Markov Chain Monte Carlo simulation methods applied for road traffic victims ocurring in federal roads of Brazil in a specified period of time. The main motivation of the present study is based on a database with information on the injury severity of each person involved in an accident occurred in federal highways in Brazil during a time period ranging from January, 2018 to April, 2019 reported by the federal highway police office of Brazil. Four types of events associated with each injured person (uninjured, minor injury, serious injury and death) are grouped for each state of Brazil in each month characterizing compositional multivariate data. Such kind of data requires specific modeling and inference approaches that differ from the traditional use of multivariate models assuming multivariate normal distributions.The proportion events associated to the accidents (uninjured, minor injuries, serious injuries and deaths) are considered as a sample of vectors of proportions adding to a value one together with some covariates such as pavement conditions in each province, regions of Brazil, months and years that may affect the severity of the injury of each person involved in an accident. From the obtained results, it is observed that the proportions of serious accidents and deaths are affected by some covariates as the different regions of Brazil and years.

Keywords: accident victims, types of injuries, deaths, federal highways, compositional data, Bayesian approach.

1 INTRODUCTION

A major world public health problem is related to traffic accidents where the death toll reached 1.35 million in 2016. With the fast increase of vehicles in circulation and the lack of monitoring infrastructure especially in third world countries the situation tends to get worse. According to a report from the World Health Organization (World Health Organization et al., 2018) as progress is made in the prevention and control of infectious diseases, the number of deaths from non-communicable diseases and injuries has increased significantly in recent years.

*Corresponding author

[1] Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes, 3900, Vila Monte Alegre, Ribeirão Preto, SP, Brasil – E-mail: rpuziol.oliveira@gmail.com – http://orcid.org/0000-0001-6134-5975

[2] Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Av. Bandeirantes, 3900, Vila Monte Alegre, Ribeirão Preto, SP, Brasil – E-mail: achcar@fmrp.usp.br – http://orcid.org/0000-0002-9868-9453

Traffic is already responsible for the eighth cause of death in all age groups, where traffic injuries are currently the leading cause of death for children and young adults aged 5 to 29 years. An improvement in traffic deaths reduction has already been observed in more developed countries, but the situation is catastrophic in most emerging and poor countries. There is a strong association between the risk of death in traffic and the income level of the countries. With an average rate of 27.5 deaths per 100,000 inhabitants, the risk of death in traffic is three times higher in low-income countries than in high-income countries, where the average rate is 8.3 deaths per 100,000 inhabitants. In addition, the number of road traffic fatalities is disproportionately high among low and middle-income countries relative to the size of their populations and the number of motor vehicles in circulation compared to the rest of the world (see Table 1).

Table 1 – Proportion of population, traffic deaths and number of registered vehicles by country in 2016 (income based on World Bank classification in 2017).

	High Income	**Average Income**	**Low Income**
% of population	15%	76%	9%
% traffic deaths	7%	80%	13%
% registered vehicles	40%	59%	1%

In many emerging countries, including Brazil, this problem gets worse by a number of factors, including low educational attainment and severe infrastructure problems on highways and urban roads (see for example, World Health Organization, 2018; Bhalla et al., 2014; Waiselfisz, 2013; Bahadorimonfared et al., 2013; Bacchieri & Barros, 2011; Jorge et al., 2009; Marín-León et al., 2012; Andrade & Mello-Jorge, 2016; Marín & Queiroz, 2000; Lyons et al., 2008). In Brazil, the high numbers of accident injuries especially with serious injuries has been a challenge for the single health system (SUS) (Malta et al., 2012; Jorge et al., 2008; Silva & Andrade, 1996; Klein, 1994; Jorge et al., 1994; Haagsma et al., 2016). It is also observed that the number of deaths at the crash site on Brazilian highways is very large compared to other emerging countries and first world countries. Many studies related to road improvement under an operational research approach are presented in the literature (see for example, Martínez et al., 2017; Castro Aragón & Leal, 2003; Novaes, 2001) but not so many related to traffic accidents. Among these studies related to road accidents under an operational research approach we could quote Baykal-Gürsoy et al. (2009); Szwed et al. (2006); Haastrup (1994); Assimizele et al. (2020); Mekker et al. (2018).

Traffic accident rates with deaths in Brazil are only surpassed by India, China, the United States and Russia (World Health Organization, 2018) where between 1980 and 2011 nearly one million people died from traffic accidents in the country, despite new laws being introduced and implemented in 1998 (Brazilian Traffic Code or CTB) establishing conduct rules, infractions and penalties for drivers and in 2008 with some changes to CTB establishing stricter penalties for drunk drivers (Abreu et al., 2018).

It is important to point out that road transport in Brazil is the country's main logistics system with a network of 1,720,700 kilometers (Boletim Estatistico do CNT, 2018) of national roads and highways (the fourth largest in the world, CIA World Factbook, Brazil), where 61.1% of all cargo handled in Brazil circulates (Boletim Estatistico do CNT, 2018). This highway system, often containing old highways, with poorly drawn roads, simple and poorly signposted roads, is the main means of transporting cargo and passengers in the country's traffic. This kind of transport system has been used since the beginning of the republic, when governments began to prioritize road transport over rail and river transport. Under the epidemiological classification, traffic accidents have been a highlight in external causes of mortality (ICD-10 codes WHO V01 to Y98, 1993), where in the period from 1977 to 1986 the traffic accident mortality rate in Brazil went from 16 to 22/100 thousand leading to a 38% increase (Barros et al., 2003).

2 METHODOLOGY

This study considered a database related to the victims of road accidents (victims of land transport accidents ICD-10 headings V01 to V89, World Health Organization, 2004) reported by the federal police (PF) of Brazil regarding all federal highways in the period ranging from January 1, 2018 to April 30, 2019 covering all states of the federation (https://www.prf.gov.br/portal/dados-abertos/acidents) where the federal police reported for each victim the type of injury (unharmed, minor injury, serious injury and death) and some important factors such as cause of the accident, type of accident, phase of the day, weather condition, type of track, road layout, age of the victim, gender of the victim and type of vehicle. This information is described in the accident reports prepared by the road police officers for each road accident. In this paper the data are grouped in the form of monthly compositional data (observed proportion of uninjured, lightly injured, severely injured and injured who died at the accident site) for each federative unit in Brazil. The data set is presented in Table A1 in an appendix at the end of the manuscript. Table 2 shows the total of casualties in each class (unharmed, mild, severe, death) from January 1, 2018 to April 30, 2019 for all units of the federation. Figure 1 shows the box-plots of each class (unharmed, mild injury, severe injury and death) considering all federative units of the Brazil federation. Figure 2 shows the time series for the proportions %unharmed, %mild, %severe and %death. Figure 3 presents time series plots of the proportions observed for the 27 federative units in Brazil.

From the box-plots of Figure 1, it is possible to see that some provinces as São Paulo state (SP) presents greater proportion of unharmed victims of the road accidents while other states as Minas Gerais (MG) presents smaller proportion of unharmed victims when compared to other federative units of Brazil. Also it is observed that the proportion of injury severity is smaller for São Paulo (SP) state in comparison to the other federative units of Brazil while for some northeast federative units as Alagoas (AL), Maranhão (MA), Sergipe (SE) and Rio Grande do Norte (RN) there are large proportions of injury severity in comparison to the other federative units of Brazil.

