IJAER/January-February 2013/Volume-2/Issue-1 ISSN: 2278-9677
International Journal of Arts & Education Research

DISTRIBUTIONS AND CONSEQUENCES FOR SIMULATED AND REAL-WORLD DATASETS

Sirajuddeen P K
Assistant Professor
Department Of Statistics
Govinda Pai Memorial Govt College, Manjeswar.
Kasaragod, Kerala

Abstract

The Exponential-H family (Ex-H) of distributions is a versatile expansion of the standard exponential distribution that makes use of the generating function H(x) to describe a wide range of data patterns. The Ex-H family is very adaptable for applications in survival analysis, reliability engineering, and risk management because of this generality, which allows it to support a variety of hazard rate behaviors, such as growing, decreasing, and bathtub-shaped hazard functions. Without having to transition between multiple distributional frameworks, the Ex-H family can be parameterized to include well-known distributions like the Weibull, Gamma, and Generalized Exponential distributions, providing versatility. It can represent uncertainty and unpredictability in complicated datasets while preserving desired characteristics like mathematical tractability thanks to its versatility. The New Flexible Exponential (NFE) distribution, a specific example of the Ex-H family, is presented. The NFE distribution takes its form from the Weibull distribution's hazard function. For both monotonically and non-monotonically growing hazard rate functions, the NFE distribution offers a versatile model. The discussion includes graphical representations of the analytical conclusions, which include the NFE distribution's cumulative distribution function (PDF), survival function, hazard rate function, and probability density function (PDF). Additionally, the quantile function, median, and r-th moments of the NFE distribution are obtained, offering helpful resources for additional statistical investigation..

Keywords: *Data Distributions, Simulated Datasets, Real-World Data, Statistical Consequences, Comparative Analysis*

1. INTRODUCTION

In the field of data analysis, knowing how simulated and real-world datasets differ from one another is essential to getting the right answers and making wise choices. It is common practice to test theories, validate models, or forecast results in controlled settings using simulated data, which is frequently produced using mathematical models or algorithms. On the other hand, real-world datasets are derived from factual observations and capture the complexity, inconsistent nature, and noise inherent in real-world situations. The two categories of datasets display unique distributions, and these distributions have important implications for how data is interpreted, how well models function, and how decisions are made.

1.1. Understanding Data Distributions

Discusses the different kinds and features of real-world and simulated dataset distributions, including as skewed, uniform, and normal distributions.

1.2. Simulated Datasets: Methodology and Benefits

Demonstrates the benefits of using simulated data for controlled studies, such as ability to manipulate variables and be repeatable and scalable.

1.3. Real-World Datasets: Complexity and Challenges

Explains the difficulties of dealing with real-world datasets, like noise, outliers, and missing values, which frequently affect how accurate models are.

1.4. Comparing Consequences of Simulated and Real-World Data

Examines the impact that variations in data distribution between simulated and real-world data can have on the results of machine learning models, statistical analysis, and predictions.

1.5. Implications for Model Development and Testing

Explores how mixing both forms of data might improve resilience and the effects of these distributions on model validation, training, and generalization.

For applied data science and decision-making, this analysis emphasizes how crucial it is to understand the underlying variations in dataset types and their implications.

2. LITERATURE REVIEW

Jiménez-Valverde et al. (2009) accepted that the presence/absence ratio in the training data, or prevalence, affects how reliable the predictions made by species distribution models are. On the other hand, little is known regarding its exact effect. In order to prevent unaccounted-for effects (false absences, non-explanatory predictors, etc.) from occurring during the modeling process, we examined its consequences using a virtual species. After sampling the virtual species' distribution to create a number of data subsets with different sample sizes and prevalence, we used logistic regressions to model the data. Our findings demonstrate that, given that the predictors are genuinely connected to the species' distribution and the training data are trustworthy, model predictions can be extremely accurate across a broad range of sample sizes and prevalence scores. Only datasets with severely imbalanced samples (<0.01 and >0.99) show a substantial effect from prevalence, while datasets with fewer than 70 data points clearly show the effect of sample size. Additionally, there is a significant interaction between sample size and prevalence,

suggesting that, contrary to what was previously believed, the most detrimental element is the sample size of each event—whether it be present or absent. We propose that the quality of the training data and the sample size of each event must interact in the actual world. We address the possibility that biased prevalences are a desired feature of the data rather than a concern that needs to be avoided, and we emphasize the significance of simulating the distribution of species with limited geographic ranges using the best absence data available.

