CAUSALITY AND PREDICTION OF ATLANTIC TROPICAL STORMS

Timothy Chen Allen
timothy.allen@udc.edu
Valbona Bejleri
vbejleri@udc.edu
University of the District of Columbia
4200 Connecticut Ave, NW, DC, 20008

Abstract: This paper discusses the prediction of Tropical Storm occurrences in a future time scale using Bayesian and frequentist approaches. Bayesian prediction limits are calculated using an informative prior based on the data on Atlantic Tropical Storms from 1851 to 1943. We adopt a prior from the Gamma family of distributions. Our sample includes occurrences of Atlantic Tropical Storms for the period of years 1944 to 2002. Bootstrap methods are used to estimate the prior distribution. Frequentist prediction limits are also derived. Bayesian Networks are used to investigate the causal relationships between storm factors and strength and damage.

Key words: Poisson; Storm; Bayesian; Prediction Intervals; Bayesian Networks

1. Introduction

Atlantic Tropical Cyclones are classified as Subtropical Storms, Tropical Storms and Hurricanes according to windspeed [1], [2]. Saffir and Simpson developed the Saffir-Simpson Hurricane Scale to further assign hurricanes into five categories based on windspeed [3] as presented in Table 1. Different levels of storm damage are associated with the different categories. Storms with windspeed below the minimum for a Category 1 Hurricane are considered Tropical Storms or Subtropical Storms [3].

In this paper, we construct prediction limits for Hurricane-level Atlantic Tropical Storms using both Bayesian and frequentist approaches. The problem is stated in section 2. In section 3 we describe the methodology using both Bayesian and frequentist approaches. In section 3.1, we address the Bayesian approach [4], [5]. Bootstrap sampling methods and numerical techniques are investigated. We restrict ourselves to prior distributions that belong to the Gamma family. Section 3.1 continues with historical data that record the number and intensity of Atlantic Tropical Cyclones per year for the period of years 1851-1943 are considered as prior information. These data pre-date aircraft reconnaissance. The predictive distribution is estimated as a conditional probability of the future outcome given the informative sample. The modified Jeffrey's and uniform distribution are considered as limiting cases of a gamma prior distribution. The Bayesian methodology applied in this paper and an algorithm for constructing prediction limits for a Poisson process using a frequentist approach is described in [6]. For further readings about predictive analysis we suggest [4] and [7].

An introduction to Bayesian Networks is presented in section 4. Conditional Probabilities of factors influencing storm strength and damage are used to investigate causality using Bayesian Networks.

2. Statement of the problem

Let F denote a future experiment, whose outcomes follow a Poisson distribution Po(tλ), and E denote an informative experiment, whose outcomes follow Po(sλ). Let Y be the random variable describing the number of occurrences of some phenomenon from experiment F during the future time interval with known length t, and X be the random variable describing the number of occurrences of some phenomenon from the informative experiment E during the given time interval s. Both random phenomena are considered independent and with the same unknown rate of occurrences λ.

We will construct a function (u(X) in frequentist approach, v^*(X) in Bayesian) that takes only integer values and that will serve as an upper bound for the values of the random variable Y. Construction of the lower prediction limit will be derived the same way.

It can be shown using probabilistic methods that Atlantic Tropical Storm occurrences from 1944 - 2002 follow a Poisson distribution [6]. This is the sample that we work with. Both frequentist and Bayesian limits are compared to the true storm occurrences in Table 3.

3. Methodology

We investigate methods of forecasting prediction intervals for annual Atlantic Storm occurrences using Bayesian, bootstrap sampling, and frequentist approaches. We also employ Bayesian Networks to evaluate causality of storm strengths based on conditional probabilities.

3.1 Prediction Using Bayesian Approach

In our application, we interpret the prior distribution as an expression of our state of knowledge about the parameter. Gelman et al [8] describe two basic interpretations about the prior distribution: the population of possible parameter values or our state of knowledge about the parameter. In setting up the prior distribution, we choose a class of distributions based on mathematical convenience, which for a Poisson distribution would be the conjugate class of gamma distributions, Gam(a,b). The predictive distribution, the distribution of the random variable Y|$\underline{X}$, is negative binomial NB(r,p) with probability of success p=(bns+1)/(bt+bns+1) and r=z+a, where an integer approximation is taken instead of a. We use this distribution to predict the annual occurrences of tropical storms.

