Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

MODELING MIXED TYPE RANDOM VARIABLES

Christopher Weld
Lawrence Leemis

Department of Applied Science
Department of Mathematics
College of William & Mary
P.O. Box 8795
Williamsburg, VA 23187, USA

ABSTRACT

Mixed type random variables contain *both* continuous and discrete components, and their role is critical in many well-studied fields. Queuing analysis, stock options, and hydrology rainfall models are among those dependent on mixed random variables to simulate event outcomes. In each of these examples, continuous positive distributions combine with a discrete spike at zero to adequately represent system uncertainty. These problems often require simulation because analytic solutions using these hybrid distributions quickly grow in complexity. Concessions are made, however, when using simulation. In addition to inherent sampling variability, perspective of discrete and continuous components is easily lost when plotting results. This paper details these challenges, and touches on the shifting line between simulations and attainable analytic results. It discusses computational probability's potential to improve model realism and accuracy, introducing MixedAPPL software prototype—an extension of Maplesoft based APPL (A Probability Programming Language)—capable of manipulating mixed type random variables.

1 INTRODUCTION

Mixed type distributions—having *both* continuous and discrete components to their probability distribution—are an inherent component to many diverse and important models. Queuing is one such example, whereas a customer can experience no wait time (a discrete likelihood), however, those with non-zero wait time are assessed on a continuous scale. Many meteorlogical models—instantaneous wind speed supporting turbine placement, or daily rainfall supporting hydrology models (Figure 1) for example—conform to an analogous paradigm (NCEI 2017). The list goes on, with applications ranging from insurance claims and actuarial science, to sports analytics—illustrated later in this paper.

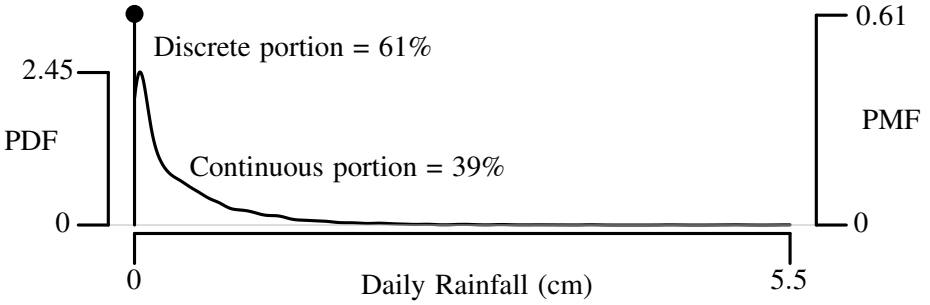

Figure 1: Daily rainfall distribution for Williamsburg, VA, for calendar years 2014 to 2016.

Effective manipulation and communication of mixed type random variables is important to adequately model and present results within the aforementioned fields. This is particularly germane to simulations. Modelers rely on simulations to represent mixed random variable outcomes for good reason. Manipulating these variables beyond trivial use quickly becomes complex, often to the point analytic solutions are unattainable. There are losses, however, in abandoning exact probabilistic result in lieu of simulation. In addition to error inherent with sample size, simulation results tend to muddy perception of underlying discrete and continuous components when communicating results. Its difficulties stem from both bookkeeping necessary to track its respective components, and challenges inherent to plotting mixed type distributions.

This paper illustrates the above-mentioned difficulties of working with mixed random variables, and presents techniques to overcome them. Specifically, following background discussion in Section 2, Section 3 details the potential graphic deception inherent in illustrating mixed random variables. An NFL starting field position model illustrates these issues in Section 4. Section 5 then identifies improved plot alternatives, adequately communicating both continuous and discrete components. Section 6 shifts focus to computational probability, and its appeal modeling mixed random variables for both its inherent book-keeping of continuous and discrete components, as well as the precision it offers. There we introduce a mixed random variable software package currently in development. It is named MixedAPPL, and is an extension of the Maplesoft based computational probability package APPL (A Probability Programming Language) enabling mixed type random variable manipulations. Finally, Section 7 concludes with a summary of the discussion.

2 BACKGROUND

This literature review is broken into three topics pertinent to this paper: mixed type distributions, statistical graphics, and computational probability. They are addressed in that sequence next.

Mixed type distributions are most heavily cited with regard to zero-inflated data. Zero-inflated data is a term introduced in the early 1990s largely in conjunction with discrete parametric distributions (e.g. zero-inflated Poisson, or ZIP distribution), and subsequently tied to continuous distributions (Lambert 1992, Tu and Liu 2002). As its name implies, zero-inflated data has a swollen discrete probability mass at the origin. The name followed decades later, but the study of such mixed type distributions in literature spans back to at least 1955 when Aitchison identified a dichotomy of some demand models, whereas a (potentially continuous) parametric distribution is followed for purchase amount of a given commodity with exception of an uncharacteristic spike for those abstaining from purchase (Aitchison 1955). A similar dichotomy is later found in environmental models such as contamination concentration in air or liquid, and precipitation forecasts (Owen and DeRouen 1980, Feuerverger 1979). In fact, rainfall is among the most frequently cited examples of mixed type distributions, and studied with interest at depth because of its important implications within the field of hydrology (Shimizu 1993; Hyndman and Grunwald 2000; Yoo, Jung, and Kim 2005; Herr and Krzysztofowicz 2005; Serinaldi 2009; Li, Singh, and Mishra 2013).

