An Empirical Copula for Modeling 3rd Party Airport Risk using Equal Probability Defined Marginal Distributions, with an External Parameter to Accommodate Agent-specific Behavior

Christian M. Salmon

Abstract

This paper describes the development of a unique empirical copula that integrates two marginal distributions via a dependent relationship through the inclusion of an external parameter. This allowed a body of sparsely distributed spatial data to be represented in a parametric model that was used to simulate the relative crash location distribution of general aviation operations inbound to and outbound from non-towered airports.

The relevance of this work is first towards developing a 3rd party external airport risk model for non-towered airports. Second is the observation that in an era of "big data", wherein there is the perception that all that is required to analysis any system, is a very large dataset, it is instructive to demonstrate that a large body of data might yield only a minute percentage of specific data points that can be validated as representative of fact. Further, any assumption pertaining to data quality, and the ability of "big data" to identify biases or distorted data through a Law of Large Numbers, though tempting, would fundamentally alter (distort) underlying model results. And lastly, the 'big data' movement can mask the modeler from the underlying nature of the system being model, increasing the likelihood that results of any analysis can be distorted.

KEYWORDS:
Aviation, 3rd Party Risk, Uncertain Data

Introduction:
This paper presents the development and use of a unique empirical copula for modeling a crash location distribution of General Aviation (GA) operations at domestic public-use, no towered airports.

The underlying data are sourced in a large objective dataset that has the potential for thousands of relevant data points that could be used to populate the model. A detailed review of individual data points, however, indicated that the majority were corrupted by poor data collection, data input and formatting errors (Salmon and Motevalli 2010). This resulted in only a small body of data being validated to within a reasonable certainty. The location of 112 validated crash sites over 20-years at 4,500 airports were spatially distributed over a 40 square kilometer area around a unit airport.

Previous research has used similar empirical crash site data in developing crash distribution sub-models that are part of larger 3rd party risk models at commercial services airports in Europe (Evans et al, 1996, Brady and Hillestad, 1995).

Initially, the sparse distribution of validated data for the current project created a challenge for fitting a copula to the data in a manner that would yield statistically significant results. Specifically, no copula or series of bivariate distributions could be fitted to the data.

A series of interviews with persons knowledgeable about the nature of GA operations suggested that an external, non-intuitive parameter might be added to the model in order to accommodate peculiarities in pilot behavior that are unique to the non-towered airport environment. This resulted in an elegant extension of the standard modeling methodologies for fitting copulas to spatial datasets in the context of external airport crash risk (Evans et al, 1996, Brady and Hillestad, 1995). The result was a statically significant representation of the objective data as an empirical copula for use in the simulating crash site distribution of GA operations.

The resulting model and modeling process is instructive for two reasons. First, it demonstrates that in an era of "Big Data", access to a very large body of data might yield only a minute percentage of data points that can be validated as representative of fact. In the case of the accident data utilized in this research, any assumption that raw data were accurate, though tempting, would have fundamentally altered (or distort) the underlying model results.

Second, access to a large dataset has the potential to function as a barrier between the modeler and the underlying system's operating characteristics. For the research presented here, it was only after a deeper understanding of the underlying system and system agents (pilots) was sought that an external parameter was identified and leveraged to successfully develop the model for simulation.

Discussion:
3rd party risk is a metric commonly used when quantifying impact of a system on society or the individual, whether as a specific metric of risk or as an assessment of impact relative to some acceptable threshold. If the aggregate risk breaches some predefined threshold, then the exposure might be deemed unacceptable, resulting in a risk management effort to reduce the risk.

In aviation, particular airport operations, 3rd party risk is defined as the probability of an individual located outside the boundaries of an airport being killed as a direct result of an aircraft accident (Evans et al, 1996). In practice, this yields to risk being the probability of an aircraft 'crash' being located at some particular location, and under the assumption of total destruction within some diameter of the crash site.

The typical 3rd party airport risk model integrates three sub-models: Crash Rate, Crash Distribution, and Crash Consequence (Evans et al, 1996). The process is well established in European civil aviation systems, and has had some implementation at US commercial service airports (Solomon et al, 1974).