Data that record the number per year of tropical cyclones that reached storm strength and hurricane strength for the period of years 1851 through 1943 are considered based on NOAA analysis with early records and with great uncertainty. The plot of the histogram with density for this data is shown in Figure 2. Based on these data, we estimate parameters of the gamma distribution and later on we update this prior with data that record the annual occurrences of Atlantic Tropical Cyclones for the period of years 1944 through 2002.

Based on the method of moments applied to the data set we adopt a gamma distribution with shape parameter a=4.583 and rate parameter 1/b = .642 (Gam(4.583, 1.558)) as our prior. Updating this information with sample data from 1944 to 2002 of size n=59, and z =

$\sum_{i=1}^{n} x_i = 598$, concludes in Gam(602.583, 0.017) as a posterior distribution for Λ, with the posterior mean of 10.244 that suggests an average of 10 tropical storms per year.

In Figure 3 are graphed 1000 simulated observations from the prior distribution that we found to follow the Gamma distribution estimated from the records over the period of years 1851-1943, while in Figure 4 are graphed 1000 simulated observations from the posterior distribution obtained after updating the prior information using the data 1944-2002. The predictive distribution would be a negative binomial NB(1017, p) with probability of success p=(0.017*59+1)/(0.017t+0.017*59+1).

3.2 Bootstrap Sampling Method

In this section we consider bootstrap methods using two numerical approaches, approximation to chi-square and approximation to negative binomial distribution, to estimate the prior distribution. In bootstrap methods the only probability mechanism considered is the one that fits data the best. This gives bootstrapping a practical advantage compare to theoretical methods that would require all the possible probability distributions for the observed data [9]. Using bootstrap, we resample samples of size m=93, same as the original dataset, and estimate from them parameters a and b of the assumed gamma prior.

Approximation to Chi-Square: When considering the approximation to Chi-Square distribution [10], parameters a and b of the assumed gamma prior are estimated as follows. The estimate of the annual rate of storms occurrences, $\hat{\lambda}$, that is computed from each of 1000 bootstrap samples of size m=93 yields a bootstrap sample of size B=1000. This is considered as a sample drawn from the population of Λ. The random variable W=2Λ/b would follow a gamma distribution with parameters a and 2, or chi-square distribution $\chi^2_{(2a)}$. The ratio of the 95% and 5% theoretical quantiles of the random variable W=2Λ/b is set equal to the ratio of the 95% and 5% empirical quantiles estimated from the bootstrap sample of Λ.

This procedure is repeated 100 times, resulting in 100 estimates of a and b, each calculated using numerical methods. Five possible prior distributions, corresponding to 0, .25, .5, .75, 1 percentiles of the calculated values for a, are graphed in Figure 5.

Approximation to Negative Binomial Distribution: The bootstrap procedure is performed and 100 samples of size m=93 are drawn. Then the sum of the 93 draws from each sample is taken. We denote these sums by Ψ_i, i={1,2,…,B}, where B is the number of bootstrap samples. The set of the Ψ_i, i={1,2,…,B} is considered as a sample drawn from the random variable Ψ, that conditioned on parameter Λ, follows a Poisson distribution with parameter λm. The rate parameter Λ is assumed to follow a gamma distribution. Hence, the marginal distribution of the random variable Ψ would be a negative binomial distribution NB(a,p), with probability of success p=1/(1+mb). Repeating this procedure 100 times and using the method of moments, we end up with 100 equations on a and b that are solved by numerical methods. In Figure 6 are graphed five possible prior distributions corresponding to 0, .25, .5, .75, 1 quantiles of the calculated values for a.

The noninformative priors that we considered are modified Jeffreys' prior and the uniform prior. Limits derived from noninformative priors are presented Table 2 [6].

3.3 Prediction Using the Frequentist Approach

We describe the method of constructing the lowest upper bound $u^*(X)$ of the future outcome Y with respect to some error probability α. Any function u(X) that takes only integer values, satisfies $u(X)\geq u^*(X)$, and has a probability of wrong prediction less than α, would be an upper limit.

The algorithm for constructing the function $u^*(X)$ follows by calculating first the joint probability function of X and Y:

$$\mathbf{p_{EF}(x,y|\lambda) = p_E(x|\lambda)p_F(y|\lambda) = \left[\frac{(\lambda s+\lambda t)^{x+y}e^{-(\lambda s+\lambda t)}}{(x+y)!}\right]\left[\frac{(x+y)!(\lambda s)^x(\lambda t)^y}{x!y!(\lambda s+\lambda t)^{x+y}}\right]} \qquad (1)$$

Conditionally, X given X+Y=r follows a binomial distribution, Bin(r,p).