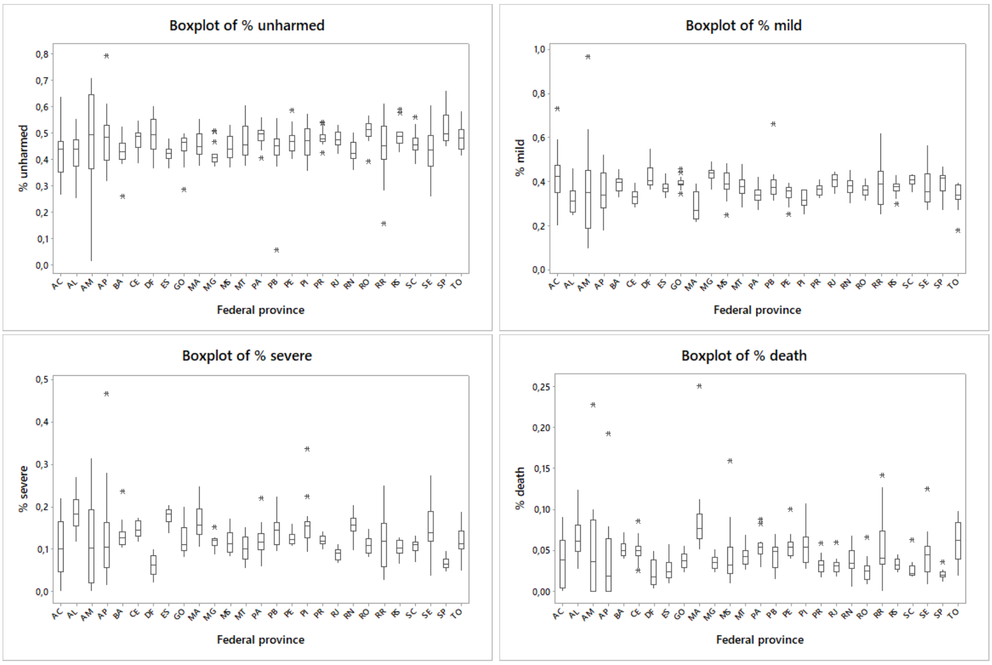

Figure 1 – Box plots for the proportions (unharmed, mild, severe, death) by each federative unit.

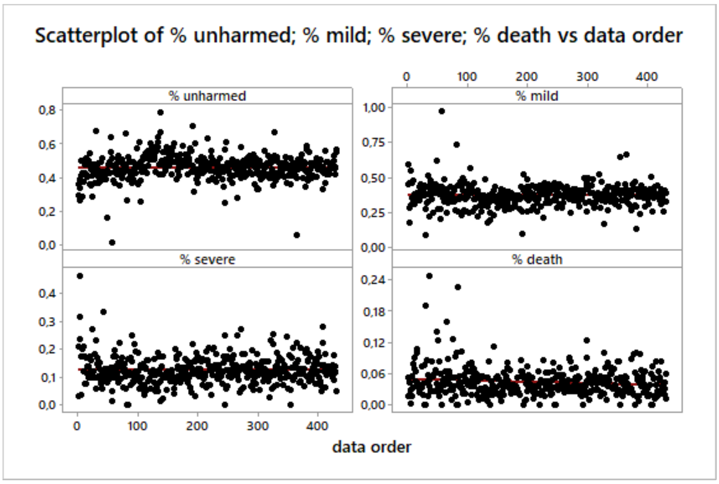

Figure 2 – Time series for %unharmed, %mild, %severe, death).

2.1 Modeling of Compositional Data

Compositional data are vectors of proportions specifying G fractions of a total. Denoting $x = (x_1, x_2, \ldots, x_G)$ to be a compositional vector, we must have $x_i > 0$, for $i = 1, \ldots, G$ and $x_1 + x_2 + \ldots + x_G = 1$. Compositional data often result when raw data is normalized or when data is obtained as proportions of a certain heterogeneous amount. These conditions are usual in geol-

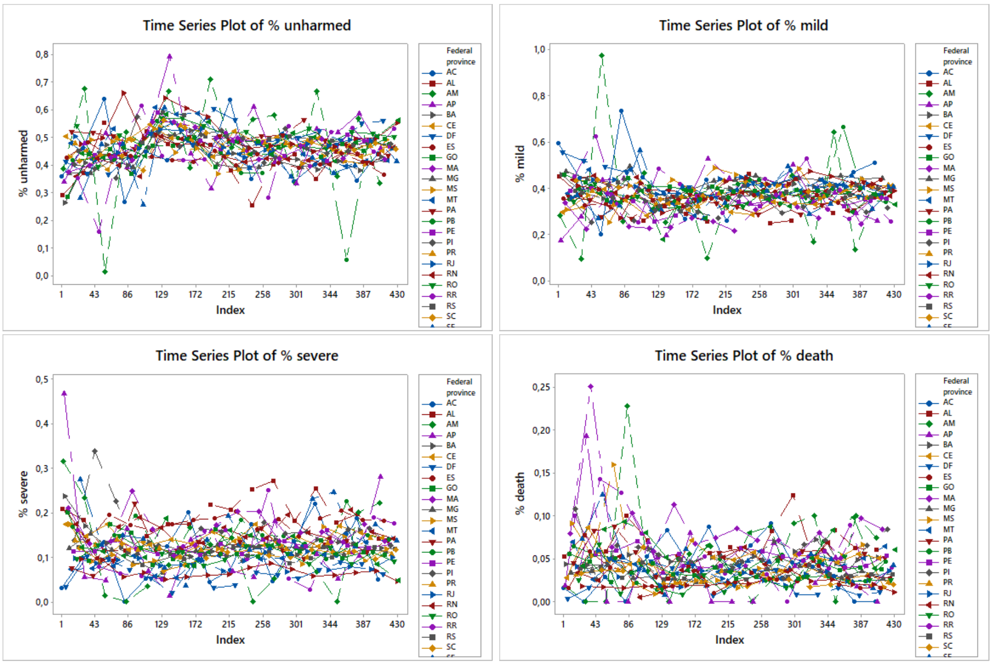

Figure 3 – Compositional proportions (unharmed, mild, severe, death) by federative unit.

ogy, economy and biology. Standard existing methods for analyzing multivariate data under the usual assumption of normal multivariate distribution (see, for example, Johnson et al., 2002) are not appropriate to analyze compositional data, since we have compositional constraints. Different modeling approaches are considered to analyze compositional data. A first model considered to analyze compositional data was based on the Dirichlet distribution, but this model requires that the correlation structure should be totally negative, an unobserved fact for compositional data where some correlations are positive (see, for example, Aitchison, 1982; Atchison & Shen, 1980).

