1. **Supplementary Empirical Application**
   1. *Data*

We use IRI Nielsen data containing 102 total weekly sales values for 32 grocery stores.

* 1. *Time Series Features for Forecast Accuracy*

STL decomposition requires more than two seasonal periods of data (our seasonal period is 52 weeks), which we do not have. We do not include features such as *Trend, Seasonality, Error ACF,* and *Spike* that require STL decomposition.

Table 1 displays the time series features selected for k-nTS and k-nTS+ swapping. The same features were also used for protecting the M3 monthly micro data in our original analysis. We refer the reader to Hyndman et al. (2022) for a detailed explanation of these features and further mathematical detail on the time series features is provided in the Appendix.

**Table 1: Time series feature descriptions and value ranges.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Description** | **Value Range** | **Selected (Literature)** | **Selected (k-nTS+)** |
| *Spectral Entropy* | Signal-to-noise ratio of the time series. | [0, 1] | X |  |
| *Hurst* | Long-range dependence (self-similarity) of a time series. | [0, 1] | X |  |
| *Skewness* | Symmetry of the distribution of time series values. |  | X |  |
| *Kurtosis* | Weight of the tails of the distribution of time series values. |  | X |  |
| *Mean* | Mean of the time series. |  | X | X |
| *Variance* | Variance of the time series. |  | X | X |
| *Max Variance Shift* | Largest variance shift between two consecutive sliding windows. |  |  | X |
| *Max Level Shift* | Largest mean shift between two consecutive sliding windows. |  |  | X |

* + 1. *Privacy Methods*

We use the differential privacy and additive noise implementations described in Section 4.3.1 and 4.3.2. For k-nTS+ we select so the features are within 5% of the minimum average prediction error from the best random forest model. For the k-nTS+ protected data, we use the four features (last column of Table 1) with the highest average rank across the RFE iterations for all forecasting models with .

* + 1. *Identification disclosure risk*

Our methodology for assessing identification disclosure risk is identical to what was described in Section 4.3.4.

* 1. *Results*

For all privacy methods, we generate one-step ahead forecasts for time T+1 using off-the-shelf models in R and Python shown in Table 2. We do not include TES here since initializing the seasonal components of the model requires more than two seasonal periods in the data. All reported forecast accuracy and standard deviation results are derived from comparing the forecasts for T + 1 to the actual data from T + 1. Reported privacy results are derived from calculating the identification disclosure risk using the protected data from time period 1 to T. We perform minimal data pre-processing and use the standard settings in the off-the-shelf packages. [[1]](#footnote-1)

**Table 2: Univariate and Multivariate Forecast Models**

|  |  |
| --- | --- |
| **Model Name** | **Variant** |
| SES | - |
| DES | Additive trend |
| Auto-ARIMA | Seasonal |
| VAR | - |
| LGBM | - |
| RNN | LSTM |

Table 3displays the average MAE of one-step ahead point forecasts across all models and series, the identification disclosure metric , and the average performance gap across all series. The percentages in parentheses are the increase in average MAE relative to the average MAE from the original data. The results show an inverse relationship between forecast accuracy and the strength of privacy protection. While strong differential privacy provides the lowest identification disclosure risk at 1.85% (random guessing is 0.6%), it nearly quintuples (+383%) the average forecast error relative to the original data resulting in unusable forecasts. Under weak differential privacy with , over 49% of series are identified correctly on average, which is poor identification disclosure risk. Protection against identification disclosure is better under additive noise with where 22.5% of series are correctly identified on average. However, this comes at a cost to forecast accuracy, which degrades by nearly 45%.[[2]](#footnote-2)

**Table 3: Identification disclosure risk, forecast accuracy, and representativeness for original and protected data sets.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Privacy Method** | **Parameter Value** | Privacy (Identification Disclosure Risk) | Accuracy (MAE) | Representativeness (Performance Gap) |
| **Original Data** | **-** | 100.0% | 685.71 (0.0%) | 42.7 |
| ***k*-nTS+** | **15** | 2.7% | 839.8  (+22.5%) | 73.0 |
| **7** | 3.5% | 822.3  (+19.9%) | 66.5 |
| **3** | 3.3% | 781.0  (+13.9%) | 62.1 |
| ***k*-nTS** | **15** | 1.6% | 1066.2  (+55.5%) | 106.3 |
| **7** | 2.1% | 987.0  (+43.9%) | 100.4 |
| **3** | 2.1% | 956.9  (+39.6%) | 92.9 |
| **Differential Privacy** | **1.0** | 1.9% | 3310.3  (+382.8%) | 1,826,437.0 |
| **4.6** | 13.6% | 1401.0  (+104.3%) | 311,037.7 |
| **10** | 49.0% | 899.4  (+31.2%) | 78,456.4 |
| **Additive Noise** | **2.0** | 5.8% | 1821.4  (+165.6%) | 503,658.2 |
| **1.5** | 10.4% | 1343.3  (+95.9%) | 326,834.5 |
| **1.0** | 22.5% | 994.0  (+45.0%) | 166,171.2 |

