Bayesian Forecasting and Decision Theory

Srini Sunil

Abstract

This paper concerns forecasting methodology and the application of new loss functions for decision analysis in studies of time series of non-negative counts. In particular, variations of the absolute percent error loss function are developed. The purpose for exploring new loss functions is to investigate, compare, and evaluate forecast performance using a Bayesian decision analysis approach. The development of novel loss functions is utilized in the context of forecasting consumer sales as there is very limited research in this space. Since common loss functions such as quadratic loss and absolute loss are not the most appropriate forms of loss to use in the consumer sales context, multivariate loss functions are developed and studied. We are particularly interested in scenarios encompassing decisions about the future values of low-count time series in consumer sales. Thus it is of interest to explore multivariate loss functions that can account for multiple days/products containing low-value count data. Simulations and applications of multi-step forecasting examples with different loss functions are explored. This is followed by a discussion of forecast accuracy and broader questions regarding other forms of loss and applications.

1 Introduction

Decision analysis is a practice in which a decision maker tries to maximize his/her expected utility after analyzing a particular decision space. A decision space is defined as all possible decisions that could be chosen. Each decision in the decision space has some corresponding utility associated with it. Thus the goal of decision analysis is to find the optimal decision that maximizes a decision maker's utility. In order to identify an optimal decision, two components are needed: a loss function and a probably distribution function. A loss function specifies the amount a decision maker will lose if some decision d is made and the future outcome is θ. The probability distribution function is on the future outcome. A decision can thus be made by taking into account the probabilities of all future outcomes and by assessing how to minimize expected loss. Using a decision rule called expected monetary value strategy, a decision maker is instructed to make some decision d^* that minimizes the expected loss (Smith, 2010). We define risk as the expected loss with respect to the probability distribution function described previously. Note that there are other forms of decision rules that can be used to optimize a risk function, but that is beyond the scope of this paper.

A loss function is parameterized by two variables, a decision and an observed outcome. From these variables, a loss metric is then calculated based on some specification of the loss function (Smith, 2010). Some commonly used loss functions include the squared error loss, absolute error loss, and 0-1 loss. Each of these loss functions also have their own optimal decisions (values), d^*, which are the mean, median, and mode, respectively. These individual d^* values minimize the risk of their corresponding loss functions.

Loss functions can be used in many contexts such as to optimize a machine learning algorithm, to assess economic risk, to estimate parameters of a model, and more. This paper focuses on the application of Bayesian decision theory and loss functions in the context of consumer sales forecasting. In consumer sales forecasting literature, many forecasting methods have been developed to predict sales outcomes over some period of days, whether it be through time series analysis and ARIMA models (Box et al., 2008), state-space models (Berry and West 2018; Berry et al. 2018 for examples), or others. The focus of this paper is on the development and investigation of novel, multivariate loss functions, which are used in the context of forecasting consumer sales. Given a multivariate forecast distribution of sales, a Bayesian decision theory approach is used to compute a series of point forecasts by minimizing the risk for a particular loss function. A loss function in general can be interpreted as a method that evaluates how well a forecast predicts actual sales. The further away a prediction is from the actual value, the greater the realized loss will be. The closer the prediction is to the actual value, the lower the realized loss. The purpose of minimizing risk in the sales context is to attain point forecasts that are as close as possible to the true sales outcome.

The motivation behind exploring new, multivariate loss functions in the context of consumer sales is that optimizing forecasts over common loss functions can lead to misleading results, particularly in the case of low-value count data (Kolassat 2016; Hyndman and Koehler 2006; Yelland 2009). With these misleading point forecasts, a decision maker would be unable to appropriately assess a current business situation and make an optimal decision. Thus new loss functions are needed. It is of interest in industry to create and better utilize multivariate loss functions so that instances when observed sales are 0 are more accounted for. In addition to developing more meaningful loss functions, there is very limited literature on a Bayesian decision analysis approach to evaluating optimal forecasts beyond traditional classes of loss functions. In this paper, Bayesian methods are used to evaluate optimal point forecasts given multivariate forecast distributions.

For the remainder of the paper, variations of the absolute percent error (APE) loss are investigated. In particular, weighted absolute percent error (WAPE), zero adjusted percent error (ZAPE), and weighted absolute forecast error (WAFE) loss functions are described and further explored in the following sections. These are three completely new loss functions that have come out of industry after close collaboration with 84.51°. WAPE and WAFE are interesting to explore because they are quasi-percent error metrics that can be used to evaluate forecasts of sales over multiple days or to evaluate forecasts of sales of multiple items at once. Thus these two loss functions provide general measures of forecast accuracy. ZAPE is interesting to explore because it is a relative error metric that can be applied to cases where observed values are 0. ZAPE could be very useful in assessing forecast accuracy for low-valued count data if the loss function is defined appropriately. Optimizing over these three novel loss functions can provide insight into forecasting from a decision theory perspective. In this paper, the decision theory and its respective analysis is new, and these newly developed loss functions are explored in both a simulated and application setting.

The traditional and new loss functions and their respective implied risk functions are introduced in Section 2. Decision analysis and methods of optimization for each loss function are described in Section 3. Simulations and properties observed for each loss function are described in Section 4. An application of the loss functions to sales data is described in Section 5. The paper concludes with a discussion on the findings and potential next steps in Section 6. An appendix is available in Section 7.

2 Loss Functions

2.1 Quadratic Error

2.1.1 Loss

Quadratic error loss is defined as the following:

$$L(y, f) = (y - f)^2$$

In this definition, y represents the sales outcome and f is the corresponding point forecast.

2.1.2 Risk

The implied risk function is defined as the following:

$$R(f) \equiv \int_{y \in Y} L(y, f) p(y) dy = \int_{y \in Y} (y - f)^2 p(y) dy$$

In this definition, $p(y)$ is the forecast p.d.f. and the risk function is minimized through direct calculus.

2.2 Absolute Error

2.2.1 Loss

Absolute error loss is defined as the following:

$$L(y, f) = |y - f|$$

In this definition, y represents the sales outcome and f is the corresponding point forecast.

2.2.2 Risk

The implied risk function is defined as the following:

$$R(f) \equiv \int_{y \in Y} L(y, f) p(y) dy = \int_{y \in Y} |y - f| p(y) dy$$

In this definition, $p(y)$ is the forecast p.d.f. and the risk function is minimized through direct calculus.

2.3 Absolute Percent Error

2.3.1 Loss

Absolute percent error (APE) loss is defined as the following:

$$L(y, f) = \frac{|y - f|}{y}$$

In this definition, y represents the sales outcome and f is the corresponding point forecast.

2.3.2 Risk

The implied risk function is defined as the following:

$$R(f) \equiv \int_{y \in Y} L(y, f) p(y) dy = \int_{y \in Y} \frac{|y - f|}{y} p(y) dy$$

In this definition, $p(y)$ is the forecast p.d.f. and the risk function is minimized through direct calculus. The following three loss functions in this section will be extensions/variations of APE loss.

2.3.3 Context

This is a less commonly known loss function that is used in the consumer sales space as a measure of forecast accuracy. It is a univariate loss function that is a slight variation of absolute loss function. Typically it is used to measure forecast loss for a single product on a single day.

2.4 Weighted Absolute Percent Error

2.4.1 Loss

Weighted absolute percent error (WAPE) loss is defined as the following:

$$L(\mathbf{y}, \mathbf{f}) = a(\mathbf{y}) \sum_{i=1:n} |y_i - f_i| \quad \text{with} \quad a(\mathbf{y})^{-1} = \sum_{i=1:n} y_i.$$

In this definition, $\mathbf{y} = y_{1:n} = (y_1, \ldots, y_n)'$ represents a vector of sales outcomes of a particular item over some period of n days or a vector of sales outcomes of n products for a particular day. $\mathbf{f} = f_{1:n} = (f_1, \ldots, f_n)'$ is the corresponding point forecast. This loss function assumes that at least one $y_i > 0$, implying that $0 < a(\mathbf{y}) \leq 1$. WAPE is very similar to APE in the sense that WAPE is more of a quasi-percent error loss calculated over some n days/products instead of over a single observation.

2.4.2 Risk

The implied risk function is defined as the following:

$$R(\mathbf{f}) \equiv \sum_{\mathbf{y}} L(\mathbf{y}, \mathbf{f}) p(\mathbf{y}) d\mathbf{y} = \sum_{i=1:n} R_i(f_i), \quad \text{where}$$

$$R_i(f_i) = \sum_{y_i=0}^{\infty} |y_i - f_i| h_i(y_i) dy_i \quad \text{with} \quad h_i(y_i) = \sum_{\mathbf{y}_{-i}} a(\mathbf{y}) p(\mathbf{y}) d\mathbf{y}_{-i}.$$

In this definition, $R_i(f_i)$ is a risk function for day/product i, $p(\mathbf{y})$ is the forecast p.d.f, and $h_i(y_i)$ is the marginal distribution for day/product i. Since each i depends only on f_i, $R(\mathbf{f})$ is simply minimized by independently minimizing and taking the sum of the n individual risk functions.

2.4.3 Context

This is the first novel loss function that is investigated. In the context of consumer sales, this loss function is a metric that the provides a quasi-percent error for a particular product over a period of n days, which could culminate into 1 week, 2 weeks, or more. This loss function can also be used in the context of n different products for a single day. For instance, it could be used for obtaining point forecasts for n pasta products, n dairy products, or others.

In the decision making context, WAPE loss extends on APE loss such that more days/products are used to calculate the loss metric. For a store manager, the WAPE loss can be more useful and informative than the APE loss because it is a more generalized metric that takes into accounts more days/products. Depending on how it is used, WAPE loss could be informative on longitudinal trends and/or product basket trends.

However, the primary downside to WAPE loss is the assumption that $a(\mathbf{y}) \neq 0$, meaning that this loss function would not be suitable in forecasting scenarios in which sales tend to be low and have a high probability of being 0.

2.5 Zero Adjusted Percent Error

2.5.1 Loss

Zero adjusted percent error (ZAPE) loss is defined as the following:

$$L_i(y_i, f_i) = c_i(f_i) I(y_i = 0) + \frac{|y_i - f_i|}{y_i} I(y_i > 0).$$

In this definition, $L_i(y_i, f_i)$ represents the loss for each day/product i, y_i represents the sales outcome for day/product i, f_i is the corresponding point forecast, $c_i(\cdot)$ is a non-negative increasing

function (which can possibly be day-dependent), and I is the indicator function that is dependent on the sales outcome of day/product i. Unlike WAPE, ZAPE does not assume that at least one $y_i > 0$ and allows for cases where all $y_i = 0$.

2.5.2 Risk

The implied risk function is defined as the following:

$$R(\mathbf{f}) \equiv \sum_{\mathbf{y}} L(\mathbf{y}, \mathbf{f}) p(\mathbf{y}) d\mathbf{y} = \sum_{i=1:n} R_i(f_i), \quad \text{where}$$

$$R_i(f_i) = c_i(f_i)\pi_{i0} + \sum_{y_i=1}^{\infty} |y_i - f_i| y_i^{-1} p_i(y_i) dy_i.$$

In this definition, $R_i(f_i)$ represents the risk function for each day/product i and $\pi_{i0} = Pr(y_i = 0)$ is the explicit forecast probability of zero sales under $p_i(y_i)$, which is the forecast probability at day/product i. Similar to WAPE, since each $R_i(f_i)$ is non-negative and depends only on f_i, $R(\mathbf{f})$ is simply minimized by independently minimizing and taking the sum of the n individual risk functions. With this formulation of ZAPE loss, the risk is more dominated by the second term for items with higher probabilities of taking the value of 0. In these scenarios, larger point forecasts are increasingly penalized.

2.5.3 Context

This is the second novel loss function that is investigated. Similar to WAPE, the ZAPE again is a quasi-percent error metric that generates an overall measure of forecast accuracy. However unlike WAPE, the primary motivation for ZAPE is to define a loss function that applies to cases in which $y_i = 0$ occurs frequently. In the consumer sales context, ZAPE loss could be applied to products that typically experience low levels of sales. Luxury branded goods, whether they be pastas or ice creams, are less likely to be sold (and can have days where they do not sell) primarily because of their relatively higher price point.

Once again, given the multivariate nature of the ZAPE loss, n is not only representative of the number of days, but can also represent a series of different products within a particular basket of goods. Depending on the forecasting context, ZAPE loss can be utilized for a single product or for multiple products that experience low levels of sales.