Statistical graphics is a well researched field. Tufte defines graphic excellence as communicating complex ideas with clarity, precision, and efficiency (Tufte 1983). A host of other books spanning into the twenty-first century also focus on techniques to practice and avoid in statistical graphics, and the field of data visualization continues to grow (Tukey 1977; Cleveland 1985, 1993; Tufte 1991, 1997, 2006). Its current popularity is not surprising, given the recent rise of data-intensive scientific discovery, and the need to visualize information often too complex and/or cumbersome to make sense of numerically or through formulas (Simoff and Böhlen 2008, LaValle et al. 2011, Chen and Zhang 2014; Ambur et al. 2016). Despite growing interest in the field, the authors found no evidence of specific attention given to the visualization of mixed type probability distributions, which this paper addresses.

Computational probability has attributes well suited for the complexities of modeling mixed random variables, and is discussed at the end of this paper. Multiple computational probability platforms exist to leverage the symbolic calculations capability of a computer algebra system. Examples include Operations Research and Probability and Statistics focused textbooks with regard to Maple (Parlar 2012, Tanis and Karian 1999), and Mathematica (Rose and Smith 2002, Hastings 2006). This research introduces software

expanding the existing capabilities of APPL, which are well documented in the textbooks *Computational Probability* and *Computational Probability Applications* (Drew et al. 2017, Glen and Leemis 2017). The authors found no evidence of other software specifically catering to mixed type distributions.

3 DECEPTIVE MIXED TYPE DISTRIBUTION PLOTS

Graphics are often best to quickly communicate statistical analysis; at-a-glance intuition of underlying system behavior is more accessible than via numeric presentation. Communicating mixed type distributions effectively, however, is non-trivial. Pulling the otherwise disjoint worlds of continuous and discrete components into focus together can lead to an unintended distortion of scale, and often deceiving and/or misleading plots. Overcoming this dynamic to leverage the utility of probability distribution plots requires deliberate attention by its designer. Before addressing these solutions, we first elaborate on its challenges, and how the simulation community is uniquely apt to fall victim to this graphic misrepresentation.

This section focuses on counter-examples to the intent that graphics are clear, concise, and effective. The first two examples illustrate difficulties inherent to plotting mixed type random variables. The third example presents and critiques a histogram of mixed type distribution outcomes.

3.1 Example 1

The first example illustrates two mixed type random variables, X and Y. Define X as

$$X \sim \begin{cases} C \text{ with probability } 0.1 \\ D \text{ with probability } 0.9, \end{cases}$$

where $C \sim \text{triangular}(0, 0.3, 0.6)$, and $f_D(x) = 1$ for $x = 0.1$. Similarly, define Y as

$$Y \sim \begin{cases} C \text{ with probability } 0.92 \\ D \text{ with probability } 0.08, \end{cases}$$

where $C \sim \text{triangular}(0, 30, 60)$, and $f_D(x) = 1$ for $x = 10$.

Ignoring axes scales, both distributions produce identical plots in Figure 2 with regard to their respective continuous and discrete components. This does not, however, imply the relative contributions of those components are proportional. In fact, our example illustrates the counterpoint; the continuous distribution occurs with probability 0.1 in X, and 0.92 in Y. Vastly different relative discrete and continuous contributions can result in strikingly similar graphics. In other words, mixed type distribution plots can distort (from a visual perception standpoint) the relative contribution of their continuous and discrete components.

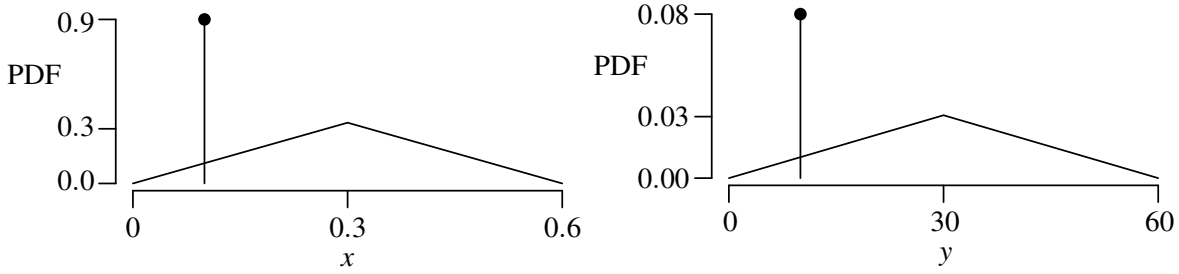

Figure 2: Similar plots, despite vastly different proportions of continuous and discrete components.

3.2 Example 2

A second example highlights how distortion of a relative scale is an inherent side-effect of differences between continuous and discrete distributions. Transforming continuous distributions to a broader sample

space spreads their probability density accordingly, whereas discrete probability maintains its respective point density mass in an analogous transformation. Consider the results of the transformation $Y = g(X) = 10X$ on random variables X_c and X_d as shown in Table 1.

Table 1: Density function impact of $Y = g(X) = 10X$ transformation on X_c and X_d.

Random Variable	$f_X(x)$	$f_Y(y)$
$X_c \sim \text{uniform}(0,1)$	$1 \quad \text{for } 0 \le x \le 1$	$1/10 \quad \text{for } 0 \le y \le 10$
$X_d \sim \text{Bernoulli}(1/2)$	$1/2 \quad \text{for } x \in \{0,1\}$	$1/2 \quad \text{for } y \in \{0,10\}$

Uniform density values for $g(X_c)$ are small in comparison to those in X_c. Conversely, $g(X_d)$ probability mass values are identical to those in X_d, albeit for different support. This scaling impact for the continuous but not discrete distribution underpins the difficulties in presenting mixed type distributions.