Brath, et al. (2004) attracted the attention of both academics and practicing hydrologists for the purpose of applying hydrological models with spatial distributions. Calibration of some model parameters is usually still required when using physically-based methods, and a comprehensive investigation of calibration-related issues has often been impossible due to the complexity of distributed models, which requires very demanding computations. Examining a series of trials that have been automatically calibrated using a distributed hydrologic model that is conceptualized and continually simulates using the Shuffled Complex Evolution method is the goal of this work. The data used for calibration and validation are actual observations of precipitation and discharge from a watershed in the Apennine Mountains of Italy that is medium-sized (1050 km2) and extensively vegetated. Significant floods that transpired throughout the 1990s–2000s are replicated using parameters derived from the rainfall–runoff model's calibration with reference to various scenarios of available historical data. An efficient parameterization requires a certain length of time for calibration, which is investigated in a first set of experiments. The second investigation, which modifies the size and dispersion of the raingauge network, focuses on the impact of the rainfall input's spatial resolution on model calibration. A third facet pertains to evaluating the dependability of model parameters in replicating the outflow in unregulated river segments. The study's objective is to give the user guidance on how to choose the historical data set that will be best used to calibrate the model. The findings show that the rainfall-runoff model performances appear to be considerably worsened when the calibration period is shortened by three months. Assuming uniform rainfall across space, model simulations are also satisfactory when calculating the mean areal rainfall intensity using a large enough number of raingauges. However, things quickly go downhill when the density of the raingauge network is drastically reduced. Lastly, it has been demonstrated that the distributed model can produce accurate simulations for ungauged internal river portions.

Zeng, X., & Martinez, T. R. (2000) has frequently used cross-validation to estimate classifier accuracies. This paper proposes an addition to this technique, distribution-balanced stratified cross-validation (DBSCV), which provides balanced intraclass distributions when splitting a data set into different folds, hence improving the estimation quality. Using the C4.5 decision trees classifier, we have tested DBSCV on nine artificial and real-world domains. The findings indicate that DBSCV outperforms ordinary

stratified cross validation in most situations (i.e., has less biases), particularly when the number of folds is modest. DBSCV is most effective when a data set has several intraclass clusters, according to the research and experiments conducted on three simulated data sets.

Jayaraman, V., & Ross, A. (2003) supplied the PLOT (Production, Logistics, Outbound, and Transportation) system for design. The system addresses a class of distribution network design problems that include multiple product families, a central manufacturing plant site, multiple distribution centers and cross-docking locations, and retail outlets (client zones) that require multiple units of various commodities. There are two primary parts to the finished system. In the first, planning, we use a strategic decision-making process to select the "best" combination of distribution centers and cross-docks to operate. The execution stage, which comprises an operationally driven decision-making process, makes up the second phase. During this stage, the model determines how many product families must be moved from the factory to distribution centers, then transferred from warehouses to cross-docks, and finally delivered to retail locations. The architecture of the distribution system under consideration is based on the way a large retailing company now manages its products for distribution across the country. A high level of user participation is possible in the solution generating process thanks to the PLOT system that was created to put the model into practice. Using the simulated annealing (SA) technique, the entire system produces globally feasible, almost optimal distribution system design and utilization plans. There are two significant additions to the SA literature that this review made. First, by examining a novel combinatorial issue that integrates cross-docking in a supply chain setting, the scope of applications was expanded. Secondly, we assess the computational performance in a methodical manner across multiple issue scenarios and SA control parameter configurations.