The probability of the wrong coverage for an upper limit u(X), should not exceed α. For any function u(X),

$$\Pr\big(Y > u(X)\big) = \sum_{r=0}^{\infty}\left\{\frac{(\lambda s+\lambda t)^r e^{-(\lambda s+\lambda t)}}{r!}\Delta_r\right\} \qquad (2)$$

where r=x+y, p=s/(s+t), 1-p=t/(s+t), and $x \epsilon Z^+$, $y \epsilon Z^+$, and $r \epsilon Z^+$, $x \epsilon \{0,1,2,\ldots,r\}$, and

$$\mathbf{\Delta_r} = \sum_{\substack{\mathbf{y>u(x)} \\ \mathbf{x+y=r}}}\mathbf{\binom{r}{x}p^x(1-p)^{r-x}} \qquad (3)$$

The sum Δ_r gives the probability over all points (x,y) that satisfy both conditions x+y=r, and y>u(x (x<r-u(x)). This is the probability of the wrong coverage. It should not exceed some predetermined error α,

$$\mathbf{\Delta_r} = \sum_{\mathbf{x<r-u^*(x)}}\mathbf{\binom{r}{x}p^x(1-p)^{r-x} \leq \alpha} \qquad (4)$$

where Δ_r depends on both r and u(·).

For the inequality $\Pr\{Y>u(X)\} \leq \alpha$ to be satisfied, it is sufficient (from (4.2)) that $\Delta_r \leq \alpha$ for every integer $r\geq 0$. Therefore, we need to find an integer valued function u(·) that makes $\Delta_r \leq \alpha$, for all $r \geq 0$.

Among all integer valued functions u(X) that satisfy the latter condition we are interested in the function $u^*(X)$, which gives the smallest upper bound for Y that satisfies $\Pr\{Y>u^*(X)\} \leq \alpha$.

Denote by F(r,p,x) the cumulative distribution function of binomial, Bin(r,p). If we take $u^*(x) = \max\{r:F(r,p,x)>\alpha\} - x$, for every integer $x\geq 0$, then $\Delta_r \leq \alpha$, for every $r \geq 0$. This is true because of the fact that the set of points over what Δ_r is calculated, $\{(x,y):x+y=r \text{ and } y>u^*(x)\}$, agrees with the set of point such that $F(r,p,x) \leq \alpha$, $\{(x,y):x+y=r \text{ and } r>x+u^*(x)\}$. Note for some values of r, the range of summation may be empty hence $\Delta_r = 0$. See Table 4 and [6]. For the observed $r \in Z^+$ we find the value of Y using numerical methods by calculating first the maximum value of r (r_{max}), such that $F(r,p,x)>\alpha$.

Example 2: In Table 2 are shown the values of r_{max} and $u^*(x)$ for the case where $0 \leq x \leq 13$, $y \geq 0$, and α=.05. When the observed value is x=4, the maximum value of r, such that $F(r,p,4)>\alpha$, is $r_{max} = 10$. Take $u^*(4)=10-4=6$. For this case $\Delta_0 = \Delta_1 = \Delta_2$ are 0 as probabilities calculated in an empty set, while Δ_1 through Δ_{16} are all calculated to be less than .05. All other Δ_r were practically zero for all $r \geq 17$. SPLUS statistical software was used for calculations [16].

Table 3 compares forecasted numbers to actual numbers of Atlantic Tropical Storms using both Bayesian and the frequentist methods described in this paper. For 2003, all methods perform reasonably well. The unusually active 2005 storm season makes the actual figures significantly higher than forecast figures using all methods. Bayesian methods could have an advantage in this situation, allowing the effects of these unusual levels in subsequent forecasts to be lessened as each year's posteriors are considered as the next forecast's priors.

4. Causation using Bayesian Network

We investigate causality of Atlantic coastal property loss and damages using Bayesian Networks. Bayesian Networks allow analysis of a set of random variables using a directed acyclic graph (DAG) to represent conditional dependencies. The additional requirement that relationships represented be causal results in a Causal Network [11].

In our application, the evaluated random variables fall into three categories: 1) factors that affect storm strength: whether the storm is Cape Verde-type, the Sea Surface Temperature of the Caribbean Sea, and whether it is an El Niño year, 2) internal factors: Storm Category and County Population in the storm path, and 3) outcome factors: population loss percentage and County damage in dollars.