Previous research sought to adapt these processes to General Aviation (GA) operations at US non-towered airports [Salmon and Motevalli 2010]. This specific sector of civil operations was of interest because the dominate number of operations, accidents and fatalities in domestic civil aviation operations involves "Part 91" GA operations (NTSB, 2014). Further, there is evidence that the accident and fatal accident rates are significantly greater for GA operations at non-towered airports relative to cohort operations at towered airports (Salmon and Motevalli, 2013).

Adapting the "European" 3rd party risk modeling approach to domestic (USA) uncontrolled GA operations was, however, a challenge because of poor data quality. Data were extracted from the National Transportation Safety Board (NTSB) dataset, which contains data collected by the Investigator-in-Charge (IIC) from all accident investigations. During a data review, substantive and systematic errors were identified. These errors negated the use of the majority of potentially relevant data points for developing a crash distribution model (Salmon and Motevalli 2010). Thus, from an initial ~5,000 accident reports for a 20-year period involving 4,500 relevant airports, 112 inbound accidents resulting in "crash sites" physically located outside the boundaries of an airport could be verified.

Modeling and Simulation Methodology

Multiple previous 3rd party airport risk models have used a Curvilinear Coordinate system to integrate unique flight paths of inbound accident aircraft as the base for defining copula model parameters (Evans et al, 1996, Brady and Hillestad, 1995). This is conceptually illustrated in Figure 1 for clarity.

The three crash site are each located at 1.2km along the accident aircrafts' unique inbound flight-path. Thus, each shares the same "longitudinal" coordinate on the curvilinear coordinate system. Crash Sites B and C are displaced 0.3km perpendicular to the right of the longitudinal coordinate (relative to the direction of flight), and thus share the "lateral" curvilinear coordinate (0.3km). Crash Site A is displaced 0.3km as well, however in this case the displacement is to the left, and thus is designated as -0.3km. Having defined each crash site relative to the unique inbound flight path of the accident aircraft, the three curvilinear coordinate systems can be easily combined as part of one set of data.

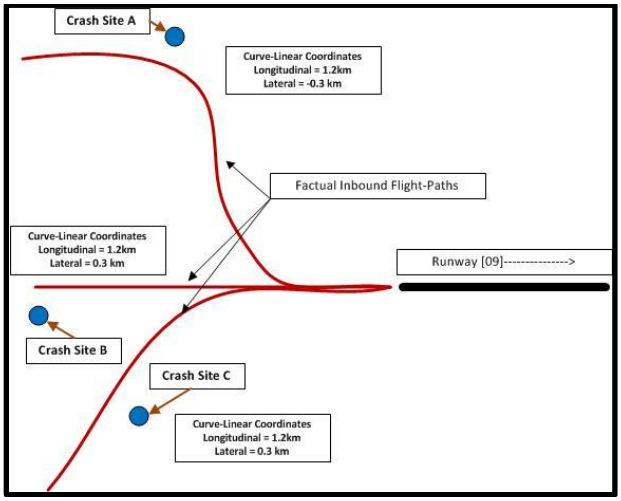

Figure 1: Illustration of Curvilinear Coordinate System.

The specific process utilized in the research presented in this paper diverged slightly from this method because of the challenges in the objective accident data noted previously. More specifically: factual inbound flight paths unique to specific accident aircraft were not available through radar track or GPS data, thus had to be assumed. For this, it was assumed that all inbound aircraft utilized one of two curvilinear coordinate systems that corresponded to the two inbound flight patterns recognized at GA airports (FAA 2003) (Left-turning and Straight-in). All validated crash sites were then located relative to the inbound runway threshold as Cartesian coordinates, latitude/longitude or polar coordinate depending on the data available in the NTSB accident report narrative. Each data point was assigned to either the

Left-turning or Straight-in curvilinear coordinates based on IIC textual descriptors archived in the accident report narrative. These data were normalized about a Unit Airport (a 1,500m runway [36]), as illustrated in Figures 2 and 3. Figure 2 plots the normalized crash site locations for all validated left-turning approaches. Figure 3 plots the same data for straight-in approach crash accidents.