*k*-nTS swapping with *k* = 3 offers a good identification disclosure risk of 2.1%, but forecast accuracy degrades by 39.6%. Our proposed method of *k-*nTS+ swapping with provides similar levels of protection against reidentification (3.3%) with a reduction in forecast accuracy of only 13.9%. Part of this improvement in forecast accuracy at a minimal tradeoff to identification disclosure risk is due to the incorporation of the accuracy feedback loop for selecting time series features. Thus, we recommend data owners to use our k-nTS+ swapping method (k=3) with the selected time series features to balance the tradeoff between privacy and forecast accuracy.

Forecasters also prefer protected data that are representative of the original time series. Representativeness improves trust between data owners and forecasters and makes it more likely for forecasters to use protected data. Table 3 displays the *performance gap* of Petropoulos & Siemsen (2022) to measure the distance between the protected and original time series values, *performance gap* , which is calculated after applying a Box-Cox transformation and scaling the original and protected series. Note that our results in Table 3 differ from Petropoulos & Siemsen (2022) where the performance gap is calculated using the fitted values of forecasting models relative to the training data (which we include in the first row of Table 3). The results show that *k*-nTS and k-nTS+ swapping produce protected time series with the smallest performance gaps by a large margin. However, we note that the average performance gap across series (62.1 for k-nTS+ with k=3) is significantly larger than the average performance gap (42.7) of the fitted values across all series and forecasting models.

Table 4 displaysthe ranks of the MAE and forecast error variance across all forecasting models using the original data and k-nTS+ swapping with ­. Past research found that complex forecasting models forecast more accurately than simple models using the monthly micro data (Koning et al., 2005). The results show that k-nTS+ swapping preserves the ranking of the best and worst models on MAE. Univariate models (SES and DES) moved up in the ranking and more complex models (Auto-ARIMA and RNN) moved down.

**Table 4: Ranks of MAE and standard deviation of forecast error for the original data and the k-nTS+ swapping (*k*=3) data.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MAE Ranks | | Standard Deviation of Forecast Error Ranks | |
| Model | Original | Protected | Original | Protected |
| TES | 1 (637.90) | 1 (731.30) | 2 (859.30) | 4 (920.57) |
| Auto-ARIMA | 2 (646.07) | 4 (764.83) | 1 (834.78) | 1 (897.67) |
| RNN | 3 (665.38) | 5 (783.15) | 5 (883.86) | 5 (966.35) |
| DES | 4 (680.54) | 2 (743.68) | 3 (866.35) | 2 (901.22) |
| SES | 5 (686.71) | 3 (752.08) | 4 (867.13) | 3 (914.20) |
| LGBM | 6 (709.48) | 6 (809.00) | 7 (919.67) | 6 (982.35) |
| VAR | 7 (773.90) | 7 (883.07) | 6 (892.62) | 7 (998.08) |

* 1. *Analysis of Time Series Features*
     1. *Importance of Time Series Features*

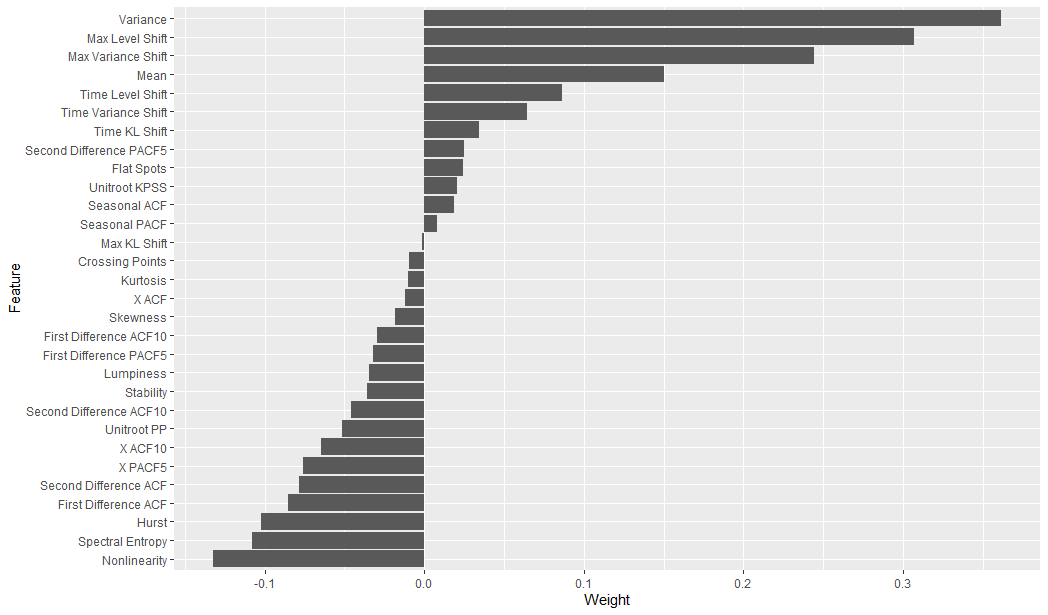
Let denote one of the nearest neighbor feature vectors to , where is the number of nearest neighbors considered by RReliefF, and let and denote the forecast errors for the corresponding time series. Let and denote the events that series and have different forecast errors and different values for feature , respectively, conditional on being nearest neighbors. The RReliefF weight for feature approximates the difference in conditional probabilities,

