

Prediction of the incidence of deaths secondary to road traffic injuries in Connecticut and Delaware

Sandra Catalina Correa Herrera^{a*}, María Alejandra González^b,
Signed Esperanza Prieto Bohórquez^a, Adiel Ruiz^b

^a *Insight Group. Insight Research Group SAS. Bogotá, D.C, Colombia*

^b *OdontopostgradosUCC Group. Universidad Cooperativa de Colombia. Bogotá, D.C, Colombia*

Abstract

Probability theory allowed the development of the probabilistic random walk, which is useful to predict the number of cases for epidemics such as the deaths caused by traffic accidents. In this investigation, it was wanted to confirm the applicability of the probabilistic random walk to predict the annual incidence of deaths secondary to road traffic injuries in Connecticut and Delaware for 2017 and 2016 respectively. For this, probability spaces that consider the incidences of deaths secondary to road traffic injuries as probabilistic events as well as the increases and decreases of annual incidences were built for Connecticut and Delaware in order to apply a quadratic equation that predicted the number of cases for both states in 2017 and 2016. It was found that the methodology predicted that for Connecticut and Delaware the incidence of deaths were 288 and 123, respectively, that when compared with the real values reported, 278 and 119, obtaining accuracies superior to 96%. Hence, it was confirmed that the probabilistic random walk can be applicable to predict with high precision the behavior of the incidence of deaths for Connecticut and Delaware, suggesting its potential role as an epidemiological surveillance tool.

Keywords: probability; incidence; traffic accidents.

Introduction

According to the World Health Organization, the leading cause of death for people between 5 to 29 years is road traffic injuries, generating 1.35 million deaths in 2016

although the proportions of deaths relative to the world population have been stabilized in recent years. Besides, road traffic injuries are the eighth cause of death worldwide, surpassing the mortality associated to the HIV infection, tuberculosis and diarrheic diseases (World Health Organization, 2018). Given the magnitude of deaths and disability associated to road traffic injuries, it has been proposed among the Sustainable Development Goals to reduce by half these events (Wangdi et al. 2017), which requires the implementation of trustworthy methodologies that monitor the fulfillment of this goal

Different methodologies have been developed to predict or simulate the behavior of deaths by road traffic injuries. Some of these models are based on nonlinear relationships, artificial neural networks or autoregressive integrated moving average modelling techniques (Huang et al 2016; García et al. 2018; Ihueze and Onwurah 2018). However, these usually rely on several retrospective data and different variables, which can make more complex the study of road traffic injuries. The results of a given experiment analyzed in the context of probability theory are grouped in specific conditions called sample space. The quantification of probability of an event in an experiment allows to establish how likely is that a certain event happens within the sample space (Feynman et al. 1964; Laplace 1995). Probability led to the development of the random walk, which is characterized for the irregularity of its behavior. This can be imagined as a walker that moves from the origin of a coordinate plane towards positive or negative values, like a zigzag. This methodology has allowed to predict the behavior of apparently chaotic phenomena in different sciences as biology and economics, among others (Pearson 1905; Cattoni et al. 2004; Van Kampen 1992; Redner 2001; Goel and Richter 1974; Doi and Edwards 1986; De Gennes 1980; Weiss 1994). Previously, in the context of theoretical physics and mathematics, a methodology based in random walk and probability was developed, being useful to predict the trajectories of infectious diseases (Rodríguez et al. 2017; Rodríguez and Correa 2009; Rodríguez et al. 2018a) such HIV (Rodríguez et al. 2018a) or dengue (Rodríguez et al. 2017) but this method has also been applied to predict the death rates secondary to traffic accidents (Rodríguez et al. 2020). These numerical predictions have been compared with the values reported by official entities, obtaining precisions greater than 90%. The purpose of this study is to predict the incidence of road traffic injuries for the states of Connecticut for 2017 and Delaware for 2016, respectively, through a probabilistic random walk.

Methods

Data source

Data was taken from the National Highway Traffic Safety Administration (NHTSA) database where the deaths secondary to traffic are submitted (NHTSA 2018). The years 2016 and 2015 were chosen as the years to predict for Connecticut and Delaware, respectively, since these values have been officially reviewed and fixed by the NHTSA, allowing to conduct more precise comparisons between the real values and the ones obtained through the methodology applied.

Procedure

According to the values of death rates reported for Connecticut and Delaware, the lengths of the annual variations were established through Equation 1. The values of the coordinates in the y axis represent the death rates caused by traffic while the variation of the values of time in the x axis represent time, and, given that the annual variation remains as a constant, its value is zero.

$$L = \sqrt{(X_f - X_0)^2 + (Y_f - Y_0)^2} \text{ Eq. 1}$$

Where X0 and Y0 are the coordinates for the initial year while Xf and Yf are the coordinates for the following year.