Consider two mixed type random variables, X and Y. The random variable X is X_c with probability 0.5 and X_d with probability 0.5. Similarly, Y is $g(X_c)$ with probability 0.5 and $g(X_d)$ with probability 0.5. Plots of X and Y are given in Figure 3.

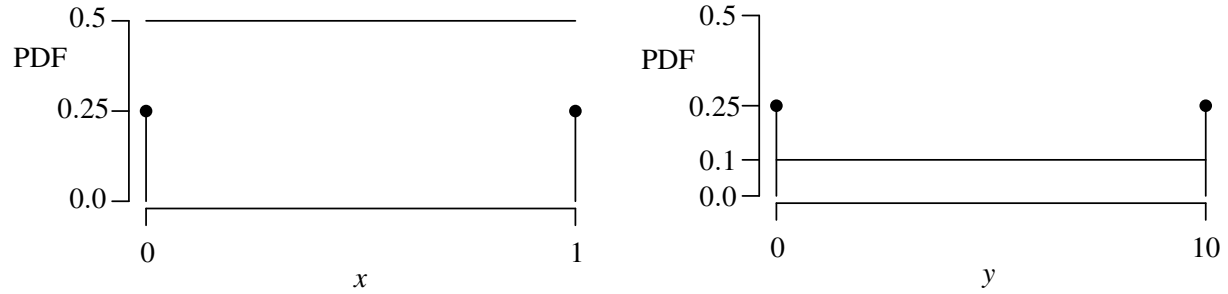

Figure 3: Dissimilar plots, despite identical relative proportions of continuous and discrete components.

Although each distribution has continuous components accounting for 0.5 probability, their dissimilar mixed type distribution plots can mislead intuition to conclude that relative proportions of continuous and discrete components vary greatly between them. As was also the case with Figure 2, on a common scale, mixed random variable distributions plots necessitate close inspection before solidifying intuition of the underlying distribution. While discrete probability mass is clearly represented through its discrete spike values, the influence contained within continuous segments can become drowned out or, conversely, overstated in a relative comparison.

3.3 The Hoodwinking Histogram

Distortion of scale is an inherent consequence of combining discrete and continuous outcomes into a common frame of reference, as seen in Sections 3.1 and 3.2. When the otherwise disjoint worlds of continuous and discrete are pulled into focus together, they are obliged to conform to shared axes scales. Unfortunately for the simulator, an analogous predicament awaits their much relied-on histograms. While histograms do enjoy consistent calibration given by way of a uniform bar height per unit probability, they also struggle with presenting results for both discrete and continuous portions on common axes. Their struggle, however, pertains to lost clarity of respective continuous and discrete components, and the insight they offer.

Clearly communicating mixed random variable circumstance necessitates distinguishing its continuous and discrete parts. Alternatively, this underlying dichotomy is lost in a plot of homogeneous components. The latter is the fate of the histogram. As future detail will prove, these shortcoming befall Figure 4, which illustrates NFL starting field position in the 2016 season, and will serve as our example moving forward (Horowitz 2017).

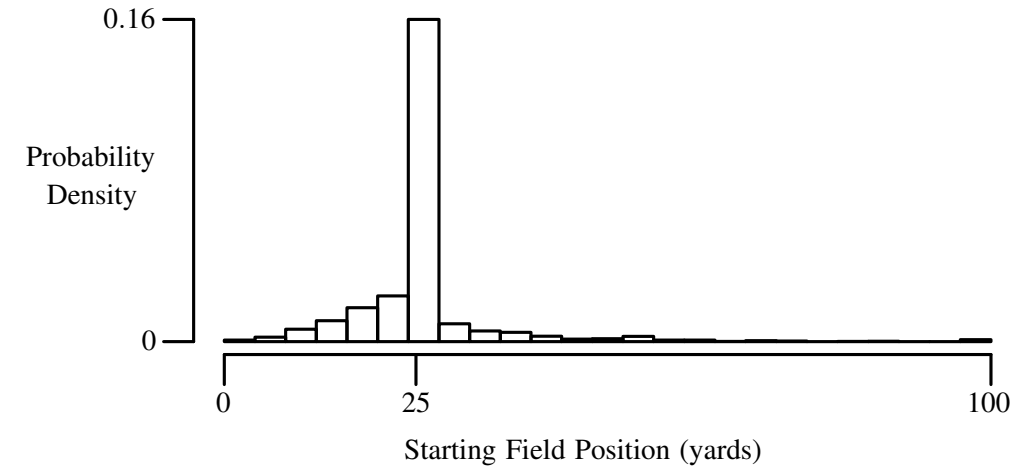

Figure 4: A probability histogram of NFL starting field position in 2016.

Valuable information is lost in Figure 4 by using the perspective given by this histogram. By muting the distinction between discrete and continuous, relevant underlying system behavior insights are lost. This is the trap awaiting mixed random variable simulation results that are traditionally best illustrated with histograms, or used to derive a kernel density function. When keying in data for use in developing an input model for a simulation, careful bookkeeping of the discrete and continuous portions is necessary. The same care is needed to adequately present results. Following an in-depth introduction to this football field position model, we will review these shortcomings in detail.