.

The RReliefF weights approximate the difference between the probability that feature discriminates between series with different forecast errors, and the probability that feature discriminates between series with the same forecast error. Features with have a higher probability of varying across series with different forecast errors than varying across series with similar forecast errors. If we swap using features with , we will maintain the values of these features throughout the swapping process and maintain forecast accuracy.

Figure 3 shows the RReliefF weights for each of the 39 features averaged across all forecasting models. RReliefF was used to predict the absolute forecast errors for each model and series across the original and protected data sets. Surprisingly, *Hurst* and *Spectral Entropy* had negative weights which implied they were not useful to maintain forecast accuracy for swapping in the protected data. On the other hand, *Spike*, *Variance*, *Linearity*, *Max Level Shift*, *Max Variance Shift,* and *Curvature* had large positive weights and were important to maintain forecast accuracy.

**Figure 3: RReliefF weights averaged across the results of each forecasting model.**

****

* + 1. *Selection of Time Series Features*

Figure 4 presents the number of features included for each forecasting model after k-nTS+ eliminated features with negative weights that were poor predictors of forecast error. Over iterations, most of the reduction in OOB MSE occurred using five or fewer features for all forecasting models.

**Fig. 4: Average OOB MSE across feature subset sizes when predicting the MAE of each forecasting model.**

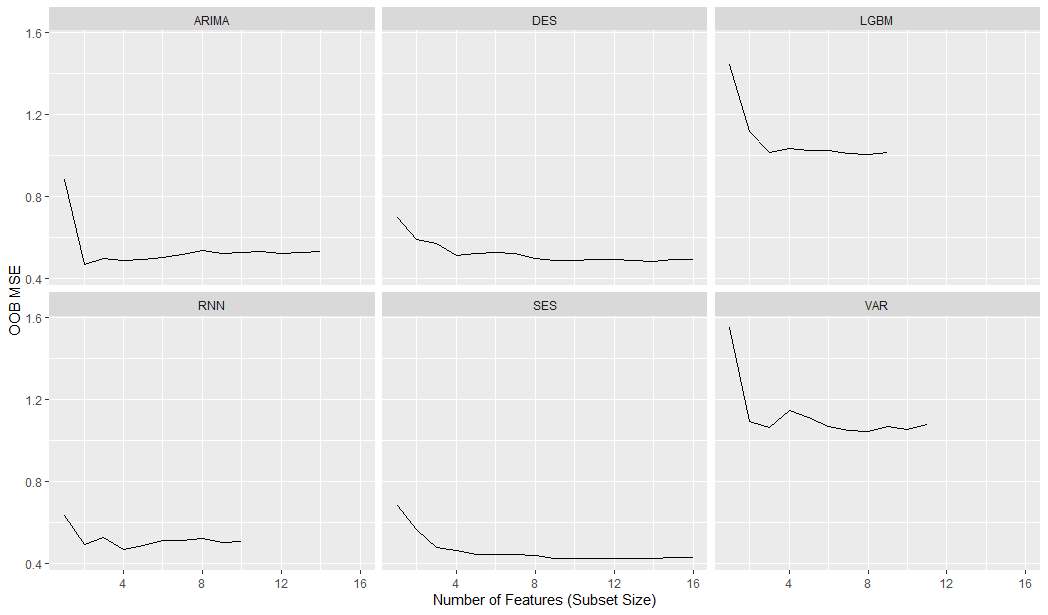
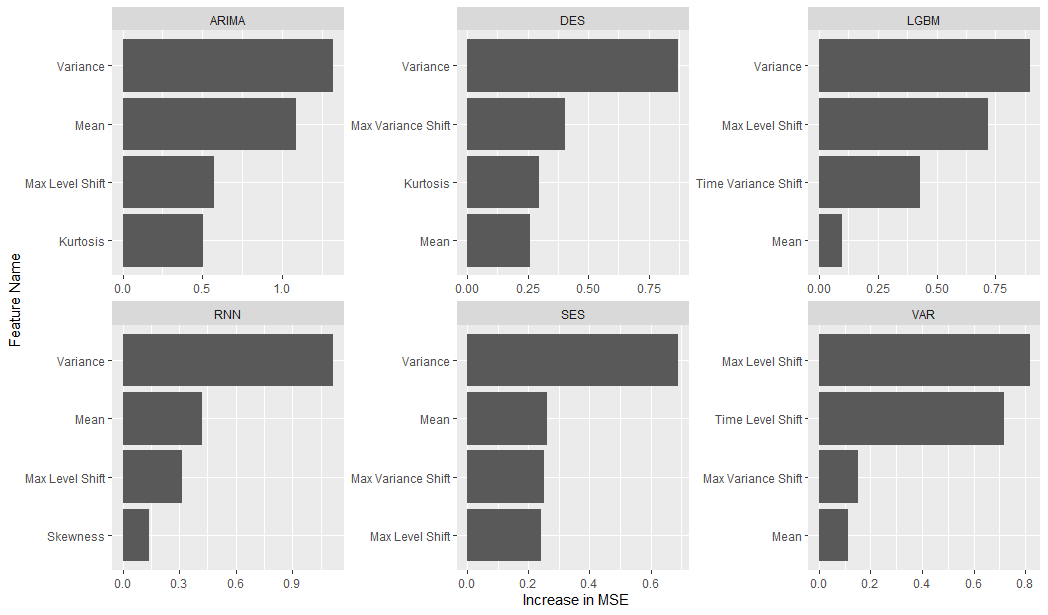