Then, a probability space was established considering each length as a probabilistic event. The probability for each length was calculated through the quotient between the value of each length and the sum of all lengths (see Equation 2).

$$P(L) = \frac{\text{Length of annual variation of deaths caused by traffic}}{\text{Totality of lengths}} \text{ Eq. 2}$$

Afterwards, another probability space was established by calculating the values of probabilities of death rates through Equation 3.

$$P(N) = \frac{\text{Annual rate of deaths caused by traffic}}{\text{Totality of death rates}} \text{ Eq. 3}$$

After this, it was fundamental to determine if the values of probability previously obtained had a tendency towards specific values, so the root mean square was calculated through Equation 4 as follows:

$$P(Rn) = \frac{\text{Annual rate of deaths caused by traffic}}{\text{Totality of death rates}} \pm \frac{1}{2\sqrt{N}} \text{ Eq. 4}$$

From the values obtained in the two previous probability spaces that were defined, a third space was established, in order to predict the death rates caused by traffic for both states. This space was built from the calculation of an arithmetic average of the three previous lengths found between 2014 to 2016 for Connecticut and between 2013 to 2015 for Delaware. These averages were taken as well as the sum of lengths were all replaced in Equation 5, which provided two results corresponding to a range where the value of the prediction is delimited.

$$Y_{(tp)} = \frac{2Y_{(p)} \pm \sqrt{(-2Y_{(p)})^2 - 4\{Y_{(p)}^2 + (X_f - X_o)^2 - [(P(L)^2 \times (TL)^2]\}}}{2} \text{ Eq. 5}$$

Where $Y_{(tp)}$ year to predict and $Y_{(p)}$ previous year

Finally, in order to establish only one value between the range of the most likely values, another probability space was built considering two probabilistic events, which are the consecutive increases and decreases of deaths with respect to the previous year in periods of two and three years.

Results

The values of the lengths for the deaths secondary to road traffic injuries in Connecticut and Delaware were between 4 to 99 and 1 to 31 respectively (table 1 and 2).

Probabilistic lengths varied between 0.006 to 0.145 and 0.003 to 0.106, respectively. The probability of the annual death rates varied between 0.032 to 0.050 and 0.037 to 0.055 while the values of root mean square varied in a range of 0.026 to 0.056 and 0.027 to 0.065, finding that a difference between the last ones and the expected values varied in a range of -0.006 to 0.006 and -0.001 to 0.001 (table 1 and 2). The difference among the values of the root mean square obtained for the death rates show that the behavior of this dynamic is not equiprobable, but it shows a loading of probability, allowing the establishment of predictions (tables 1 and 2).

Based on the analysis performed to the probability space of the last three years and the application of Eq. 5, two values of the prediction were determined for the death rates, which were 273 to 335 and 115 to 147 (data not shown). The determination of the most likely event for 2017 and 2016 for Connecticut and Delaware was made through the study of the behavior of the frequency and probability of consecutive augments (A) and decreases (D) between 1994 to 2016 and 2015, respectively (Table 3).

The analysis of this probability space showed that for two or more consecutive years, the number of consecutive A or D varied between 2 and 7, finding that the augments for two years have a probability of 0.5 for Delaware, while the number of consecutive A or D varied between 2 and 6 with decreases having an overall probability of 0.52 (Table 3). When evaluating the possible combinations of A and D for a period of three consecutive years it was found that there is a higher probability of finding DAD for both Connecticut and Kentucky with probability values of 0.556 and 0.3, respectively (tables 4 and 5).

In order to establish the most likely events for 2017 and 2016 for Connecticut and Delaware, respectively, two new arithmetic averages between the values of the ranges obtained with equation 5 were obtained, which were 288 and 123. On the other hand, the real values were 278 and 119, which means that the methodology had a predictive success of 96% and 97% for Connecticut and Delaware, respectively.

Discussion

This is the first investigation where the capability of a methodology based a probabilistic random walk to analyze and predict the incidence of deaths secondary to road traffic injuries for Connecticut and Delaware for 2017 and 2016, respectively, was confirmed. The percentages of predictive precision were equal to or greater than 96%, which suggest that the phenomenon studied is complex and irregular, but it also highlights the existence of an underlying mathematical order that allows to obtain accurate predictions.