Atchison & Shen (1980) introduced the lognormal distribution to analyze compositional data, transforming the vector of G components **x** into a vector **y** defined in the real coordinate space R_{G_1} considering an additive ratio log (ALR) function. Rayens & Srinivasan (1991) extended the ALR transformation considering Box & Cox (1964) transformations as a generalization of the log-ratio function. Another possibility is to consider the isometric log-ratio (ILR) transformation (Egozcue et al., 2003; Martín Fernández et al., 2015), but the inverse transformation to get the proportions in each class are more complex in the computational work and the obtained results are very similar to the obtained results assuming the ALR transformation (see for example, Martinez et al., 2020). Usually we have great difficulty to get classical inference results for these models, especially in the presence of a covariate vector. Alternatively, the use of Bayesian methods (see, for example, Gelman et al., 2013) is a good alternative to analyze compositional data (see, for example, Iyengar & Dey, 1996, 1998; Tjelmeland & Lund, 2003; Shimizu et al., 2015), especially

Table 2 – Total count of ocurrences in each class (unharmed, mild, severe, death) from January 1, 2018 to April 30, 2019 for all federation units (FU).

FU	Unharmed	Mild	Severe	Death
AC	422	416	88	30
AL	1335	1017	528	201
AM	261	214	63	36
AP	429	290	132	38
BA	7363	7160	2660	890
CE	4255	3031	1367	444
DF	2027	2058	327	114
ES	4439	3813	1845	309
GO	7493	7173	2410	671
MA	2553	1686	966	553
MG	17461	18713	5013	1509
MS	4052	3171	984	572
MT	5038	4184	997	473
PA	2444	1651	568	285
PB	3494	3064	1183	373
PE	5580	4342	1498	741
PI	2718	1865	976	354
PR	17438	13128	4557	1302
RJ	9050	7743	1716	641
RN	2339	2204	878	192
RO	4277	3117	989	254
RR	431	453	120	72
RS	10871	8786	2338	792
SC	16326	14863	4004	1133
SE	1214	1080	454	174
SP	11266	8046	1469	398
TO	1331	960	325	176

considering Markov Chain Monte Carlo (MCMC) methods (see, for example, Gelfand & Smith, 1990; Smith & Roberts, 1993).

Thus, the compositional data introduced in Table A.1 are denoted by x_{1i} = % unharmed, x_{2i} = % mild injuries, x_{3i} = % severe injuries and x_{4i} = % deaths. Let us assume a model with additive ratio log (ALR) transformation given by $y_{1i} = log(x_{2i}/x_{1i})$, $y_{2i} = log(x_{3i}/x_{1i})$ and $y_{3i} = log(x_{4i}/x_{1i})$ given by,

$$\begin{aligned} y_{1i} &= g(\boldsymbol{\beta}_1, \mathbf{z}_i) + w_i + \varepsilon_{ji} \\ y_{2i} &= g(\boldsymbol{\beta}_2, \mathbf{z}_i) + w_i + \varepsilon_{ji} \\ y_{3i} &= g(\boldsymbol{\beta}_3, \mathbf{z}_i) + w_i + \varepsilon_{ji} \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are vectors of regression parameters, $\mathbf{z}_i$ is a covariate vector associated to the i^{th} observation $i = 1, 2, \ldots, 432$, w_i is a random effect (latent unobserved variable) that captures the dependency between the proportions for each province/month and ε_{ji} are errors (non-observed variables) assumed to be independent random variables with normal distributions $N(0, \sigma_j^2)$. Different distributions could be assumed for the random effects w_i; in study, it is assumed a normal distribution $N(0, \sigma_w^2)$.

For a hierarchical Bayesian analysis of the model, it is assumed normal prior distributions for the regression parameters with known hyperparameter values. For the second stage of the hierarchical Bayesian analysis, it is assumed a gamma prior distribution for the inverse of the variance σ_w^2 of the latent variable w_i, that is,

$$\tau_w \sim G(a_w, b_w) \tag{2}$$

where $G(a,b)$ denotes a gamma distribution with mean a/b and variance a/b_2; $\tau_j = 1/\sigma_w^2$; a_w and b_w are known hyperparameters. Further, it is assumed prior independence among the parameters.

Posterior summaries of interest for model (1) are obtained using simulated samples of the joint posterior distribution for the model parameters using MCMC methods. The simulation algorithm to generate samples of the joint posterior distribution for the model parameters is obtained from the complete conditional posterior distributions for each parameter required in the MCMC simulation algorithm. A great simplification in the simulation procedure is to use some existing Bayesian simulation software. One such software is the Openbugs software (see, for example, Lunn et al., 2009), where it is only needed to specify the joint distribution for the observations and the prior distributions for the parameters of the assumed model.

Associated with the compositional data, there are some covariates such as month, year and region of Brazil where the accident occurred. In addition to these covariates, other independent variables of interest may also be associated with the compositional responses, such as road condition, road layout, weather condition, and accident time. An important covariate in the occurrence of road accidents is given by the condition of the pavement. Table 3 presents road pavement conditions considering samples of a few kilometers of highways in each federal unit of Brazil presented in the site related to the year 2018 “CNT 2018 Highways Survey”.

For the analysis of the compositional data given in Table A.1, it is assumed the following covariates: month, year, percentage of pavement good, fair, bad, very bad (the optimum percentage is not considered due to restriction %optimum + %good + regular% +%bad + %very bad = 1) and the dummy variables related to the northeast (1 for NE and 0 otherwise), midwest (1 for CO and

0 otherwise), southeast (1 for SE and 0 otherwise) and south (1 for S and 0 otherwise) regions where the northern region (N) is considered as a reference.

In the data analysis, it is first assumed a regression model with compositional data (1) not considering the presence of the latent factor W denoted as "model 1", that is, assuming independence among the responses in the additive log-ratio (ALR) transformation $y_{1i} = log(x_{2i}/x_{1i})$, $y_{2i} = log(x_{3i}/x_{1i})$ and $y_{3i} = log(x_{4i}/x_{1i})$ where x_{1i} = % unharmed, x_{2i} = % minor injuries, x_{3i} = % severe injuries and x_{4i} = % deaths. Thus, it is assumed the linear regression models:

$$\begin{aligned} y_{1i} &= g(\boldsymbol{\beta}_1, \mathbf{z}_i) + w_i + \varepsilon_{ji} \\ y_{2i} &= g(\boldsymbol{\beta}_2, \mathbf{z}_i) + w_i + \varepsilon_{ji} \\ y_{3i} &= g(\boldsymbol{\beta}_3, \mathbf{z}_i) + w_i + \varepsilon_{ji} \end{aligned} \tag{3}$$

where,

$$\begin{aligned} g(\boldsymbol{\beta}_1, \mathbf{z}_i) &= \beta_{11} + \beta_{12} month_i + \beta_{13} year_i + \beta_{14} \%good.pav_i + \beta_{15} \%regular.pav_i \\ &+ \beta_{16} \%bad.pav_i + \beta_{17} \%lousy.pav_i + \beta_{18} region.NE_i + \beta_{19} region.CO_i \\ &+ \beta_{110} region.SE_i + \beta_{111} region.S_i, \\ g(\boldsymbol{\beta}_2, \mathbf{z}_i) &= \beta_{21} + \beta_{22} month_i + \beta_{23} year_i + \beta_{24} \%good.pav_i + \beta_{25} \%regular.pav_i \\ &+ \beta_{26} \%bad.pav_i + \beta_{27} \%lousy.pav_i + \beta_{28} region.NE_i + \beta_{29} region.CO_i \\ &+ \beta_{210} region.SE_i + \beta_{211} region.S_i, \\ g(\boldsymbol{\beta}_3, \mathbf{z}_i) &= \beta_{31} + \beta_{32} month_i + \beta_{33} year_i + \beta_{34} \%good.pav_i + \beta_{35} \%regular.pav_i \\ &+ \beta_{36} \%bad.pav_i + \beta_{37} \%lousy.pav_i + \beta_{38} region.NE_i + \beta_{39} region.CO_i \\ &+ \beta_{310} region.SE_i + \beta_{311} region.S_i \end{aligned} \tag{4}$$

and ε_{ji} are independent assumed errors with normal distributions $N(0, \sigma_j^2), j = 1,2,3$.