Storms that originate off the coast of Cape Verde tend to be stronger [12]. Higher Sea Surface Temperatures in the Caribbean Sea and Gulf of Mexico tend to strengthen storms as well [13]. Figure 7 shows the path followed by Cape Verde-type storms. It has been observed that El Niño winds tend to weaken Atlantic storms [10], [14], [15].

The factors to be evaluated are represented as conditional random variables in the nodes of a Bayesian Network, represented using the Netica Bayesian belief software package [17] shown in Figure 8. Initial conditional probabilities are associated with each node. To test beliefs represented in the Bayesian Network, we set one or more nodes to true and update beliefs. In Figure 9 [17], we fix characteristics of an individual storm (Cape Verde-type is TRUE, Caribbean Sea Surface Temperature is 29.4C, and El Niño year=FALSE) and then evaluate the conditional probabilities that the storm will have a particular category on the Saffir-Simpson Hurricane Wind Scale. Resulting probabilities are shown in Table 5.

5. Discussion

In this paper, we investigated methods of forecasting prediction intervals for annual Atlantic Storm occurrences using Bayesian and frequentist approaches. We utilized Bayesian Networks to evaluate causality of storm strengths based on conditional probabilities.

The influence of the data in the calculation of the Bayesian upper prediction limits is considerable, because of the fact that we have a relatively big sample size (n=59) for a rare event, and the sum of the observations is also large ($z = \sum_{i=1}^{59} X_i$), such that no difference appears among corresponding prediction limits derived from each noninformative prior (see Table 2), while we could show that this is not always true.

We constructed two different priors. The difference that appears in priors estimated based on Chi-Square and Negative Binomial approach results in different prediction limits at the end; It is of interest to mention that neither of them agrees with the frequentist ones. All prediction limits calculated so far, based on both frequentist and Bayesian approaches, are presented in Table 3.

This is ongoing work related to the first author's MS thesis. He intends to investigate further the use of Bayesian Networks to show causality between environmental factors and storm strength, with the possibility of extension to show other causal relationships [18].

References

[1] G. J. Holland, “Ready Reckoner” – Chapter 9, *Global Guide to Tropical Cyclone Forecasting*, World Meteorological Organization, Geneva, Switzerland, 1993

[2] “Hurricane Research Division”, Retrieved 30 September 2013 from http://www.aoml.noaa.gov/hrd

[3] Saffir, Herbert and Simpson, Robert. “Saffir-Simpson Hurricane Scale”, Retrieved 30 September 2013 from http://www.aoml.noaa.gov/general/lib/laescae.html, 1969.

[4] J. Aitchison and I. R. Dunsmore, *Statistical Prediction Analysis*, Cambridge University Press, 1975.

[5] E. L. Lehmann and George Casella, *Theory of Point Estimation* (Second Edition), Springer Texts in Statistics, Springer-Verlag, 1998

[6] Valbona Bejleri, *Prediction Intervals for the Poisson Model with Application to Atlantic Storms Data*, American University, 2005.

[7] G. J. Hahn and W. Nelson, “A Survey of Prediction Intervals and Their Applications”, Journal of Quality Technology, 5(1973) 178-188.

[8] Gelman, Andrew, Carlin, B. John, Stern, S. Hal, and Rubin, B. Donald (2004), *Bayesian Data Analysis*, (Second Edition), Chapman and Hall/CRC.

[9] Efron, Bradley (1987), "Better Bootstrap Confidence Intervals: Rejoinder", *Journal of the American Statistical Association*, Vol. 82, No. 397. pp. 198-200.

[10] Elsner, James B. and Bossak, Brian H. (2001) “Bayesian Analysis of U.S. Hurricane Climate”, *Journal of Climate*, Volume 14, 4341-4350.

[11] Korb, Kevin B and Nicholson, Ann E, *Bayesian Artificial Intelligence*, Chapman and Hall/CRC, 2004.

[12] “Hurricane Research Division Frequently Asked Questions”, Retrieved 2 October 2013 from http://www.aoml.noaa.gov/hrd/tcfaq/A2.html

[13] “National Hurricane Center Reynolds SST Analysis”, Retrieved 2 October 2013 from http://www.nhc.noaa.gov/aboutsst.shtml

[14] “Interaction with El Niño”, Retrieved 2 October 2013 from http://ww2010.atmos.uiuc.edu/(Gh)/guides/mtr/hurr/enso.rxml

[15] Bell, G. D., and M. Chelliah, 2006: Leading tropical modes associated with interannual and multi-decadal fluctuations in North Atlantic hurricane activity. *Journal of Climate*. 19, 590-612.