Different approaches have come up in recent years in order to predict the behavior of road traffic injuries and its outcomes (Huang et al. 2016; García et al. 2018; Ihueze and Onwurah 2018). For instance, a modified Smeed equation has been used to establish a nonlinear relationship between mortality and the level of motorization in a region with the possibility of obtaining simulations for this phenomenon. This

method has been useful in countries like China to simulate this variable and observe that there are discrepancies between the values reported by the health services and the police. However, the high inputs of data and specialized analysis make these methodologies hard to implement in some scenarios, although they have been proven to be very effective and precise.

This method, instead, relies only on values for deaths secondary to road traffic injuries for the last three years which greatly simplifies the analysis of mortality and allows this to be an easily applicable complementary method to predict its tendency in time, providing quick and useful information for governments, aside from epidemiological considerations such as endemic channels (Hernández et al. 2016) or adjustments for sex given the acausal and theoretical thinking that led to the development of this methodology. Nevertheless, the precision of this method can vary according to the quality of data reported, since it has been noticed that primary data sources tend to underreport these events (Samuel et al, 2012).

This acausal perspective in biomedical research has led to the development of methodologies that achieve to forecast the binding of peptides to HLA class II through entropy and probability (Rodríguez 2008) and predictions of CD4+ counts in ranges for patients infected with HIV through probability and set theory (Rodríguez et al, 2013). Similarly, diagnostics of heart dynamics have been achieved through the Zipf/Mandelbrot law (Rodríguez et al. 2018b) and generalizations of the arterial re-stenosis have been developed (Rodríguez et al. 2010).

Conclusion

The probabilistic random walk allows to predict with high accuracy the incidence of deaths secondary to traffic, suggesting its possible role as a public health tool.

Acknowledgements

We extend our thanks to Insight Research Group SAS for all their support to our investigations by providing us their facilities and resources.

Compliance with Ethical Standards

Funding: This investigation was funded by Universidad Cooperativa de Colombia through the project INV 2466.

Disclosure of potential conflicts of interest: none.

Ethics approval: This is an observational study where data made publicly available was analyzed, thus, no ethical approval is required

Informed consent: not applicable.

Table 1. Values of the annual deaths rates caused by traffic in Connecticut between 1994 to 2016

Year	Incidence	L		Avg	P(L)	P(N)	RMS +	RMS -	RMS + P	RMS- P
1994	310					0.046	0.052	0.039	0.006	-0.006
1995	317	7	A	1.75	0.010	0.047	0.053	0.041	0.006	-0.006
1996	310	7	D	1.75	0.010	0.046	0.052	0.039	0.006	-0.006
1997	339	29	A	7.25	0.042	0.050	0.056	0.044	0.006	-0.006
1998	329	10	D	2.50	0.015	0.048	0.054	0.042	0.006	-0.006
1999	301	28	D	7.00	0.041	0.044	0.050	0.038	0.006	-0.006
2000	341	40	A	10.00	0.058	0.050	0.056	0.044	0.006	-0.006
2001	318	23	D	5.75	0.034	0.047	0.053	0.041	0.006	-0.006
2002	325	7	A	1.75	0.010	0.048	0.054	0.042	0.006	-0.006
2003	298	27	D	6.75	0.039	0.044	0.050	0.038	0.006	-0.006
2004	294	4	D	1.00	0.006	0.043	0.049	0.037	0.006	-0.006
2005	278	16	D	4.00	0.023	0.041	0.047	0.035	0.006	-0.006
2006	311	33	A	8.25	0.048	0.046	0.052	0.040	0.006	-0.006
2007	296	15	D	3.75	0.022	0.043	0.050	0.037	0.006	-0.006
2008	302	6	A	1.50	0.009	0.044	0.050	0.038	0.006	-0.006
2009	224	78	D	19.50	0.114	0.033	0.039	0.027	0.006	-0.006
2010	320	96	A	24.00	0.140	0.047	0.053	0.041	0.006	-0.006
2011	221	99	D	24.75	0.145	0.032	0.039	0.026	0.006	-0.006
2012	264	43	A	10.75	0.063	0.039	0.045	0.033	0.006	-0.006
2013	286	22	A	5.50	0.032	0.042	0.048	0.036	0.006	-0.006
2014	248	38	D	9.50	0.056	0.036	0.042	0.030	0.006	-0.006
2015	270	22	A	5.50	0.032	0.040	0.046	0.034	0.006	-0.006
2016	304	34	A	8.50	0.050	0.045	0.051	0.039	0.006	-0.006

L: length; Avg: average; P: probability; RMS: root mean square

Table 2: Values of the annual deaths rates caused by traffic in Delaware between 1994 to 2015