From the ALR transformations assuming the real proportions p_{1i}, p_{2i}, p_{3i} and p_{4i} where, $p_{1i} + p_{2i} + p_{3i} + p_{4i} = 1$, we have, $y_{1i} = log(x_{2i}/x_{1i})$, $y_{2i} = log(x_{3i}/x_{1i})$ and $y_{3i} = log(x_{4i}/x_{1i})$, and the inverse estimated proportions in each class are easily obtained from the expressions,

$$\begin{aligned} \widehat{p_{1i}} &= 1/[1 + exp(\widehat{y_{1i}}) + exp(\widehat{y_{2i}}) + exp(\widehat{y_{3i}})], \\ \widehat{p_{2i}} &= exp(\widehat{y_{1i}})/[1 + exp(\widehat{y_{1i}}) + exp(\widehat{y_{2i}}) + exp(\widehat{y_{3i}})], \\ \widehat{p_{3i}} &= exp(\widehat{y_{2i}})/[1 + exp(\widehat{y_{1i}}) + exp(\widehat{y_{2i}}) + exp(\widehat{y_{3i}})], \\ \widehat{p_{4i}} &= exp(\widehat{y_{3i}})/[1 + exp(\widehat{y_{1i}}) + exp(\widehat{y_{2i}}) + exp(\widehat{y_{3i}})] \end{aligned} \tag{5}$$

where $\widehat{y_{1i}}, \widehat{y_{2i}}, \widehat{y_{3i}}$ and $\widehat{y_{4i}}$ are predicted values based on the estimated model.

Assuming normal independent prior distributions N(0,1) for all regression parameters and gamma distributions G(1,1) for the variances of the errors ε_{1i}, ε_{2i} and ε_{3i}, Table 4 shows the posterior summaries of interest (Monte Carlo estimators given by the posterior parameter means, posterior standard deviations of the parameters and 95% credibility intervals) based on 1000 simulated Gibbs samples (every 100th simulated sample among 100,000 generated Gibbs samples

Table 3 – Condition of the pavement – total length evaluated.

FU	Optimum	Good	Regular	Bad	Very Bad	Total
AC	0	10	480	192	651	1333
AL	635	62	71	20	0	788
AM	10	0	429	282	363	1084
AP	127	10	306	60	0	503
BA	4218	849	3021	576	278	8942
CE	1127	423	1396	493	142	3581
DF	214	24	153	51	30	472
ES	908	100	611	88	24	1731
GO	2434	257	3575	580	617	7463
MA	1841	355	1540	362	579	4677
MG	5346	1441	5922	2322	205	15236
MS	2257	182	1711	190	70	4410
MT	1725	309	2063	633	80	4810
PA	1226	242	1978	235	222	3903
PB	842	98	466	240	62	1708
PE	1704	161	869	374	56	3164
PI	1572	161	1542	36	79	3390
PR	2815	163	2718	516	118	6330
RJ	1486	198	486	313	71	2554
RN	561	133	807	252	103	1856
RO	574	133	532	491	155	1885
RR	657	0	391	44	10	1102
RS	3776	814	3449	724	92	8855
SC	1277	276	1104	486	91	3234
SE	320	15	54	165	94	648
SP	6851	849	1836	419	28	9983
TO	708	50	2154	61	546	3519

to get an approximately uncorrelated sample) of the joint posterior distribution for all model parameters obtained using the Openbugs software and considering a burn-in sample of size 11,000 discarded to eliminate the effect of the initial parameter values needed for the MCMC algorithm. Convergence of the MCMC simulated samples was monitored by traceplots of the generated Gibbs samples (see Gelman et al., 2013)

From the results presented in Table 4, it is observed that the significative effects (zero not included in the 95% credibility intervals) are:

- Response $y_2 = log(x_3/x_1)$ where x_1 = % unharmed and x_3 = % serious injury: poor pavement (regression parameter β_{27} is estimated by a negative value) and NE (northeast) region where regression parameter β_{28} is estimated by a positive value indicating that the difference between x_3 = % of serious injuries and x_1 = % unharmed increases in the NE region when compared to the N region (north considered as reference).

- Response $y_3 = log(x_4/x_1)$ where x_1 = % unharmed and x_4 = % deaths: covariate year (regression parameter β_{33} is estimated by a negative value indicating a decrease in the death/unharmed difference in the year 2019); NE (northeastern) region where the regression parameter β_{38} is estimated by a positive value indicating that the difference between x_3 = % deaths and x_1 = % unharmed increases in the NE region as compared to the N region (north considered as reference); CO region (midwest) where the regression parameter β_{39} is estimated by a positive value indicating that the difference between x_3 = % deaths and x_1 = % unharmed increases in the CO region when compared to the N region (north considered as reference); SE region (southeast) where the regression parameter β_{310} is estimated by a positive value indicating that the difference between x_3 = % deaths and x_1 = % unharmed increases in the SE region when compared to the N region (north considered as reference); and region S (south) where the regression parameter β_{311} is also estimated by a positive value indicating that the difference between x_3 = % deaths and x_1 = % unharmed increases in region S when compared with region N (north considered as reference).

Now assuming a regression model with compositional data defined by (1) and (4) in the presence of the latent factor W denoted by “model 2”, that is, assuming dependence between the responses assuming a gamma distribution G(1,1) for the variance σ_w^2 of the random factor w_i with a normal distribution $N(0,\sigma_w^2)$ included in model (4), we have in Table 5, the posterior summaries of interest assuming the MCMC simulation method based on 1000 simulated Gibbs samples (every 400^{th} simulated samples among 400,000 generated Gibbs samples to get an approximately uncorrelated sample) of the joint posterior distribution for all model parameters obtained using Openbugs software and considering a burn-in sample of size 111,000 discarded to eliminate the effect of the initial parameter values needed for the MCMC algorithm. Convergence of the MCMC simulated samples was monitored by traceplots of the generated Gibbs samples.

From the results presented in Table 5, it is observed that the significative effects (zero not included in the 95% credibility intervals) are the same as those obtained using “model 1”.