2.6 Weighted Absolute Forecast Error

2.6.1 Loss

Weighted absolute forecast error (WAFE) loss is defined as the following:

$$L(\mathbf{y}, \mathbf{f}) = a(\mathbf{y}, \mathbf{f}) \sum_{i=1:n} |y_i - f_i| \quad \text{with} \quad a(\mathbf{y}, \mathbf{f})^{-1} = \sum_{i=1:n} (y_i + f_i)/2.$$

In this definition, $\mathbf{y} = y_{1:n} = (y_1, \ldots, y_n)'$ represents a vector of sales outcomes of a particular item over some period of n days or a vector of sales outcomes of n products for a particular day. $\mathbf{f} = f_{1:n} = (f_1, \ldots, f_n)'$ is the corresponding point forecast. Similar to WAPE, this loss function

assumes that at least one $y_i > 0$, implying that $0 < a(\mathbf{y}, \mathbf{f}) \leq 1$. The difference in WAFE is that $a(\cdot)^{-1}$ is dependent on the full forecast vector whereas WAPE does not, which creates complications in evaluating the optimal forecasts.

2.6.2 Risk

The implied risked function is defined as the following:

$$R(\mathbf{f}) \equiv \sum_{\mathbf{y}} L(\mathbf{y}, \mathbf{f}) p(\mathbf{y}) d\mathbf{y} = \sum_{i=1:n} R_i(f_i), \quad \text{where}$$

$$R_i(f_i) = \sum_{y_i=0}^{\infty} |y_i - f_i| h_i(y_i|\mathbf{f}) dy_i \quad \text{with} \quad h_i(y_i|\mathbf{f}) = \sum_{\mathbf{y}_{-i}} a(\mathbf{y}, \mathbf{f}) p(\mathbf{y}) d\mathbf{y}_{-i}.$$

In this definition, $R_i(f_i)$ is a risk function for day/product i, $p(\mathbf{y})$ is the forecast p.d.f, and $h_i(y_i|\mathbf{f})$ is the marginal distribution for day i conditional on $\mathbf{f}$. Since each $R_i(f_i)$ is dependent on the full forecast vector, $R(\mathbf{f})$ cannot simply be minimized by minimizing each individual risk function. An more detailed explanation of WAFE optimization is described in the next section.

2.6.3 Context

This is the third novel loss function that is investigated. WAFE is defined similarly to WAPE, meaning that WAFE produces a quasi-percent error metric that aggregates over multiple items/days (depending on the consumer sales forecasting context). However, the difference between the two loss functions lies in how $a(\cdot)^{-1}$ is defined. WAFE depends on the full forecast vector $\mathbf{f}$ and WAPE does not. But similar to WAPE, the primary downside is the assumption that $a(\mathbf{y}, \mathbf{f}) \neq 0$, meaning that this loss function may not be suitable in sales scenarios in which sales are low and have a high probability of being 0.

Just like WAPE and ZAPE, WAFE can be used for forecasting over n days or for forecasting n products.

3 Decision Analysis and Optimization

3.1 Quadratic Error

Through direct calculus, it can be calculated that the optimal f is the mean of the forecast distribution $p(y)$.

3.2 Absolute Error

Through direct calculus, it can be calculated that the optimal f is the median of the forecast distribution $p(y)$.

3.3 Absolute Percent Error

Through direct calculus, it can be calculated that the optimal f is the median of the forecast distribution $p(y)/y$. The derivation is found in section 7.1 of the appendix.

3.4 Weighted Absolute Percent Error

In order to optimize WAPE, first define the joint p.d.f $g(\mathbf{y}) = ca(\mathbf{y})p(\mathbf{y})$, where c is the implied normalizing constant. For each day/product i, define $g_i(y_i)$ as the univariate marginal p.d.f. for y_i. Based on this definition, it can be seen that $h_i(y_i) = g_i(y_i)/c$. And since c is just a positive constant, $R_i(f_i)$ is minimized by minimizing $g_i(y_i)$ under the absolute loss error, which is the median of this joint p.d.f. Thus, each f_i is the median of the corresponding marginal distribution $g_i(y_i)$ at day i. It is important to note that because of the way WAPE is defined, the optimal f_i will always be less than or equal to the median of $p(y_i)$

3.5 Zero Adjusted Percent Error

First define the case $c_i(f_i) = f_i$. This is the first special case that is explored. Thus in order to optimize ZAPE, direct calculus shows that $R_i(f_i)$ is minimized at f_i where

$$f_i = \begin{cases} 0 & q_i \leq 0 \\ G_i^{-1}(q_i) & q_i > 0 \end{cases}$$

$G_i(y)$ is the c.d.f on integer values where $x \geq 1$. The derivation of the optimal f_i can be found in section 7.2 of the appendix. Its p.d.f. is defined as $g_i(y) = c_i x^{-1} p_i(y) I(y \geq 1)$. c_i is the implied normalizing constant for the corresponding day i and $q_i = (1 - c_i \pi_{i0})/2$. Note that the value of q_i will always be less than 0.5, implying that f_i will always be below the median of $g_i(y)$. Thus as π_{i0} increases, q_i decreases, which causes f_i to become closer and closer to zero.

3.6 Weighted Absolute Forecast Error

Given that the marginal distribution $h_i(y_i|\mathbf{f})$ is conditional on the entire forecast vector $\mathbf{f}$, complications arise and WAFE is not amenable to the same analysis as WAPE or ZAPE. This kind of complication can be generalized to any loss function that depends on a full forecast vector. Since there is no straightforward way of calculating the optimal forecast values for the full forecast vector, an algorithm to numerically approximate the WAFE-optimal forecasts is explored. This is a new form of evaluation and whether or not this algorithm is proven to converge is still an open question.

The optimization approximation is based on the following algorithm:

1. First set $\mathbf{f} = \mathbf{f}^0$ at an initial guess at the optimal solution. An obvious starting value is the MAPE-optimal forecast. However, other initial starting values could include the mean, median, or a random series of points.

2. For each day/product i, update the value f_i^0 to f_i^1 where the latter is the WAPE-optimal forecast replacing $a(\mathbf{y})$ in the WAPE framework with $a(\mathbf{y}, \mathbf{f}^0)$.

3. Then iterate: for some steps $t > 1$, now define $\mathbf{f}^t$ based on $\mathbf{f}^{t-1}$ to iteratively update the forecast vector at $t = 1$ in the above step.

4. Finally, stop at some chosen final iterate t and/or when the changes in the resulting risk $R(\mathbf{f}^t)$ stabilize to some defined level of accuracy/resolution. In general, the closer the starting values are to the optimal forecast values, a fewer the number of iterations will be required.

With this process, each iteration of the optimization process will generally decrease the value of the risk function $R(\mathbf{f}^t)$. This is based on the observation that $R_i(\mathbf{f}^{ti}) \leq R_i(\mathbf{f}^{t-1})$ for each item i and each iterate t where $\mathbf{f}^{ti}$ is the $\mathbf{f}^{t-1}$ vector changed only by inserting f_i^t in the element for item i. As a result, $R(\mathbf{f}^t) \leq R(\mathbf{f}^{t-1})$ at each iterate, and then– since $R(\mathbf{f}) \geq 0$ everywhere– the algorithm will converge to some $\mathbf{f}$ vector.

4 Computation and Implementation: Synthetic Data

4.1 Context

This section explores simulated examples for WAPE, ZAPE, and WAFE. The purpose for investigating the results of a simulation is to better understand each loss function's respective properties. In addition to simple simulated scenarios, additional hypothetical scenarios are explored with WAPE, ZAPE, and WAFE as a way to evaluate and compare how optimal forecasts vary for each loss function.

4.2 Importance Sampling

In order to optimize WAPE, ZAPE, and WAFE, importance sampling is used. Importance sampling is a Monte Carlo sampling method in which samples from an original distribution are taken and then transformed to simulate the desired distribution. Refer to section 7.2 of the appendix for more mathematical/technical details on importance sampling.

The main motivation for using importance sampling over a brute force method is scalability. Given some large number n, the probabilities of all the combinations of sales over n would need to be computed, and this becomes computationally infeasible as n becomes larger. Importance sampling resolves this issue since not all combinations of sales over n days/products is necessary to optimize WAPE, ZAPE, and WAFE.

4.3 Weighted Absolute Percent Error

4.3.1 Set-Up

Since WAPE requires that at least one $y_i > 0$, a shifted Poisson distribution is used to ensure this requirement is met. The shifted Poisson is defined as the following:

$$y_i \sim 1 + Poisson(\mu),$$

where each y_i is an independently and identically distributed component of the vector $\mathbf{y}$. Instead of having at least one $y_i > 0$, using shifted Poisson results in every $y_i \geq 1$. This definition of the shifted Poisson also assumes that μ remains the same for all n days.

As defined before, $g(\mathbf{y}) = ca(\mathbf{y})p(\mathbf{y})$, where c is the implied normalizing constant and $p(\mathbf{y})$ is the forecast p.d.f. Given that $p(y_i)$ is an i.i.d. shifted Poisson distribution, the likelihood distribution, $p(\mathbf{y})$, is defined as the following:

$$p(\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\mu}\lambda^{y_i-1}}{(y_i-1)!}.$$

This looks almost identical to the Poisson likelihood. However, since $p(y_i)$ is shifted by +1, it is necessary to subtract each y_i value by 1 in order to attain the appropriate probability.

4.3.2 Optimization

Importance sampling is used to optimize WAPE. Firstly $N = 1{,}000{,}000$ samples of $\mathbf{y}$ are drawn from the shifted Poisson distribution. Then for each sample, the corresponding $a(\mathbf{y})$ value is computed, which is termed as the weight of that particular sample. Each weight is then normalized such that $w_i = a(\mathbf{y}_i)/\sum_{i=1}^{N} a(\mathbf{y}_i)$, implying that $\sum_{i=1}^{N} w_i = 1$. From each of the w_i's, the marginal distribution $g_i(y_i)$ is computed for each day, and then the median of each $g_i(y_i)$ is calculated, which is the optimal forecast f_i.

4.3.3 Results

With this approach, every marginal distribution of $g_i(y_i)$ is identical because the sales outcome for each day i is simulated from the same shifted Poisson distribution. After running many simulations with this approach for various values of μ and n, it is observed that the optimal f_i is either μ or $1+\mu$. Table 1 shows a brief distribution of how f_i varies with μ and n. From the table, it is seen that $f_i = \mu$ when $n = 1$ or $n = 2$ and $f_i = 1 + \mu$ when $n \geq 3$.

The weights and the ESS of the simulated WAPE importance sampling are also evaluated. Figure 1 is an example histogram of the weights for a sample size of N=1,000,000 with $\mu = 4$. The ESS is 98.82%.

μ	n=1	n=2	n=3	... n=14
1	1	1	2	2
2	2	2	3	3
3	3	3	4	4
4	4	4	5	5
5	5	5	6	6
6	6	6	7	7
7	7	7	8	8
8	8	8	9	9
9	9	9	10	10
10	10	10	11	11
11	11	11	12	12
12	12	12	13	13
13	13	13	14	14
14	14	14	15	15
15	15	15	16	16

Table 1: WAPE Simulation Results at f_i

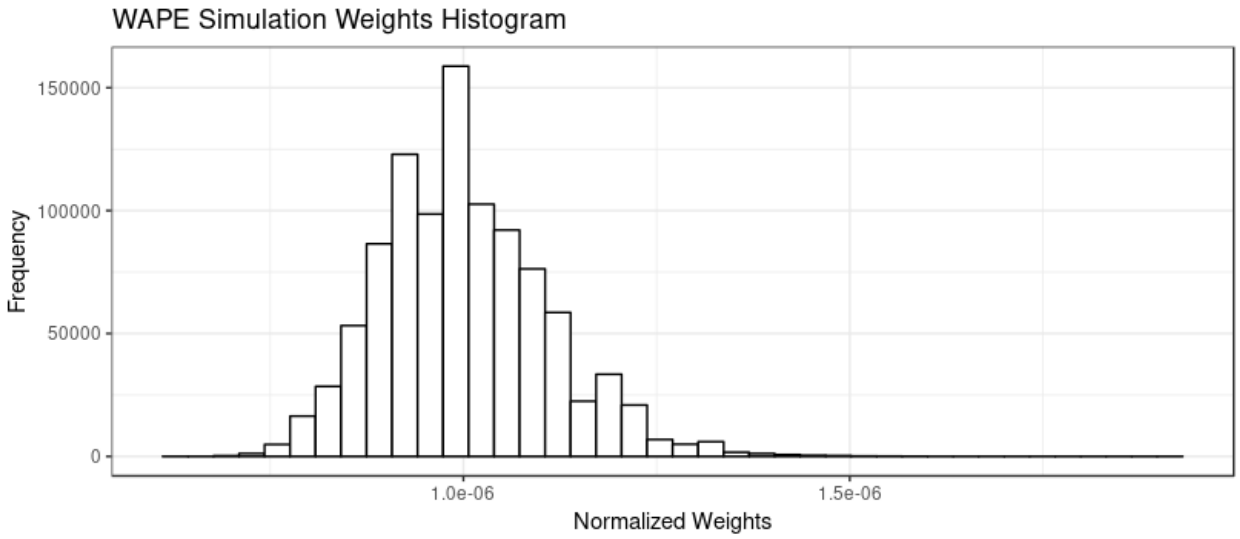

Figure 1: Example Importance Sampling Weights

4.3.4 Additional Simulated Scenario #1

The first hypothetical scenario that is explored is one in which the first set of 7 days of sales are relatively low and the second set of 7 days of sales are relatively high. Specifically, the first 7 days are simulated from a Poisson distribution with $\mu = 3$ and the next 7 days are simulated from a Poisson distribution with $\mu = 20$. A shifted Poisson distribution is not used because drawing a series of 14 0-sales has a near 0 probability. Figure 2 shows the mean, median, and WAPE optimal forecasts. The median and the WAPE optimal forecasts are the same over the 14-day period, which is 3 in the first 7 days and 14 in the second 7 days. The mean optimal forecasts are very close to the median and WAPE forecasts. When the first 7 days are of relatively high sales and when the second 7 days are of relatively low sales, similar results are observed.