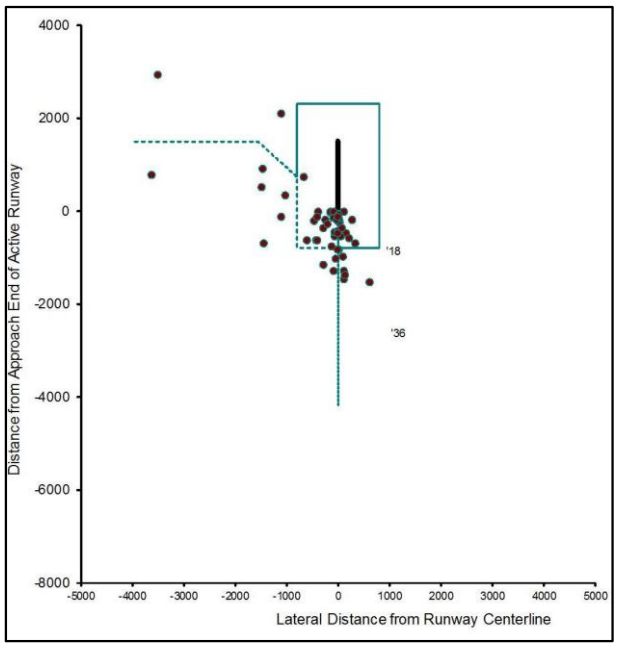

Figure 2: Left-turning approach crash site locations

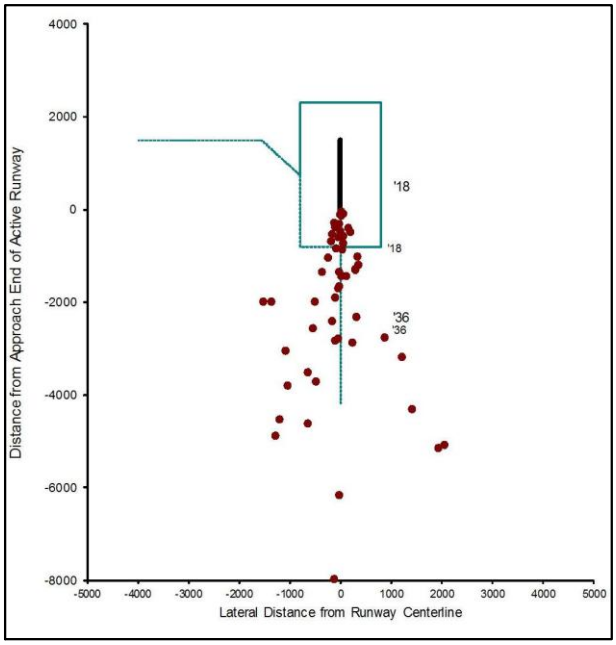

Figure 3: Straight-in approach crash site locations

These data were integrated as a single dataset by integrating the two curvilinear coordinate systems as illustrated in Figure 4. By convention, crash sites located to the left of (below) the longitudinal coordinate relative to the direction of travel are negative, and those to the right (above) are positive.

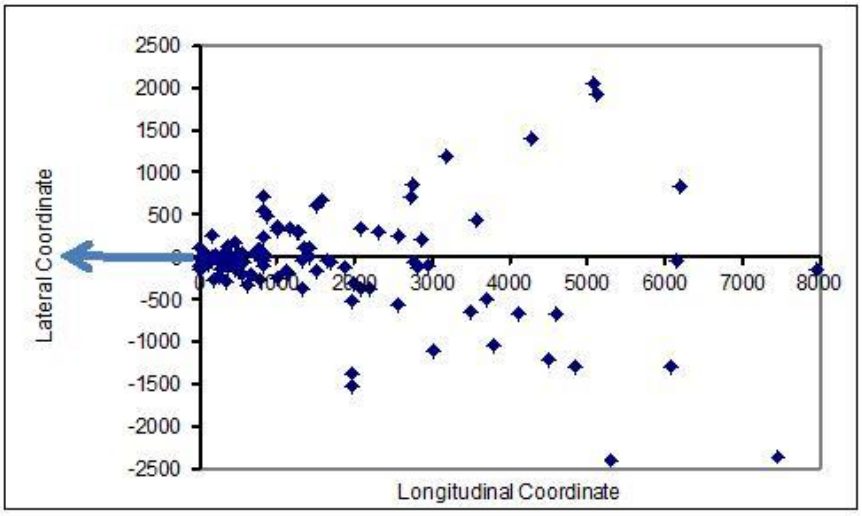

Figure 4: 112 inbound crash site locations using a curvilinear coordinate system

Copulas and Sparse Data

Figure 4 offers substantive visual evidences of a relationship between longitudinal and lateral coordinates, with the lateral coordinate increasing as some function of the longitudinal coordinate. Such a relationship can be exploited for simulation proposes if it can be represented as a parametric model.