3. EXPONENTIAL- H FAMILY (EX-H) OF DISTRIBUTIONS

A more extensive range of forms is incorporated into the basic exponential distribution through the use of a generator function, resulting in the extensible and generalized Exponential-H family (Ex-H) of statistical distributions. Modeling a wide range of data types is made possible by the fact that this family comprises as special cases a number of well-known distributions. When the failure rates or hazard function need to be represented with more flexibility than the conventional exponential distribution provides, the Ex-H family is frequently employed in survival analysis, reliability engineering, and risk management. A function, H(x), which acts as the generating function, is applied to transform a baseline exponential distribution in this extended form. Thanks to this modification, the Ex-H family may support several types of tail behavior and hazard rates, including bathtub-shaped, declining, and increasing hazard functions. Depending on the generator function selected, the Ex-H family can be parameterized to include the

Weibull, Gamma, or Generalized Exponential distributions as special cases. Since it can simulate various patterns in data without requiring the user to transition between completely distinct distributional frameworks, the Ex-H family is highly flexible. Thanks to its versatility, it is especially helpful in the health sciences, engineering, and economics—fields that demand accurate modeling of uncertainty and unpredictability. Additionally, the Ex-H family retains a number of advantageous characteristics, such as mathematical analytical tractability, which makes it valuable for inferential tasks like maximum likelihood estimation, goodness-of-fit testing, and Bayesian model creation. To further improve its usefulness in real-world circumstances where the assumptions of simpler distributions, such as the basic exponential, may not apply, the family additionally allows a wide range of forms for probability density and cumulative distribution functions. For statisticians and academics wishing to apply adaptable, yet mathematically sound, models to complicated datasets, the Ex-H family has thus grown in importance.

Allocations Mainly connected to the Weibull-G family of distributions is the Exponential-H family (Ex-H). The Exponential-H family (Ex-H)'s cumulative distribution function (or CDF) looks like this.

$$G(x, a, \zeta) = 1 - exp(-aL(x;\zeta)), \quad x, a > 0$$

Where, $L(x;\zeta) = H(x;\zeta)exp(x)$, and $H(x;\zeta)$ and In terms of the parameter vector z, the non-decreasing function hazard rate function is denoted as H(x;z). Probability density function that corresponds.

$$g(x, a, \zeta) = aexp(-aL(x;\zeta))l(x;\zeta), \quad x, a > 0$$

4. NEW FLEXIBLE EXPONENTIAL DISTRIBUTION (NFE)

In this section, the hazard function of the Weibull distribution is used to show the Ex-H family special case. The Weibull distribution's hazard function is determined by

$$H(x;\zeta) = ax^{b-1}$$

We were able to derive the NFE distribution's CDF and PDF by utilizing the previously mentioned outcome.

$$F(x, a, b) = 1 - exp\left(-a^2 bx^{(b-1)}exp(x)\right), \qquad x > 0, b > 1, a > 0$$

$$f(x) = a^2 bx^{(b-2)}(x + b - 1)exp\left(x - a^2 bx^{b-1}exp(x)\right), \qquad x > 0$$