WAPE importance sampling normalized weights from this scenario is shown in Figure 3. The ESS is 99.37%.

4.3.5 Additional Simulated Scenario #2

The second hypothetical scenario that is explored is one in which the level of sales alternates from low to high every day. The low level of sales are simulated from a Poisson distribution with $\mu = 3$ and the high level of sales are simulated from a Poisson distribution with $\mu = 20$. Figure 4 is a plot of the mean, median, and WAPE optimal forecasts. For days with low levels of sales, all forms of optimal forecasts are all equal to 3. For days with high levels of sales, mean and median optimal forecasts are 20 whereas the WAPE optimal forecast is 19, which is interesting to note.

WAPE importance sampling normalized weights from this scenario is shown in Figure 5. The ESS is 94.91%.

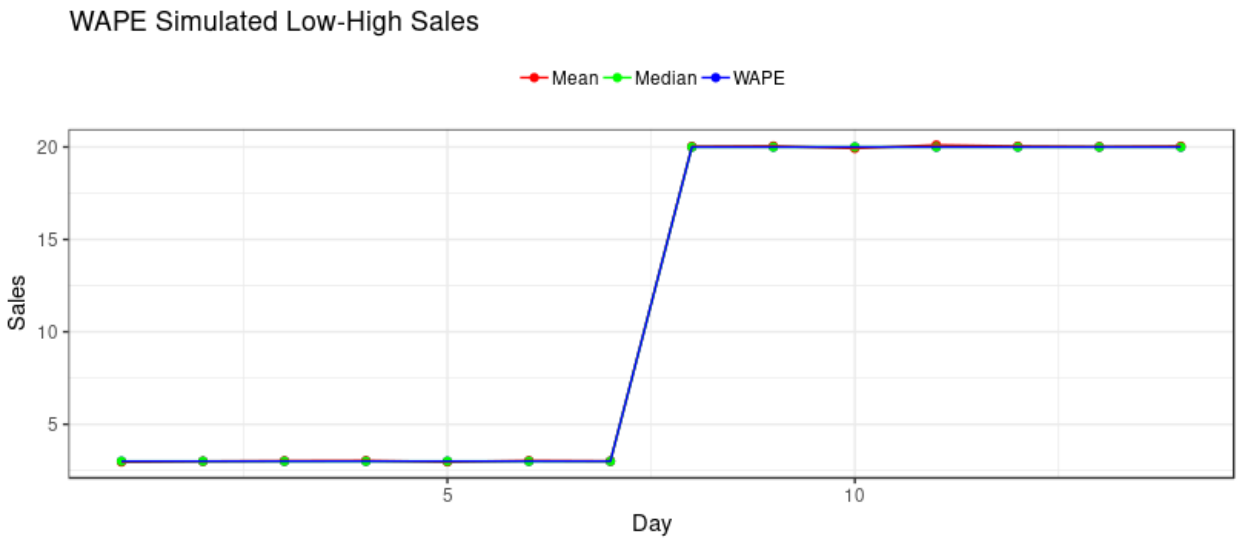

Figure 2: Scenario 1 Simulation

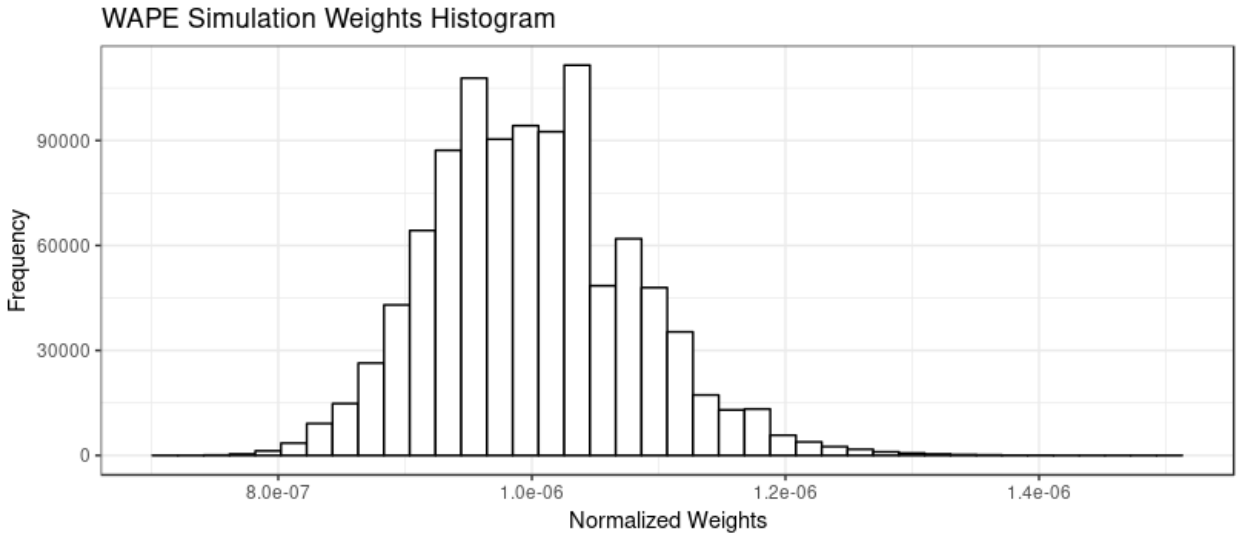

Figure 3: Scenario 1 Importance Sampling Weights

4.3.6 Additional Simulated Scenario #3

The third hypothetical scenario that is explored is one in which the level of sales progresses from low to medium to high and back down for over a period 14 days. The low level of sales are simulated from a Poisson distribution with $\mu = 3$, the medium level of sales are simulated from a Poisson distribution with $\mu = 11$, and the high level of sales are simulated from a Poisson distribution with $\mu = 20$. Unlike the previous hypothetical scenario, all the optimal forecasts are equal to the respective μ's of the low, medium, and high levels of sales as shown in Figure 6. It is interesting to note that the addition of the medium level of sales results in the WAPE optimal forecast for high level of sales to be 20 instead of 19.

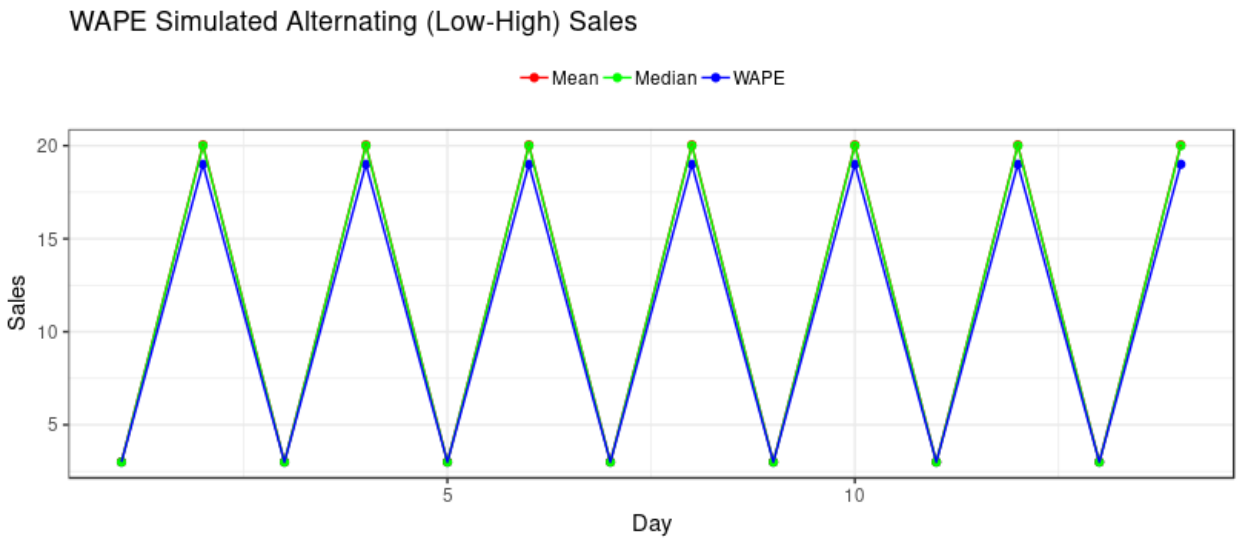

Figure 4: Scenario 2 Simulation

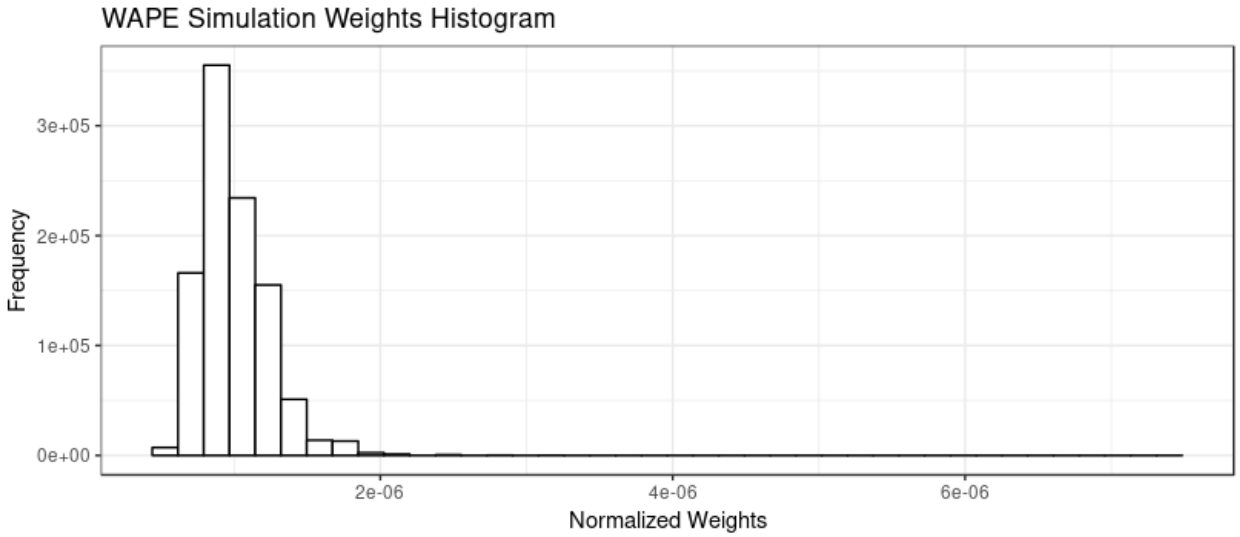

Figure 5: Scenario 2 Importance Sampling Weights

WAPE importance sampling normalized weights from this scenario is shown in Figure 7. The ESS is 99.31%.