For the discrimination of the best model, it is used the Deviance Information Criterion (DIC). The DIC criterion (Spiegelhalter et al., 2014) is based on the posterior average of the deviance. Deviance is defined by,

$$D(\boldsymbol{\theta}) = -2\log L(\boldsymbol{\theta}) + C \tag{6}$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters of the model; $L(\boldsymbol{\theta})$ is the likelihood function and C is a constant (not always known) when comparing two models. The DIC criterion is defined by,

$$DIC = D(\widehat{\boldsymbol{\theta}}) + 2p_D \tag{7}$$

Table 4 – Posterior summaries - “model 1”.

			95% Cred. Int.	
	Mean	**S.D.**	**Lower**	**Upper**
β_{11}	-0.1200	1.0130	-1.9990	2.1180
β_{110}	0.1473	0.0865	-0.0078	0.3300
β_{111}	0.0635	0.0816	-0.0992	0.2222
β_{12}	0.0077	0.0058	-0.0038	0.0193
β_{13}	-0.7870	0.5030	-0.0011	0.8560
β_{14}	-0.3231	0.6715	-1.6350	0.9870
β_{15}	-0.1707	0.1967	-0.5356	0.2120
β_{16}	0.2969	0.3370	-0.3835	0.9469
β_{17}	0.1121	0.2347	-0.3668	0.5830
β_{18}	0.0281	0.0706	-0.1086	0.1624
β_{19}	0.0997	0.0685	-0.0277	0.2317
β_{21}	0.0616	0.9568	-1.7390	2.0130
β_{210}	0.2155	0.2147	-0.2029	0.6464
β_{211}	0.1927	0.2102	-0.2015	0.5921
β_{22}	0.0114	0.0163	-0.0223	0.0436
β_{23}	-0.0009	0.0004	-0.0018	0.0098
β_{24}	0.0462	0.9390	-1.7410	1.8570
β_{25}	0.3911	0.4695	-0.5950	1.2790
β_{26}	-0.9501	0.6877	-2.2780	0.3853
β_{27}	-1.6230	0.5235	-2.6020	-0.5539
β_{28}	0.6654	0.1736	0.3120	1.0080
β_{29}	0.1341	0.1809	-0.2256	0.4783
β_{31}	0.0197	1.0190	-1.9200	2.0560
β_{310}	0.7747	0.3104	0.1547	1.3910
β_{311}	0.9365	0.3458	0.2847	1.6620
β_{32}	-0.0025	0.0273	-0.0532	0.0526
β_{33}	-0.0016	0.5410	-0.0027	-0.5880
β_{34}	0.5721	0.9496	-1.3120	2.3590
β_{35}	-0.9367	0.6656	-2.2780	0.4054
β_{36}	-1.1290	0.8710	-2.7460	0.6063
β_{37}	-1.0830	0.6919	-2.4720	0.2858
β_{38}	1.5261	0.2702	1.0070	2.0890
β_{39}	1.2461	0.2984	0.6302	1.8620
$1/\sigma_1^2$	5.6140	0.3881	4.8290	6.3680
$1/\sigma_2^2$	0.7283	0.0493	0.6323	0.8277
$1/\sigma_3^2$	0.2558	0.0176	0.2233	0.2904

Table 5 – Posterior summaries – “model 2”.

			95% Cred. Int.	
	Mean	**S.D.**	**Lower**	**Upper**
β_{11}	-0.2007	0.9045	-1.8150	1.8190
β_{110}	0.1602	0.0907	-0.0156	0.3379
β_{111}	0.0663	0.0877	-0.1007	0.2466
β_{12}	0.0076	0.0061	-0.0039	0.0191
β_{13}	-0.0470	0.4500	-0.0010	0.7650
β_{14}	-0.4424	0.6514	-1.7060	0.8840
β_{15}	-0.1485	0.2154	-0.5578	0.2642
β_{16}	0.3774	0.3595	-0.3213	1.1020
β_{17}	0.1272	0.2441	-0.3475	0.6037
β_{18}	0.0359	0.0735	-0.1082	0.1924
β_{19}	0.1068	0.0693	-0.0324	0.2394
β_{21}	0.0264	1.0140	-2.0100	2.1040
β_{110}	0.1952	0.2043	-0.1968	0.5986
β_{111}	0.1679	0.2109	-0.2214	0.5542
β_{22}	0.0114	0.0157	-0.0190	0.0421
β_{23}	-0.0088	0.5280	-0.0018	0.2000
β_{24}	0.0379	0.9090	-1.6840	1.8520
β_{25}	0.3794	0.4785	-0.5885	1.3240
β_{26}	-0.8914	0.7033	-2.2470	0.4895
β_{27}	-1.5260	0.5123	-2.5010	-0.5200
β_{28}	0.6415	0.1645	0.3257	0.9473
β_{29}	0.1151	0.1722	-0.2182	0.4574
β_{31}	-0.0455	1.0170	-2.1460	1.9760
β_{310}	0.6979	0.3211	0.0693	1.2940
β_{311}	0.8629	0.3189	0.2144	1.4810
β_{32}	-0.0031	0.0259	-0.0545	0.0473
β_{33}	-0.0015	0.5360	-0.0025	-0.3890
β_{34}	0.5301	0.9555	-1.3520	2.3830
β_{35}	-0.9200	0.6549	-2.2080	0.3785
β_{36}	-1.1600	0.8287	-2.8000	0.4658
β_{37}	-0.9826	0.6951	-2.3480	0.2872
β_{38}	1.4450	0.2563	0.9511	1.9600
β_{39}	1.1620	0.2731	0.6319	1.6870
$1/\sigma_w^2$	12.550	1.9490	9.2210	17.320
$1/\sigma_1^2$	8.8500	1.1930	6.9640	11.630
$1/\sigma_2^2$	0.8561	0.0608	0.7379	0.9810
$1/\sigma_3^2$	0.2930	0.0208	0.2555	0.3362

where $D(\widehat{\boldsymbol{\theta}})$ is the posterior averaged deviation $\widehat{\boldsymbol{\theta}} = E(\widehat{\boldsymbol{\theta}}/y)$ and p_D is the number of model parameters, given by $p_D = \bar{D} - D(\widehat{\boldsymbol{\theta}})$ where $\bar{D} = E(D(\boldsymbol{\theta}/y)$ is the posterior mean of the deviation that measures the quality of data fit for each model.

Table 6 shows the DIC values obtained from the generated Gibbs samples using the Openbugs software for both models considered in the data analysis.

Table 6 – DIC estimates for model 1 and model 2

Model	$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\mathbf{Y}_3$
Model 1	483.2	1373.0	1826.0
Model 2	448.9	1318.0	1772.0

From the results of Table 4, it can be observed that the "model 2" is better fitted by the data. Assuming "model 2", the estimated proportions for the four classes given by (5) and the observed proportions are presented in Figure 4. From the plots of Figure 4, it is observed good fit of model 2 to the compositional data associated to accident victims in Brazilian federal roads.