4 FOOTBALL STARTING FIELD POSITION MODEL

Football starting field position following a kickoff will serve as our mixed type distribution example for the remainder of the paper. This section introduces this model in depth, to include its data and an initial presentation of results. The National Football League (NFL) 2016 season official rulebook governs descriptions within this section (Goodell 2016).

4.1 Description and Parameters

Starting field position depends on the result of the kickoff from the kicking team's 35-yard line, and feasible possibilities are well understood. The ball is spotted wherever forward progress of the return is stopped, specifically at the furthest point of advance of the football. Measured as the distance from the receiving team's end zone goal-line, the tip of the ball can spot at any point along the length of the 100-yard field. In addition to this continuous support, multiple discrete possibilities are also feasible. Discrete scenarios account for unique circumstance which are adjudicated according to the rules of the game, and are broken into two categories below: a ball kicked out-of-bounds, and a return where the ball ends up in either end zone of the field.

1. If kicked out-of-bounds within the field of play, the ball is spotted 25 yards from the kick (the 40-yard line of the return team under normal circumstances), assuming that is better field position than where it crossed out-of-bounds.
2. A ball in either end zone at the end of the kickoff takes on one of three forms: the opposing team's end zone, the 25-yard line, or the return team's end zone. These are detailed below.
    - If the ball returner successfully traverses the entire field in-front of them it is possible they cross into their opponents end zone for a touchdown. For this circumstance "starting" field position is assumed in the opposing team's end zone.

- If the football is kicked into the returning team's end zone (the one at their back if they were to progress up-field) and purposely downed (the receiver kneels to the ground) or exits the end zone out-of bounds then it is ruled a "touchback," and the ball is placed at the 25-yard line. This strategy is often adopted when the receiver assesses their likelihood of attaining field position better than the 25-yard line is not worth the risk given the circumstances they face ahead.
- In the rarest circumstance, the return-team is either inadvertently stopped in their own end zone (ruled a "safety") or the kicking team is able to regain control of the ball and score a touchdown themselves. For these circumstance "starting" field position is assumed in the return team's end zone.

4.2 Data

A play-by-play account assembled by Horowitz of 2016 NFL games comprise our dataset (Horowitz 2017). It contains 45,736 entries, 2,593 of which are kickoffs with implications to our model. Kickoffs are further subdivided into those with discrete categorical outcomes (touchback, out-of-bounds, and end zone touchdowns or safeties) and those returned within the field of play. Table 2 shows highlights of this division.

Table 2: Summary data for 2016 NFL kickoff starting field position outcomes (Horowitz, 2017).

Type	Category	Frequency	Probability
Continuous	Return within the field-of-play	1047	$\frac{1047}{2593} \cong 0.404$
Discrete	End Zone (return team)	3	$\frac{3}{2593} \cong 0.001$
	Touchback	1518	$\frac{1518}{2593} \cong 0.585$
	Out-of-bounds	18	$\frac{18}{2593} \cong 0.007$
	End Zone (kicking team)	7	$\frac{7}{2593} \cong 0.003$
	Total	2593	$\frac{2593}{2593} = 1.000$

4.3 Model and Results

Let X be the starting field position following an NFL kickoff, measured in yards from the return team's end zone (regardless if a turn-over occurs). Using 2016 play-by-play data (Table 2) X is modeled as

$$X \sim \begin{cases} C \text{ with probability } \dfrac{1047}{2593} \\ D \text{ with probability } \dfrac{1546}{2593} \end{cases}$$

where

$$f_C(x) = \begin{cases} \text{kernel density function} \\ \text{of outcomes returned} \\ \text{in the field-of-play} \end{cases} \qquad \text{and} \qquad f_D(x) = \begin{cases} \frac{3}{2593}\left(\frac{2593}{1546}\right) & x = 0 \\ \frac{1518}{2593}\left(\frac{2593}{1546}\right) & x = 25 \\ \frac{18}{2593}\left(\frac{2593}{1546}\right) & x = 40 \\ \frac{7}{2593}\left(\frac{2593}{1546}\right) & x = 100, \end{cases}$$

assuming "starting" field position for a touchdown or safety is defined at those respective end zones. We use a Gaussian smoothing kernel with a bandwidth of 1.84 for smoothing the continuous data. Figure 5

illustrates a resulting mixed type probability distribution. Isolating its continuous component produces Figure 6.

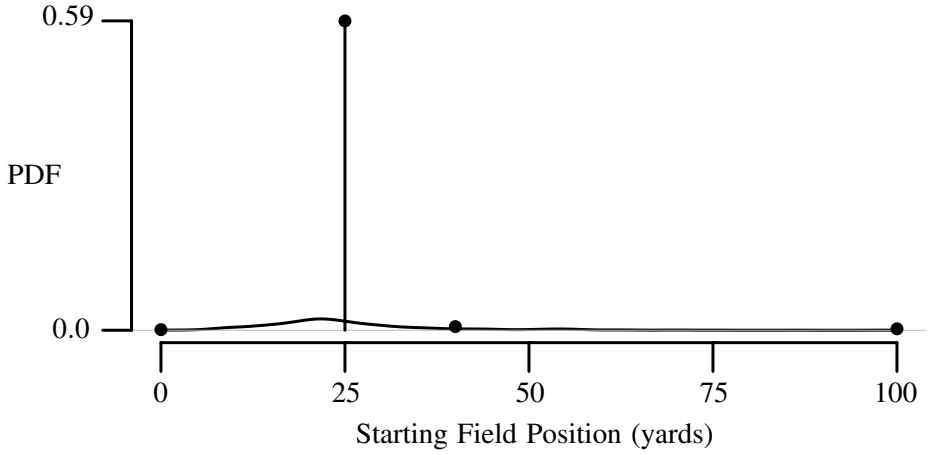

Figure 5: X mixed type distribution representing 2016 NFL starting field position.