4.4 Zero Adjusted Percent Error

4.4.1 Set-Up

Unlike WAPE, ZAPE does not require that at least one $y_i > 0$, and therefore an ordinary Poisson distribution is used:

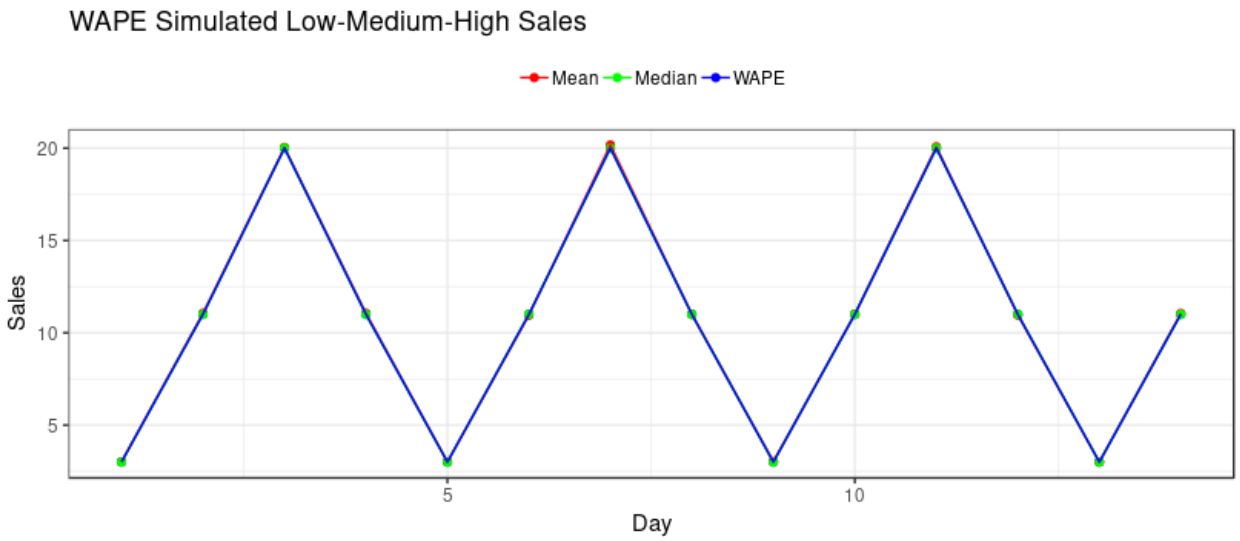

Figure 6: Scenario 3 Simulation

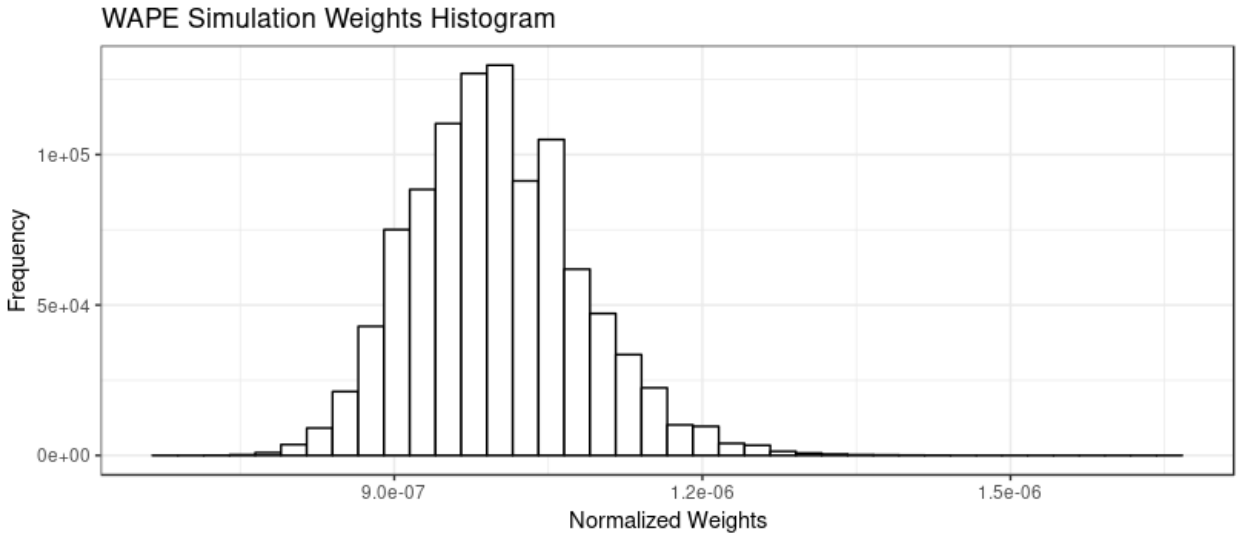

Figure 7: Scenario 3 Importance Sampling Weights

$$y_i \sim Poisson(\mu),$$

where each y_i is an independently and identically distributed component of the vector $\mathbf{y}$. This means that there is the possibility that $\mathbf{y}$ may consist of all zeros.

As defined before, $g_i(y) = c_i y^{-1} p_i(y) I(y \geq 1)$, where c is the implied normalizing constant. $G_i(y)$ is the corresponding c.d.f. and $G_i^{-1}(q_i) \leq \sum_{i=1:n} g_i^{-1}(q_i)$ if $q_i > 0$, where i is the first value that satisfies this criteria.

μ	n=1	n=2	n=3	... n=14
0	0	0	0	0
1	1	1	1	1
2	1	1	1	1
3	2	2	2	2
4	3	3	3	3
5	4	4	4	4
6	5	5	5	5
7	6	6	6	6
8	7	7	7	7
9	8	8	8	8
10	9	9	9	9
11	10	10	10	10
12	11	11	11	11
13	12	12	12	12
14	13	13	13	13
15	14	14	14	14

Table 2: ZAPE Simulation Results at f_i

4.4.2 Optimization

Importance sampling is used to optimize ZAPE, a process similar to WAPE. Firstly, $N = 1{,}000{,}000$ samples of $\mathbf{y}$ are drawn from the Poisson distribution. Next the inverse of every simulated sales outcome for each day is computed. In other words, the inverse of each y_i in every simulated $\mathbf{y}$ is calculated, and the inverse of each $\mathbf{y}$ is denoted as $\mathbf{y}^{-1}$. For all y_i's that had a value of 0, the inverse was set as 0. For all of the inverses with $y_i^{-1} > 0$, the corresponding ith day is then normalized amongst all $\mathbf{y}^{-1}$, which is proportional to $g_i(y)$ as defined previously. Then the normalizing constant, c_i and q_i are calculated. q_i is then used to find the optimal forecast f_i.

4.4.3 Results

With this approach, every marginal distribution of $g_i(y_i)$ is identical because the sales outcome for each day i is simulated from the same Poisson distribution. This implies that every q_i value is the same, meaning that the optimal forecast, f_i, is also the same. After running many simulations with this approach for various values of μ and n, it is observed that the optimal f_i is either μ or $\mu - 1$. Table 2 shows a brief distribution of how f_i values with μ and n. From the table, it is seen that f_i = μ when $\mu = 0$ or $\mu = 1$ and f_i = $\mu - 1$ when $\mu \geq 2$.

The weights and the ESS of the simulated ZAPE importance sampling are also evaluated. Because of the way ZAPE is optimized, there is a set of 14 ESS values and histograms. The ESS for all 14 days for a sample size of N=1,000,000 with $\mu = 4$ ranges from 68.82% to 68.92%.

4.4.4 Additional Simulated Scenario #1

This hypothetical scenario is the same one from section 4.3.4. Figure 8 is a plot of the mean, median, and ZAPE optimal forecasts. The median and the mean optimal forecasts are the same over the 14-day period, which is 3 in the first 7 days and 14 in the second 7 days. The ZAPE optimal forecasts are $\mu - 1$ for each day. The ESS for the first 7 days ranges from 72.03% to 72.10% and the ESS for the last 7 days ranges from 93.96% to 94.01%.

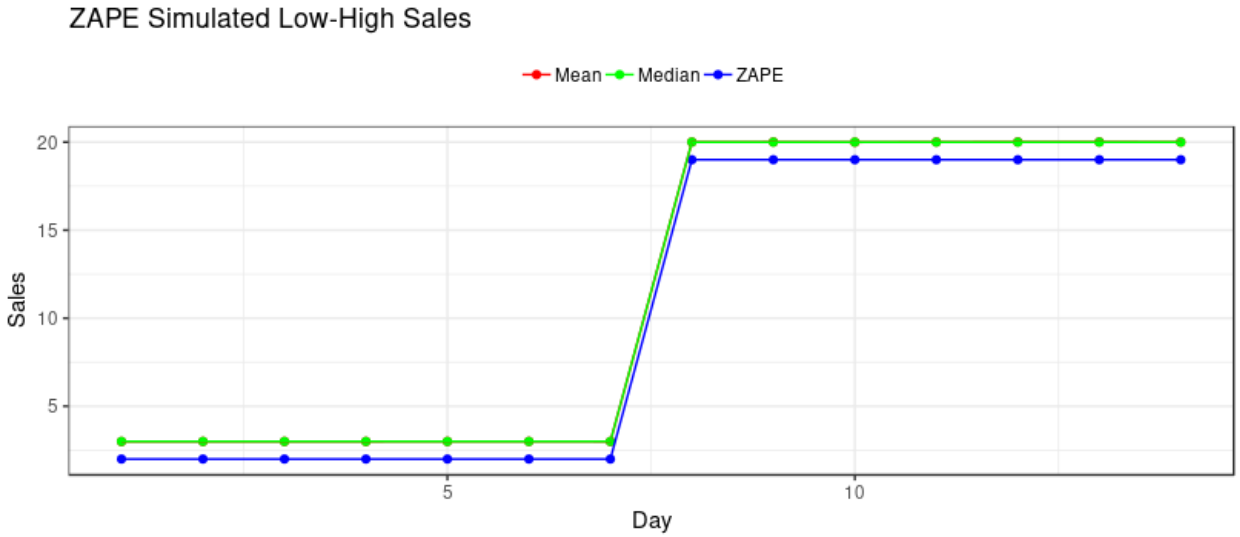

Figure 8: Scenario 1 Simulation

4.4.5 Additional Simulated Scenario #2

This hypothetical scenario is the same one from section 4.3.5 Figure 9 is a plot of the mean, median, and ZAPE optimal forecasts. Similar to the previous scenario, the mean and median optimal forecasts are the same over the 14-day period and are equivalent to the μ of the respective Poisson distribution. Once again, the ZAPE optimal forecasts are $\mu - 1$ for each day. The ESS for the low-sale days is 72.08% and the ESS for the high-sale days is 93.96%.

4.4.6 Additional Simulated Scenario #3

This hypothetical scenario is the same one from section 4.3.6. Once again, the mean and median optimal forecasts are the same over the 14-day period and are equivalent to the μ of the respective Poisson distribution. The ZAPE optimal forecasts are $\mu - 1$ for each day as shown in Figure 10. The ESS for the low-sale days range from 72.06% to 72.08%. The ESS for the medium-sale days range from 85.64% to 85.93%. The ESS for the high-sale days range from 93.97% to 94.01%.