[16] Tibco Corporation (2005). S-Plus statistical computing software v7.0. URL http://www.tibco.com/.

[17] Norsys Software Corp (2013). Netica Bayesian belief software v5.12. URL http://www.norsys.com/netica.html/.

[18] Stewart et al. “To fund or not to fund: Using Bayesian Networks to Make Decisions About Conserving Our World’s Endangered Species ”, *Change Magazine*. Retrieved from http:/chance.amstat.org/2013/09/2-pitchforth.

Figures

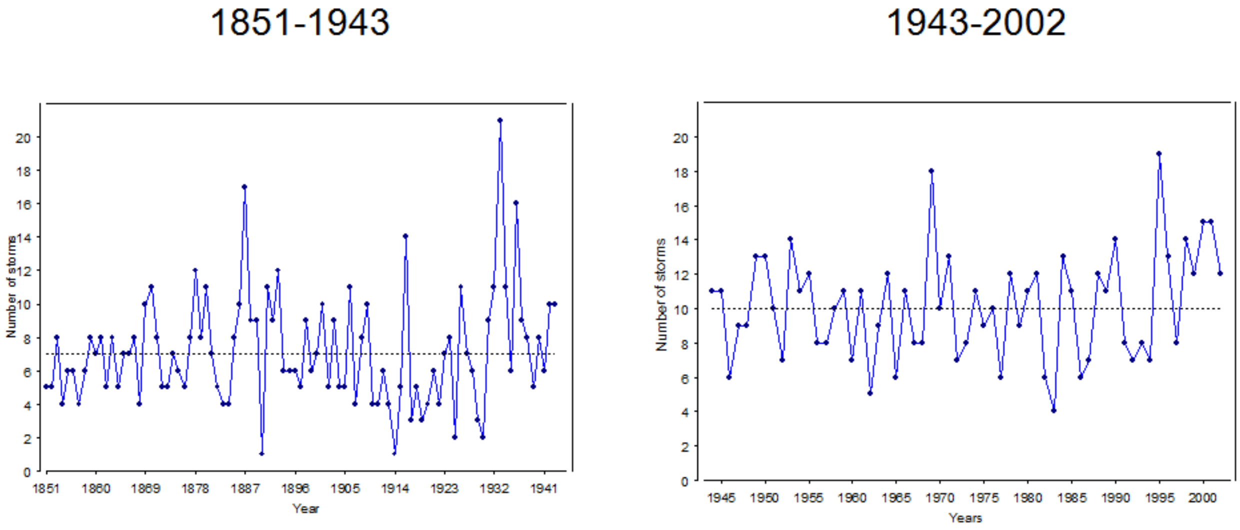

Figure 1. North Atlantic tropical storm occurrences 1851-2002

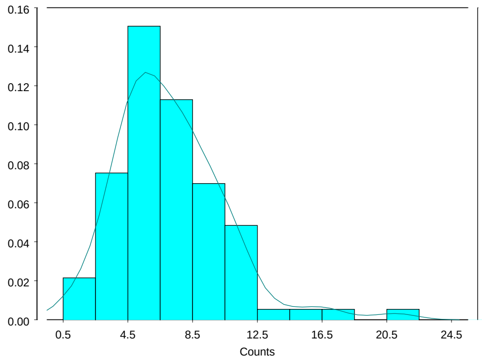

Figure 2. Histogram with density for annual occurrences of Atlantic Tropical Cyclones during 1851-1943

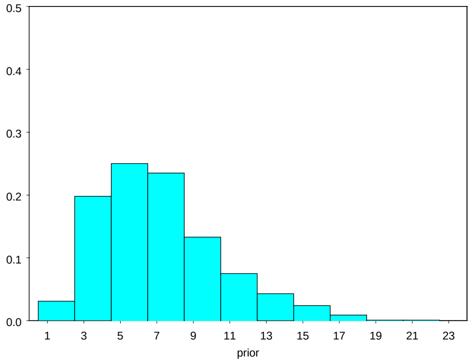

Figure 3. Histogram of the 1000 simulated observations from gamma prior

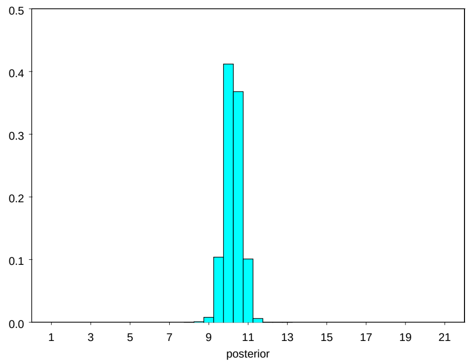

Figure 4. Histogram of the 1000 simulated observations from posterior distribution