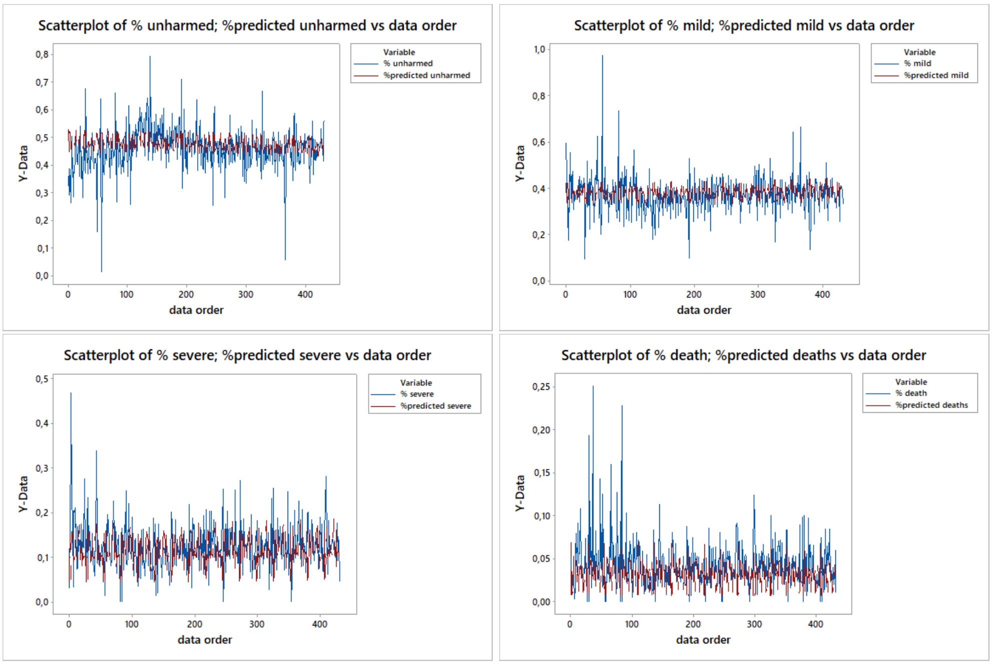

Figure 4 – Estimated and observed proportions (unharmed, mild, severe, deaths).

3 DISCUSSION OF THE RESULTS AND CONCLUDING REMARKS

From the obtained results usig ALR compositional models it is possible to get important conclusions on the study. Since the significative covariates affecting the responses $y_{2i} = log(x_{3i}/x_{1i})$ and $y_{3i} = log(x_{4i}/x_{1i})$, where x_{1i} = % unharmed, x_{2i} = % minor injuries, x_{3i} = % severe injuries and

x_{4i} = % deaths are given by poor pavement, NE region and year, Figures 5, 6, 7 and 8 show the scatter plots associated to each response and covariate from where it is possible to get important interpretations for the compositional multivariate dataset.

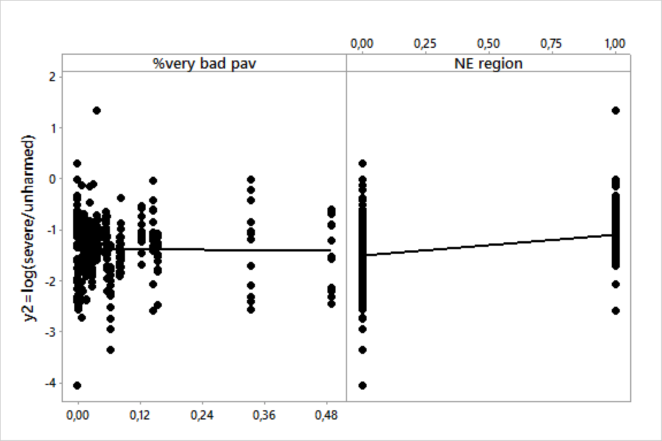

Figure 5 – Graphs of $y_2 = \log(severe/unharmed)$ versus %lousy pavement and NE region.

From the graphs of Figure 5, it is possible to observe that although there is great variability in the response y_{2i} there is a slight decreasing in the response with %lousy pavement (pavement in very poor condition) and an increasing in the response y_{2i} in the NE region when compared to the other regions of Brazil.

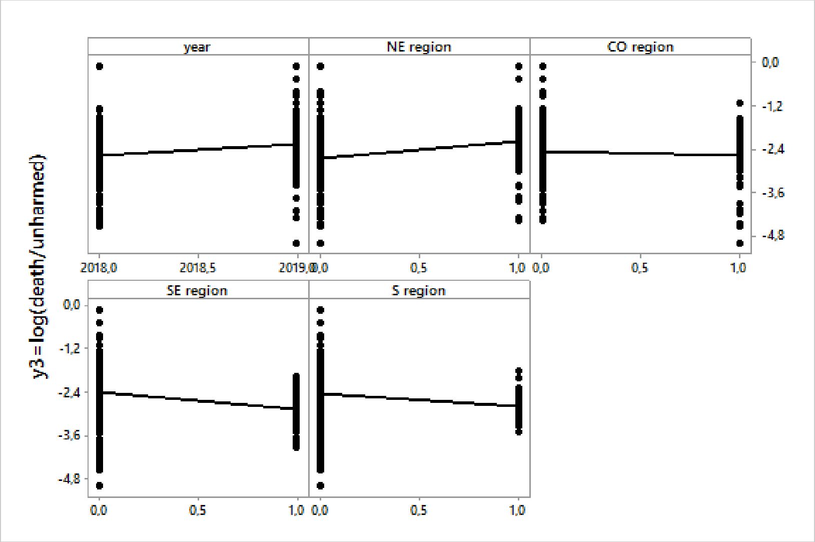

Figure 6 – Graphs of $y_3 = \log(death/unharmed)$ versus year, NE region, CO region, SE region and S region.

From the graphs of Figure 6, it is possible to see an increasing in the response y_{3i} in the year 2019 when compared to the year 2018; an increasing in the response y_{3i} in the NE region when compared to the other regions of Brazil; a decreasing in the response y_{3i} in the CO region when compared to the other regions of Brazil; an apparently decreasing in the response y_{3i} in the SE region when compared to the other regions of Brazil and an apparently decreasing in the response y_{3i} in the S region when compared to the other regions of Brazil.

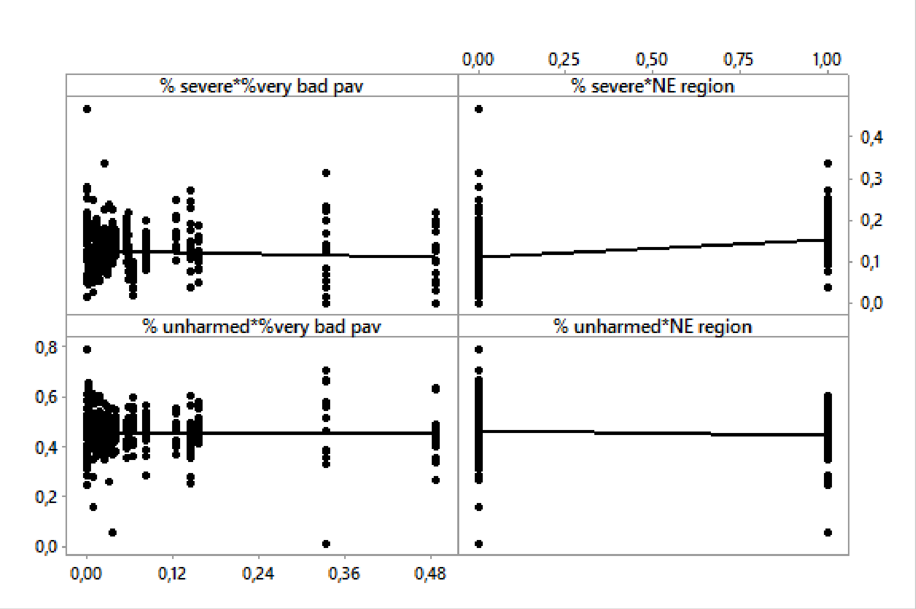

Figure 7 – Graphs of %severe injury and %unharmed injury versus very bad pavement and NE region.