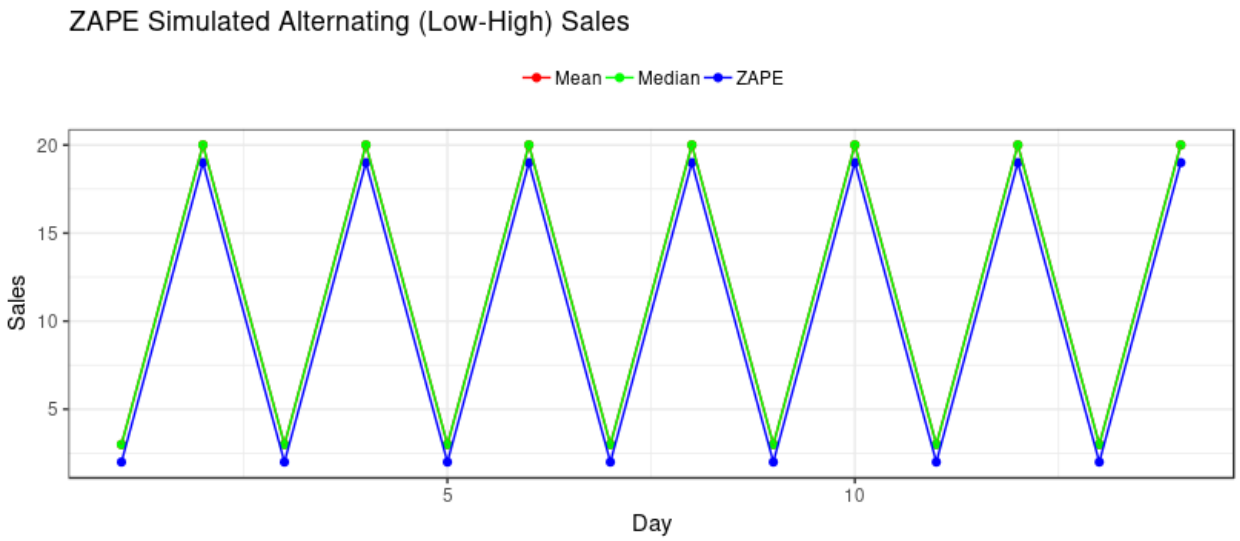

Figure 9: Scenario 2 Simulation

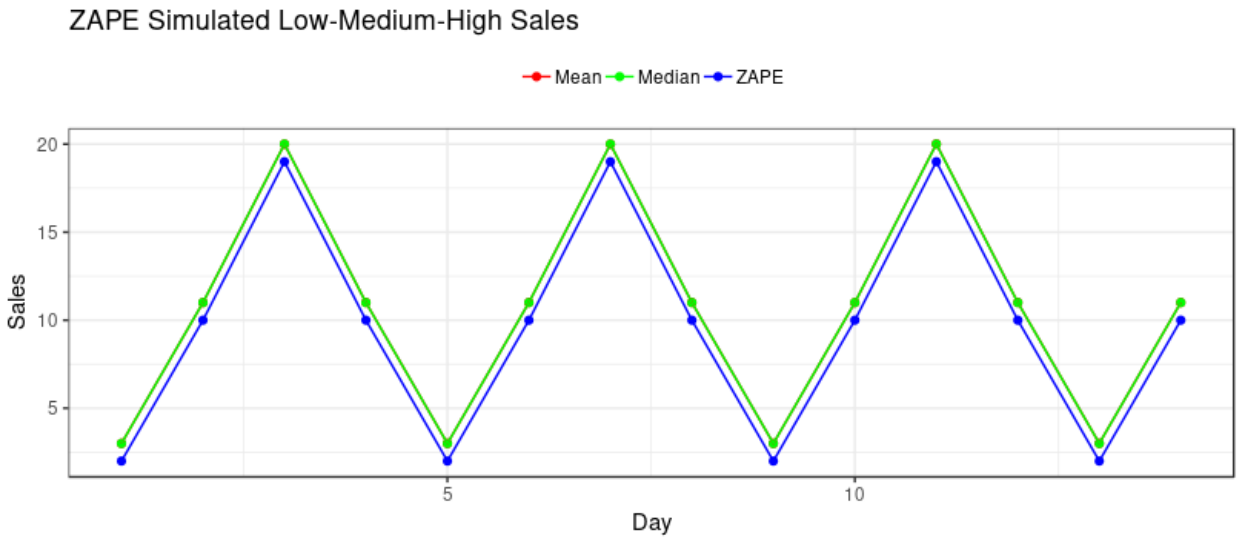

Figure 10: Scenario 3 Simulation

4.5 Weighted Absolute Forecast Error

4.5.1 Set-Up

Similar to WAPE, $a(\mathbf{y}, \mathbf{f})^{-1}$ needs to be greater than 0, but there is no requirement that $y_i > 0$. But to ensure $a(\mathbf{y}, \mathbf{f})^{-1} > 0$, a shifted Poisson distribution is used to ensure this requirement is met, just as it was implemented in the WAPE simulation.

4.5.2 Optimization

Importance sampling is used to optimize WAFE, a process similar to WAPE and ZAPE. Firstly, 1,000,000 samples of $\mathbf{y}$ are drawn from the shifted Poisson distribution. Next, the inverse of every simulated sales outcome for each day is computed. In other words, the inverse of each y_i in every simulated $\mathbf{y}$ is calculated, and the inverse of each $\mathbf{y}$ is denoted as $\mathbf{y}^{-1}$. For all y_i's that had a value of 0, the inverse was set as 0. For all of the inverses with $y_i^{-1} > 0$, the corresponding ith day is then normalized amongst all $\mathbf{y}^{-1}$. MAPE optimization is then implemented for each day, and this initial forecast vector is defined as $\mathbf{f}^0$. Using $\mathbf{f}^0$ to begin, the WAPE framework is applied with $a(\mathbf{y}, \mathbf{f}^i)^{-1}$, where i corresponds with the ith iteration of the optimal forecast vector $\mathbf{f}^i$. Iterations are stopped once $\mathbf{f}^i$ no longer changes.

4.5.3 Results

With this approach, every marginal distribution of $g_i(y_i)$ is identical because the sales outcome for each day i is simulated from the same shifted Poisson distribution. After running many simulations for various values of μ and n, it is observed that the optimal f_i is always $1 + \mu$ in this simulated scenario. It is interesting to note that only two iterations of $\mathbf{f}^i$ are needed in order to optimize WAFE under these circumstances.

The weights and the ESS of the simulated WAFE importance sampling are also evaluated. Figure 11 is an example histogram of the weights for a sample size of N=10,000 with $\mu = 4$ over 14 days. The ESS is 99.71%. The reason a smaller sample size is used is due to computation time.

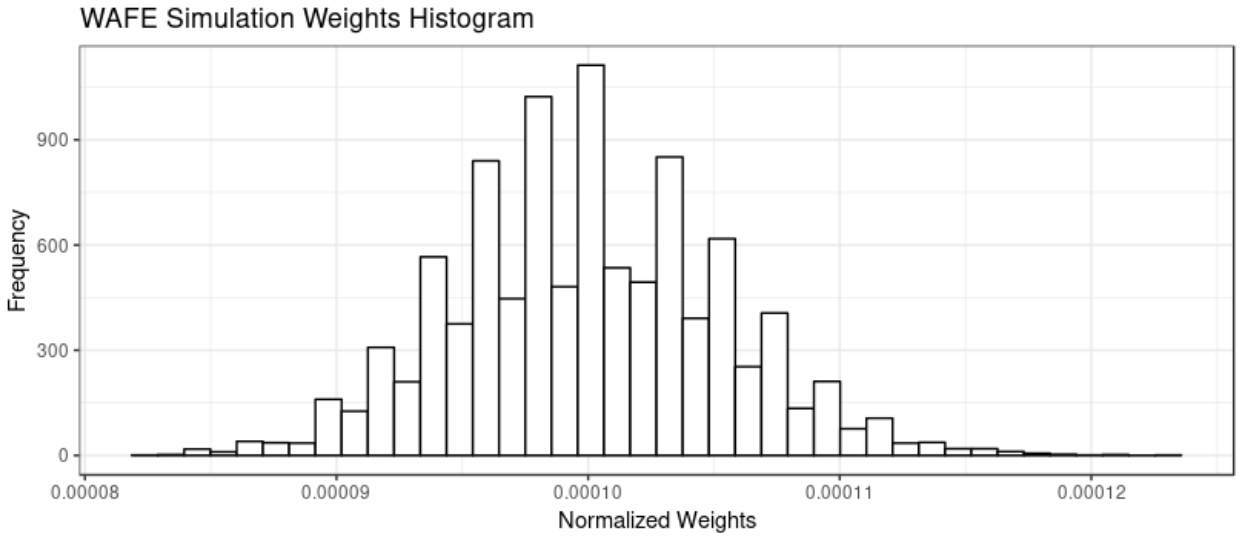

Figure 11: Example Importance Sampling Weights

4.5.4 Additional Simulated Scenario #1

This hypothetical scenario is the same one from section 4.3.4, but with $N = 10{,}000$. Figure 12 is a plot of the mean, median, and WAFE optimal forecasts. All forms of optimal forecasts are the same over the 14-day period, which is 3 in the first 7 days and 14 in the second 7 days.

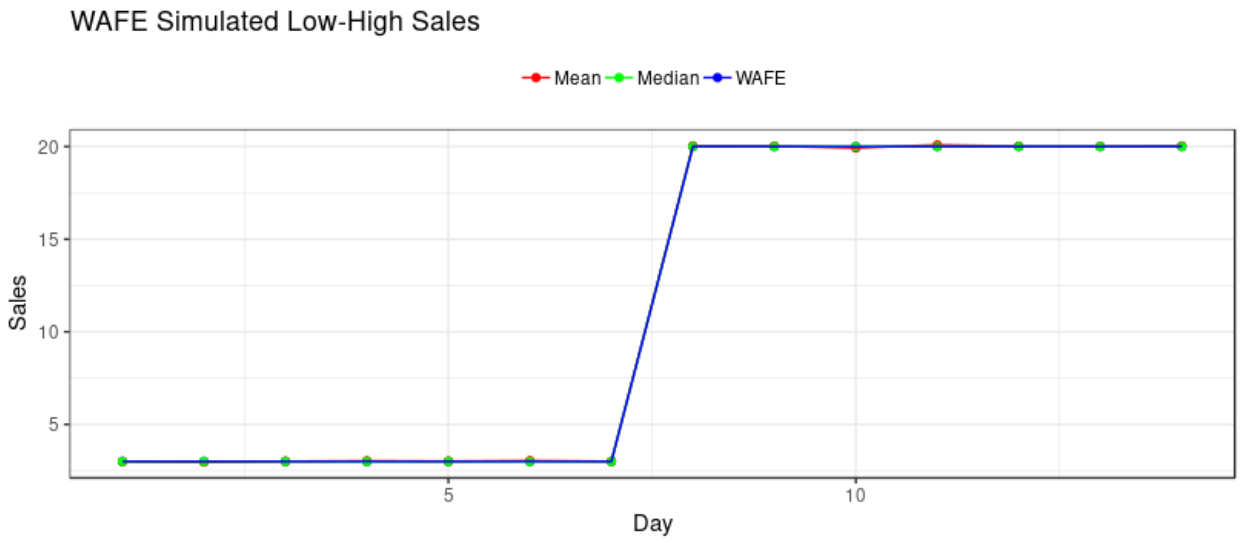

Figure 12: Scenario 1 Simulation

Figure 13 is the histogram of the WAFE importance sampling normalized weights from this scenario. The ESS is 99.85%.

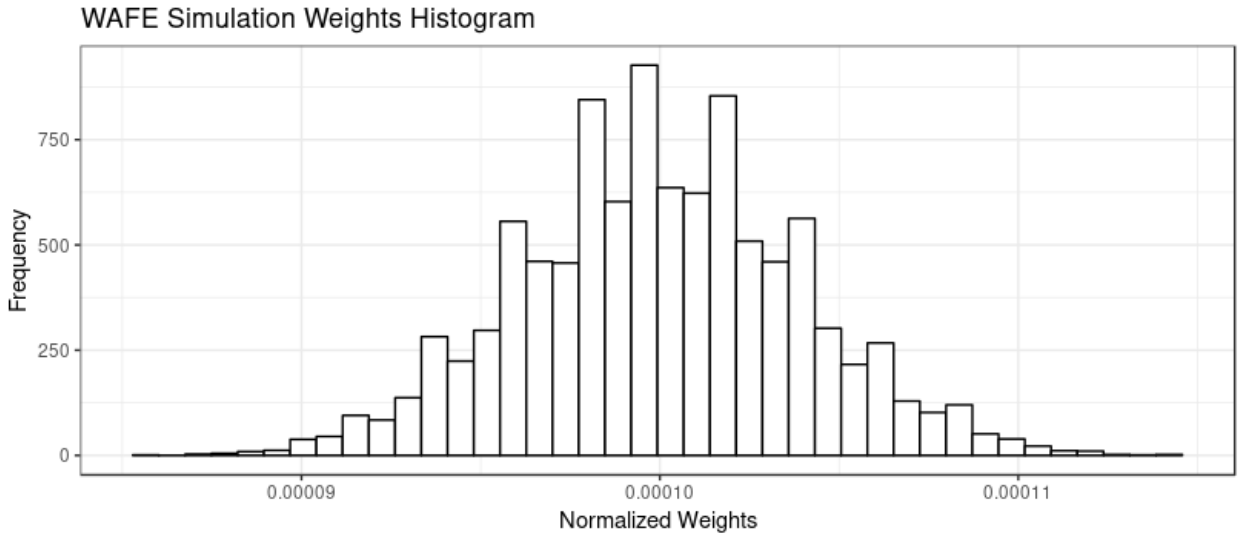

Figure 13: Scenario 1 Importance Sampling Weights

4.5.5 Additional Simulated Scenario #2

This hypothetical scenario is the same one from section 4.3.5, but with $N = 10{,}000$. Figure 14 is a plot of the mean, median, and WAFE optimal forecasts. Similar to the previous scenario, all forms of optimal forecasts are the same over the 14-day period.

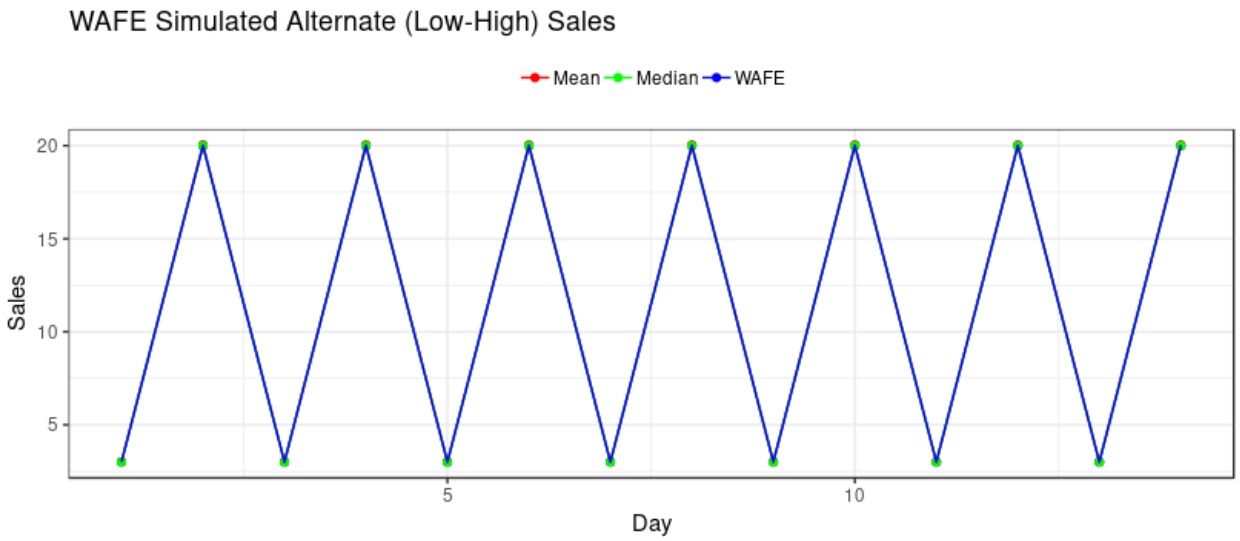

Figure 14: Scenario 2 Simulation

Figure 15 is the histogram of the WAFE importance sampling normalized weights from this scenario. The ESS is 98.92%.