Developing this model proved difficult due to the sparse nature of the data at longitudinal coordinates greater than 3,000 meters along the longitudinal axis. Specifically, the scarcity of data in the tails negated easy representation of the data by a copula. Therefore, the possibility of the data being represented as an empirical copula was investigated through a process outlined below.

Longitudinal Marginal Distribution

All crash site data were transformed to the "longitudinal" axis of the curvilinear coordinate system in order to fit a marginal distribution under the assumption that the longitudinal coordinate was the independent random variable. This is illustrated in Figure 5. A common method for fitting parametric models (probability distributions) to empirical data is to discretize the data into a relevant number of histogram bins. A parametric model can then be fitted to the data via minimizing sum-squared-error (SSE) between the observed frequency and expected frequency of observation within each bin. A simple Chi Squared test then indicates whether the chosen parametric model offers statistical significance for representation.

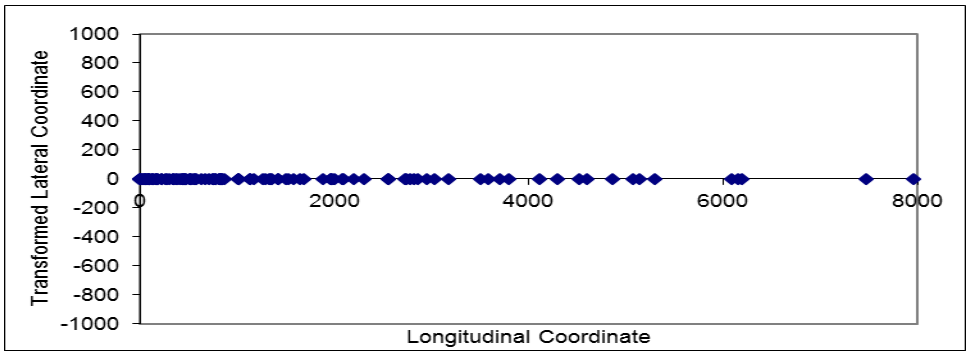

Figure 5: Illustration of independent coordinate data transformed to common axis

This was accomplished with the data in Figure 5 through use of "equal probability" bins, wherein the physical boundaries defining the width of each bin were defined such that 0.083 of all observations would be found within the bin (derived by 1/12 = 0.083). These bins boundaries, with expected and factual count are illustrated in Table 1. Notable is that each bin under the equal probability paradigm contains at least 6% of all observations, with a minimum of 7 observations. The 12 bin paradigm was chosen because more bins yielded low frequency in some, decreasing the likelihood of parametric model fit.

These data were fitted with a Gamma(0.7, 2315) distribution. A Goodness of Fit test yielded a $X^2_{critical}$ = 16.9 for a two parameter model at a significant $\boldsymbol{\alpha}$ = 0.05. The $X^2_{calculated} = 4.1$. It should be noted that further analysis would demonstrate multiple parametric models would suffice as representative of the data; however, the Gamma(0.7, 2315) distribution was used for convenience during simulation.