Figure 5 summarizes the results of RFE in Figure 4 and displays the permutation-based importance values for each forecasting model's six most highly ranked features. Some features, such as *spike,* *max variance shift, max level shift, mean,* and *variance* are highly ranked across most or all forecasting models. Other features appear to be highly important only for specific forecasting models. Examples include *trend*, which is required for DES and TES, *seasonal strength*, which is required for TES, and *X ACF* (the first autocorrelation coefficient of the time series)*,* which is important for Auto-ARIMA and RNN.

**Figure 5: Permutation-based importance for the top six features for each forecasting model.**

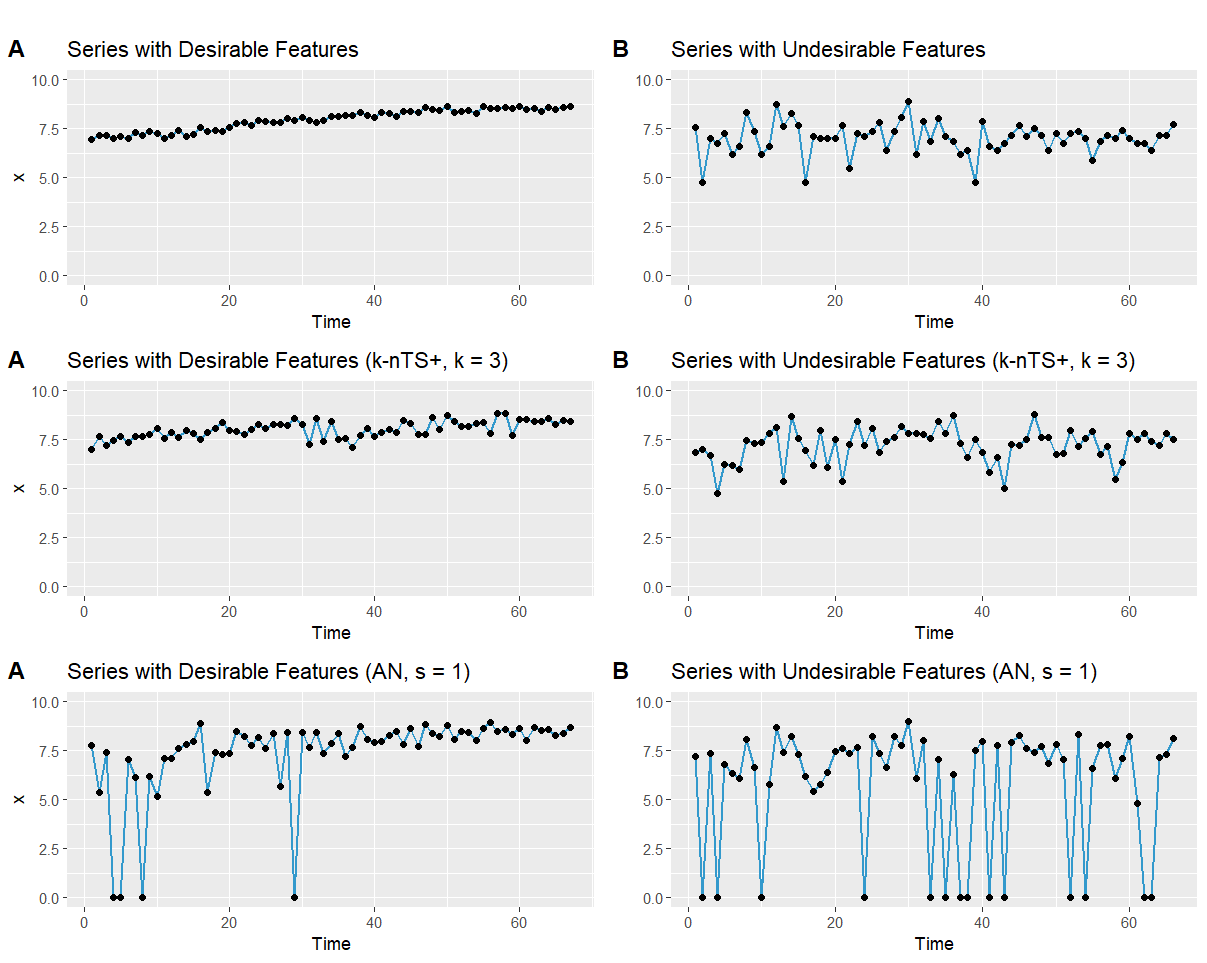


* + 1. *Illustration of Changes in Time Series Features After Protection*

Figure 6 displays two monthly time series from the M3 monthly micro data with desirable and undesirable features. After applying the privacy methods to the original time series, Figure 6 illustrates the results using k-nTS+ with and additive noise with . We can see that for k-nTS+ with k=3, there is little visual change for the undesirable series. For additive noise, there are drastic changes to both series.

Table 5 displays the values of the time series features before and after protection. Table 5 shows that the low spectral entropy and high Hurst coefficient values of the desirable time series indicate good forecastability. Table 5 shows that the undesirable series is essentially a random walk as indicated by the 0.50 value of the Hurst coefficient. Furthermore, the undesirable series has a spectral entropy of 1 indicating a low signal-to-noise ratio. When comparing the two series, the variance of the desirable series is due to a forecastable trend, whereas the variance of the undesirable series is due to randomness. The desirable series also has low *Kurtosis* with a light tailed distribution compared to the undesirable series. One interesting finding is that the k-nTS+ (k=3) version of the desirable series has a lower *Variance* than the original series. However, the higher (long run) variance of the original series is due to the strong trend. Figure 7 shows the short run month-to-month variance of the k-nTS+ protected series is higher than the original series, as indicated by the values of *Max Variance Shift* in Table 5.

**Figure 6: Comparison of original, AN (s = 1), and *k*-nTS+ (*k* = 3) protected series with desirable and undesirable features.**