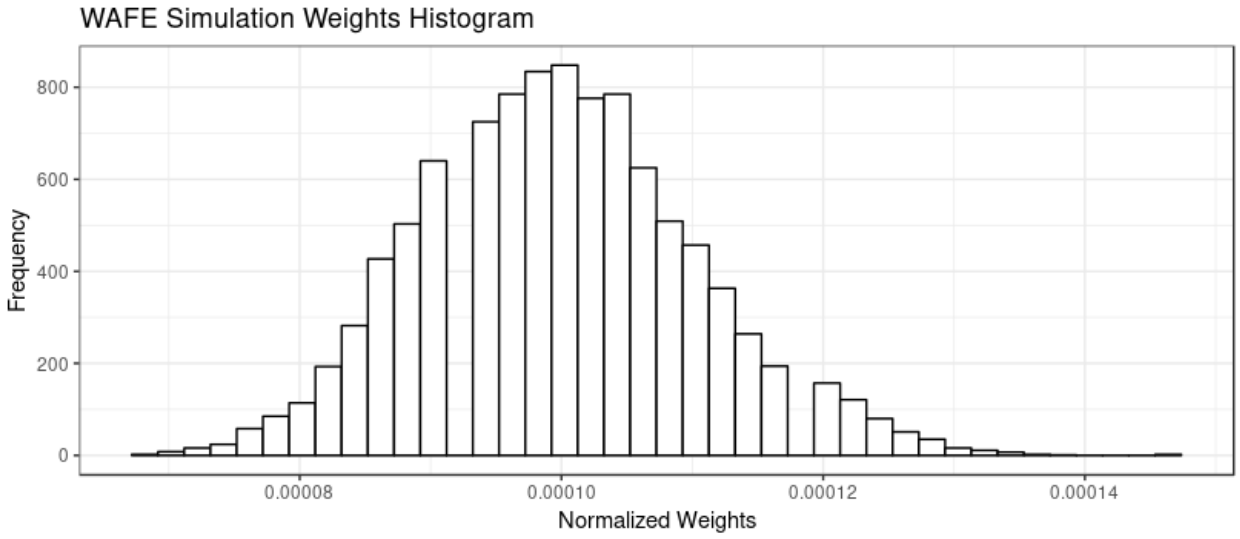

Figure 15: Scenario 2 Importance Sampling Weights

4.5.6 Additional Simulated Scenario #3

This hypothetical scenario is the same one from section 4.3.6, but with $N = 10{,}000$. Once again, all forms of optimal forecasts are the same over the 14-day period as shown in Figure 16.

Figure 17 is the histogram of the WAFE importance sampling normalized weights from this scenario. The ESS is 99.83%

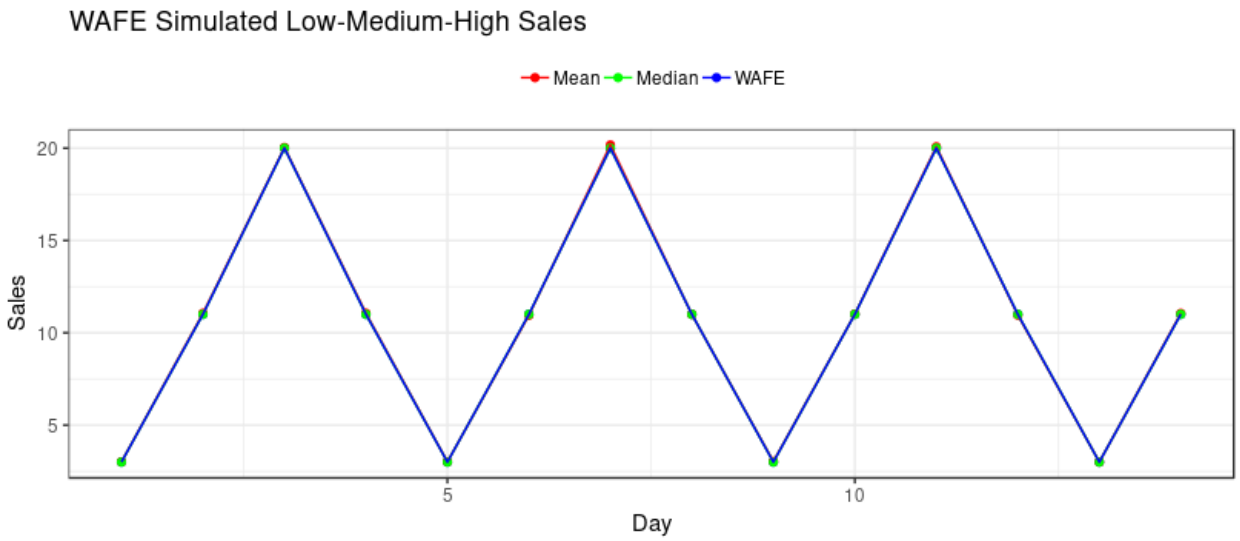

Figure 16: Scenario 3 Simulation

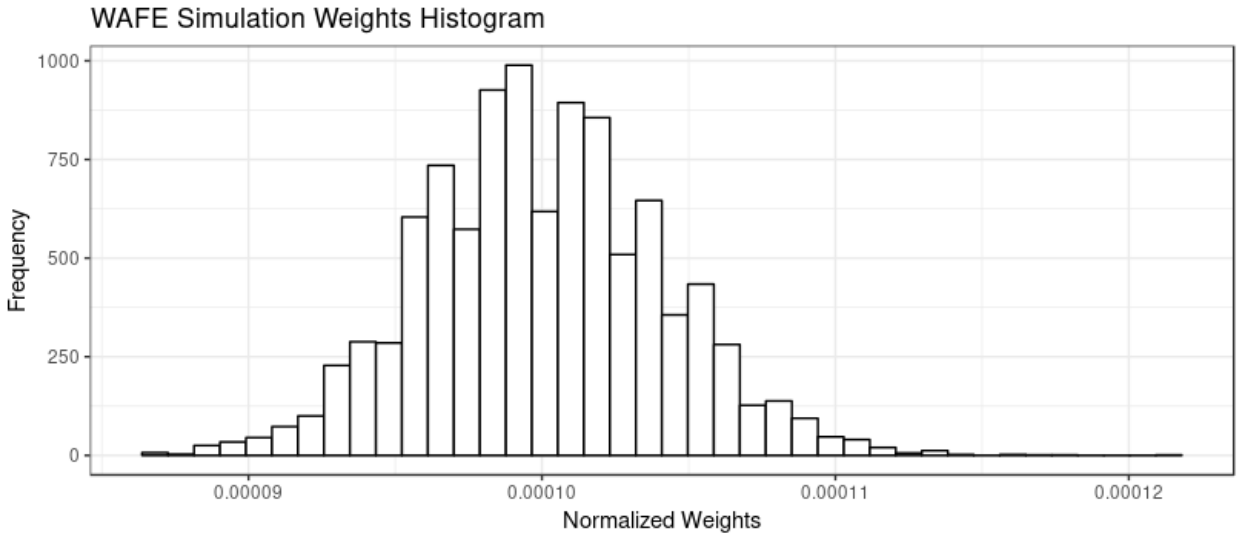

Figure 17: Scenario 3 Importance Sampling Weights

5 Real Data Application

5.1 Context

This case study involves using sales data from 84.51°and implementing WAPE, ZAPE, and WAFE to assess how well these forecasts perform. Two data sets are provided. The first data set is of observed daily sales of a spaghetti item over a 2-week period, from December 30, 2010 to January 12, 2011. The second data set is of forecast simulations over a forecast horizon of 14 days. There are 5,000 draws from the joint forecast distribution, where each row corresponds to a single joint forecast draw from the 14-dimensional forecast distribution. Over this time period, the forecasted

Date	Proportion 0's	Mean	Median
Dec 30	0.08	3.69	4
Dec 31	0.10	3.19	3
Jan 1	0.09	3.47	3
Jan 2	0.12	2.66	3
Jan 3	0.13	2.72	3
Jan 4	0.09	5.24	5
Jan 5	0.07	7.21	7
Jan 6	0.06	8.12	8
Jan 7	0.07	6.79	7
Jan 8	0.07	7.42	7
Jan 9	0.08	5.53	6
Jan 10	0.09	5.58	6
Jan 11	0.08	5.27	5
Jan 12	0.06	7.31	7

Table 3: Daily Sales Characteristics

probability of zero sales ranges from about 6-13%. Figure 18 is a graph showing true daily sales outcomes over the 14-day time period.

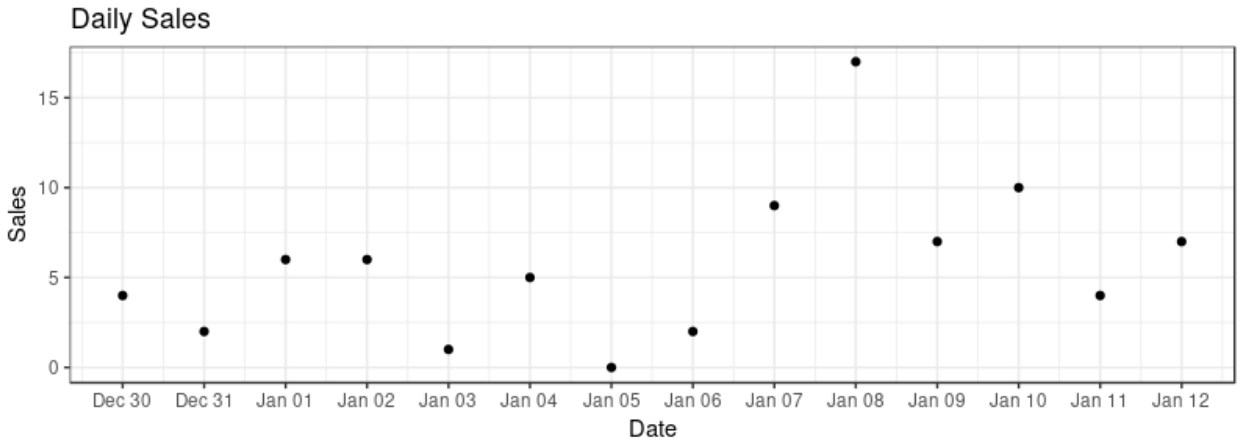

Figure 18: Daily Sales

Table 3 depicts the proportion of the 5,000 simulations that result in zero sales, the mean of the 5,000 simulations per day, and the median of the 5,000 simulations per day. A general trend that is observed is that as the proportion of 0 sales increases, the mean and median of the simulation get closer to 0.

5.2 Weighted Absolute Percent Error

WAPE optimization is first used to find the optimal forecasts from the simulated data. Figure 19 is a visualization of observed sales alongside three forecasting methods: WAPE, mean, and median

forecasting. From the graph, it is clear that all three forecasting methods follow almost identical trajectories over the 14 day period, but slightly diverge on January 9 and 10. Something to note is that WAPE and median forecasts are exactly the same for all days except for January 9 and 10.