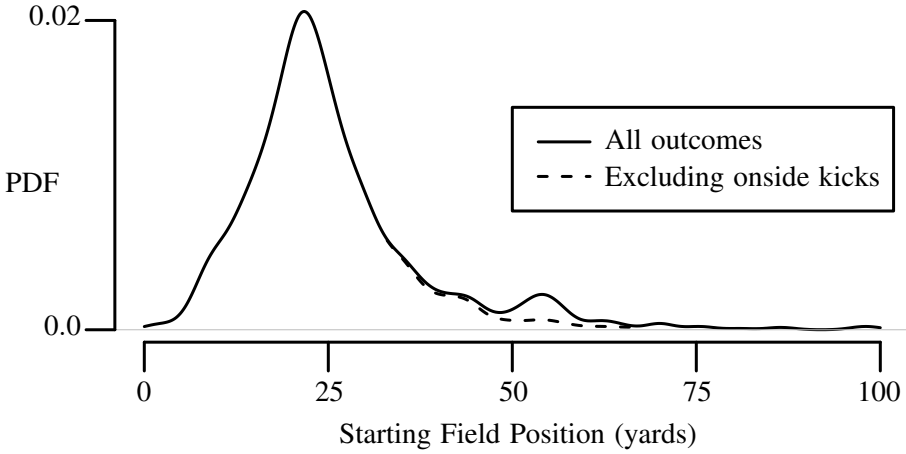

Figure 6: Kernel density function for the continuous component of mixed type distribution, X.

Starting field position is most often the 25 yard-line per touchback rules. Those attempting a return have varying degrees of success, with few reaching distances past midfield. This continuous shape has two well pronounced modes (see Figure 6). The first occurs at roughly the 22 yard-line, and is a consequence of the kicking team's ability to challenge the returner with many tacklers by the time they reach those distances. The second occurs just past mid-field and is a consequence of the 54 onside kick attempts during the 2016 season, as confirmed by the dotted line illustrating its kernel density function without those attempts.

5 EFFECTIVE MIXED TYPE DISTRIBUTION PLOTS

Both Figures 5 and 6 provide evidence why the histogram in Figure 4 is insufficient to portray these results. Figure 5 clearly illustrates four discrete outcomes, whereas only one (around the 25 yard line) is clearly evident (or suspected, rather) by inspection of the histogram. Parsing discretized components from mixed type random variable simulations is essential to properly represent those components within the final distribution. Additionally, Figure 6 illuminates nuanced and noteworthy peaks and troughs within its continuous component that are also muted in the histogram portrayal of Figure 4 due to scale.

The histogram in Figure 4 is clearly not an ideal vehicle to represent the intricacies of the mixed type distribution results. The plot in Figure 5 has its challenges as well; while it accurately captures both continuous and discrete components, it inadequately presents its continuous function due to issues of scale. Its low continuous profile also inhibits understanding of its rich landscape illustrated in Figure 6. The relative influence of the continuous component—representing over 40% of all outcomes—also appears understated in contrast to its discrete outcomes.

Before presenting a better mixed type distribution plot alternative, first consider its cumulative distribution plot, shown in Figure 7. Although it does well portraying relative discrete and continuous component sizes, it does have flaws. Small discrete outcomes at $0, 40$, and 100 yards are nearly imperceptible as a jump in the cumulative distribution. Also, details regarding the peaks and troughs of its continuous component (see Figure 6) are difficult to extract from Figure 7. In addition, onside kick attempts are not visible in the plot of the cumulative distribution function.

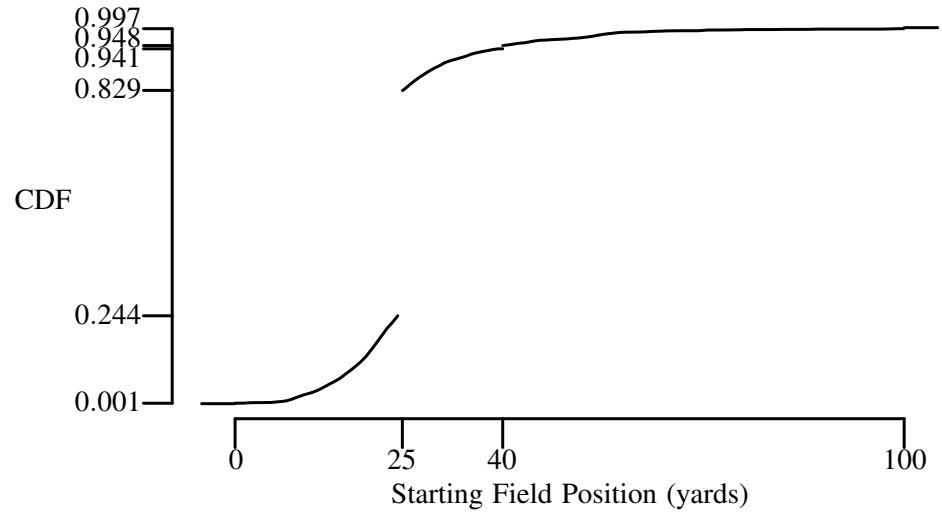

Figure 7: Cumulative distribution function of mixed type distribution, X.