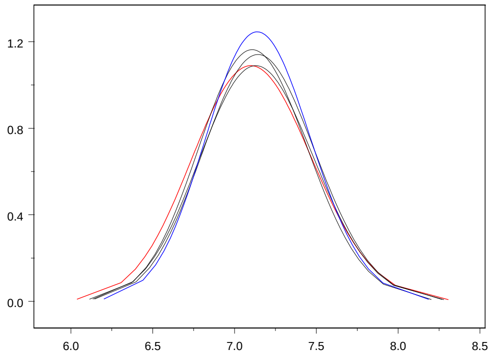

Figure 5. Density graph of possible priors with parameters estimated using the Chi-Square approach

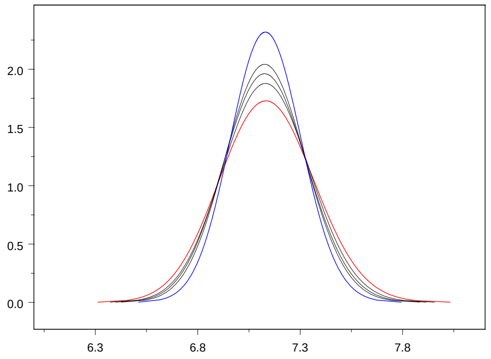

Figure 6. Density graph of possible priors with parameters estimated by using the Negative Binomial approach

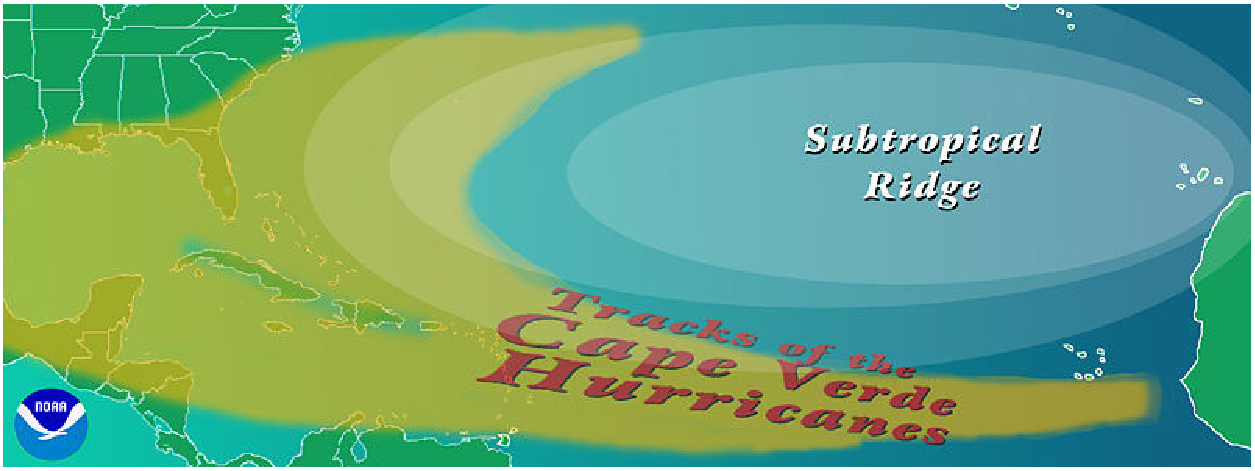

Figure 7. Tracks of the Cape Verde Hurricanes [12]