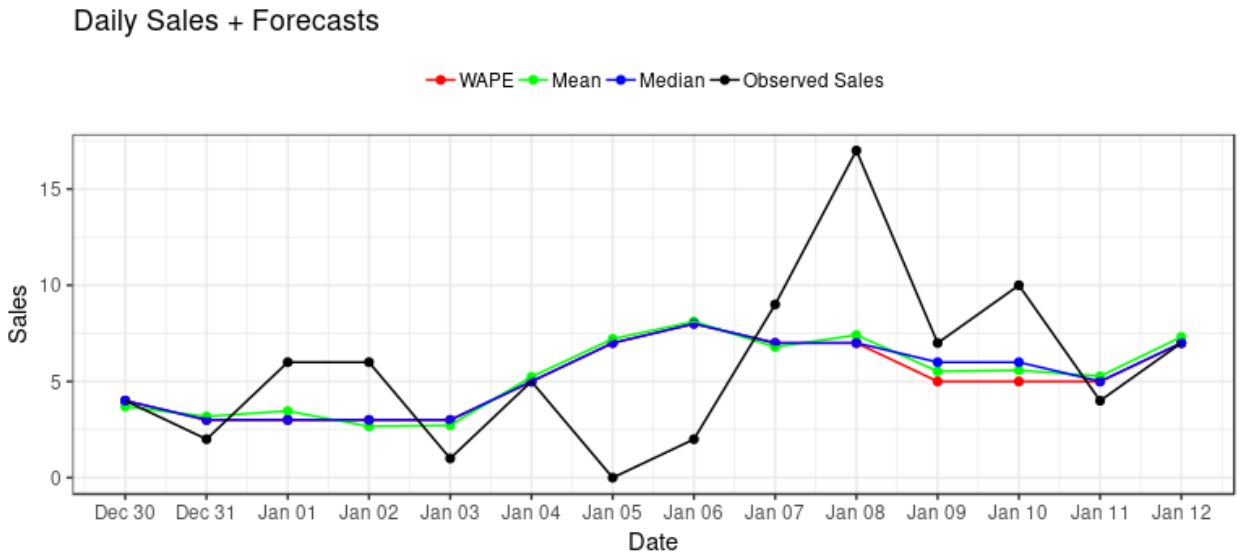

Figure 19: WAPE Forecasting

Table 4 depicts the observed sales, WAPE forecasts, mean forecasts, and median forecasts for the 14-day period. For the three forecasts vectors, the loss for each vector is calculated as shown in Table 5. It is observed that WAPE forecasts actually perform worse than the mean and median forecasts, but by marginal amounts. This discrepancy can probably be attributed to the differences that took place on January 9 and 10, else WAPE and median forecasts would have had the same loss. Figure 20 is a histogram of the WAPE importance sampling. The ESS is 97.11%.

5.3 Zero Adjusted Percent Error

ZAPE optimization was then used to find the optimal forecasts from the simulated data. Figure 21 is a visualization of observed sales alongside three forecasting methods: ZAPE, mean, and median forecasting. From the graph, it is clear that the median and mean forecasts are similar whereas the ZAPE forecasts are typically 1-2 sales fewer over the entire 14-day period. This is no surprise since the structure of ZAPE inherently causes forecasts to be lower.

Table 6 depicts the observed sales, WAPE forecasts, mean forecasts, and median forecasts for the 14-day period. After calculating the ZAPE loss for the forecasts over the 14-day period, it is observed that the ZAPE forecasts have the smallest loss by almost 2 units as shown in Table 7. This is attributed to the instance where January 5 has 0 sales. Since $c_i(f_i) = f_i$ in the loss function, the median and mean forecasts are penalized more than the WAPE forecast, thus resulting in a relatively large difference in loss. Because of the way ZAPE is optimized, there is a series of 14 sets of normalized weights where each set i corresponds to day i. Table 8 depicts the ESS for each day for the 14-day period.

Date	Observed	WAPE	Mean	Median
Dec 30	4	4	3.69	4
Dec 31	2	3	3.19	3
Jan 1	6	3	3.47	3
Jan 2	6	3	2.66	3
Jan 3	1	3	2.72	3
Jan 4	5	5	5.24	5
Jan 5	0	7	7.21	7
Jan 6	2	8	8.12	8
Jan 7	9	7	6.79	7
Jan 8	17	7	7.42	7
Jan 9	7	5	5.53	6
Jan 10	10	5	5.58	6
Jan 11	4	5	5.27	5
Jan 12	7	7	7.31	7

Table 4: WAPE Forecast Results

Forecast	Loss
WAPE	0.525
Mean	0.524
Median	0.500

Table 5: WAPE Loss

Date	Observed	ZAPE	Mean	Median
Dec 30	4	2	3.69	4
Dec 31	2	2	3.19	3
Jan 1	6	2	3.47	3
Jan 2	6	2	2.66	3
Jan 3	1	2	2.72	3
Jan 4	5	3	5.24	5
Jan 5	0	5	7.21	7
Jan 6	2	6	8.12	8
Jan 7	9	4	6.79	7
Jan 8	17	5	7.42	7
Jan 9	7	3	5.53	6
Jan 10	10	3	5.58	6
Jan 11	4	3	5.27	5
Jan 12	7	5	7.31	7

Table 6: ZAPE Forecast Results

Forecast	Loss
ZAPE	13.30191
Mean	15.51169
Median	15.10331

Table 7: ZAPE Loss

Date	ESS
Dec 30	72.70%
Dec 31	73.56%
Jan 1	72.45%
Jan 2	75.02%
Jan 3	73.89%
Jan 4	75.14%
Jan 5	80.28%
Jan 6	80.62%
Jan 7	78.71%
Jan 8	80.71%
Jan 9	75.55%
Jan 10	76.14%
Jan 11	74.07%
Jan 12	79.26%

Table 8: ZAPE Effective Sample Sizes

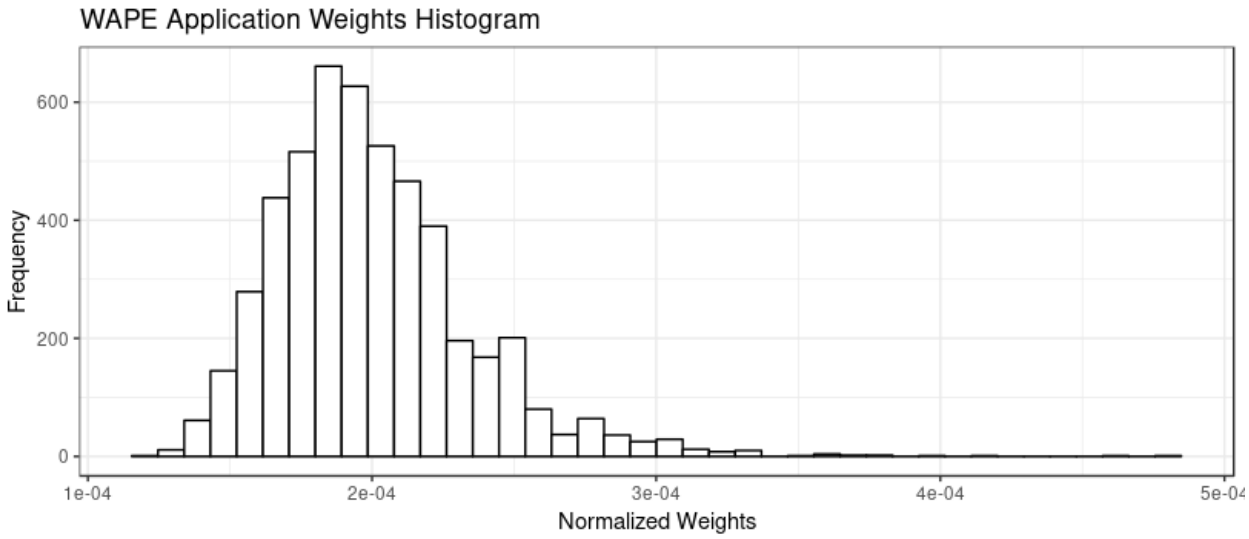

Figure 20: WAPE Importance Sampling Weights

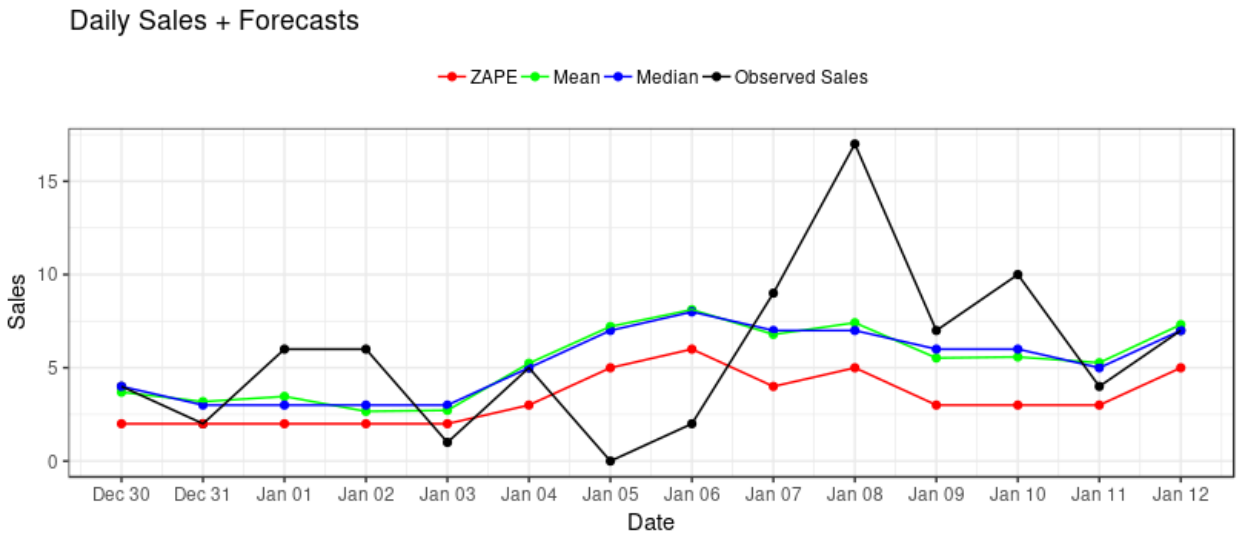

Figure 21: ZAPE Forecasting

5.4 Weighted Absolute Forecast Error

WAFE optimization was then used to find the optimal forecasts from the simulated data. Figure 22 is a visualization of the observed sales alongside three forecasting methods: WAFE, mean, and median forecasting. From the graph, it is clear that all three forecasting methods follow almost identical trajectories over the 14 day period, but diverge slightly on January 8 as a result of the large sale that day. Something to note is that WAFE and median forecasts are exactly the same for all days except for January 8.

Table 9 depicts the observed sales, WAFE forecasts, mean forecasts, and median forecasts for the 14-day period. After calculating the WAFE loss for the forecasts over the 14-day period, it is

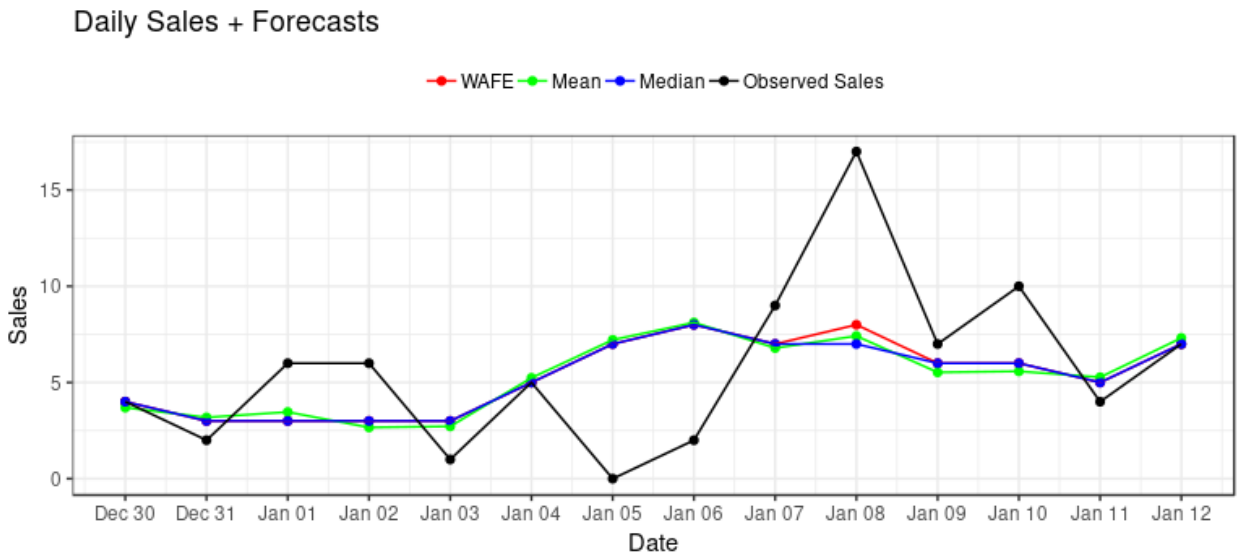

Figure 22: WAFE Forecasting

observed that the WAFE forecasts have the smallest loss as shown in Table 10. This is most likely attributed to the WAFE forecast on January 8, where this loss had somewhat accounted for the spike in sales. Figure 23 is a histogram of the distribution of the WAFE importance sampling after its final iteration of optimization. The ESS is 99.38%.