Year	Incidence	L		Avg	P(L)	P(N)	RMS +	RMS -	RMS + P	RMS- P
1994	112					0.042	0.052	0.032	0.010	-0.010
1995	121	9	A	9.00	0.031	0.045	0.055	0.036	0.010	-0.010
1996	116	5	D	5.00	0.017	0.043	0.053	0.034	0.010	-0.010
1997	143	27	A	27.00	0.092	0.054	0.063	0.044	0.010	-0.010
1998	115	28	D	28.00	0.096	0.043	0.053	0.033	0.010	-0.010
1999	100	15	D	15.00	0.051	0.037	0.047	0.028	0.010	-0.010
2000	123	23	A	23.00	0.078	0.046	0.056	0.036	0.010	-0.010

2001	136	13	A	13.00	0.044	0.051	0.061	0.041	0.010	-0.010
2002	124	12	D	12.00	0.041	0.046	0.056	0.037	0.010	-0.010
2003	142	18	A	18.00	0.061	0.053	0.063	0.044	0.010	-0.010
2004	134	8	D	8.00	0.027	0.050	0.060	0.041	0.010	-0.010
2005	133	1	D	1.00	0.003	0.050	0.060	0.040	0.010	-0.010
2006	148	15	A	15.00	0.051	0.055	0.065	0.046	0.010	-0.010
2007	117	31	D	31.00	0.106	0.044	0.054	0.034	0.010	-0.010
2008	121	4	A	4.00	0.014	0.045	0.055	0.036	0.010	-0.010
2009	116	5	D	5.00	0.017	0.043	0.053	0.034	0.010	-0.010
2010	101	15	D	15.00	0.051	0.038	0.048	0.028	0.010	-0.010
2011	99	2	D	2.00	0.007	0.037	0.047	0.027	0.010	-0.010
2012	114	15	A	15.00	0.051	0.043	0.052	0.033	0.010	-0.010
2013	99	15	D	15.00	0.051	0.037	0.047	0.027	0.010	-0.010
2014	124	25	A	25.00	0.085	0.046	0.056	0.037	0.010	-0.010
2015	131	7	A	7.00	0.024	0.049	0.059	0.039	0.010	-0.010

L: length; Avg: average; P: probability; RMS: root mean square

Table 3. Frequency and probability of consecutive augments (A) and decreases (D) of the deaths rates caused by traffic

Consecutive years	Connecticut					Delaware			
	1994 A	A		D		A		D	
	2014	V	P	V	P	V	P	V	P
1		7	0.31	6	0.27	6	0.28	4	0.19
2		2	0.18	1	0.09	2	0.19	2	0.19
3		0	0	1	0.13	0	0	1	0.14
4		0	0	0	0	0	0	0	0
5		0	0	0	0	0	0	0	0
6		0	0	0	0	0	0	0	0
7		0	0	0	0	0	0	0	0
8		0	0	0	0	0	0	0	0
9		0	0	0	0	0	0	0	0
Total	22	11	0.50	11	0.50	10	0.476	11	0.52

Table 4. Values of incidence, lengths and probability for the three previous years to predict.

Connecticut	Year	Incidence	Length	Probability
	2014	248	38	0.40
	2015	270	22	0.23
	2016	304	34	0.36
			94	0.33
Delaware	Year	N°	L	P
	2013	99	15	0.31
	2014	124	25	0.53
	2015	131	7	0.14
			47	0.33

Table 5. Number of possible combinations of augments (A) and decreases (D) for a period of three consecutive years of the deaths rates

Connecticut			Delaware		
Combinations	V	P	Combinations	V	P
DDD	1	0.111	DDD	1	0.1
DDA	0	0	DDA	0	0
DAD	5	0.56	DAD	3	0.3
DAA	1	0.1	DAA	2	0.2
ADD	1	0.1	ADD	2	0.2
ADA	1	0.1	ADA	2	0.2
AAD	0	0	AAD	0	0
AAA	0	0	AAA	0	0
Total	9	1	Total	10	1

References

- [1] Cattoni, D., Ozu, M., Chara, O. (2004). Ruidos en la naturaleza. *ANALES AFA*, 16: 294-99.
- [2] De Gennes, P.G. (1980). *Scaling Concepts in Polymer Physics*. London: Cornell University Press.
- [3] Doi, M., & Edwards, S.F. (1986). *The Theory of Polymer Dynamics*. Oxford: Clarendon Press.
- [4] Feynman, R.P., Leighton, R.B., & Sands M. (1964). Probabilidad. In: R. P. Feynman, R. B. Leighton & M. Sands. *Física* (pp. 1-16) Wilmington: Addison-Wesley Iberoamericana.