The definition of the NFE's survival and hazard rate function is

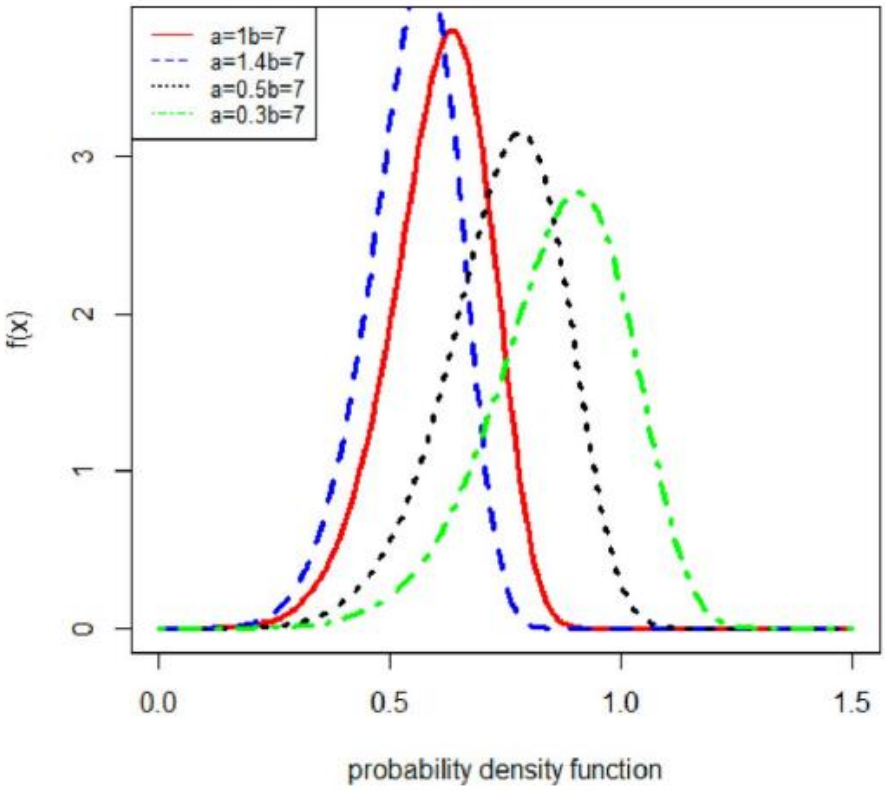
Fig 1 displays the cumulative distribution function and probability density function graphically, with various parameter values.

Theorem 1. As a function of NFE (a,b), the hazard rate h(x) behaves as defined by

a. Increasing when a>0, b>1,
b. Decreasing when a>0, b<1,

Proof. The derivative of Eq (3.)