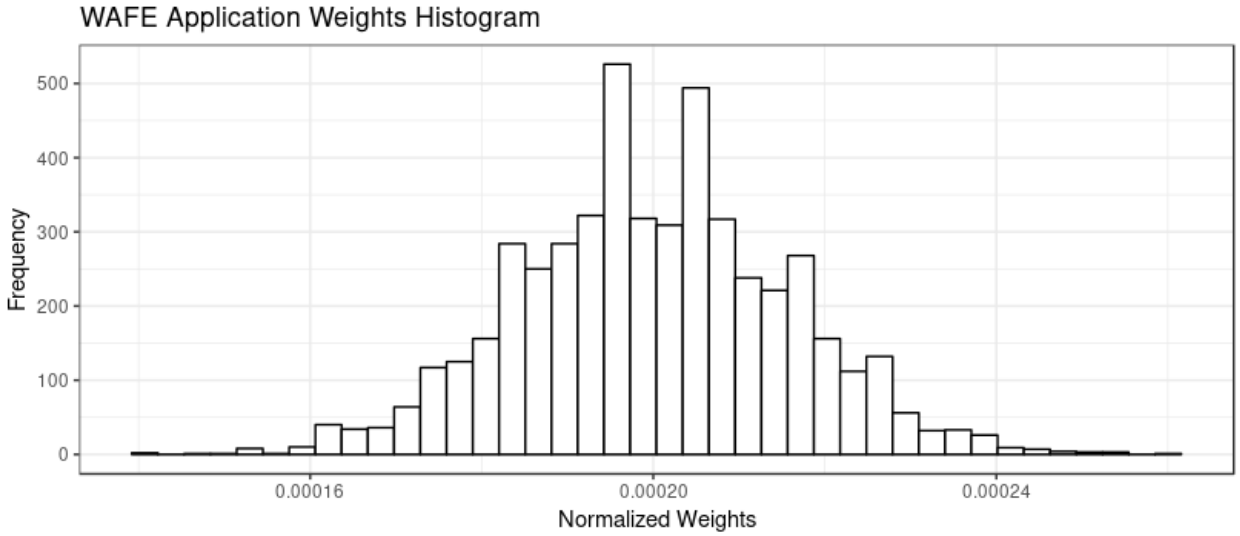

Figure 23: WAFE Importance Sampling Weights

Date	Observed	WAFE	Mean	Median
Dec 30	4	4	3.69	4
Dec 31	2	3	3.19	3
Jan 1	6	3	3.47	3
Jan 2	6	3	2.66	3
Jan 3	1	3	2.72	3
Jan 4	5	5	5.24	5
Jan 5	0	7	7.21	7
Jan 6	2	8	8.12	8
Jan 7	9	7	6.79	7
Jan 8	17	8	7.42	7
Jan 9	7	6	5.53	6
Jan 10	10	6	5.58	6
Jan 11	4	5	5.27	5
Jan 12	7	7	7.31	7

Table 9: WAFE Forecast Results

Forecast	Loss
WAFE	0.503
Mean	0.544
Median	0.519

Table 10: WAFE Loss

6 Summary/Discussion

In the context of the motivating case study and the application sales forecasting, the WAPE, ZAPE, and WAFE loss functions are introduced and investigated for time series of non-negative counts. This research stems from the recognition that common loss functions such as quadratic loss, absolute loss, 0-1 loss, and absolute percent error loss are not sufficient and may not be the most appropriate forms of loss to use for obtaining point forecasts in the consumer sales context. Thus multivariate loss functions are explored. The main motivations for developing WAPE and WAFE is to aggregate forecasts over some period of n days by generating a meaningful quasi-percent error loss function. The main motivation for developing ZAPE is so that the aggregation present in WAPE and WAFE can be applied to cases of $y_i = 0$ or $\mathbf{y} = \mathbf{0}$.

After applying and optimizing these loss functions to a series of simulated data, different properties for each loss function are observed. In general, WAPE and WAFE tend to optimize at relatively higher values than ZAPE does. Something interesting to note is that the optimal values of WAPE and WAFE changes as the number of days, n, varies and the optimal values of ZAPE changes as μ of the Poisson distribution changes. Simulated examples provide a lot of insight into how actual applications of these loss functions perform with actual sales data.

A case study analysis also provides insight into how these loss functions perform in forecasting sales outcomes. In general, it appears that ZAPE optimizes at values that are lower than WAPE and WAFE, which allows ZAPE to better account for instances of 0 sales for a given day. It is still possible for WAPE and WAFE to forecast a sale of 0 for any particular day, but in the explored forecasting example, it did not take place.

Further studies and research on this topic can be tailored to exploring other forms of loss functions for different types of products with different kinds of sales variations. In particular, more variations of APE can be developed and explored in similar simulated/application settings. Another area of exploration could be delving more into ZAPE, specifically focusing on different definitions of $c_i(\cdot)$. Another topic of exploration could be assessing the dependency of the individual loss functions $L(\mathbf{y}, \mathbf{f})$ on the forecast distribution $p(\mathbf{y})$. In the simulated scenarios, only Poisson distributions were used. However, exploring discrete bimodal/multimodal forecast distributions may also be of interest. Another area to study is computing the Monte Carlo standard error for the median of a discrete probability distribution. It is still an open question as to how the standard error for a median is defined, and additional technical aspects of Monte Carlo sampling could be further investigated. Lastly, applying multivariate loss functions to other applications involving time series of non-negative is worth exploring. Forecasting money supply to make decisions on monetary policy or forecasting the number of ER patients to make decisions on resource allocation are examples of other applications.

7 Appendix

7.1 APE Optimization Derivation

APE Loss:

$$L(y, f) = \frac{|y - f|}{y}$$

Note:

$$\frac{|y - f|}{y} = \begin{cases} 1 - \frac{f}{y} & y > f \\ 0 & y = f \\ \frac{f}{y} - 1 & y < f \end{cases}$$

Risk:

$$\begin{aligned} R(f) &= \int_{-\infty}^{\infty} \frac{|y - f|}{y} p(y) dy \\ &= \int_{-\infty}^{f} (\frac{f}{y} - 1) p(y) dy + \int_{f}^{\infty} (1 - \frac{f}{y}) p(y) dy \end{aligned}$$

Risk Minimization:

$$\frac{\partial R(f)}{\partial f} = \int_{-\infty}^{f} \frac{p(y)}{y} dy - \int_{f}^{\infty} \frac{p(y)}{y} dy = 0$$

Define the p.d.f.: $g(y) = kp(y)/y$, where $k > 0$ is the implied normalizing constant. The implied c.d.f. is $G(y)$

$$\frac{\partial R(f)}{\partial f} = G(f) - (1 - G(f)) = 0$$

$$\implies G(f) = \frac{1}{2}$$

Thus the optimal f_i is,

$$\implies f = G^{-1}(\frac{1}{2})$$

7.2 ZAPE Optimization Derivation

ZAPE Loss:

$$L_i(y_i, f_i) = c_i(f_i)I(y_i = 0) + \frac{|y_i - f_i|}{y_i} I(y_i > 0),$$

where

$$c_i(f_i) = f_i.$$

Note:

$$\frac{|y_i - f_i|}{y_i} = \begin{cases} 1 - \frac{f_i}{y_i} & y_i > f_i \\ 0 & y_i = f_i \\ \frac{f_i}{y_i} - 1 & y_i < f_i \end{cases}$$

Risk:

$$\begin{aligned} R(f_i) &= c_i(f_i)p(y_i = 0) + \int_1^\infty \frac{|y_i - f_i|}{y_i} p(y_i)dy_i \\ &= f_i p(y_i = 0) + \int_1^{f_i} (\frac{f_i}{y_i} - 1)p(y_i)dy_i + \int_{f_i}^\infty (1 - \frac{f_i}{y_i})p(y_i)dy_i \end{aligned}$$

Risk Minimization:

$$\frac{\partial R(f_i)}{\partial f_i} = p(y_i = 0) + \int_1^{f_i} \frac{p(y_i)}{y_i} dy_i - \int_{f_i}^\infty \frac{p(y_i)}{y_i} dy_i = 0$$

Define the p.d.f.: $g_i(y_i) = k_i p(y_i)/y_i$ on $y \geq 1$, where $k_i > 0$ is the implied normalizing constant. The implied c.d.f. is $G_i(y_i)$ By multiplying the above differential by k_i, the following is attained:

$$\frac{\partial R(f_i)}{\partial f_i} = k_i p(y_i = 0) + G_i(f_i) - (1 - G_i(f_i)) = 0$$

$$\implies G_i(f_i) = \frac{(1 - k_i p(y_i = 0))}{2} = q_i$$

Thus the optimal f_i is,

$$f_i = \begin{cases} 0 & q_i \leq 0 \\ G_i^{-1}(q_i) & q_i > 0 \end{cases}$$

7.3 Importance Sampling

Importance sampling is a Monte Carlo integration technique used to estimate a particular distribution where simulated draws are taken from a distribution that is not the one of interest. Importance sampling is particularly useful when the distribution of interest can easily be simulated, but is difficult to sample from directly. The idea behind importance sampling is that there are certain values of the random variable in a simulation are "more important" in estimating a parameter. This importance is characterized by the frequency at which a particular sample is drawn, which can result in a biased estimator. In order to correct for this bias, simulation outputs are weighted to ensure that the new importance sampling estimator is unbiased.

Take for example a Bayesian inference scenario in which the distribution $g(\mu)$ is a posterior distribution based on a model analysis and fit to observed data. Given that this p.d.f. is able to be evaluated at any point, it can be assumed there is some importance sampling distribution with p.d.f. $p(\mu)$. It is often the case where $p(\mu)$ is derived as an analytic approximation to $g(\mu)$.

In this scenario, the expectation of $g(\mu)$ can be written as the following:

$$H = \int h(\mu)g(\mu)d\mu = \int h(\mu)w(\mu)p(\mu)d\mu,$$

where

$$w(\mu) = g(\mu)/p(\mu).$$

Using direct Monte Carlo integration,

$$\overline{h} = m^{-1}\sum_{i=1}^{m} w(\mu_i)h(\mu_i),$$

where samples of μ_i are drawn at random from $p(\mu_i)$ for $i = 1, ..., m$. Each are these random samples are i.i.d.

At each simulated μ_i, $w(\mu_i) = g(\mu_i)/p(\mu_i)$ is known as the importance weight. Each $w(\mu_i)$ weights the sample estimates of $h(\mu_i)$ to adjust for the bias caused by "important samples". Note that by taking the sum of all the $w(\mu)$'s for some unique sample, that particular sample's probability is then computed and this results in a distribution that is proportional to the desired probability distribution. By multiplying the newly attained distribution by some normalizing constant c, the desired distribution, $g(\mu)$ is then observed.

There are many important properties that arise from using importance sampling. Firstly, the Monte Carlo estimate $\overline{h}$ has the expectation H, meaning that it is an unbiased estimate of the true H. And given the law of large number and the central limit theorems, $\overline{h}$ will converge to H. Additionally, it can be estimated that $Var(\overline{h}) = \int w(\mu)g(\mu)d\mu = \int w(\mu)^2p(\mu)d\mu = \int g(\mu)^2p(\mu)d\mu$.

Another important aspect of importance sampling is effective sample size (ESS), which is defined as:

$$ESS = W_m^{-1}(\sum_{i=1}^{m} w(\mu_i))^2 \quad \text{with} \quad W_m = \sum_{i=1}^{m} w(\mu_i)^2$$

.

When the weights are normalized to $w_i = w(\mu_i)/c$ for some constant c,

$$ESS = \frac{1}{\sum_{i=1}^{m} w_i^2}$$

With importance sampling and other Monte Carlo integration techniques, samples are drawn from some specified probability distribution in order to get to some desired probability distribution. Since the desired distribution is specified from some original probability distribution, it is difficult to estimate the quality of the desired distribution after sampling. This is where ESS plays an important role. ESS provides a quantitative measure of how many samples are needed in order to attain the desired probability distribution. For instance, if a sample size of $n = 5000$ were used and ESS = 2500, this means that the quality of the desired probability distribution estimate is about the same as if 2500 direct samples were used.

References

Berry, L. R., P. Helman, and M. West (2018). Probabilistic forecasting of heterogeneous consumer transaction-sales time series. *International Journal of Forecasting*. (Invited Revision), arXiv:1808.04698.

Berry, L. R. and M. West (2018). Bayesian forecasting of many count-valued time series. *Journal of Business and Economic Statistics*. (Invited Revision), arXiv:1805.05232.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (2008). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons. 4th ed.

Hyndman, R. J. and A. B. Koehler (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting 22*, 679–688.

Kolassat, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting 32*, 788–803.

Smith, J. Q. (2010). *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press. 1st ed.

Yelland, P. M. (2009). Bayesian forecasting for low-count time series using state-space models: An empirical evaluation for inventory management. *International Journal of Production Economics 8*, 95–103.