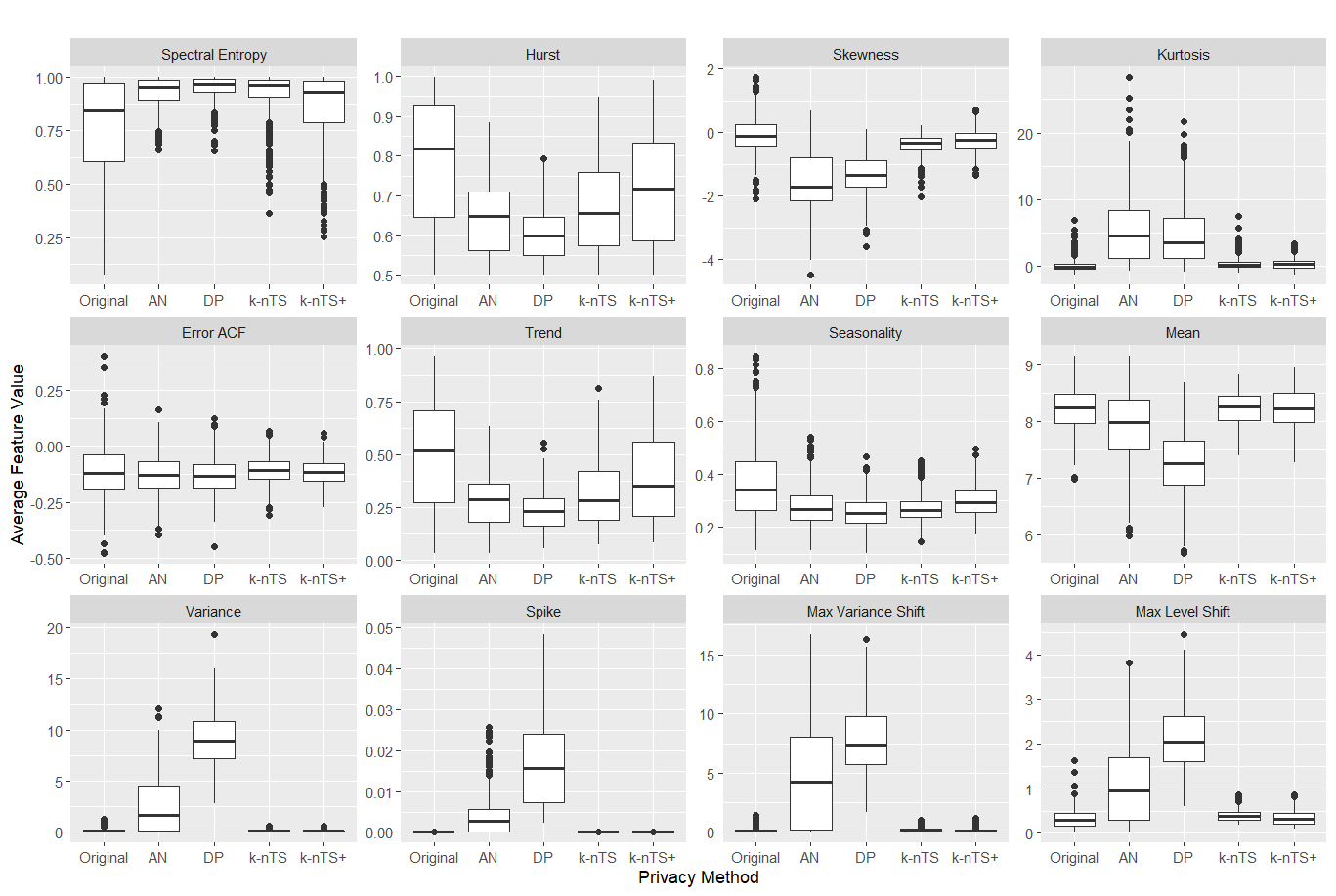


**Table 5: Time series feature values from undesirable and desirable time series.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature** | **Desirable Time Series (left Fig. 7)** | | | **Undesirable Time Series (right Fig. 7)** | | |
|  | **Original** | **k-nTS+ (k=3)** | **AN (s=1)** | **Original** | **k-nTS+ (k=3)** | **AN (s=1)** |
| *Spectral Entropy* | 0.07 | 0.89 | 0.92 | 1.00 | 0.98 | 1.00 |
| *Hurst* | 0.99 | 0.81 | 0.76 | 0.50 | 0.66 | 0.50 |
| *Skewness* | -0.41 | -0.18 | -2.74 | -0.57 | -0.71 | -1.17 |
| *Kurtosis* | -1.24 | -0.74 | 6.99 | 1.16 | 0.25 | -0.37 |
| *Error ACF* | -0.09 | -0.22 | -0.20 | -0.19 | -0.06 | -0.21 |
| *Trend* | 0.97 | 0.58 | 0.49 | 0.12 | 0.22 | 0.11 |
| *Seasonality* | 0.16 | 0.25 | 0.39 | 0.23 | 0.24 | 0.13 |
| *Mean* | 7.96 | 8.02 | 7.41 | 7.01 | 7.21 | 5.73 |
| *Variance* | 0.29 | 0.19 | 4.27 | 0.65 | 0.76 | 9.57 |
| *Spike* | 0.0000 | 0.0000 | 0.0037 | 0.0001 | 0.0001 | 0.0268 |
| *Max Variance Shift* | 0.05 | 0.24 | 9.37 | 1.10 | 1.12 | 11.44 |
| *Max Level Shift* | 0.57 | 0.51 | 2.77 | 0.70 | 0.84 | 3.29 |

Figure 7 displays boxplots of the time series feature values before and after protection across all time series in our application. Random noise privacy methods (AN and DP) increase the randomness and significantly change distributional characteristics of all features except *Error ACF*, leading to poor forecast accuracy. Random noise also produces a negative bias in the means of the protected series and significantly increases the variance. On the other hand, the *k*-nTS swapping method increases the spectral entropy but better preserves most feature distributions. The feature distributions of *k*-nTS+ swapping are much closer to the original distributions for those features important for forecast accuracy (*Spike, Max Variance Shift, Max Level Shift, Mean, Variance,* and *Trend*), which led to improved forecast accuracy results.