$$h'(x) = a^2 b x^{b-2}(b^2 + b(2x-3) + x^2 - 2x + 2)exp(x)$$



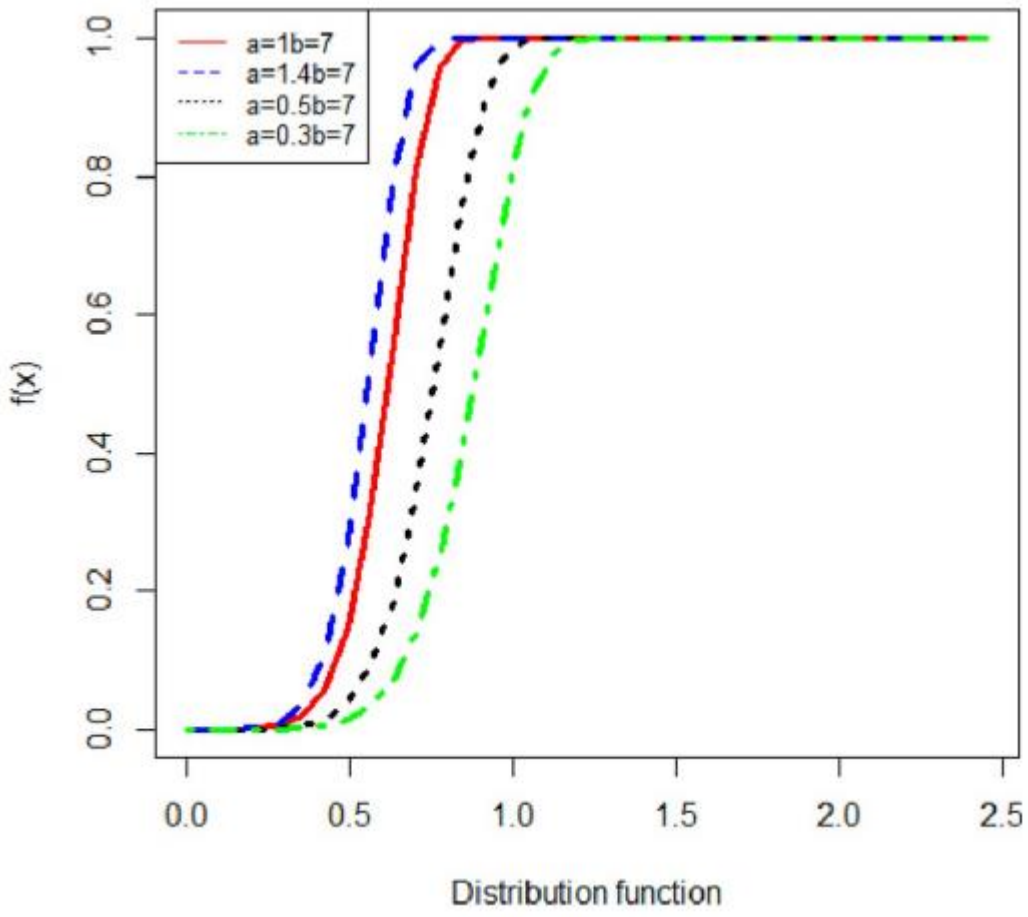


Fig 1. The Pdf and Cd of NFE

Let b1=h0(x) for every a>0. Once a>0 and b>1, we find that h(x) is a declining function. If h0 (x) = 0 for all a and b greater than 0, then the greatest value of h(x) is at

X= 1-b $\pm\sqrt{b} - \sqrt{1}$

and, the function h(x) is increasing for a>0, b>1,

As a result, the hazard rate function can represent hazard rate functions that are monotonically and nonmonotonically different.

In Figure 2, we can see the hazard function plotted against several parameter values for the New Flexible Exponential distribution.

4.1. Quantile function and median

The true solution to the following equation is the quantile function Q(NFE)(x) of the NFE(a,b).

$$F(x) = u$$

$$1 - exp\left(-a^2 b x^{b-1} exp(x)\right) = u$$

where u~Uniform (0,1).

Solving (4.1) for x, we have

$$x = (b-1)W\left(\frac{\left(\frac{-\log(1-u)}{a^2 b}\right)^{1/(b-1)}}{b-1}\right)$$