Table 1: Equal probability defined bins, with frequency of observations

Bin Boundaries	Bin Width (m)	Frequency	Relative Fequency
0 - 62	62	10.0	9%
62 - 169	107	9.0	8%
169 - 310	141	7.0	6%
310 - 486	176	11.0	10%
486 - 701	215	8.0	7%
701 - 963	263	12.0	11%
963 - 1289	326	7.0	6%
1289 - 1704	415	11.0	10%
1704 - 2258	554	8.0	7%
2258 - 3064	806	11.0	10%
3064 - 4489	1425	7.0	6%
> 4488	infinite	11.0	10%

Lateral Marginal Distribution

A similar process was utilized for the lateral coordinate data, wherein each data point was transformed to the central longitudinal coordinate at 4,000m from the runway threshold as a linear function based on the

assumption that the lateral coordinate of each crash site is dependent on the longitudinal coordinate, and that this dependency is linear in nature. The transformation function is presented as Equation 1, and illustrated in Figure 6 for clarity.

$$t_{\bar{s}} = \ t_i * \frac{\bar{s}}{s_i} \quad [1]$$

Where:
$t_{\bar{s}}$ = transformed lateral coordinate
t_i = original lateral coordinate
$\bar{s}$ = average longitudinal coordinate
s_i = original longitudinal coordinate

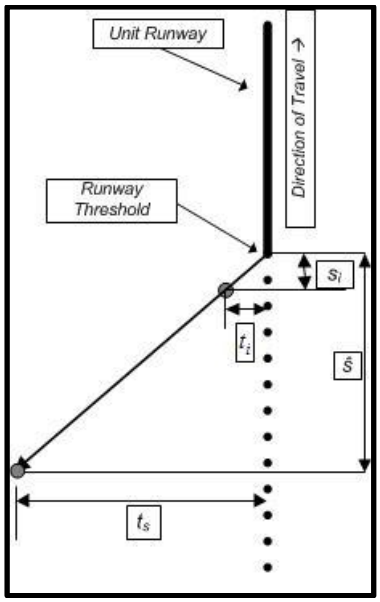

Figure 6: Linear transformation of lateral coordinates to marginal axis

This simple linear transformation yielded results that did not lend themselves to any parametric model. This was because the nature of the linear assumption dictated that objective crash sites with low longitudinal coordinate values (close to the runway threshold) and relatively high lateral coordinates transformed to extreme lateral coordinates that no parametric model could represent. In the extreme case of the crash site located at coordinate [2, -91], the transformed coordinate at the defined marginal axis by Equation 1 was [4,000, 73,921], or a 74 km lateral displacement from the centerline at 4km away from the runway threshold. This is clearly an absurd result that cannot be reasonably represented in any model or simulation.

Further, a review of the remainder of the data suggested that any crash site data point with a lateral coordinate 10-times greater than the corresponding longitudinal coordinate would transform beyond any reasonable value at the marginal axis located at 4,000m. In the objective data there were 14 such data points. Since this represents 11% of the validated dataset, it was deemed impractical to consider these data points as statistical anomalies.

What followed was an investigating into the nature of operations and pilot behavior at non-towered airports, which yielded insight to the assumption that the focal point of all inbound operations was the threshold of the approach end of the runway. While it was originally assumed that all inbound pilots used the threshold of the runway as a focal point during approach and landing, it proved adventitious to redevelop Equation 1 such that this assumption could be tested.

For this, Equation 2 was developed to include a parameter that allow for the common focal point for all inbound accident aircraft to be located at some point beyond the runway threshold. The potential impact

of this parameter is visually depicted in Figure 7, wherein it can be immediately recognized that dispersion in the transformed lateral data would be substantively reduced depending on final determination of the additional parameter.

$$t_{\bar{s}} = \ t_i * \frac{\bar{s}+s_x}{s_i+s_x} \quad [2]$$

Where:
s_x = common focal point for all inbound operations

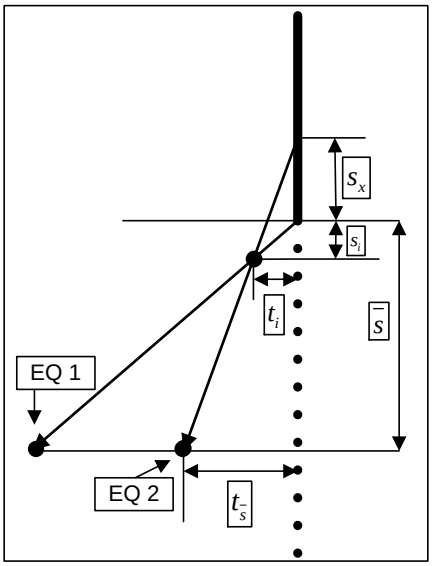

Figure 7: Comparison of transformation equations 1 and 2

The challenge that remained was in selecting an adequate parametric model that could be fitted to the transformed data under constraints of the additional parameter. Common and logical parametric models that were rejected due to poor fit included the Triangular, Laplace, Normal and related distributions.