**Fig 7: Distributions of time series features for each privacy method.**



We note that while *k*-nTS performed swapping based on the values of *Spectral Entropy, Hurst*, and *Seasonality*, it does not preserve these feature distributions as well as *k*-nTS+. One reason could be that all three of these features are based on autocorrelation and we found that k-nTS swapping degrades *Seasonal ACF*, *X ACF*, *X ACF10*, and *X PACF5* more than k-nTS+ (mathematical details and a figure can be found in the Appendix). *k*-nTS+ did not explicitly swap based on the values of *Spectral Entropy, Hurst*, and *Seasonality* which demonstrates the importance of the k-nTS+ feedback loop. Although *Spectral Entropy* and *Hurst* were correlated with forecast accuracy across series, they were eliminated in the first stage of the *k*-nTS+ feature selection process using RReliefF.

* + 1. *Privacy Adjusted Forecasts*

Similar to Fildes et. al. (2009) and Khosrowabadi et al. (2022), we compare the percentage of forecast adjustments that improved accuracy across adjustment direction, magnitude, and the coefficient of variation of the original time series. We use the adjusted forecasts using the k-nTS+ (k = 3) protected data set which was the top performing privacy method in our application.

To compute adjustment magnitude, we normalize the absolute difference between the adjusted and original forecasts using the mean of the original series,

where the and superscripts denote a forecast based using the original and protected data, respectively. Using the approach of Khosrowabadi et al. (2022), we bin the magnitudes into high ( quantile), low ( quantile) and medium ( quantile and quantile) intervals.

We also compute the average relative absolute error (AvgRelAE, see Fildes et al., (2013)) to compare the relative accuracy of the adjusted and original forecasts. The AvgRelAE of the adjusted forecasts is computed as

,

where and are the absolute forecast error for the protected and original versions of series , respectively. An less than one indicates an average improvement in accuracy and an greater than one indicates an average reduction in accuracy.[[3]](#footnote-3) We remove the forecasts with the 5% smallest and 5% largest ratios ( to prevent extreme outliers from affecting AvgRelAE (Fildes et al., 2013).

Using the k-nTS+ (k=3) protected data, we find that less than half (43%) of the adjusted forecasts improved forecast accuracy (lower absolute error), which is less than the reported 49.9% of judgmentally adjusted forecasts that improved accuracy in Khosrowabadi et al. (2022). Table 6 breaks down the results by adjustment magnitude and direction and displays the AvgRelAE and percentage of adjusted forecasts that improved accuracy. The results show that most privacy adjusted forecasts degraded accuracy and the AvgRelAE is greater than one in five out of six cases. Also, our results are contrary to the findings in the judgmental literature which shows that large adjustments and negative adjustments improve forecast accuracy. We find that small privacy adjustments improved (47.9% of cases) forecast accuracy more frequently than large privacy adjustments (35.6% of cases). Furthermore, positive adjustments improved (44.6% of cases) forecast accuracy more than negative adjustments (40.2% of cases). However, none of these cases improved forecast accuracy overall which is expected due to privacy protection.

**Table 6: (and percentage of adjustments that improved accuracy) by magnitude and direction.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Direction | |  |
|  |  | Positive | Negative | **Total** |
| Magnitude | Large | 1.35 (40.5%) | 1.47 (30.4%) | **1.41 (35.6%)** |
| Medium | 1.12 (44.4%) | 1.17 (41.9%) | **1.14 (43.2%)** |
| Small | 0.99 (49.1%) | 1.06 (46.8%) | **1.03 (47.9%)** |
|  | **Total** | **1.14 (44.6%)** | **1.21 (40.2%)** | **1.17 (42.5%)** |

One issue with our data is that 73% of the series have negative slopes, which could cause positive adjustments to have a dampening effect on forecasts, and negative adjustments to overestimate the impact of the trend (Hyndman & Athanasopoulos, 2021). Table 7 displays the AvgRelAE and the percentage of adjustments for time series with positive slopes vs. negative slopes. To measure the slope, we calculate the slope coefficient of a simple linear regression that regresses the time series values on a continuous time variable. Our results show that time series with negative slopes and negative adjustments (32% of all time series) tended to degrade forecast accuracy the most.

**Table 7: (and the percentage of adjustments that improved accuracy) by slope and direction.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Direction | |  |
|  |  | Positive | Negative | **Total** |
| Slope | Positive | 1.13 (42.9%) | 1.14 (41.6%) | **1.14 (42.2%)** |
| Negative | 1.14 (45.0%) | 1.24 (39.6%) | **1.18 (42.6%)** |
|  | **Total** | **1.14 (44.6%)** | **1.21 (40.2%)** | **1.17 (42.5%)** |

Table 8 measures the percentage of adjustments that improved accuracy and AvgRelAE categorized by the coefficient of variation of the original series and whether k-nTS+ (k=3) swapping increased, decreased, or maintained (within five percent) the coefficient of variation. We measure the coefficient of variation using the original time series values since there was only one forecast horizon. We bin the coefficients of variation into high ( quantile), low ( quantile) and medium ( quantile and quantile). We find that none of the coefficient of variation categories improve forecast accuracy compared to the original data. However, forecast accuracy degraded the most when k-nTS+ (k=3) was applied to time series with small coefficients of variation.