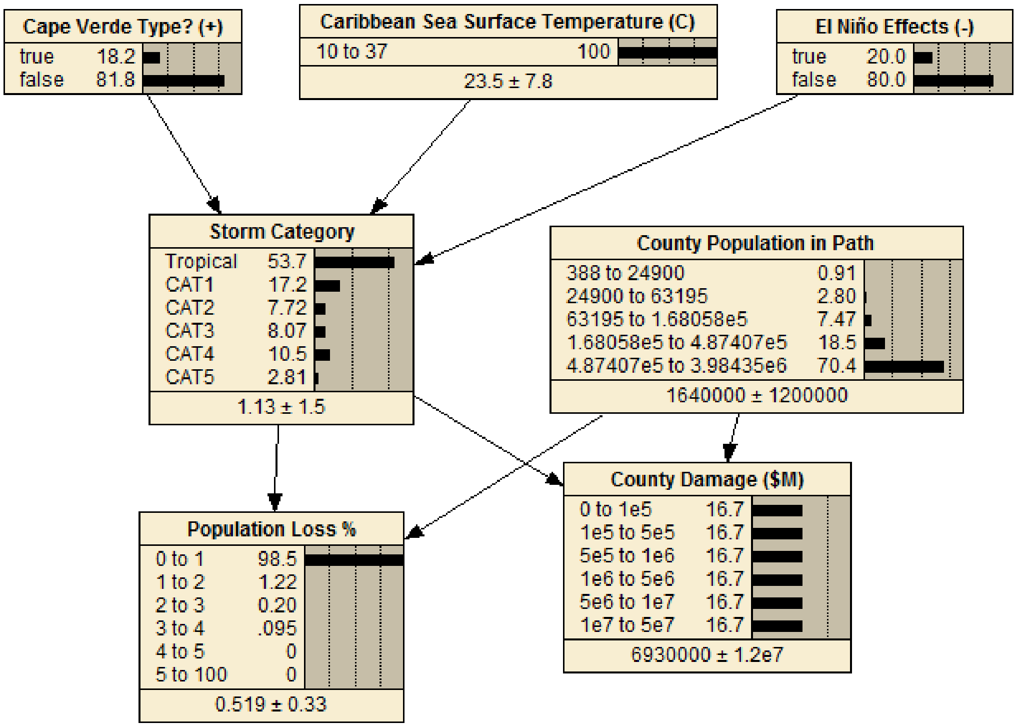

Figure 8. Bayesian Network (initial state)

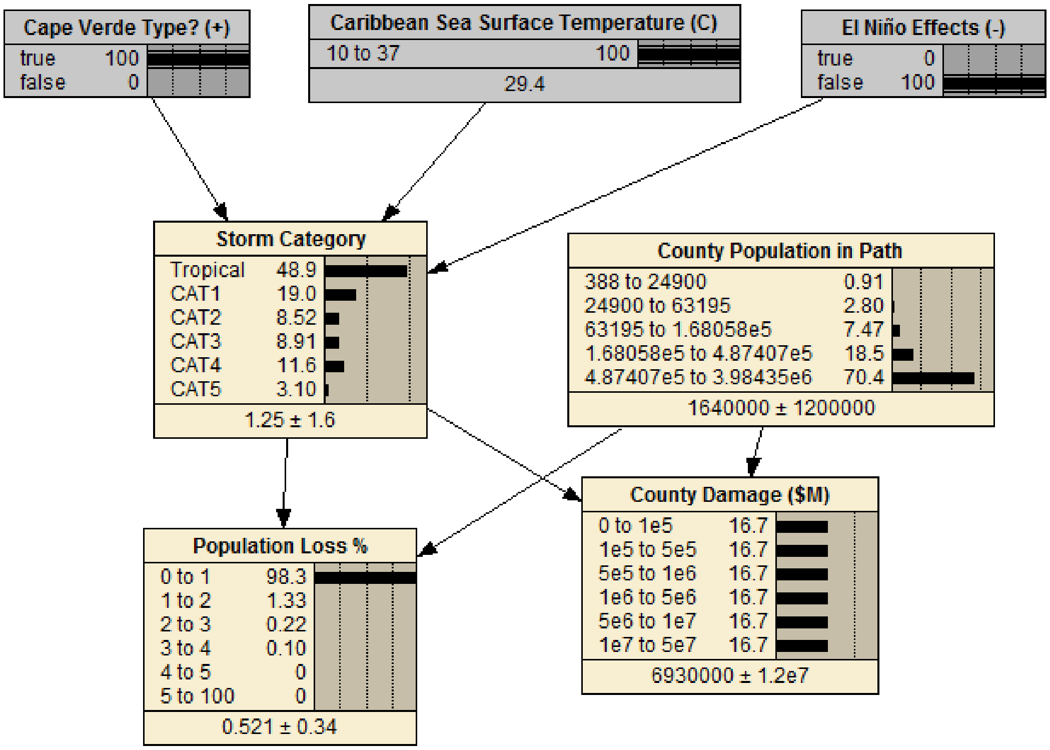

Figure 9. Bayesian Network with updated beliefs

Tables

Table 1. Saffir-Simpson Hurricane Scale

Category	**Winds**
One	74-95 mph
Two	96-110 mph
Three	111-130 mph
Four	131-155 mph
Five	greater than 155 mph