From the graphs of Figure 7, it is possible to see an increasing of %severe injuries in the NE region when compared to the other regions of Brazil and a decreasing of %unharmed persons in the NE region; in relation to the factor very bad (lousy) pavement, it is difficult to see the effect in %unharmed and %severe injuries.

From the graphs of Figure 8, it is possible to observe that although there is great variability in the responses %deaths and %unharmed, we see a small increasing of %deaths related to the year 2019 when compared to the year 2018; similarly apparently there is an increasing of %deaths in the NE region when compared to the other regions of Brazil; a decreasing of %deaths in the SE and S regions.

In summary, from the obtained results, it is concluded that the rates of serious accidents and deaths are affected by some covariates as the regions of Brazil (especially the NE region where the rates for accidents with severe injuries and deaths are higher than the rates for the other regions of Brazil), years and some sligh effect of pavement conditions of the roads, which could be important for the road managers to take decisions to improve the road conditions in Brazil.

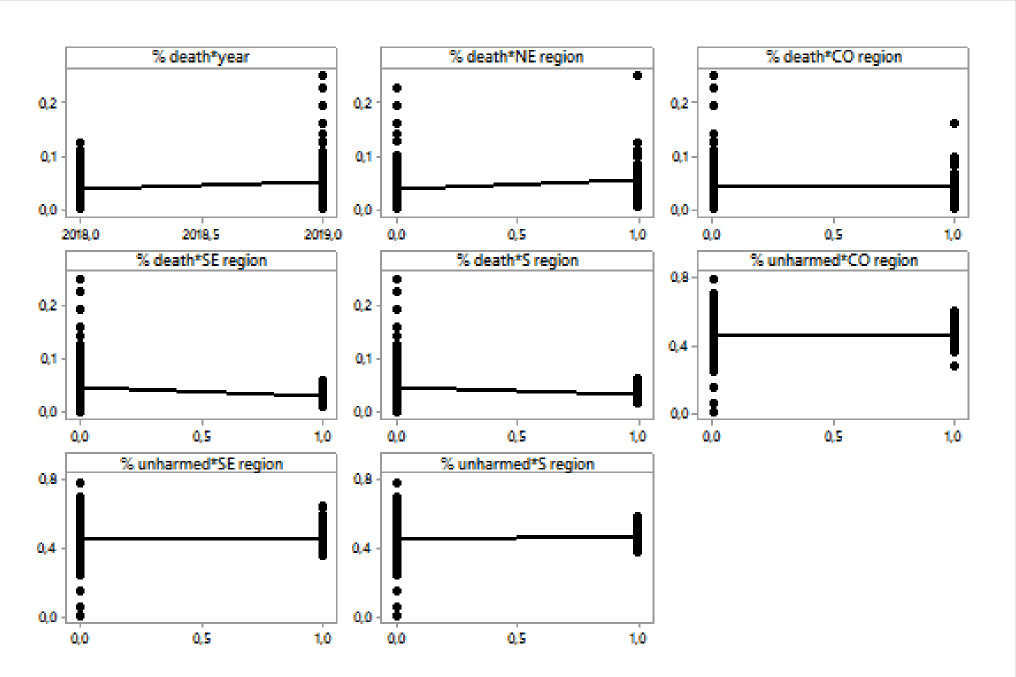

Figure 8 – Graphs of %death and %unharmed injury versus year, NE region, CO region, SE region and S region.

This is an important result which could help in future decreasing of the high rates of severe injuries and deaths in the Brazilian federal roads.

As concluding remarks, it is possible to point out that the use of existing compositional Bayesian models could be of great interest in the data analysis of road accidents as seen in this study. It is important to point out that other prior distributions could be considered for the parameters of the model possibly incorporating with prior opinions of engineer experts in road traffic. The use of MCMC methods to get the posterior summaries of interest using free existing simulation softwares like the OpenBugs software could be a great option in the data analysis under a hierarchical Bayesian data analysis which only requires the specification of the likelihood function and the prior distributions for the parameters of the model. It is important to point out that other dependence structures could be assumed for the ALR transformed data, like a multivariate normal distribution for the errors in the compositional model (see for example, Shimizu et al., 2015).

In a future work the results of this study could be extended to the presence of other covariates as weather conditions, type of road (double lane and single lane), roads with and without tolls, period of day, speed of the vehicle at the moment of the accident and many other possible covariates that could affect the responses given by a proportion vector (x_{1i} = %unharmed, x_{2i} = %mild injuries , x_{3i} = %severe injuries and x_{4i} = %deaths).

Acknowledgments

The authors are very grateful for the reviewers' comments that led to a great improvement of the manuscript.

References

[1] ABREU DROM, SOUZA EM & MATHIAS TAF. 2018. Impacto do Código de Trânsito Brasileiro e da Lei Seca na mortalidade por acidentes de trânsito. *Cadernos de Saúde Pública*, **34**: e00122117.

[2] AITCHISON J. 1982. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(2): 139–160.

[3] ANDRADE SSCA & MELLO-JORGE MHP. 2016. Mortality and potential years of life lost by road traffic injuries in Brazil, 2013. *Revista de saude publica*, **50**: 59.

[4] ASSIMIZELE B, BYE RT ET AL. 2020. Minimizing the Environmental Risk from Oil Tanker Grounding Accidents in the High North. *American Journal of Operations Research*, **10**(03): 83.

[5] ATCHISON J & SHEN SM. 1980. Logistic-normal distributions: Some properties and uses. *Biometrika*, **67**(2): 261–272.

[6] BACCHIERI G & BARROS AJ. 2011. Acidentes de trânsito no Brasil de 1998 a 2010: muitas mudanças e poucos resultados. *Revista de Saúde Pública*, **45**(5): 949–963.

[7] BAHADORIMONFARED A, SOORI H, MEHRABI Y, DELPISHEH A, ESMAILI A, SALEHI M & BAKHTIYARI M. 2013. Trends of Fatal Road Traffic Injuries in Iran (2004–2011).

[8] BARROS AJ, AMARAL RL, OLIVEIRA MSB, LIMA SC & GONÇALVES EV. 2003. Acidentes de trânsito com vítimas: sub-registro, caracterização e letalidade. *Cadernos de Saúde Pública*, **19**: 979–986.

[9] BAYKAL-GÜRSOY M, XIAO W & OZBAY K. 2009. Modeling traffic flow interrupted by incidents. *European Journal of Operational Research*, **195**(1): 127–138.