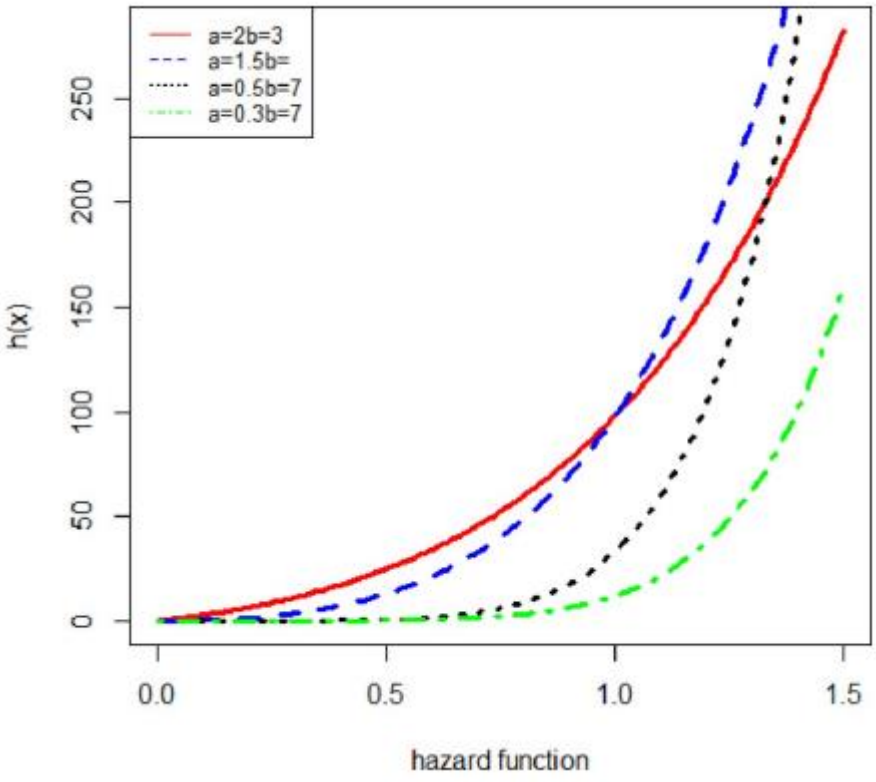


Fig 2. Hazard function of NFE

which is defined as the Lambert W function, denoted as W (Z)

$$W(z) = \frac{\sum_{n=1}^{\infty} (-1)^n n^{n-2}}{(n-1)!} z^n.$$

For the median, put u = 0.5

R^th moments

Theorem 2: A random variable X with parameters a and b and an NFE distribution is characterized by its r-th moments, which are specified around the origin, as

$$u_r' = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} (a^2 b)^{k+1} \left[\frac{\Gamma(r + b + bk - k)}{(k+1)^{r+b+bk-k}} + (b-1) \frac{\Gamma(r + b + bk - k - 1)}{(k+1)^{r+b+bk-k-1}} \right]$$

Proof. We know that

$$u_r' = E(x^r) = \int_0^{\infty} x^r f(x) dx$$

The following form was derived by substituting (3.2) into the above formula.

$$u_r' = \int_0^{\infty} \left(x^r a^2 b x^{(b-2)} (x + b - 1) exp\left(x - a^2 b x^{b-1} exp(x)\right)\right) dx$$

$$= \int_0^{\infty} \left(x^{r+b-1} a^2 b exp(x - a^2 b x^{b-1} exp(x))\right) dx + b - 1 \int_0^{\infty} \left(x^{r+b-2} a^2 b exp(x - a^2 b x^{b-1} exp(x))\right) dx \quad (1)$$

Solving the aforementioned expression's first component

$$= \int_0^{\infty} \left(x^{r+b-1} a^2 b exp\left(x - a^2 b x^{b-1} exp(x)\right)\right) dx$$

$$= \int_0^{\infty} \left(x^{r+b-1} a^2 b exp(x) exp\left(-a^2 b x^{b-1} exp(x)\right)\right) dx$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} (a^2 b)^{k+1} \int_0^{\infty} \left(x^{r+b+bk+k-1} exp((k+1)x)\right) dx$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} (a^2 b)^{k+1} \frac{\Gamma(r + b + bk - k)}{(k+1)^{r+b+bk-k}} \quad (2)$$

Now solving the second part

$$= b - 1 \int_0^{\infty} \left(x^{r+b-2} a^2 b exp\left(x - a^2 b x^{b-1} exp(x)\right)\right) dx$$

$$= \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} (a^2 b)^{k+1} (b-1) \frac{\Gamma(r+b+bk-k-1)}{(k+1)^{r+b+bk-k-1}}$$

Combining the two expressions (1) and (2) yielded the desired outcome.

$$u_r' = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} (a^2 b)^{k+1} \left[\frac{\Gamma(r+b+bk-k)}{(k+1)^{r+b+bk-k}} + (b-1) \frac{\Gamma(r+b+bk-k-1)}{(k+1)^{r+b+bk-k-1}} \right]$$

5. CONCLUSION

The New Flexible Exponential (NFE) distribution, in particular, belongs to the Exponential-H family (Ex-H) of distributions, which provides a flexible and strong framework for modeling a variety of real-world data with different hazard rate patterns. With the use of a generator function, the Ex-H family expands on the fundamental exponential distribution, making it possible to depict intricate data patterns in a flexible way without having to switch between different distribution families. A special instance of the Ex-H family, the NFE distribution offers an improved means of modeling both decreasing and growing hazard rate functions, which makes it very useful in survival analysis and reliability engineering. The generated quantile function, median, and r-th moments enhance the NFE distribution's mathematical usefulness and increase its suitability for a variety of statistical investigations. The Ex-H family's adaptability and mathematical tractability guarantee that it will always be relevant for analyzing complicated datasets in a variety of fields.