- [5] García, B., Bumbacher, A., Deublein, M. & Adey, B.T. (2018). Predicting road traffic accidents using artificial neural network models. *Infrastructure Asset Management* 5(4):132–144, <https://doi.org/10.1680/jinam.17.00028>
- [6] Goel, N.W., & Richter, N. (1974). *Stochastic Models in Biology*. New York: Academic Press.
- [7] Hernández, M., Arboleda, D., Arce, S., Benavides, A., Tejada, P., Ramírez S, et al. (2016). Methodology to develop endemic channels and notification trends for dengue in Valle del Cauca, Colombia, 2009-2013. *Biomedica*, 36:98–107
- [8] Huang, H., Yin, Q., Schwebel, D.C., Li, L., Hu, G. (2016). Examining Road Traffic Mortality Status in China: A Simulation Study. *PLoS ONE*, 11(4): e0153251. <https://doi.org/10.1371/journal.pone.0153251>
- [9] Ihueze, C.C., & Onwurah, U.O. (2018). Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. *Accident Analysis & Prevention*, 112:21-29.
- [10] Laplace, P. (1995). *Ensayo filosófico sobre las probabilidades*. Barcelona: Altaza.
- [11] NHTSA. (2018). Fatality Analysis Reporting System (FARS) Encyclopedia. Resource document. <http://www-fars.nhtsa.dot.gov/Main/index.aspx>. Accessed 06 February 2019.
- [12] Pearson, K. (1905) The Problem of the Random Walk. *Nature*. 1905;72:294.
- [13] Redner, S. (2001). *A Guide to First-Passage Process*. Cambridge: Cambridge University Press.
- [14] Rodríguez, J. (2008). Binding to class II HLA theory: probability, combinatory and entropy applied to peptide sequences. *Inmunología*, 27:151–166.
- [15] Rodríguez, J., & Correa, C. (2009). Predicción temporal de la epidemia de dengue en Colombia: dinámica probabilista de la epidemia. *Rev. Salud pública*, 11(3): 443-453.
- [16] Rodríguez, J., Jattin, J., Soracipa, Y. (2020). Probabilistic temporal prediction of the deaths caused by traffic in Colombia. Mortality caused by traffic prediction. *Accident Analysis & Prevention*, 135:105332.
- [17] Rodríguez, J., Oliveros, D., Rodríguez, D., Sosa, J., Prieto, S., & Correa, C. (2018b). Diagnostic methodology of cardiac dynamics based on the Zipf-Mandelbrot law: evaluation with 50 patients. *Rev. Mex. Cardiol*, 29: 83–89
- [18] Rodríguez, J., Oliveros, D., Soracipa, Y., Bernal, L., Correa, C., Abraham, L., et al. (2018a). Análisis probabilista con caminata al azar del número de personas viviendo con VIH mundialmente. *Rev. Fac. Nac. Salud Pública*, 36(1): 27-33.
- [19] Rodríguez, J., Prieto, S., Correa, C., Bernal, P., Puerta, G., & Vitery S. (2010). Theoretical generalization of normal and sick coronary arteries with fractal dimensions and the arterial intrinsic mathematical harmony. *BMC Med. Phys*, 10:1–6.
- [20] Rodríguez, J., Prieto, S., Correa, C., Pérez, C., Mora, J., & Bravo, J. (2013). Predictions of CD4 lymphocytes' count in HIV patients from complete blood count. *BMC Med. Phys*, 2013;13:3.

- [21] Rodríguez, J., Prieto, S.E., Correa, S.C., Pérez, C.E., & Soracipa, M.Y. (2017). Dinámica de la epidemia de malaria en Colombia: Predicción probabilística temporal. *Rev. Salud. Pública*, 19(1):52-59.
- [22] Samuel, J., Sankhulani, E., Qureshi, J., Baloyi, P., Thupi, C., Lee, C., et al. (2012). Under-reporting of road traffic mortality in developing countries: application of a capture-recapture statistical model to refine mortality estimates. *PLoS One*, 7,e31091.
- [23] Van Kampen, N.G. (1992). *Stochastic Processes in Physics and Chemistry*. Amsterdam: North-Holland Publishing Company.
- [24] Wangdi, C., Gurung, M.S., Duba, T., Wilkinson, E., Tun, Z.W., et al. (2017). Burden, pattern and causes of road traffic accidents in Bhutan, 2013–2014: a police record review. *International Journal of Injury Control and Safety*, <https://doi.org/10.1080/17457300.2017.1341930>
- [25] Weiss, G. (1994). *Aspects and Applications of the Random Walk*. Amsterdam: North-Holland Publishing Company.
- [26] World Health Organization. (2018). Global Status Report on Road Safety. Resource document. https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/. Accessed 27 January 2019.