A search for more an exotic distribution yielded the Asymmetric Laplace Distribution(-43, 4, 53, 34) (van Dorp and Kotz, 2002), which return a $X^2_{calculated} = 5.1$. The ALD is represented by Equation 3

$$F(x:\theta,\sigma,\kappa) = \begin{cases} \frac{\kappa^2}{1+\kappa^2} e^{\frac{\sqrt{2}}{\sigma\kappa}(\theta-x)} & for\ x < \theta \\ \frac{\kappa^2}{1+\kappa^2} + \frac{1}{1+\kappa^2}\left[1 - e^{\frac{\sqrt{2}\kappa}{\sigma}(x-\theta)}\right] & for\ \theta \le x \end{cases} \quad [3]$$

In the absence of the additional parameter s_x, fitting the data transformed by Equation 1 would have been the simple optimization of the parametric model parameters to the data towards minimization of the sum of squared error (SSE) between the model and empirical data. Under the paradigm of Equation 2, the optimization process also minimized the SSE, though at the same time testing different assumptions about s_x. This resulted in the Asymmetric-Laplace(-43, 4, 53, 34), with s_x = 168m. For demonstration purposes results of Equation 1 ($s_x = 0$) and Equation 2 ($s_x = 168$) paradigms are each included in Table 2 for comparison.

Table 2: Asymmetric Laplace Distribution Results

	s(x) = 0	s(x) not = 0
Calculated Chi-Squared Statistic	13.4	5.0
Critical Chi-Squared Statistic - Alpha .05	12.6	12.6
Critical Chi-Squared Statistic - Alpha 0.1	10.6	10.6
Results	Reject	Fail to Reject

Results

The results of the modeling process outlined above enabled a relative frequency crash location distribution to be simulated that incorporated airport-specific operating conditions, and ultimately 3rd party exposure conditions. This is illustrated in Figure 8 as four contours (shaded areas) within which it can be expected that if a GA aircraft does crash while inbound to a non-towered airport, the area wherein there is a 80%, 60%, 40% and 20% can be determined. Set to the spatial scaling of the axis, the relative crash risk for any resident or sub-development can be determined for the specific exposure conditions modeled.

The specific exposure conditions illustrated in Figure 8 assumed 100% of all inbound operations utilize runway '36. Of these, 75% utilize the left-turning approach, while the remaining 25% utilize the straight-in approach. Also captured in Figure 7 are similar analysis results for outbound operations, though detail descriptors of the analysis was not included above for reasons of brevity and clarity. Similar to inbound operations, it was assumed that 100% of all outbound operations utilizes runway '36, with 75% utilizing the left-turning departure, and 25% utilizing the straight-out departure.

The assignment of these values are arbitrary and used here for demonstration purposes only, though any real-world distribution of inbound and outbound operations across the base and reciprocal runways of the model airport can be inputted to the model. Figure 9 demonstrates this with a equal distribution of 50% of inbound and outbound operations utilizing the Base and Reciprocal runways, with 50% of inbound operations utilizing the left-turning approach, and 50% the straight-in approach.