Many plot alternatives are possible, but we present Figure 8 as a best alternative. It combines the favorable attributes of our preceding graphics, plus some additional customization to provide clearer perspective and intuition of the resulting distribution.

The most notable change in Figure 8 is its secondary vertical axis. Secondary axes often draw criticism for their ability to mislead, but mixed type distributions appear uniquely able to accommodate them. They inherently combine two paradigms of probability—density and mass—which have proven through previous examples to have unfavorable graphic consequences under a single pair of axes. When a single set of axes proves deceptive, incorporation of a second axes with deliberate attention to its implications of scale can improve clarity.

Proportions of continuous to discrete components should align with intuition. Without considering axes scales, equal maximum height for a Gaussian shape continuous component and a single discrete spike arguably promotes a sense of equal proportion between them. Assuming this calibration is accurate, Figure 8 adjusts those relative sizes in a 2:3 ratio to incite intuition consistent with the 40%:60% share of continuous to discrete probability within the distribution. This subjective scaling appears to work well for the unique circumstances of this model.

Figure 8 also makes provisions to enable the relative comparison of small discrete spikes. Those discrete spikes—indistinguishable under a common scale as seen in Figure 5—are amplified using a nonlinear vertical PMF axis scale to enable their relative comparison. A square root scale is chosen over a logarithmic one to mitigate the risk of exaggerating the influence of these infrequent discrete events.

Two final touches complete the graphic. First, mid-plot labels state respective continuous and discrete component contributions. Next, two football player silhouettes—a kicker and a returner—provide perspective by orienting to the context of the underlying scenario responsible for the distribution (Freepik 2017).

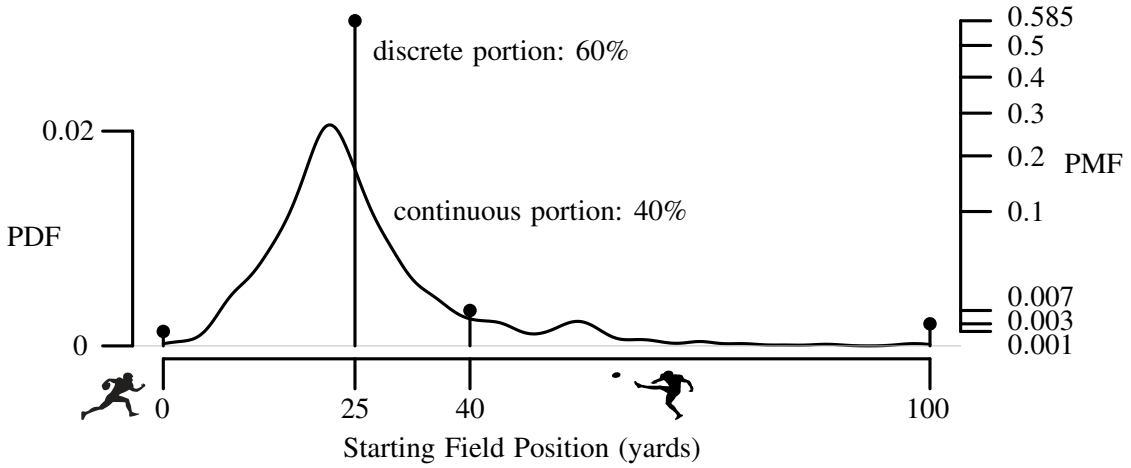

Figure 8: NFL starting field position in 2016 with a discrete square root scale secondary vertical axis.

An analogous technique is possible using a histogram approach in lieu of the kernel density function to represent its continuous component. Identical concessions are necessary with regard to a secondary axes and its relative scale. This result is shown in Figure 9.

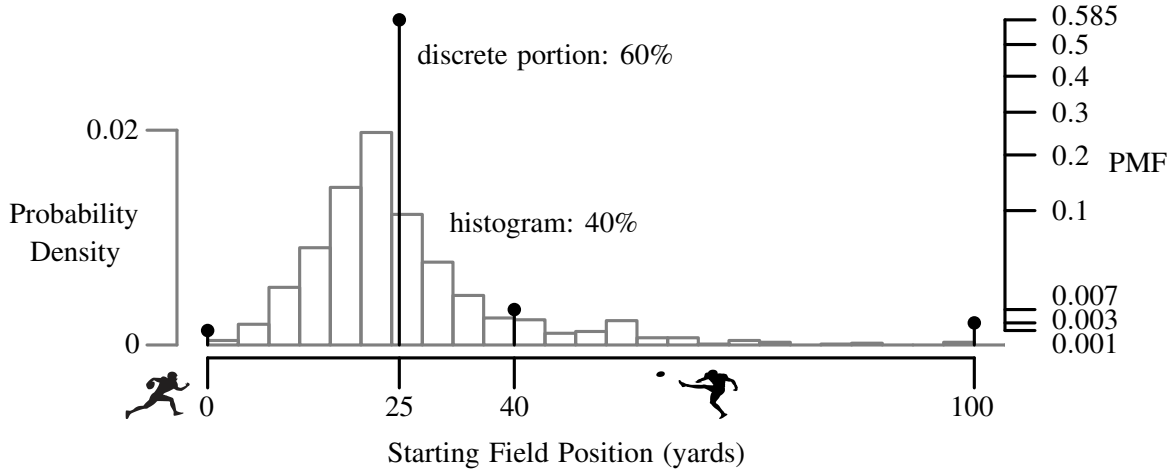

Figure 9: NFL 2016 starting field position with a histogram representing continuous outcomes.