REFERENCES

1. *Abdinnour-Helm, S. (2001). Using simulated annealing to solve the p-Hub Median Problem. International Journal of Physical Distribution & Logistics Management, 31(3), 203-220.*
2. *Barreto-Souza W., & Cribari-Neto F. A generalization of the exponential-Poisson distribution. Statistics & Probability Letters, 79(24), 2009, 2493–2500.*

3. *Brath, A., Montanari, A., & Toth, E. (2004). Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model. Journal of Hydrology, 291(3-4), 232-253.*

4. *Chen F., & Chen S. (2002). Injury severities of truck drivers in single-and multi-vehicle accidents on rural highways. Accident Analysis & Prevention, 43(5), 1677–1688.*

5. *Chen F., Chen S., & Ma X. (2006). Crash frequency modeling using real-time environmental and traffic data and unbalanced panel data models. International journal of environmental research and public health, 13(6), 609.*

6. *Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009, July). Nus-wide: a real-world web image database from national university of singapore. In Proceedings of the ACM international conference on image and video retrieval (pp. 1-9).*

7. *Dong B., Ma X., Chen F., & Chen S. (2008). Investigating the differences of single-vehicle and multivehicle accident probability using mixed logit model. Journal of Advanced Transportation, 2008.*

8. *Gonder, J., Markel, T., Simpson, A., & Thornton, M. (2007). Using GPS travel data to assess the real-world driving energy use of plug-in hybrid electric vehicles (PHEVs) (No. NREL/CP-540-40858). National Renewable Energy Lab. (NREL), Golden, CO (United States).*

9. *Jayaraman, V., & Ross, A. (2003). A simulated annealing methodology to distribution network design and management. European Journal of Operational Research, 144(3), 629-645.*

10. *Jiménez-Valverde, A., Lobo, J., & Hortal, J. (2009). The effect of prevalence and its interaction with sample size on the reliability of species distribution models. Community Ecology, 10(2), 196-205.*

11. *Liu, X., Shackleton, M. B., Taylor, S. J., & Xu, X. (2007). Closed-form transformations from risk-neutral to real-world distributions. Journal of Banking & Finance, 31(5), 1501-1520.*

12. *McAuliffe, J. D., Blei, D. M., & Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. Statistics and Computing, 16, 5-14.*

13. *McComas. (2003, December). How the ExpertFit distribution-fitting software can make your simulation models more valid. In Proceedings of the 2003 Winter Simulation Conference, 2003. (Vol. 1, pp. 169-174). IEEE.*

14. *Newman, M. E., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. Physical review E, 64(2), 026118.*

15. *Shiode, N., & Batty, M. (2000). Power law distributions in real and virtual worlds.*

16. *Stuart, J. S., & Binzel, R. P. (2004). Bias-corrected population, size distribution, and impact hazard for the near-Earth objects. Icarus, 170(2), 295-311.*

17. *Wild, C. H. R. I. S. (2006). The concept of distribution. Statistics Education Research Journal, 5(2), 10-26.*

18. *Zeng Q., Guo Q., Wong S. C., Wen H., Huang H., & Pei X. (2005). Jointly modeling area-level crash rates by severity: a Bayesian multivariate random-parameters spatio-temporal Tobit regression. Transportmetrica A: Transport Science, 15(2), 1867–1884.*

19. *Zeng, X., & Martinez, T. R. (2000). Distribution-balanced stratified cross-validation for accuracy estimation. Journal of Experimental & Theoretical Artificial Intelligence, 12(1), 1-12.*

20. *Zhang, C., Florêncio, D., Ba, D. E., & Zhang, Z. (2008). Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings. IEEE Transactions on Multimedia, 10(3), 538-548.*

****