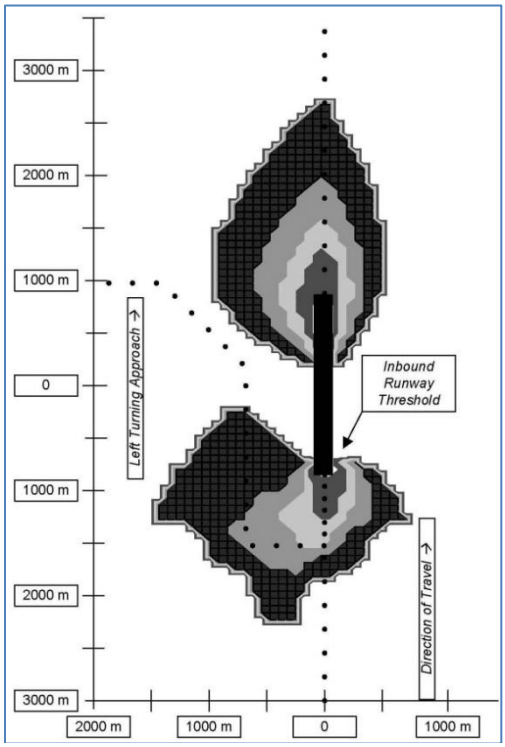

Figure 8: Simulated crash distribution results using parametric model

Summary and Future Research

This paper made three primary observations that are instructive when modeling complex systems. First, modeling and simulation methodologies successful in one sector of an industry are not necessarily directly transferable to a seemingly related sector. In this case, a well-established process for developing a crash location distribution at large commercial service airports was not directly applicable to the non-towered GA operations because of: 1) quality of the objective data record, 2) scarcity of the data, and 3) behavior of individual pilots.

Second, unconventional, yet equally valid modeling methodologies, such as the equal probability assumption used here for fitting parametric models to data, and the use of the Asymmetric Laplace distribution, can yield too successful models that might otherwise gone unobtainable if the unconventional was not sought.

Third, in an era wherein there is a trend towards reliance on "big data" to yield meaningful results, this paper demonstrated that irrespective of the ever-increasing capacity of machine computation, it is through a deep review of the data, coupled with an investigative approach to the underlying system being modeled that biased and distorted results can be avoided.

Future research will extend the capacity of the crash location distribution model presented here by integrating results with Crash Consequence and Crash Rate models that are under development. The results will be a fully integrated 3rd party risk model for operations at non-towered airports. The flexibility of the model presented above will enable airport-specific operations and exposure characteristics to be incorporated, including: 1) number of operations, 2) distribution of operations across inbound/outbound traffic patterns and runways, and 3) population densities of at risk communities.

Works Cited

"Recommended Standard Traffic Patterns and Practices for Aeronautical Operations at Airports without Operating Control Towers," Federal Aviation Administration, AC No. 90-66A, Washington, D.C., 1993.

"Annual Review of Accident Statistics", National Transportation Safety Board, https://www.ntsb.gov/data/aviation_stats.html (accessed on June 1, 2014)

A. Singh, J. Rene van Dorp and T.A.Mazzuchi, "A Novel Asymmetric Distribution with Power Tails", Communications in Statistics, Theory and Methods, Vol. 36, No. 2, pp. 235-252.

Brady, S.B., Hillestad, R.J., "Modeling the External Risks of Airports for Policy Analysis," RAND European-American Center for Policy Analysis, Santa Monica, CA, 1995.

Evans, A.W., Foot, P.B., Mason, I.G., Slater, K., "Third Party Risk Near Airports and Public Safety Zone Policy: R&D Report 9636," National Air Traffic Service Limited, 8CS/091/03/10, London, England, 1996.

J.R van Dorp and S. Kotz (2002). "A Novel Extension of the Triangular Distribution and its Parameter Estimation", The Statistician, Vol. 51, No. 1, pp. 63-79.

Piers, M.A., Loog, M.P., "The Development of a Method for the Analysis of Societal and Individual Risk Due to Aircraft Accidents in the Vicinity of Airports," National Aerospace Laboratory, NLR CR 93372, Amsterdam, The Netherlands, 1993.

Salmon C. M., Harrald J., van Dorp J.R., Motevalli V. (2010), "Exposure at the Airport Community Interface - Quantifying Metrics of Exposure in the Vicinity of Public-Use, Non-Towered Airports", Journal of the Transportation Research Board , No. 21 84, Transportation Research Board of the National Academies, Washington, D.C.,2010, pp.31-40. DOT. 10.3141/2184.D4

Salmon C.S., Motevalli V. "Characteristics of General Aviation Accident Data, with Implications for Modeling Third-Party Risk in Vicinity of Public-Use, Non-Towered Airports". 54th Annual Transportation Research Forum. March 21-23, 2013.

Solomon, K.A., Erdmann, R.C., Hicks, T.E., Okrent, D., “Airplane Crash Risk to Ground Population”, University of California, UCLA-ENG-7424, Los Angeles, California, March, 1974.