6 COMPUTATIONAL PROBABILITY

Computational probability is an attractive alternative to model mixed type random variable scenarios. Prior to computational probability advancements, models fit a dichotomy of solution possibilities: relatively simple circumstance permitted development of exact solutions, and those beyond the reach of analytic solutions relied on approximation techniques or simulation. Computational probability now extends the subset of problems with attainable *exact computations* by using a symbolic computer algebra systems. The examples in this section will use the Maplesoft based APPL software and its associated MixedAPPL prototype extension.

The unique appeal to modeling mixed type random variables using computational probability is due in-part to its data structure. Book-keeping between continuous and discrete components is necessary to perform mixed type random variables manipulations within computational probability algorithms. Using the prototype MixedAPPL software, the general format for storing mixed type distributions is

```
[[f(x), f(y)], [support X, support Y], ["Mixed", "PDF"]]
```

where X is its continuous component, and Y is its discrete component. An example in APPL is

```
> X := [[[x -> 0.1], [0.3, 0.3]], [[0, 4], [0, 1]], ["Mixed", "PDF"]];
```

where X is uniform$(0,4)$ with probability 0.4 and Bernoulli(0.5) with probability 0.6.

The natural division of continuous and discrete components enables intuition, and is useful for follow-on analysis and/or presentation. This contrasts with most simulation techniques, whereas outcomes of mixed random variables are stored in a manner indistinguishable from their continuous counterparts, assuming custom provisions to delineate those results are not made.

Additional appeal accompanies computational probability with regard to its precision. Its solutions are exact, rather than bound by confidence intervals inherent to simulation sample size. Its computational power also enables exact solutions to problems otherwise impractical to complete. The introductory chapter of *Computational Probability* (pages 3–11) illustrate this comparison well using APPL by way of two examples (Drew et al. 2017).

MixedAPPL procedures enable users to manipulate mixed random variables in a variety of ways. For example, the `Mixture` procedure permits the combination of multiple random variables in specified proportions. The code below illustrates its use creating a mixed type random variable combining a triangular$(-1,0,1)$ distribution and a discrete spike at 0 in a 2:1 ratio.

```
> X := Triangular(-1, 0, 1);
> Y := [[1], [0], ["Discrete", "PDF"]];
> W := Mixture[[X, Y], [2 / 3, 1 / 3]];
```

The `Truncate` procedure permits options to redistribute, amass, or ignore probability outside a specified interval (given as the optional input parameters `"-"`, `"|"`, and `"x"`, respectively). The amass option inherently produces a mixed type distribution result given at least one of its endpoints lies within a continuous segment. Table 3 shows results when applying Truncate to a uniform$(0,1)$ random variable X.

Table 3: Implementation options for the MixedAPPL `Truncate` procedure.

Syntax	Result
`Truncate(X, 0.25, 0.75, "-");`	a uniform$(0.25,0.75)$ random variable
`Truncate(X, 0.25, 0.75, "|");`	a mixed random variable with discrete spikes of 0.25 at 0.25 and 0.75
`Truncate(X, 0.25, 0.75, "x");`	isolates the segment $f(x) = 1$ for $0.25 < x < 0.75$

A suite of other MixedAPPL procedures exist, and additional are in development. These include the sum and product of mixed type random variables, whose categorization results are given in Table 4. Table 4 is symmetric since the commutative property holds for both the sum and product of random variables.

Table 4: Sum and product results for combinations of continuous (C), discrete (D), and mixed (M) random variables. † applies if and only if $f_D(0) > 0$ or $f_M(0) > 0$.

+	C	D	M
C	C		
D	C	D	
M	C	M	M

*	C	D	M
C	C		
D	C or M†	D	
M	C or M†	M	M

Manipulating mixed random variable sum and product calculations are two examples of symbolic calculations in MixedAPPL. These dynamics underscore the benefit of exploring computational probability's reach, in addition to simulation alternatives which remain an accessible avenue to model such complexities.

7 SUMMARY

The incorporation of mixed type random variables into a model has non-trivial implications to its associated calculations and presentation of results. Differentiating continuous and discrete components is necessary to illustrate an accurate perspective of the model's underlying circumstance, but in doing so misleading graphics are both possible and probable. In this regard, a single scale plot or histogram of results are poorly equipped for graphic excellence. Steps are necessary to highlight respective continuous and discrete components, and ensure their associated noteworthy trends are detectable within the given scale(s). These objectives are attainable using a custom plot. A secondary vertical axis can parse out probability mass from density using scales appropriate to elicit accurate intuition of relative proportions. Incorporating computational probability is an alternative to facilitate with both book-keeping of these respective continuous and discrete components, and improve precision of results. Its utility is expanding for mixed type distributions using software such as MixedAPPL, an emerging extension to the Maple based APPL probability and statistics package.

ACKNOWLEDGMENTS

This work is supported in part by the Omar Nelson Bradley Fellowship.

REFERENCES

Aitchison, J. 1955. "On the Distribution of a Positive Random Variable having a Discrete Probability Mass at the Origin". *Journal of the American Statistical Association* 50 (271): 901–908.