[10] BHALLA K, SHOTTEN M, COHEN A, BRAUER M, SHAHRAZ S, BURNETT R, LEACH-KEMON K, FREEDMAN G & MURRAY C. 2014. *Transport for health: the global burden of disease from motorized road transport.*

[11] BOX GE & COX DR. 1964. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2): 211–243.

[12] CASTRO ARAGÓN FR & LEAL JE. 2003. Alocação de fluxos de passageiros em uma rede de transporte público de grande porte formulado como um problema de inequações variacionais. *Pesquisa Operacional*, **23**(2): 235–264.

[13] EGOZCUE JJ, PAWLOWSKY-GLAHN V, MATEU-FIGUERAS G & BARCELO-VIDAL C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**(3): 279–300.

[14] GELFAND AE & SMITH AF. 1990. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, **85**(410): 398–409.

[15] GELMAN A, CARLIN JB, STERN HS, DUNSON DB, VEHTARI A & RUBIN DB. 2013. *Bayesian data analysis*. CRC press.

[16] HAAGSMA JA, GRAETZ N, BOLLIGER I, NAGHAVI M, HIGASHI H, MULLANY EC, ABERA SF, ABRAHAM JP, ADOFO K, ALSHARIF U ET AL. 2016. The global burden of injury: incidence, mortality, disability-adjusted life years and time trends from the Global Burden of Disease study 2013. *Injury prevention*, **22**(1): 3–18.

[17] HAASTRUP P. 1994. Overview of problems of risk management of accidents with dangerous chemicals in Europe. *European journal of operational research*, **75**(3): 488–498.

[18] IYENGAR M & DEY DK. 1996. Bayesian analysis of compositional data. *Department of Statistics, University of Connecticut, Storrs, CT*, pp. 06269–3120.

[19] IYENGAR M & DEY DK. 1998. Box–Cox transformations in Bayesian analysis of compositional data. *Environmetrics: The official journal of the International Environmetrics Society*, **9**(6): 657–671.

[20] JOHNSON RA, WICHERN DW ET AL. 2002. *Applied multivariate statistical analysis*. vol. 5. Prentice hall Upper Saddle River, NJ.

[21] JORGE M, KOIZUMI MS, TUONO VL ET AL. 2008. *Acidentes de trânsito no Brasil: a situação nas capitais*.

[22] JORGE M, KOIZUMI MS ET AL. 2009. Acidentes de trânsito causando vítimas: possível reflexo da lei seca nas internações hospitalares. *Revista ABRAMET*, **27**(2): 16–25.

[23] JORGE M, LATORRE MR ET AL. 1994. Acidentes de trânsito no Brasil: dados e tendências. *Cadernos de Saúde Pública*, **10**: S19–S44.

[24] KLEIN CH. 1994. Mortes no trânsito do Rio de Janeiro, Brasil. *Cadernos de Saúde Pública*, **10**: S168–S176.

[25] LUNN D, SPIEGELHALTER D, THOMAS A & BEST N. 2009. The BUGS project: Evolution, critique and future directions. *Statistics in medicine*, **28**(25): 3049–3067.

[26] LYONS RA, WARD H, BRUNT H, MACEY S, THOREAU R, BODGER O & WOODFORD M. 2008. Using multiple datasets to understand trends in serious road traffic casualties. *Accident Analysis & Prevention*, **40**(4): 1406–1410.

[27] MALTA DC, SILVA MMAD & BARBOSA J. 2012. Violências e acidentes, um desafio ao Sistema Único de Saúde. *Ciência & Saúde Coletiva*, **17**(9): 2220–2220.

[28] MARÍN L & QUEIROZ MS. 2000. A atualidade dos acidentes de trânsito na era da velocidade: uma visão geral. *Cadernos de Saúde Pública*, **16**: 7–21.

[29] MARÍN-LEÓN L, BELON AP, BARROS MBDA, ALMEIDA SDDM & RESTITUTTI MC. 2012. Tendência dos acidentes de trânsito em Campinas, São Paulo, Brasil: importância crescente dos motociclistas. *Cadernos de Saúde Pública*, **28**(1): 39–51.

[30] MARTÍN FERNÁNDEZ JA, DAUNIS I ESTADELLA J & MATEU I FIGUERAS G. 2015. On the interpretation of differences between groups for compositional data. *SORT: statistics and operations research transactions, 2015, vol. 39, núm. 2, p. 231-252*, .

[31] MARTINEZ EZ, ACHCAR JA, ARAGON DC & BRUNHEROTTI MA. 2020. A Bayesian analysis for pseudo-compositional data with spatial structure. *Statistical Methods in Medical Research*, **29**(5): 1386–1402.

[32] MARTÍNEZ F, BALDOQUÍN MG & MAUTTONE A. 2017. And solution method to a simultaneous route design and frequency setting problem for a bus rapid transit system in Colombia. *Pesquisa Operacional*, **37**(2): 403–434.

[33] MEKKER M, LI H, COX E, BULLOCK D ET AL. 2018. Dashboards for Monitoring Congestion and Crashes in Interstate Work Zones. *American Journal of Operations Research*, **9**(1): 15–30.

[34] NOVAES AG. 2001. Rapid-transit efficiency analysis with the assurance-region DEA method. *Pesquisa Operacional*, **21**(2): 179–197.

[35] RAYENS WS & SRINIVASAN C. 1991. Box–Cox transformations in the analysis of compositional data. *Journal of Chemometrics*, **5**(3): 227–239.

[36] SHIMIZU TK, LOUZADA F, SUZUKI AK & EHLERS RS. 2015. Modeling Compositional Regression with uncorrelated and correlated errors: a Bayesian approach. *arXiv preprint arXiv:1507.00225*, .

[37] SILVA S & ANDRADE S. 1996. Acidentes de trânsito: Problema prioritário de saúde. *A Construção do SUS a partir do Município*, pp. 95–99.

[38] SMITH AF & ROBERTS GO. 1993. Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 3–23.

[39] SPIEGELHALTER DJ, BEST NG, CARLIN BP & VAN DER LINDE A. 2014. The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(3): 485–493.

[40] SZWED P, VAN DORP JR, MERRICK JR, MAZZUCHI TA & SINGH A. 2006. A Bayesian paired comparison approach for relative accident probability assessment with covariate information. *European Journal of Operational Research*, **169**(1): 157–177.

[41] TJELMELAND H & LUND KV. 2003. Bayesian modelling of spatial compositional data. *Journal of Applied Statistics*, **30**(1): 87–100.

[42] WAISELFISZ JJ. 2013. *Mapa da violência 2013: acidentes de trânsito e motocicletas*. Rio de Janeiro.

[43] WORLD HEALTH ORGANIZATION. 2004. *International statistical classification of diseases and related health problems*. vol. 1. World Health Organization.

[44] WORLD HEALTH ORGANIZATION. 2018. *Global status report on road safety 2018*. World Health Organization.

[45] WORLD HEALTH ORGANIZATION ET AL. 2018. *Noncommunicable diseases country profiles 2018*. World Health Organization.

How to cite

OLIVEIRA RP & ACHCAR JA. 2020. Compositional statistical models under a Bayesian approach: an application to traffic accident data in federal highways in Brazil. *Pesquisa Operacional*, **40**: e230462. doi: 10.1590/0101-7438.2020.040.00230462.