**Table 8: AvgRelAE (and percentage of adjustments that improved accuracy) by coefficient of variation of the original series and the change in coefficient of variation in the protected series.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Change in Coefficient of Variation | | |  |
|  |  | Decreased | Maintained (+/- 5%) | Increased | Total |
| Original Coefficient of Variation | Large | 1.06 (49.1%) | 1.14 (40.5%) | 1.30 (44.7%) | **1.09 (46.2%)** |
| Medium | 1.18 (42.5%) | 1.19 (40.3%) | 1.12 (49.1%) | **1.16 (44.1%)** |
| Small | 1.15 (38.2%) | 1.36 (31.6%) | 1.27 (36.1%) | **1.27 (35.7%)** |
|  | **Total** | **1.13 (45.0%)** | **1.19 (39.4%)** | **1.21 (41.7%)** | **1.17 (42.5%)** |
|  |  | Proportion of Time Series | | |  |
|  |  | Decreased | Maintained (+/- 5%) | Increased | Total |
| Original Coefficient of Variation | Large | 15.9% | 7.8% | 1.3% | **25.0%** |
| Medium | 22.7% | 11.7% | 15.6% | **50.0%** |
| Small | 1.1% | 2.5% | 21.4% | **25.0%** |
|  | **Total** | **39.7%** | **22.0%** | **38.3%** | **100.0%** |

Overall, our empirical results show that privacy adjustments affect forecast accuracy differently than judgmental adjustments. Specifically, we found that privacy adjustments had better forecast accuracy when the adjustments were small or positive, or when the coefficient of variation of the original series was large. However, on average, forecast accuracy worsened for nearly every combination of magnitude, direction, and coefficient of variation. This is not surprising since a major motivation of judgmental adjustments is to improve forecast accuracy (Fildes et al., 2019) and judgmental adjustments have been shown to improve forecast accuracy by 5-10% on average (Davydenko & Fildes, 2013; Khosrowabadi et al., 2022). For our application, privacy adjustments blur the data for privacy reasons and are expected to reduce forecast accuracy. The secondary goal of our proposed privacy method is to maintain forecast accuracy, which the top performing method (k-nTS+ (k=3) swapping) did with only a +13.9% average degradation. Furthermore, the average coefficient of variation of k-nTS+ (k=3) protected data is approximately 2% less than the average coefficient of variation in the original data. However, the average coefficients of variation under DP and AN were 18% and 35% larger than the average from the original data, respectively.

**6.5 Autocorrelation Feature Results**

In Section 4.5.3, we noted that while *k*-nTS performed swapping based on the values of *Spectral Entropy, Hurst*, and *Seasonality*, it does not preserve these feature distributions as well as *k*-nTS+. To help explain this difference, in Figure 8 we plot the distributions of four autocorrelation-based features, *Seasonal ACF* (first coefficient of the seasonal autocorrelation function), *X ACF* (first coefficient of the autocorrelation function), *X ACF10* (sum of the first ten coefficients of the autocorrelation function), and *X PACF5* (sum of the first five coefficients of the partial autocorrelation function). While neither k-nTS or k-nTS+ used these autocorrelation features for swapping, k-nTS+ preserves the distributions of these features much better than the other privacy methods which again demonstrates the importance of the k-nTS+ feedback loop.

**Figure 8: distributions of autocorrelation-based features in the original and protected data sets.**

Chart, box and whisker chart

Description automatically generated

1. Implementation details can be found in the appendix. [↑](#footnote-ref-1)
2. The averages for additive noise and differential privacy excludes the VAR model error for AN (s = 1) and DP () since the errors were over 1000% larger than the error of any other model. Due to the large noise infused from these privacy methods, the VAR could not fit small enough coefficients to smooth out the noise, resulting in extremely poor forecast accuracy. For example, the magnitude of the first lag coefficient for an AN (s = 1) protected time series increased from -0.372 in the original data to -0.679 in the protected data. This coefficient was multiplied by an extreme outlier at time causing the forecast at time to explode and skew the overall average forecast error. This problem did not occur for the other forecasting models, which did a better job smoothing out the random noise. [↑](#footnote-ref-2)
3. AvgRelAE can be generalized to accommodate multiple forecasts for each series. See Fildes et al. (2013) for the AvgRelMAE. [↑](#footnote-ref-3)