Ambur, M. Y., J. J. Yagle, W. Reith, and E. McLarney. 2016. "Big Data Analytics and Machine Intelligence Capability Development at NASA Langley Research Center: Strategy, Roadmap, and Progress". Technical report, Hampton, VA.

Chen, C. P., and C.-Y. Zhang. 2014. "Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data". *Information Sciences* 275:314–347.

Cleveland, W. S. 1985. *The Elements of Graphing Data*. Wadsworth Advanced Books and Software Monterey, CA.

Cleveland, W. S. 1993. *Visualizing Data*. Hobart Press.

Drew, J. H., D. L. Evans, A. G. Glen, and L. M. Leemis. 2017. *Computational Probability*. Springer.

Feuerverger, A. 1979. "On some Methods of Analysis for Weather Experiments". *Biometrika* 66 (3): 655–658.

Freepik 2017. "American Football Player Silhouettes Collection". http://www.freepik.com/free-vector/american-football-player-silhouettes-collection722363.htm.

Glen, A. G., and L. M. Leemis. 2017. *Computational Probability Applications*. Springer.

Goodell, R. 2016. "2016 Official Playing Rules of the National Football League". *New York: National Football League*.

Hastings, K. J. 2006. *Introduction to the Mathematics of Operations Research with Mathematica®*, Volume 279. CRC Press.

Herr, H. D., and R. Krzysztofowicz. 2005. "Generic Probability Distribution of Rainfall in Space: The Bivariate Model". *Journal of Hydrology* 306 (1): 234–263.

Horowitz, M. 2017. "nflscrapR: R Package for Scraping NFL Data off their JSON API". https://github.com/maksimhorowitz/nflscrapR.

Hyndman, R. J., and G. K. Grunwald. 2000. "Applications: Generalized Additive Modelling of Mixed Distribution Markov Models with Application to Melbourne's Rainfall". *Australian & New Zealand Journal of Statistics* 42 (2): 145–158.

Lambert, D. 1992. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing". *Technometrics* 34 (1): 1–14.

LaValle, S., E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. 2011. "Big Data, Analytics and the Path from Insights to Value". *MIT Sloan Management Review* 52 (2): 21.

Li, C., V. P. Singh, and A. K. Mishra. 2013. "A Bivariate Mixed Distribution with a Heavy-Tailed Component and its Application to Single-Site Daily Rainfall Simulation". *Water Resources Research* 49 (2): 767–789.

NCEI 2017. "Williamsburg, VA, Daily Percipitation for 2014 to 2016". https://www.ncdc.noaa.gov/data-access.

Owen, W., and T. DeRouen. 1980. "Estimation of the Mean for Lognormal Data Containing Zeroes and Left-Censored Values, with Applications to the Measurement of Worker Exposure to Air Contaminants". *Biometrics* 36:707–719.

Parlar, M. 2012. *Interactive Operations Research with Maple: Methods and Models*. Springer Science & Business Media.

Rose, C., and M. D. Smith. 2002. *MathStatica: Mathematical Statistics with Mathematica*. Springer.

Serinaldi, F. 2009. "A Multisite Daily Rainfall Generator Driven by Bivariate Copula-Based Mixed Distributions". *Journal of Geophysical Research: Atmospheres* 114 (D10).

Shimizu, K. 1993. "A Bivariate Mixed Lognormal Distribution with an Analysis of Rainfall Data". *Journal of Applied Meteorology* 32 (2): 161–171.

Simoff, S., M. H. Böhlen, and A. Mazeika. 2008. *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, Volume 4404. Springer Science & Business Media.

Tanis, E., and Z. Karian. 1999. *Probability and Statistics: Explorations with Maple*. Prentice–Hall.

Tu, W., and H. Liu. 2002. "Zero-Inflated Data". *Wiley StatsRef: Statistics Reference Online*.

Tufte, E. R. 1983. *The Visual Display of Quantitative Information*. Graphics Press Cheshire, CT.

Tufte, E. R. 1991. *Envisioning Information.* Graphics Press Cheshire, CT.

Tufte, E. R. 1997. *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press Cheshire, CT.

Tufte, E. R. 2006. *Beautiful Evidence*. Graphics Press Cheshire, CT.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading, MA.

Yoo, C., K.-S. Jung, and T.-W. Kim. 2005. "Rainfall Frequency Analysis using a Mixed Gamma Distribution: Evaluation of the Global Warming Effect on Daily Rainfall". *Hydrological Processes* 19 (19): 3851–3861.

AUTHOR BIOGRAPHIES

CHRISTOPHER WELD is a Ph.D. candidate in the Department of Applied Science at The College of William & Mary, Williamsburg, VA. He holds a B.S. in mechanical engineering (2000) from Cornell University, NY, and an M.S. in computational operations research (2010) from The College of William & Mary, VA. He has formerly held a faculty position as an Assistant Professor in The Department of Mathematical Sciences at the United States Military Academy, West Point, NY. His e-mail address is ceweld@email.wm.edu.

LAWRENCE LEEMIS is a Professor in the Department of Mathematics at the College of William & Mary. He has formerly held faculty positions at The University of Oklahoma and Baylor University. He holds a B.S. in mathematics (1978), M.S. in mathematics (1980), and Ph.D. in industrial engineering (1984) from Purdue University. His email address is leemis@math.wm.edu.