Table 2. Prediction Using the Frequentist Approach

x	0	1	2	3	4	5	6	7	8	9	10	11	12	13
r_{max}	2	4	6	8	10	12	4	15	17	19	21	22	24	26
$u^*(x)$	2	3	4	5	6	7	8	8	9	10	11	11	12	13

Table 3. Prediction limits derived based on both frequentist and Bayesian approaches

Method	Period	2003	2004	2005	2005-2006
	t	1	2	5	10
Actual Atlantic Hurricane Occurrences		**16**	**15**	**28**	**38**
Frequentist	Lower	6	9	9	19
	Upper	16	21	22	40
Hist.Dat.Inf Method of Moments	Lower	6	7	8	18
	Upper	16	20	21	39
1/λ	Lower	6	7	7	19
	Upper	16	20	21	39
Uniform prior	Lower	6	7	7	18
	Upper	16	20	21	39

Table 4. Poisson upper prediction limit

r\x	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	NA	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0.333	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0.111	0.556	1	0	0	0	0	0	0	0	0	0	0	0
3	0.037	0.259	0.704	1	0	0	0	0	0	0	0	0	0	0
4	0.012	0.111	0.407	0.802	1	0	0	0	0	0	0	0	0	0
5	0.004	0.045	0.21	0.539	0.868	1	0	0	0	0	0	0	0	0
6	0.001	0.018	0.1	0.32	0.649	0.912	1	0	0	0	0	0	0	0
7	0	0.007	0.045	0.173	0.429	0.737	0.941	1	0	0	0	0	0	0
8	0	0.003	0.02	0.088	0.259	0.532	0.805	0.961	1	0	0	0	0	0
9	0	0.001	0.008	0.042	0.145	0.35	0.623	0.857	0.974	1	0	0	0	0
10	0	0	0.003	0.02	**0.077**	0.213	0.441	0.701	0.896	0.983	1	0	0	0
11	0	0	0.001	0.009	**0.039**	0.122	0.289	0.527	0.766	0.925	0.988	1	0	0
12	0	0	0.001	0.004	**0.019**	0.066	0.178	0.368	0.607	0.819	0.946	0.992	1	0
13	0	0	0	0.002	**0.009**	0.035	0.104	0.241	0.448	0.678	0.861	0.961	0.995	1
14	0	0	0	0.001	**0.004**	0.017	0.058	0.149	0.31	0.524	0.739	0.895	0.973	0.997
15	0	0	0	0	**0.002**	0.009	0.031	0.088	0.203	0.382	0.596	0.791	0.921	0.981
16	0	0	0	0	**0.001**	0.004	0.016	0.05	0.127	0.263	0.453	0.661	0.834	0.941
17	0	0	0	0	**0**	0.002	0.008	0.027	0.075	0.172	0.326	0.522	0.719	0.87
18	0	0	0	0	**0**	0.001	0.004	0.014	0.043	0.108	0.223	0.391	0.588	0.769
19	0	0	0	0	**0**	0	0.002	0.007	0.024	0.065	0.146	0.279	0.457	0.648
20	0	0	0	0	**0**	0	0	0.004	0.013	0.038	0.092	0.191	0.339	0.521
21	0	0	0	0	**0**	0	0	0	0.007	0.021	0.056	0.125	0.24	0.399
22	0	0	0	0	**0**	0	0	0	0	0.012	0.033	0.079	0.163	0.293
23	0	0	0	0	**0**	0	0	0	0	0	0.019	0.048	0.107	0.206
24	0	0	0	0	**0**	0	0	0	0	0	0	0.028	0.068	0.14
25	0	0	0	0	**0**	0	0	0	0	0	0	0	0.042	0.092
26	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0.058

Table 5. Bayesian Network updated beliefs

Category	P(Category | CapeVerde=T, SST=29.4C, ElNiño=F)
TS	0.48873
CAT1	0.18979
CAT2	0.085211
CAT3	0.089085
CAT4	0.1162
CAT5	0.030986